LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-439511

Lecture Notes on Multigrid Methods

P. S. Vassilevski

July 1, 2010

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Lecture Notes on Multigrid Methods

Panayot S. Vassilevski

Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA.

E-mail address: panayot@llnl.gov

Preface

The Lecture Notes [1] are primarily based on a sequence of lectures given by the author while been a Fulbright scholar at "St. Kliment Ohridski" University of Sofia, Sofia, Bulgaria during the winter semester of 2009-2010 academic year. The notes are somewhat expanded version of the actual one semester class he taught there. The material covered is slightly modified and adapted version of similar topics covered in the author's monograph "Multilevel Block–Factorization Preconditioners" published in 2008 by Springer.

The author tried to keep the notes as self-contained as possible. That is why the lecture notes begin with some basic introductory matrix-vector linear algebra, numerical PDEs (finite element) facts emphasizing the relations between functions in finite dimensional spaces and their coefficient vectors and respective norms.

Then, some additional facts on the implementation of finite elements based on relation tables using the popular compressed sparse row (CSR) format are given. Also, typical condition number estimates of stiffness and mass matrices, the global matrix assembly from local element matrices are given as well.

Finally, some basic introductory facts about stationary iterative methods, such as Gauss–Seidel and its symmetrized version are presented.

The introductory material ends up with the smoothing property of the classical iterative methods and the main definition of two–grid iterative methods.

From here on, the second part of the notes begins which deals with the various aspects of the principal TG and the numerous versions of the MG cycles. At the end, in part III, we briefly introduce algebraic versions of MG referred to as AMG, focusing on classes of AMG specialized for finite element matrices.

Sofia, Bulgaria
January 30, 2010

[1] This work was in part performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Contents

List of Figures

Part 1

Motivation and Preliminaries

CHAPTER 1

Matrix-vector linear algebra and some basic finite elements facts

This lecture contains a brief summary of results about matrix-vector notation, elliptic boundary value problems, their weak formulation, Galerkin method and some preliminary facts about finite element Galerkin discretization.

1. Notation

Vectors and matrices. Vector quantities are denoted in boldface, i.e., $\mathbf{u}$, $\mathbf{v}$, ... We use vector-columns, i.e.,

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \in \mathbb{R}^n.$$

The transpose of $\mathbf{v}$ denoted $\mathbf{v}^T$ is the vector-row $(v_1, \ldots, v_n)$. Given a $m \times n$ matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1,n} \\ \vdots & \ddots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}$$

the product $A\mathbf{v}$ equals the vector $\begin{bmatrix} \sum_{j=1}^{n} a_{1j}v_j \\ \vdots \\ \sum_{j=1}^{n} a_{ij}v_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj}v_j \end{bmatrix} \in \mathbb{R}^m$. Sometimes we write for short $A = (a_{ij})$.

More generally, given two matrices, an $m \times n$ matrix $A = (a_{ik})$ and an $n \times \ell$ matrix $B = (b_{kl})$ the product $C = AB$ is the $m \times \ell$ matrix with entries c_{il} given by the expressions $\sum_{k=1}^{n} a_{ik}b_{kl}$. In short, matrices are multiplied "row-times-column".

Symmetric and positive definite matrices. A $n \times n$ (square) matrix $A = (a_{ij})$ is called symmetric if $\mathbf{v}^T A\mathbf{w} = \mathbf{w}^T A\mathbf{v}$ for any two vectors $\mathbf{v}$ and $\mathbf{w}$. It is clear that this is equivalent to $a_{ij} = a_{ji}$.

A square matrix A is called positive definite if $\mathbf{v}^T A\mathbf{v} > 0$ for any non-zero vector $\mathbf{v}$.

For symmetric matrices the following extreme values of the Rayleigh quotient

$$\max_{\mathbf{v}} \frac{\mathbf{v}^T A\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \text{ and } \min_{\mathbf{v}} \frac{\mathbf{v}^T A\mathbf{v}}{\mathbf{v}^T\mathbf{v}}$$

characterize the minimal and maximal eigenvalues of A. Note that symmetric matrices have real eigenvalues.

By definition, for an $n \times n$ matrix A, λ is an eigenvalue of A if there is a non-zero vector $\mathbf{q}$ such that

$$A\mathbf{q} = \lambda\mathbf{q}.$$

For symmetric matrices both λ and $\mathbf{q}$ are real.

More over, for symmetric matrices the following spectral decomposition of A holds. There is an orthogonal $n \times n$ matrix Q, that is, $Q^T = Q^{-1}$ and a diagonal matrix $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_{n-1} & 0 \\ 0 & 0 & \dots & 0 & \lambda_n \end{bmatrix}$ such that

$$A = Q\Lambda Q^T.$$

Equivalently, $AQ = Q\Lambda$, that is

$$A[\mathbf{q}_1, \dots, \mathbf{q}_n] = [\mathbf{q}_1, \dots, \mathbf{q}_n]\Lambda = [\mathbf{q}_1\lambda_1, \dots, \mathbf{q}_n\lambda_n].$$

The latter written componentwise read:

$$A\mathbf{q}_k = \lambda_k\mathbf{q}_k, \ k = 1, \dots, n.$$

That is, $\mathbf{q}_k$ and λ_k are an eigenvector and a corresponding eigenvalue of A.

Based on the spectral decomposition for positive definite matrices (in that case all $\lambda_k > 0$), we can define functions of A, for example, we can define square root of A by letting $A^{\frac{1}{2}} = Q\Lambda^{\frac{1}{2}}Q^T$. The $\Lambda^{\frac{1}{2}}$ is the diagonal matrix with entries on the main diagonal equal to $\sqrt{\lambda_k}$. It is clear that $A^{\frac{1}{2}}A^{\frac{1}{2}} = A$.

Scalar and vector–functions. We consider scalar functions $u = u(\mathbf{x})$ where $\mathbf{x} \in \Omega \subset \mathbb{R}^d$. For the most part, we consider $d = 2$, however the results are general and hold for $d = 3$ as well. Also, the domain Ω is a bounded planar polygon ($d = 2$).

Also, we consider vector functions, for example,

$$\mathbf{u} = \mathbf{u}(\mathbf{x}) = \begin{bmatrix} u_1(\mathbf{x}) \\ u_2(\mathbf{x}) \\ \vdots \\ u_n(\mathbf{x}) \end{bmatrix}.$$

Dot–product of vector-functions. Let $\mathbf{u} = (u_i)$ and $\mathbf{v} = (v_i)$ be two vector functions. The following dot product is often used

$$\mathbf{u} \cdot \mathbf{v} = u_1v_1 + \cdots + u_nv_n.$$

For any fixed $\mathbf{x}$ this is simply the inner (scalar) product of the vectors $\mathbf{u}(\mathbf{x}), \ \mathbf{v}(\mathbf{x}) \in \mathbb{R}^n$.

Gradient of scalar function. The gradient of a scalar function is a vector-function, i.e., we have

$$\nabla u = \begin{bmatrix} \frac{\partial u}{\partial x_1} \\ \vdots \\ \frac{\partial u}{\partial x_d} \end{bmatrix}.$$

Divergence of vector–function. For a vector-function $\mathbf{u} = (u_i)_{i=1}^d$ we can define divergence

$$\mathrm{div}\mathbf{u} = \frac{\partial u_1}{\partial x_1} + \cdots + \frac{\partial u_d}{\partial x_d}.$$

Laplace operator. The Laplace operator is defined by

$$\Delta u = \mathrm{div}\nabla u = \frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_d^2}.$$

Normal vector to a domain boundary. For a given polygonal domain Ω, we can define unit outward normal vector $\mathbf{n}$ that is piecewise constant (and not defined at the corners (vertices) of Ω).

2. Boundary–value problems

Let $\Omega \subset \mathbb{R}^d$ $(d = 2)$ be a planar polygon. Also, let $\Gamma = \partial\Omega$ be the boundary of Ω and $\mathbf{n}$ its unit normal pointing outward Ω.

Integration by-parts formula. For any sufficiently smooth scalar function φ and vector function $\mathbf{v}$ the following formula for integration by parts holds:

$$\int\limits_{\Omega} \varphi \mathrm{div}\mathbf{v}\ d\mathbf{x} = -\int\limits_{\Omega} \mathbf{v}\cdot\nabla\varphi\ d\mathbf{x} + \int\limits_{\partial\Omega} \varphi\mathbf{v}\cdot\mathbf{n}\ d\boldsymbol{\sigma}.$$

It is a simple consequence from the following formula of Gauss

$$\int\limits_{\Omega} \frac{\partial w}{\partial x_i}\ d\mathbf{x} = \int\limits_{\partial\Omega} w\ \cos(\mathbf{n},\ \mathbf{e}_i)\ d\sigma.$$

Simply, for a given $\mathbf{v} = (v_i)_{i=1}^d$ and φ, apply the above formula for $w := v_i\varphi$ using $\frac{\partial(v_i\varphi)}{\partial x_i} = v_i\frac{\partial\varphi}{\partial x_i} + \varphi\frac{\partial v_i}{\partial x_i}$. We arrive at the desired result after summing up the formulas

$$\int\limits_{\Omega} \varphi\frac{\partial v_i}{\partial x_i}\ d\mathbf{x} = -\int\limits_{\Omega} v_i\frac{\partial\varphi}{\partial x_i}\ d\mathbf{x} + \int\limits_{\partial\Omega} \varphi v_i\ \cos(\mathbf{n},\ \mathbf{e}_i)\ d\sigma.$$

for $i = 1,\ 2,\ \ldots,\ d$, using the decomposition

$$\mathbf{n} = \sum_{i=1}^{d} \cos(\mathbf{n},\ \mathbf{e}_i)\mathbf{e}_i,$$

which implies

$$\mathbf{v}\cdot\mathbf{n} = \sum_{i=1}^{d}\sum_{i=1}^{d} v_i\cos(\mathbf{n},\ \mathbf{e}_i).$$

Poincaré–Steklov operators and traces of functions. Let Δ_F be the $d-1$-dimensional Laplace operator and consider its L_2–orthogonal system of eigen-functions

$$-\Delta_F \psi_k = \lambda_k \, \psi_k.$$

The functions ψ_k vanish on ∂F and satisfy $\int\limits_F \psi_k \psi_l \, d\mathbf{y} = \delta_{k,l}$. For each $\psi_k = \psi_k(\mathbf{y})$ solve the following 1–d boundary value problem

$$\lambda_k \varphi_k - \varphi_k^{\text{“}} = 0,$$

subject to $\varphi_k(-1) = 0$ and $\varphi_k(0) = 1$. The solution reads

$$\varphi_k(x) = \frac{e^{\sqrt{\lambda_k}(x+1)} - e^{-\sqrt{\lambda_k}(x+1)}}{e^{\sqrt{\lambda_k}} - e^{-\sqrt{\lambda_k}}}.$$

It is clear then that $u_k = \varphi_k(x)\psi_k(\mathbf{y})$ solves the homogeneous PDE $-\Delta u_k = 0$.

Given now a $g = g(\mathbf{y})$ expanded in terms of the basis of the eigenfunctions $\{\psi_k\}$

$$g = \sum_k c_k \psi_k,$$

the following function

$$u(x, \ \mathbf{y}) = \sum_k c_k \varphi_k(x) \psi_k(\mathbf{y}),$$

solves the Dirichlet boundary value problem

$$-\Delta u = 0 \text{ in } \Omega = (-1, 0) \times F,$$

subject to $u(0, y) = g(y), \ y \in F$ and $u = 0$ on $\partial\Omega \setminus \{(0, \ F)\}$.

The latter boundary value problem defines the so-called Poincaré–Steklov operator via the relation

$$g \in L_2(F) \mapsto Sg = \left. \frac{\partial u}{\partial x} \right|_{x=0}.$$

We have

$$Sg = \sum_k \sqrt{\lambda_k} \, c_k \coth\left(\sqrt{\lambda_k}\right) \psi_k.$$

The latter expression imposes some restrictions on the growth rate of the Fourier coefficients $\{c_k\}$ of g. Namely, we assume that $g = g(\mathbf{y})$ is such that

$$\|g\|^2_{H_0^{\frac{1}{2}}(F)} \equiv (Sg, \ g) = \sum_k \sqrt{\lambda_k} \, c_k^2 \coth\left(\sqrt{\lambda_k}\right) \ \simeq \sum_k \sqrt{\lambda_k} \, c_k^2 < \infty.$$

REMARK 2.1. *It can be shown that $\lambda_k \simeq k^2$ where the equivalence constants depend on the diameter of F.*

Hence above and in what follows, we can replace $\sqrt{\lambda_k}$ with k.

The integration by parts formula (valid for sufficiently smooth functions) can be extended by continuity to give the following variational definition of S

$$(Sg, \ \varphi)_F = \int\limits_\Omega \nabla u \cdot \nabla \varphi \, d\,\mathbf{x}.$$

Therefore

$$\|g\|^2_{H_0^{\frac{1}{2}}(F)} = (Sg,\ g)_F = |u|_1^2.$$

More generally, we can define for any $s \in R$ the fractional order Sobolev spaces on domain boundary $F \subset \partial\Omega$,

$$\|g\|_{s,\,F} \equiv \|g\|^2_{H_0^s(F)} = \sum_k \lambda_k^s c_k^2,$$

as long as the above series is convergent.

PROPOSITION 2.1. *Let $u \in H^1(\Omega)$, $\Omega = (-1,0)\times F$ and let u vanish on $\partial\Omega \setminus \{0\}\times F$. Then $g = u|_{x=0} \in H_0^{\frac{1}{2}}(F)$ and $Sg \in H_0^{-\frac{1}{2}}(F)$ and the following trace inequalities hold:*

$$\|g\|_{H_0^{\frac{1}{2}}(F)} \simeq \|Sg\|_{H_0^{-\frac{1}{2}}(F)} \leq |u|_1.$$

PROOF. For any given harmonic function φ (i.e., $\Delta\varphi = 0$) vanishing on $\partial\Omega \setminus (\{x = 0\}\times F)$we denote φ_F its trace on F. Next, we use the fact that S is symmetric, i.e., $(Sg,\ \varphi_F)_F = (g,\ S\varphi_F)_F$. Indeed, let for two functions g and g' defined on F with Fourier expansions $g = \sum_k c_k\psi_k$ and $g' = \sum_k c_k'\psi_k$, we have

$$(Sg,\ g')_F = \sum_k \sqrt{\lambda_k}\coth(\sqrt{\lambda_k})\ c_k c_k' = (g,\ Sg')_F.$$

The rest follows from the formula $(Sg,\varphi) = \int\limits_\Omega \nabla u\cdot\nabla\varphi\ d\mathbf{x}$ (where now $g = u|_F$) and the definition of fractional order Sobolev norms. More specifically, using the duality definition, for any harmonic function φ vanishing on $\partial\Omega\setminus\{0\}\times F$, we obtain

$$\|g\|_{\frac{1}{2},\,F} \simeq \|Sg\|_{-\frac{1}{2}} \simeq \sup_\varphi \frac{(Sg,\ \varphi_F)_F}{\|\varphi_F\|_{\frac{1}{2},\,F}} = \sup_\varphi \frac{(g,\ S\varphi_F)_F}{\|\varphi_F\|_{\frac{1}{2},\,F}} = \sup_\varphi\ \frac{\int\limits_\Omega \nabla u\cdot\nabla\varphi}{\|\nabla\varphi\|_0} \leq \|\nabla u\|_0.$$

□

As a corollary of the above proof, we obtain the following characterization result for S.

COROLLARY 2.1. *We have the following minimization property of S:*

$$(Sg,\ g)_F = \inf_{\{u\in H^1(\Omega),\ u|_F = g\ and\ u=0\ on\ \partial\Omega\setminus F\}} \int\limits_\Omega |\nabla u|^2\ d\mathbf{x}.$$

REMARK 2.2. *The results in this sub-section hold for general polyhedral domains Ω not necessarily being of the tensor product form $(-1,0)\times F$ assumed here using more general definitions of Sobolev spaces on (parts of) the boundary $\partial\Omega$.*

Boundary value problems. Let part of Γ be Γ_D and the remainder be $\Gamma_N = \Gamma \setminus \Gamma_D$. We consider Γ_D to be non–empty. For a given function $f = f(\mathbf{x}) \in L_2(\Omega)$, i.e., $\int\limits_\Omega f^2(\mathbf{x})\, d\mathbf{x} < \infty$, and a function $g_N \in L_2(\Gamma_N)$, we are interested in the following boundary–value problem:

Find a sufficiently smooth function $u = u(\mathbf{x})$ such that

$$-\Delta u = f(\mathbf{x}) \text{ in } \Omega, \tag{1.1}$$

such that

$$u = 0 \text{ on } \Gamma_D \text{ and } \nabla u \cdot \mathbf{n} = g_N \text{ on } \Gamma_N. \tag{1.2}$$

If Γ_N is empty, i.e., $\Gamma_D = \Gamma = \partial\Omega$, the above problem is referred to as the Dirichlet boundary value problem. Note that if Γ_D is empty set, then we have a Neumann boundary value problem that may not have a solution for any f and g_N. Also, for the Neumann problem if u is a solution, then u + const is also a solution, that is, the solution is determined up to a constant.

Weak formulation of boundary value problems. Introduce the Sobolev space $H^1(\Omega)$ of functions $u \in L_2(\Omega)$ such that their first partial derivatives $\frac{\partial u}{\partial x_i}$ also belong to $L_2(\Omega)$. If the functions vanish on $\partial\Omega$ the corresponding subspace is denoted by $H^1_0(\Omega)$.

Let u solve the Laplace equation $-\Delta u = f$ for a given $f \in L_2(\Omega)$. Introduce the vector function $\mathbf{v} = -\nabla u$. For any smooth function φ using the integration by part formula, we have

$$\begin{aligned}\int\limits_\Omega f\varphi\, d\mathbf{x} = -\int\limits_\Omega \varphi \Delta u\, d\mathbf{x} = \int\limits_\Omega \varphi\, \mathrm{div}\mathbf{v}\, d\mathbf{x} \;&= -\int\limits_\Omega \mathbf{v}\cdot\nabla\varphi\, d\mathbf{x} + \int\limits_{\partial\Omega} \varphi\mathbf{v}\cdot\mathbf{n}\, d\boldsymbol{\sigma} \\ &= \int\limits_\Omega \nabla\varphi\cdot\nabla u\, d\mathbf{x} - \int\limits_{\partial\Omega} \varphi\nabla u\cdot\mathbf{n}\, d\boldsymbol{\sigma}.\end{aligned}$$

Assume now that $u = u(\mathbf{x})$ satisfies the boundary conditions (1.2). Choosing then φ vanishing on Γ_D (the same as u), the following identity is obtained

$$\int\limits_\Omega \nabla u\cdot\nabla\varphi\, d\mathbf{x} = \int\limits_\Omega f\varphi\, d\mathbf{x} + \int\limits_{\partial\Omega_N} \varphi g_N\, d\boldsymbol{\sigma}. \tag{1.3}$$

The above identity is referred as the *weak formulation* of the boundary value problem (1.1)-(1.2). In this form the minimal requirement on u is to have only first partial derivatives in L_2–sense, that is, $u \in H^1(\Omega)$ and vanishing on Γ_D. Similarly, it is sufficient to choose $\varphi \in H^1(\Omega)$ also vanishing on Γ_D.

Bilinear form and solution of boundary value problem. Introduce the bilinear form

$$a(u,\ v) = \int\limits_\Omega \nabla u\cdot\nabla v\, d\mathbf{x},$$

defined for functions in $H^1(\Omega)$ vanishing on $\Gamma_D \subset \partial\Omega$.

The bilinear form is symmetric and positive definite for the above class of functions further denoted as V. Namely, if $a(u,u) = 0$ it follows that $\nabla u = 0$, hence u = const. However u vanishes on Γ_D hence $u = 0$. In short, $a(.\ .)$ defines an inner product on V. By definition, V is a Hilbert space in that inner product.

Consider $g_N \in H^{-\frac{1}{2}}(\Gamma_N)$ and $f \in L_2(\Omega)$. The following expression

$$\ell(\varphi) \equiv \int_\Omega f\varphi \, d\mathbf{x} + \int_{\Gamma_N} g_N \, \varphi \, d\boldsymbol{\sigma},$$

defines a linear functional for $\varphi \in V$. Based on the trace estimates in Proposition 2.1 and Friedrich's inequality, we have that $\ell(\varphi)$ is bounded, i.e.,

$$|\ell(\varphi)| \leq \|f\|_0 \|\varphi\|_0 + \|g_N\|_{-\frac{1}{2},\,\Gamma_N} \|\varphi\|_{\frac{1}{2},\,\Gamma_N} \leq C \, \left(\|f\|_0 + \|g_N\|_{-\frac{1}{2},\,\Gamma_N} \right) \, \|\nabla\varphi\|_0.$$

Using the Riesz representation theorem for bounded linear functionals in Hilbert spaces, it follows that the above linear functional $\ell(\varphi)$ defined for $\varphi \in V$, can be represented based on the inner product of the Hilbert space, in our case the one given by the bilinear form $a(.,.)$. That is, there is an unique element $u = u_\ell \in V$ such that for all $\varphi \in V$

$$a(u, \, \varphi) = \ell(\varphi).$$

This shows that the weak formulation (1.3) of the boundary value problem (1.1)-(1.2) has a (unique) solution.

3. The Galerkin method

Let $\{\varphi_i\}_{i=1}^n$ be a finite set of linearly independent functions in V. The Galerkin method constructs the best approximation to a function $u \in V$ from the finite dimensional space spanned by the functions $\{\varphi_i\}_{i=1}^n$. That is, we are looking for the coefficients $\{u_i\}_{i=1}^n$ such that

$$\|u - \sum_{i=1}^n u_i \varphi_i\| \mapsto \min .$$

Here, $\|v\| = \sqrt{a(v, \, v)}$ is the norm induced by the inner product in V, that is, by our bilinear form $a(.,.)$. If $u = u(\mathbf{x})$ is the (unknown) solution for the weak form of the boundary value problem (or b.v.p.) (1.3), it turns out that even though u is not actually known, the coefficients $\{u_i\}$ are computationally feasible and uniquely determined. Indeed, we get the following quadratic functional to minimize

$$J(u_1, \, \ldots, \, u_n) \equiv a(u - \sum_i u_i \varphi_i, \, u - \sum_i u_i \varphi_i).$$

By looking at $\frac{\partial J}{\partial u_j} = 0$, we obtain

$$0 = a(u - \sum_i u_i \varphi_i, \, \varphi_j).$$

Therefore,

$$\sum_{i=1}^n u_i a(\varphi_i, \, \varphi_j) = a(u, \, \varphi_j).$$

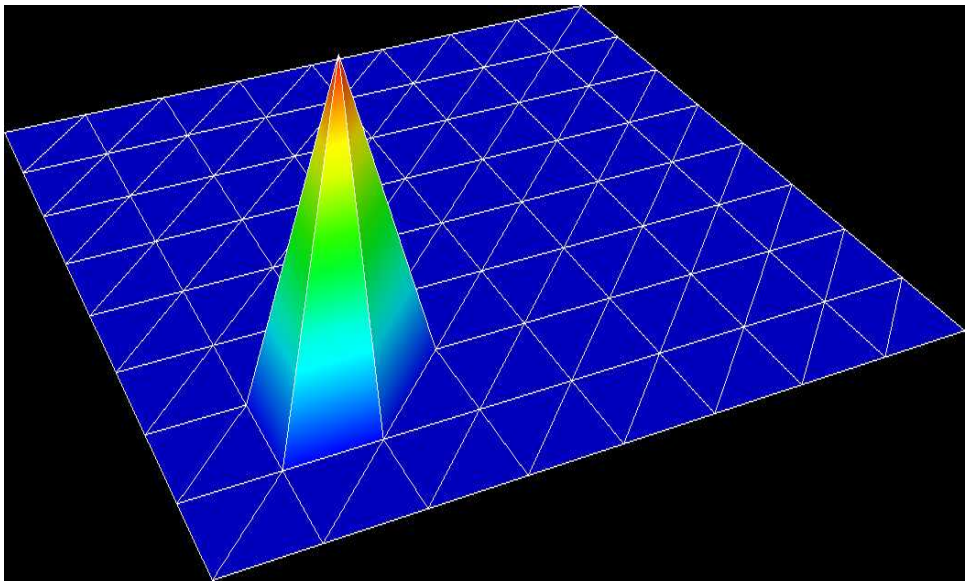

FIGURE 1. *Piecewise linear basis function on triangular elements.*

Since $a(u,\ \varphi_j) = \int\limits_\Omega\ f,\ \varphi_j\ d\mathbf{x} + \int\limits_{\Gamma_N}\ \varphi_j\ g_N\ d\boldsymbol{\sigma}$, we obtain at the end the linear system of n equations for n unknowns

$$\sum_i^n a(\varphi_i,\ \varphi_j)u_i = \int\limits_\Omega\ f\ \varphi_j\ d\mathbf{x} + \int\limits_{\Gamma_N}\ \varphi_j\ g_N\ d\boldsymbol{\sigma},\ \text{for all } j = 1, 2,\ \ldots,\ n.$$

This system has a unique solution since the functions φ_i are linearly independent which implies that the $n \times n$ "Gram " matrix A with (i,j)th entry $a(\varphi_j,\ \varphi_i)$ is invertible. If we form the right-hand side vector $\mathbf{f} = (f_j)_{j=1}^n$ with $f_j = \int\limits_\Omega\ f\ \varphi_j\ d\mathbf{x} + \int\limits_{\Gamma_N}\ \varphi_j\ g_N\ d\boldsymbol{\sigma}$, the Galerkin system for the vector of unknown coefficients $\mathbf{u} = (u_i)_{i=1}^n$ can be written in the matrix-vector form

$$A\mathbf{u} = \mathbf{f}.$$

Again, since A is a Gram matrix, it is symmetric and positive definite. In general, it inherits in our case the properties of the bilinear form $a(.,.)$ (such as symmetry and positive definiteness).

The finite element method. The finite element method is a special case of the Galerkin method corresponding to a specific choice of the linearly independent test functions $\{\varphi_i\}_{i=1}^n$.

A typical local basis function in 2D is illustrated in Fig. 1.

A partition of the given polygonal domain Ω into simple elements τ (typically triangles, or quadrilaterals such as rectangles)such that any pair of elements share a common edge or a common vertex or are non-intersecting is referred to as a triangulation $\mathcal{T}$. The elements are assumed to have diameter of order h which is meant to tend to zero. We sometimes write $\mathcal{T} = \mathcal{T}_h$.

Consider the case of triangular elements. The set of vertices, called nodes $\{\mathbf{x}_i\}$, is denoted by $\mathcal{N} = \mathcal{N}_h$. For each vertex $\mathbf{x}_i$ we associate a basis function φ_i that is linear function of two variables when restricted to an individual element τ. The function φ_i is such that it is locally supported in the neighborhood of elements sharing the vertex $\mathbf{x}_i$. This implies that $\varphi_i(\mathbf{x}_j) = \delta_{i,j} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$.

Consider an element τ with its vertices

$$\tau = (\mathbf{x}_{i_1},\ \mathbf{x}_{i_2},\ \mathbf{x}_{i_3}).$$

One of the vertices $\mathbf{x}_{i_s}$ equals $\mathbf{x}_i$. Let $\mathbf{x}_{i_s}$ have coordinates $(x_{i_s},\ y_{i_s})$. Then the following equation can be used to define $\varphi_i(\mathbf{x})$ for $\mathbf{x} = (x, y) \in \tau$

$$0 = \begin{vmatrix} x & y & \varphi_i & 1 \\ x_{i_1} & y_{i_1} & \delta_{i,\,i_1} & 1 \\ x_{i_2} & y_{i_2} & \delta_{i,\,i_2} & 1 \\ x_{i_3} & y_{i_3} & \delta_{i,\,i_3} & 1 \end{vmatrix}$$

The derivatives of φ_i are similarly computed. For example, $\frac{\partial \varphi_i}{\partial x}$ is computed from the equation

$$0 = \begin{vmatrix} 1 & 0 & \frac{\partial \varphi_i}{\partial x} & 0 \\ x_{i_1} & y_{i_1} & \delta_{i,\,i_1} & 1 \\ x_{i_2} & y_{i_2} & \delta_{i,\,i_2} & 1 \\ x_{i_3} & y_{i_3} & \delta_{i,\,i_3} & 1 \end{vmatrix}.$$

Similarly, $\frac{\partial \varphi_i}{\partial y}$ satisfies

$$0 = \begin{vmatrix} 0 & 1 & \frac{\partial \varphi_i}{\partial y} & 0 \\ x_{i_1} & y_{i_1} & \delta_{i,\,i_1} & 1 \\ x_{i_2} & y_{i_2} & \delta_{i,\,i_2} & 1 \\ x_{i_3} & y_{i_3} & \delta_{i,\,i_3} & 1 \end{vmatrix}.$$

The Galerkin Gram matrix A in the case of finite elements is referred to as the stiffness matrix. Due to the choice of locally supported basis functions φ_i, the entries $a(\varphi_i,\ \varphi_j)$ of A are zero if the supports of φ_i and φ_j do not overlap. In the case of triangular elements described above an entry $a_{ij} = a(\varphi_j,\ \varphi_i)$ is nonzero if the nodes $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to a same element τ. Thus, the number of nonzero entries in a row i of the matrix A is bounded by $\kappa_i + 1$, where κ_i is the number of elements τ that share the vertex $\mathbf{x}_i$. If the angles of the triangles is kept bounded away from zero when $h \mapsto 0$, then we have that $\kappa = \max\limits_{\mathbf{x}_i \in \mathcal{N}_h} \kappa_i < \infty$.

This shows that the total number of nonzero entries of A is $\mathcal{O}(n)$, where n is the number of basis functions or equivalently the number of nodes $\mathcal{N}_h$.

Since the basis functions φ_i are Lagrangian (nodal), we have that the finite element Galerkin approximation

$$u_h = \sum_{i=1}^{n} u_i \varphi_i,$$

satisfies

$$u_h(\mathbf{x}_i) = u_i.$$

That is, the coefficients u_i are actually nodal values of the finite element approximation function u_h.

Finally, since φ_i restricted to any element is a polynomial function (linear in the above setting), it follows that the finite element Galerkin approximation u_h has some approximation properties. More specifically, let V_h stand for the finite element space spanned by the basis functions $\{\varphi_i\}_{i=1}^n$, then the following error estimate holds (see next Chapter)

$$\sqrt{a(u-u_h,\ u-u_h)} = \min_{v_h \in V_h} \sqrt{a(u-v_h,\ u-v_h)} \leq Ch\ |u|_2.$$

Here, we assume that the solution of the b.v.p. u is sufficiently smooth, that is, u has all partial derivatives up to order two in $L_2(\Omega)$. The term $|u|_2$ stands for the semi–norm defined by

$$|u|_2^2 = \int\limits_{\Omega} \left(\left(\frac{\partial^2 u}{\partial x^2} \right)^2 + \left(\frac{\partial^2 u}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 u}{\partial y^2} \right)^2 \right) d\mathbf{x}.$$

CHAPTER 2

Further results on finite elements and stationary iterative methods

This lecture contains a brief summary of further results on the finite element method, its computational aspects, element matrices, sparsity of assembled matrices, condition number estimates and some preliminary results about preconditioned stationary iterative methods.

1. The finite element method: further results

Computational aspects. The finite element method is a special case of the Galerkin method with specific choice of the basis (test) functions $\{\varphi_i\}_{i=1}^n$. A finite element method is characterized with a set of elements $\tau \in \mathcal{T}_h$ and the set of nodes $\mathbf{x}_i \in \mathcal{N}_h$. Typically, for piecewise linear basis functions φ_i and triangular elements τ, $\mathcal{N}_h$ is the set of vertices. In general, the indices i run over so called degrees of freedom (or dofs) that specify the basis of the finite element space V_h.

Relation tables. One way to specify the topology of a finite element mesh is via the so–called relation tables implemented as Boolean sparse matrices. For example, the incidence element i has a vertex j can be represented by the rectangular matrix where in row i we have nonzero entry at column j if the node $\mathbf{x}_j$ is a vertex of element i. It is clear that for triangular elements such a relation table will have exactly three non–zero entries per row. We denote this Boolean matrix as "element_vertex". Similarly, we can form the Boolean matrices "element_edge", "edge_vertex" etc. Utilizing operation between Boolean sparse matrices we can form transposed relation or transient relations. For example, the transposed matrix

$$\text{"vertex\_element"} = \text{"(element\_vertex)}^T\text{"}$$

has a non-zero entry at position (i, j) which represents the relation "node i is a vertex of element j".

The product of the Boolean sparse matrices

$$\text{"vertex\_vertex"} = \text{"vertex\_element"} \times \text{"element\_vertex"}$$

has non-zero entry at position (i, j) which represents the fact that node i and node j are vertices of a same element. If the degrees of freedom are associated with the vertices of the elements, the latter "vertex_vertex" relation shows exactly the sparsity (non–zero) pattern of the finite element stiffness matrix A.

Sparse matrices. From a finite element prospective a matrix $M = M_h$ is called sparse if it has bounded number of non–zero entries both per row and per column. "Bounded" means with respect to $h \mapsto 0$.

A characteristic feature of the finite element method (as we demonstrated earlier in Lecture # 1) is that the stiffness matrix A has at most $\kappa + 1$ nonzero entries per row

(and due to symmetry, per column) where κ stands for the maximum number of elements that share a common vertex. This is the case of finite element space V_h with degrees of freedom the vertices (nodes) $\mathcal{N}_h$ of elements $\tau \in \mathcal{T}_h$.

The same sparsity property holds for the finite element mass (Gram) matrix G with entries $g_{ij} = \int\limits_{\Omega} \varphi_i \varphi_j \, d\mathbf{x}$, $i,\ j = 1, \ldots,\ n$.

This fact shows that in order to store A (and G), we need $\mathcal{O}(n)$ memory.

Sparse matrix storage in CSR format. CSR stands for "compressed sparse row". The CSR format is a popular way to store finite element sparse matrices. For an $n \times m$ sparse matrix A, the CSR format exploits two one-dimensional integer arrays I and J and if the matrix is not Boolean (as the relation tables discussed previously) a real array "Data" is needed in addition to store the actual entries of A.

Let A have at row i, $m_i \geq 1$ non–zero entries at positions $(i, j_1^{(i)}),\ \ldots,\ (i, j_{m_i}^{(i)})$.

The one-dimensional array I has length $n+1$. With $I[0] = 0$, we set

$$I[i] = I[i-1] + m_i \text{ for } i \geq 1.$$

The array J has length $I[n]$. Similarly the data array has the same length $I[n]$.

For each row $i = 1,\ \ldots,\ n$ of A, we list consecutively in the one–dimensional array J the indices $j_s^{(i)}$, $s = 1,\ \ldots,\ m_i$ starting at position $I[i-1]$ till position $I[i] - 1$, that is

$$J[I[i-1] + s - 1] = j_s^{(i)},\ \text{for } s = 1,\ \ldots,\ m_i.$$

The data array is filled-in similarly, i.e., we let

$$\text{Data}[I[i-1] + s - 1] = a_{i,\, j_s^{(i)}} \text{ for } s = 1,\ \ldots,\ m_i.$$

Having sparse matrices stored in CSR format in practice it is useful to have algorithms that implement matrix operations such as A^T, matrix–matrix multiply $C = AB$. I.e., if A is stored in CSR format we need to store A^T in CSR format using only $\mathcal{O}(n)$ operations. Similarly, if the sparse matrices A and B are represented in CSR format with $\mathcal{O}(n)$ non–zero entries, we want to find an algorithm that computes and stores C in CSR format for $\mathcal{O}(n)$ storage and operations. All this is feasible for finite element sparse matrices.

Element matrices and matrix assembly. Since the entries of A (and G) are evaluation of certain integrals, these integrals can be split over the individual elements. In this way, we define element matrices. For example, let $\tau \in \mathcal{T}_h$ then for basis functions φ_i and φ_j such that their support and τ have non–empty intersection, we can compute the integrals

$$a_{ij}^{(\tau)} = \int\limits_{\tau} \nabla\varphi_j \cdot \nabla\varphi_i \, d\mathbf{x} \text{ and } g_{ij}^{(\tau)} = \int\limits_{\tau} \varphi_i \varphi_j \, d\mathbf{x}.$$

It is clear that for triangular elements τ and linear basis functions the respective (i,j) entries form 3×3 symmetric element stiffness and element mass matrices A_τ and G_τ, respectively.

The fact that the entries a_{ij} of A (the global stiffness matrix) can be computed from the respective entries of the element stiffness matrices A_τ using the formula

$$a_{ij} = \sum_{\tau:\, \mathbf{x}_i,\, \mathbf{x}_j \in \tau} a_{ij}^{(\tau)},$$

is referred to as *matrix assembly*.

A useful observation is that the diagonal entries of A, a_{ii}, can be assembled from the diagonal entries $a_{ii}^{(\tau)}$ of the respective element matrices A_τ.

Local and global quadratic forms. Denote by $\mathbf{v}_\tau$ the restriction of a given vector $\mathbf{v}$ to τ. More specifically, let $\mathbf{x}_{i_1},\ \mathbf{x}_{i_2},\ \mathbf{x}_{i_3}$ be the dofs (in our case vertices) associated with τ. Then, $\mathbf{v}_\tau = (v_{i_s})_{s=1}^3$ if $\mathbf{v} = (v_i)_{i=1}^n$.

By construction, the following identities hold

$$\mathbf{w}^T A\mathbf{v} = \sum_{\tau\in\mathcal{T}_h} \mathbf{w}_\tau^T A_\tau \mathbf{v}_\tau. \tag{1.4}$$

Similarly,

$$\mathbf{w}^T G\mathbf{v} = \sum_{\tau\in\mathcal{T}_h} \mathbf{w}_\tau^T G_\tau \mathbf{v}_\tau. \tag{1.5}$$

Given a vector $\mathbf{v} = (v_i)_{i=1}^n$ we can identify it with the finite element function $v = \sum_{i=1}^n v_i\varphi_i \in V_h$. It is clear then that

$$\mathbf{w}^T A\mathbf{v} = a(v,\ w) \text{ and } \mathbf{w}^T G\mathbf{v} = (v,\ w) \equiv \int_\Omega vw\ d\mathbf{x}.$$

Similarly, for every element τ, we have

$$\mathbf{w}_\tau^T A_\tau \mathbf{v}_\tau = \int_\tau \nabla v\cdot\nabla w\ d\mathbf{x} \text{ and } \mathbf{w}_\tau^T G_\tau \mathbf{v}_\tau = \int_\tau vw\ d\mathbf{x}.$$

The latter two representations and the fact that the integrals over Ω are sums of integrals over all $\tau\in\mathcal{T}_h$ show the relations (1.4)-(1.5).

2. Condition number estimates

The main result of this section is the estimates (spectral relations) between $A = (a_{ij})$ and the diagonal matrix D with non-zero entries a_{ii}, $i = 1,\ \ldots,\ n$. Also, we show that the mass matrix G is spectrally equivalent to the identity matrix scaled by the factor h^d. More specifically, the following main result holds.

THEOREM 2.1. *The following estimates hold for the stiffness matrix A and mass matrix G computed by a finite element space V_h on a polygonal domain $\Omega\subset\mathbb{R}^d$ ($d = 2$):*

$$\gamma_0\ h^2\ \mathbf{v}^T D\mathbf{v} \le \mathbf{v}^T A\mathbf{v} \le \max_{\tau\in\mathcal{T}_h}\ \kappa_\tau\ \mathbf{v}^T D\mathbf{v}.$$

Here, κ_τ stands for the number of dofs in an element τ. For triangular elements and linear functions, we have $\kappa_\tau = 3$. The constant $\gamma_0 > 0$ is independent of $h\mapsto 0$.

For the mass matrix G, we have for two positive mesh–independent constants c_0 and c_1, the equivalence relations

$$c_0\ h^d \mathbf{v}^T\mathbf{v} \le \mathbf{v}^T G\mathbf{v} \le c_1\ h^d\ \mathbf{v}^T\mathbf{v}.$$

This theorem implies that the symmetrically scaled stiffness matrix $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ based on the Rayleigh quotient estimates

$$\gamma_0 h^2 \leq \frac{\mathbf{v}^T D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \leq \max_{\tau\in\mathcal{T}_h}\kappa_\tau = \mathcal{O}(1),$$

has minimal eigenvalue of order $\mathcal{O}(h^2)$ and a maximal eigenvalue of order $\mathcal{O}(1)$. That is, the condition number of $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ is $\mathcal{O}(h^{-2})$. This shows that A becomes very ill-conditioned when $h \mapsto 0$.

The result in Theorem 2.1 is general. However, in what follows, we consider our model case of triangular elements and linear basis functions.

For a triangle τ with vertices $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \mathbf{x}_{i_3}$ let the angles associated with vertices $\mathbf{x}_{i_s}$ be θ_s. Then the following formulas hold for A_τ and G_τ:

$$A_\tau = \frac{1}{2}\begin{bmatrix} \cot\theta_2+\cot\theta_3 & -\cot\theta_3 & -\cot\theta_2 \\ -\cot\theta_3 & \cot\theta_3+\cot\theta_1 & -\cot\theta_1 \\ -\cot\theta_2 & -\cot\theta_1 & \cot\theta_1+\cot\theta_2 \end{bmatrix} \text{ and } G_\tau = \frac{|\tau|}{12}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

Here $\cot\theta = \frac{\cos\theta}{\sin\theta}$ and $|\tau| = \mathcal{O}(h^d)$ $(d = 2)$ stands for the area of τ. If the angles of τ stay bounded away from zero when $h \mapsto 0$, it is clear that the diagonal entries of A_τ are uniformly bounded from above and below for $h \mapsto 0$. Thus, we have for two h– and τ–independent positive constants γ_1 and γ_2

$$\gamma_1\ \mathbf{v}_\tau^T\mathbf{v}_\tau \leq \mathbf{v}_\tau^T D_\tau\mathbf{v}_\tau \leq \gamma_2\ \mathbf{v}_\tau^T\mathbf{v}_\tau. \tag{1.6}$$

Next, we compute the eigenvalues λ of the matrix $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ coming from the element mass matrix G_τ. We have

$$0 = \begin{vmatrix} 2-\lambda & 1 & 1 \\ 1 & 2-\lambda & 1 \\ 1 & 1 & 2-\lambda \end{vmatrix} = -(\lambda-1)^2(\lambda-4).$$

Therefore, $\lambda_{\min}(G_\tau) = \frac{|\tau|}{12}$ and $\lambda_{\max}(G_\tau) = \frac{|\tau|}{3}$, which implies based on the bounds for the Rayleigh quotient for G_τ,

$$\frac{|\tau|}{12}\mathbf{v}_\tau^T\mathbf{v}_\tau \leq \mathbf{v}_\tau^T G_\tau\mathbf{v}_\tau \leq \frac{|\tau|}{3}\ \mathbf{v}_\tau^T\mathbf{v}_\tau. \tag{1.7}$$

As a corollary, by comparing (1.6) and (1.7), we have

$$\overline{\gamma}_0 h^2\ \mathbf{v}_\tau^T D_\tau\mathbf{v}_\tau \leq \mathbf{v}_\tau^T G_\tau\mathbf{v}_\tau, \tag{1.8}$$

for some h– and τ–independent positive constant $\overline{\gamma}_0$.

The desired estimates for G follow from the local estimates (1.7) after summation over $\tau \in \mathcal{T}_h$, the fact that $\mathbf{v}^T G\mathbf{v} = \sum_\tau \mathbf{v}_\tau^T G_\tau\mathbf{v}_\tau$ and the estimate

$$\mathbf{v}^T\mathbf{v} \leq \sum_\tau \mathbf{v}_\tau^T\mathbf{v}_\tau \leq \kappa\ \mathbf{v}^T\mathbf{v},$$

where κ stands for the maximal number of elements that share any given vertex $\mathbf{x}_i$.

The uniform upper estimate for A is seen as follows. We have, using Cauchy–Schwarz inequality

$$\mathbf{v}_\tau^T A_\tau \mathbf{v}_\tau = \int_\tau \left| \sum_{\mathbf{x}_{i_s}\in\tau} v_{i_s} \nabla\varphi_{i_s} \right|^2 \, d\mathbf{x} \le \kappa_\tau \sum_{\mathbf{x}_{i_s}\in\tau} v_{i_s}^2 \int_\tau |\nabla\varphi_{i_s}|^2 \, d\mathbf{x} = \kappa_\tau \, \mathbf{v}_\tau^T D_\tau \mathbf{v}_\tau. \tag{1.9}$$

By summation over $\tau \in \mathcal{T}_h$, we obtain the desired upper bound

$$\mathbf{v}^T A \mathbf{v} = \sum_{\tau\in\mathcal{T}_h} \mathbf{v}_\tau^T A_\tau \mathbf{v}_\tau \le \max_{\tau\in\mathcal{T}_h} \kappa_\tau \sum_{\tau\in\mathcal{T}_h} \mathbf{v}_\tau^T D_\tau \mathbf{v}_\tau = \max_{\tau\in\mathcal{T}_h} \kappa_\tau \, \mathbf{v}^T D \mathbf{v}.$$

Inverse inequality. The inequalities (1.9) and (1.8) imply the so-called local inverse inequalities

$$\mathbf{v}_\tau^T A_\tau \mathbf{v}_\tau \le \frac{\kappa_\tau}{\overline{\gamma}_0} \, h^{-2} \, \mathbf{v}_\tau^T G_\tau \mathbf{v}_\tau.$$

The latter, after summation over $\tau \in \mathcal{T}_h$, lead to the global "inverse inequality"

$$\mathbf{v}^T A \mathbf{v} \le \max_\tau \kappa_\tau \, \overline{\gamma}_0^{-1} \, h^{-2} \mathbf{v}^T G \mathbf{v}.$$

The same result, rewritten in terms of functions and norms, reads

$$a(v, \, v) \le C_I \, h^{-2} \, \|v\|_0^2, \tag{1.10}$$

where $C_I = \max_\tau \kappa_\tau \, \overline{\gamma}_0^{-1}$.

Friedrich's inequality. For functions $v \in H^1(\Omega)$ vanishing on Γ_D a subset of $\partial\Omega$ with positive measure, the following Friedrich's inequality holds:

$$\|v\|_0^2 \equiv \int_\Omega v^2(\mathbf{x}) \, d\mathbf{x} \le C_F \, |v|_1^2, \qquad |v|_1^2 \equiv \int_\Omega |\nabla v|^2 \, d\mathbf{x}.$$

For the proof of the desired lower bound for A formulated in Theorem 2.1, we first use the Friedrich's inequality for the finite element function v corresponding to the vector $\mathbf{v}$. We have,

$$C_F \, \mathbf{v}^T A \mathbf{v} = C_F \, |v|_1^2 \ge \|v\|_0^2 = \mathbf{v}^T G \mathbf{v} = \sum_{\tau\in\mathcal{T}_h} \mathbf{v}_\tau^T G_\tau \mathbf{v}_\tau.$$

The remainder of the result follows from inequality (1.8) with $\gamma_0 = \frac{\overline{\gamma}_0}{C_F}$. I.e., we have

$$\mathbf{v}^T A \mathbf{v} \ge h^2 \frac{\overline{\gamma}_0}{C_F} \sum_{\tau\in\mathcal{T}_h} \mathbf{v}_\tau^T D_\tau \mathbf{v}_\tau = h^2 \gamma_0 \, \mathbf{v}^T D \mathbf{v}.$$

The Poincaré inequality. For any polygonal domain, the following Poincaré inequality holds for any $v \in H^1(\Omega)$ and its average value $\overline{v} = \frac{1}{|\Omega|} \int_\Omega v(\mathbf{x}) \, d\mathbf{x}$, where $|\Omega| = \int_\Omega 1 \, d\mathbf{x}$,

$$\|v - \overline{v}\|_0 \le C_\Omega \, \|\nabla v\|_0.$$

3. Stationary preconditioned iterative methods

Our ultimate goal is to solve the system of equations

$$A\mathbf{x} = \mathbf{b},$$

where A is the $n \times n$ ill-conditioned sparse finite element stiffness matrix obtained after discretizing the b.v.p. of interest using finite element space V_h corresponding to a triangulation $\mathcal{T}_h$, for $\mathcal{O}(n)$ operations where the constant in the $\mathcal{O}$-symbol is h–independent.

We first remark that direct methods cannot achieve this goal asymptotically for $h \mapsto 0$. That is why we focus our attention on iterative methods.

We begin with some standard iterative methods (like Gauss–Seidel).

Let M be an $n \times n$ matrix such that systems with M, $M\mathbf{y} = \mathbf{g}$, are easy to solve (i.e., in $\mathcal{O}(n)$ operations). Examples of such matrices are diagonal, lower (or upper) triangular sparse matrices, banded matrices etc. An example that we will be frequently using is $M = D + L$, where D is the diagonal of $A = (a_{ij})$ and L is the strictly lower triangular part of A. That is $L = (\ell_{ij})$ where

$$\ell_{ij} = \begin{cases} 0, & \text{if } i \leq j, \\ a_{ij}, & \text{if } i > j. \end{cases}$$

We have the decomposition

$$A = D + L + L^T.$$

For a given M, let $\mathbf{x}_0$ be a given initial approximation (guess), for example, $\mathbf{x}_0 = 0$, and consider the iteration process

$$M(\mathbf{x}_{k+1} - \mathbf{x}_k) = \mathbf{b} - A\mathbf{x}_k, \text{ for } k = 0, 1, \ldots \tag{1.11}$$

The matrix M is called preconditioner and the above iteration process (or method) preconditioned iteration process (or method). The term $\mathbf{x}_{k+1} - \mathbf{x}_k$ is called correction whereas the right hand-side (or r.h.s.) $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$ is called residual (or defect).

We are interested in the convergence of the error $\mathbf{e}_k = \mathbf{x} - \mathbf{x}_k$ to zero in some norm $\|.\|$. We note that

$$\mathbf{r}_k = A\mathbf{e}_k.$$

We have the following relation between two consecutive errors:

$$M(-\mathbf{e}_{k+1} + \mathbf{e}_k) = M(\mathbf{x}_{k+1} - \mathbf{x} + \mathbf{x} - \mathbf{x}_k) = \mathbf{b} - A\mathbf{x}_k = A\mathbf{e}_k.$$

That is,

$$\mathbf{e}_{k+1} = (I - M^{-1}A)\mathbf{e}_k. \tag{1.12}$$

The matrix $E = I - M^{-1}A$ is called *iteration matrix*.

Thus the method is convergent if for some norm $\|.\|$, $\|E\| < 1$, that is

$$\|I - M^{-1}A\| < 1.$$

We are interested in the convergence of the above stationary iterative method in energy norm $\|\mathbf{v}\|_A = \sqrt{\mathbf{v}^T A\mathbf{v}}$. Since A is symmetric positive definite (or s.p.d.) A defines an inner–product, hence $\|.\|_A$ is indeed a norm. Also, for s.p.d. A, we can define $A^{\frac{1}{2}}$ which is also s.p.d.. Then,

$$\|\mathbf{v}\|_A^2 = \mathbf{v}^T A\mathbf{v} = \mathbf{v}^T A^{\frac{1}{2}} A^{\frac{1}{2}}\mathbf{v} = (A^{\frac{1}{2}}\mathbf{v})^T A^{\frac{1}{2}}\mathbf{v} = \|A^{\frac{1}{2}}\mathbf{v}\|^2.$$

From (1.12), we have

$$A^{\frac{1}{2}}\mathbf{e}_{k+1} = A^{\frac{1}{2}}\left(I - M^{-1}A\right)A^{-\frac{1}{2}}\left(A^{\frac{1}{2}}\mathbf{e}_k\right) = \left(I - A^{\frac{1}{2}}M^{-1}A^{\frac{1}{2}}\right)\left(A^{\frac{1}{2}}\mathbf{e}_k\right).$$

Thus

$$\|\mathbf{e}_{k+1}\|_A \leq \|I - A^{\frac{1}{2}}M^{-1}A^{\frac{1}{2}}\|\|\mathbf{e}_k\|_A.$$

That is, we need to estimate the norm of the transformed iteration matrix

$$\mathcal{E} = I - A^{\frac{1}{2}}M^{-1}A^{\frac{1}{2}}.$$

For this purpose, we consider $\mathcal{E}^T\mathcal{E}$. We have,

$$\begin{aligned}\mathcal{E}^T\mathcal{E} &= (I - A^{\frac{1}{2}}M^{-T}A^{\frac{1}{2}})(I - A^{\frac{1}{2}}M^{-1}A^{\frac{1}{2}}) \\ &= I - A^{\frac{1}{2}}M^{-T}A^{\frac{1}{2}} - A^{\frac{1}{2}}M^{-1}A^{\frac{1}{2}} + A^{\frac{1}{2}}M^{-T}AM^{-1}A^{\frac{1}{2}} \\ &= I - \left(A^{\frac{1}{2}}M^{-T}\right)\left(M + M^T - A\right)\left(M^{-1}A^{\frac{1}{2}}\right) \\ &= I - Y^T\left(M + M^T - A\right)Y.\end{aligned}.$$

The matrix Y is invertible (it equals $M^{-1}A^{\frac{1}{2}}$). Hence, $Y^T\left(M + M^T - A\right)Y$ is s.p.d. if and only if $M + M^T - A$ is s.p.d. We need to investigate when $\|\mathcal{E}\| = \max_{\mathbf{w}} \frac{\|\mathcal{E}\mathbf{w}\|}{\|\mathbf{w}\|} < 1$, that is, when for any non–zero $\mathbf{w}$,

$$\|\mathcal{E}\mathbf{w}\|^2 = \mathbf{w}^T\mathcal{E}^T\mathcal{E}\mathbf{w} = \mathbf{w}^T\mathbf{w} - \mathbf{w}^TY^T(M + M^T - A)Y\mathbf{w} < \mathbf{w}^T\mathbf{w}.$$

Equivalently, we need to establish when for any non–zero vector $\mathbf{z} = Y\mathbf{w}$,

$$\mathbf{z}^T(M + M^T - A)\mathbf{z} > 0.$$

Thus, we showed that to have $\|\mathcal{E}\| < 1$ it is equivalent to have $Y^T\left(M + M^T - A\right)Y$ and hence $M + M^T - A$ s.p.d. In conclusion, we proved the following main result.

THEOREM 3.1. *A necessary and sufficient condition for the iteration process* (1.11) *to be A–convergent, i.e. convergent in A–norm, is*

$$M + M^T - A$$

to be s.p.d.

Applying this result to the forward Gauss–Seidel iteration matrix $M = D + L$, we find

$$M + M^T - A = D + L + D + L^T - (D + L + L^T) = D,$$

which is s.p.d. Thus, we have the following result.

COROLLARY 3.1. *The (forward or backward) Gauss–Seidel iteration method is convergent in energy norm.*

In some cases it is useful to have iteration process with s.p.d. preconditioner. If M is not symmetric, we can run the following composite iteration using both M and M^T. Given $\mathbf{x}_0$, for $k \geq 0$ compute

$$\begin{aligned}M(\mathbf{x}_{k+\frac{1}{2}} - \mathbf{x}_k) &= \mathbf{b} - A\mathbf{x}_k, \\ M^T(\mathbf{x}_{k+1} - \mathbf{x}_{k+\frac{1}{2}}) &= \mathbf{b} - A\mathbf{x}_{k+\frac{1}{2}},\end{aligned}$$

To implement this composite method we need to solve systems with both M and M^T. If $M + M^T - A$ is s.p.d. it is easy to see that the composite iteration is also convergent. We have

$$\overline{M}(\mathbf{x}_{k+1} - \mathbf{x}_k) = \mathbf{b} - A\mathbf{x}_k,$$

where

$$\overline{M} = M(M + M^T - A)^{-1}M^T.$$

Indeed, $\mathbf{x}_{k+\frac{1}{2}} = \mathbf{x}_k + M^{-1}\mathbf{r}_k$ and

$$\mathbf{x}_{k+1} = \mathbf{x}_{k+\frac{1}{2}} + M^{-T}\mathbf{r}_{k+\frac{1}{2}} = \mathbf{x}_k + M^{-1}\mathbf{r}_k + M^{-T}(\mathbf{b} - A\mathbf{x}_k - AM^{-1}\mathbf{r}_k).$$

Hence,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \left(M^{-1} + M^{-T} - M^{-T}AM^{-1}\right)\mathbf{r}_k = \mathbf{x}_k + M^{-T}\left(M + M^T - A\right)M^{-1}\mathbf{r}_k.$$

Therefore, we showed that the composite iteration with M and M^T reduces to a standard iteration with the symmetric preconditioner $\overline{M}$,

$$\overline{M}\,(\mathbf{x}_{k+1} - \mathbf{x}_k) = \mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k.$$

We have,

$$\begin{aligned} I - \overline{M}^{-1}A &= I - \left(M^{-1} + M^{-T} - M^{-T}AM^{-1}\right)A \\ &= (I - M^{-T}A)(I - M^{-1}A). \end{aligned}$$

Hence, $A^{\frac{1}{2}}(I - \overline{M}^{-1}A)A^{-\frac{1}{2}} = \mathcal{E}^T\mathcal{E}$. Thus, the composite iteration is A–convergent if and only if the original iteration with M is A–convergent.

Convergence factor. The norm of the iteration matrix is called iteration (or convergence) factor. We proved above that the convergence factor of the composite iteration with M and M^T is the square of the convergence factor of the iteration method with M (and M^T).

CHAPTER 3

Stationary iterative methods as smoothers and the TG method

This lecture introduces some facts about matrix inequalities, convergence of stationary preconditioned methods and comparison between two preconditioners. It also gives an illustration of the smoothing property of iteration methods such as Gauss–Seidel that leads to the multigrid idea to continue the iteration on a coarser version of the problem. The lecture ends up with a formal definition of a two–grid iteration method.

1. Matrix norms

Unless otherwise specified, we use the standard Euclidean vector-norm $\|\mathbf{v}\| = \sqrt{\mathbf{v}^T\mathbf{v}}$.

DEFINITION 1.1 (Symmetric definition of matrix norm).
For any $n \times m$ (rectangular) matrix B, the symmetric expression

$$\max_{\mathbf{v}\in\mathbb{R}^n,\, \mathbf{w}\in\mathbb{R}^m} \frac{\mathbf{w}^T B\mathbf{v}}{\|\mathbf{v}\|\|\mathbf{w}\|},$$

defines a matrix norm $\|B\|$.

From the identities,

$$\max_{\mathbf{v}\in\mathbb{R}^n,\, \mathbf{w}\in\mathbb{R}^m} \frac{\mathbf{w}^T B\mathbf{v}}{\|\mathbf{v}\|\|\mathbf{w}\|} = \max_{\mathbf{v}\in\mathbb{R}^n} \frac{1}{\|\mathbf{v}\|} \left(\max_{\mathbf{w}\in\mathbb{R}^m} \frac{\mathbf{w}^T B\mathbf{v}}{\|\mathbf{w}\|} \right) = \max_{\mathbf{v}\in\mathbb{R}^n} \frac{1}{\|\mathbf{v}\|} \|B\mathbf{v}\| = \|B\|,$$

we conclude that Definition 1.1 is equivalent to the more traditional one

$$\|B\| = \max_{\mathbf{v}\in\mathbb{R}^n} \frac{\|B\mathbf{v}\|}{\|\mathbf{v}\|}.$$

Since $\mathbf{w}^T B\mathbf{v} = \mathbf{v}^T B^T\mathbf{w}$, from Definition 1.1 it immediately follows that

$$\|B\| = \|B^T\|. \tag{1.13}$$

2. Inequalities between s.p.d. matrices

We will very often use the following result.

PROPOSITION 2.1. *Let A and B be two s.p.d. matrices. Then the inequality*

$$\mathbf{v}^T A\mathbf{v} \le \mathbf{v}^T B\mathbf{v} \text{ for all } \mathbf{v},$$

implies that

$$\mathbf{v}^T B^{-1}\mathbf{v} \le \mathbf{v}^T A^{-1}\mathbf{v} \text{ for all } \mathbf{v}.$$

.

PROOF. Since A and B are s.p.d. then the s.p.d. square root of A and B is well-defined. The given inequality used for $\mathbf{v} := B^{-\frac{1}{2}}\mathbf{v}$ implies

$$\mathbf{v}^T B^{-\frac{1}{2}} A B^{-\frac{1}{2}} \mathbf{v} \leq \mathbf{v}^T \mathbf{v} \text{ for all } \mathbf{v}.$$

That is, for $X = A^{\frac{1}{2}} B^{-\frac{1}{2}}$ we have $\mathbf{v}^T X^T X \mathbf{v} \leq \mathbf{v}^T \mathbf{v}$, or equivalently $\|X\| \leq 1$. Since $\|X\| = \|X^T\|$, we also have

$$\mathbf{v}^T \mathbf{v} \geq \mathbf{v}^T X X^T \mathbf{v} = \mathbf{v}^T A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}} \mathbf{v}.$$

Using this inequality for $\mathbf{v} := A^{-\frac{1}{2}}\mathbf{v}$ the desired result follows. □

Some conditions for spectral equivalence. In what follows we will need the following result.

LEMMA 2.1. *Let M and the s.p.d. matrix D satisfy the estimates*

$$\mathbf{v}^T (M + M^T - A)\mathbf{v} \geq \delta_0 \, \mathbf{v}^T D \mathbf{v} \text{ for all } \mathbf{v}, \tag{1.14}$$

and

$$\mathbf{w}^T M \mathbf{v} \leq \delta_1 \sqrt{\mathbf{w}^T D \mathbf{w}} \sqrt{\mathbf{v}^T D \mathbf{v}} \text{ for all } \mathbf{v}, \ \mathbf{w}. \tag{1.15}$$

Then, for $\overline{M} = M \left(M + M^T - A\right)^{-1} M^T$, we have

$$\frac{\delta_0}{4} \, \mathbf{v}^T D \mathbf{v} \leq \mathbf{v}^T \overline{M} \mathbf{v} \leq \frac{\delta_1^2}{\delta_0} \, \mathbf{v}^T D \mathbf{v}.$$

PROOF. Consider $X = D^{-\frac{1}{2}} M D^{-\frac{1}{2}}$. Condition (1.14) implies the following coercivity of X,

$$2\mathbf{v}^T X \mathbf{v} = \mathbf{v}^T (X + X^T)\mathbf{v} \geq \mathbf{v}^T (X + X^T - D^{-\frac{1}{2}} A D^{-\frac{1}{2}})\mathbf{v} \geq \delta_0 \mathbf{v}^T \mathbf{v}.$$

That is, $\mathbf{v}^T X \mathbf{v} \geq \frac{\delta_0}{2} \mathbf{v}^T \mathbf{v}$. Using this inequality for $\mathbf{v} := X^{-1}\mathbf{v}$ we obtain $\|X^{-1}\mathbf{v}\|^2 \leq \frac{2}{\delta_0} \, \mathbf{v}^T X^{-T} \mathbf{v} = \frac{2}{\delta_0} \, \mathbf{v}^T X^{-1} \mathbf{v} \leq \frac{2}{\delta_0} \, \|\mathbf{v}\| \|X^{-1}\mathbf{v}\|$. That is, we showed that $\|X^{-1}\mathbf{v}\| \leq \frac{2}{\delta_0} \, \|\mathbf{v}\|$ or equivalently,

$$\|X^{-1}\| \leq \frac{2}{\delta_0}. \tag{1.16}$$

Estimate (1.15) on the other hand is equivalent to $\|X\| \leq \delta_1$. Thus, as an intermediate result we showed that the symmetrically scaled matrix M (that is, X) is well-conditioned ($\|X\| \|X^{-1}\| \leq \frac{2\delta_1}{\delta_0}$).

To bound $\overline{M}$ form above in terms of D we proceed as follows. Estimate (1.14) implies

$$\mathbf{w}^T (M + M^T - A)^{-1} \mathbf{w} \leq \frac{1}{\delta_0} \, \mathbf{w}^T D^{-1} \mathbf{w} \text{ for all } \mathbf{w}.$$

Hence

$$\mathbf{v}^T \overline{M} \mathbf{v} \leq \frac{1}{\delta_0} \, \mathbf{v}^T M D^{-1} M^T \mathbf{v} = \frac{1}{\delta_0} \, (D^{\frac{1}{2}}\mathbf{v})^T X X^T (D^{\frac{1}{2}}\mathbf{v}) \leq \frac{1}{\delta_0} \, \|X^T\|^2 \, \mathbf{v}^T D \mathbf{v}.$$

From (1.13) we have $\|X^T\| = \|X\| \leq \delta_1$, hence the upper bound

$$\mathbf{v}^T \overline{M} \mathbf{v} \leq \frac{\delta_1^2}{\delta_0} \, \mathbf{v}^T D \mathbf{v},$$

follows. For the estimate from below, we obtain

$$\begin{aligned}\mathbf{v}^T D^{\frac{1}{2}} \overline{M}^{-1} D^{\frac{1}{2}} \mathbf{v} &= \mathbf{v}^T D^{\frac{1}{2}} M^{-T} (M + M^T - A) M^{-1} D^{\frac{1}{2}} \mathbf{v} \\ &= \mathbf{v}^T D^{\frac{1}{2}} (M^{-T} + M^{-1} - M^{-T} A M^{-1}) D^{\frac{1}{2}} \mathbf{v} \\ &\leq \mathbf{v}^T (X^{-T} + X^{-1}) \mathbf{v} \\ &= 2\ \mathbf{v}^T X^{-1} \mathbf{v} \\ &\leq 2\ \|X^{-1}\| \|\mathbf{v}\|^2.\end{aligned}$$

Using estimate (1.16), we obtain $\mathbf{v}^T D^{\frac{1}{2}} \overline{M}^{-1} D^{\frac{1}{2}} \mathbf{v} \leq \frac{4}{\delta_0}\ \|\mathbf{v}\|^2$, or equivalently $\mathbf{v}^T \overline{M}^{-1} \mathbf{v} \leq \frac{4}{\delta_0}\ \mathbf{v}^T D^{-1} \mathbf{v}$. The latter estimate, based on Proposition 2.1, implies the desired lower bound

$$\frac{\delta_0}{4}\ \mathbf{v}^T D \mathbf{v} \leq \mathbf{v}^T \overline{M} \mathbf{v}.$$

□

3. Convergence of classical (relaxation) iterative methods

We showed that a necessary and sufficient condition for the A-convergence of the stationary preconditioned iterative method

$$M(\mathbf{x}_k - \mathbf{x}_{k-1}) = \mathbf{b} - A\mathbf{x}_{k-1},\ k = 1, 2,\ \ldots,$$

for solving $A\mathbf{x} = \mathbf{b}$ and any given initial iterate $\mathbf{x}_0$, is the positive definiteness of the matrix $M + M^T - A$. This result implies that the forward Gauss–Seidel matrix $M = D + L$ coming from A decomposed as $D + L + L^T$ provides an A–convergent iteration. A simpler method is the scaled Jacobi iteration matrix $M = \omega\ D$. We also showed that

$$\kappa\ \mathbf{v}^T D \mathbf{v} \geq \mathbf{v}^T A \mathbf{v},$$

where $\kappa = 3$ for triangular elements (and linear basis functions). Hence, if we choose $\omega > \frac{\kappa}{2}$, then we ensure that $M + M^T - A$ is s.p.d. (for $M = \omega D$).

It turns out that Gauss–Seidel method is not much faster than the weighted Jacobi method asymptotically with respect to $h \mapsto 0$. In the case of D being the diagonal of A and $M = D + L$ the forward Gauss–Seidel, we can apply Lemma 2.1 with $\delta_0 = 1$ and $\delta_1 \leq \kappa$ (with $\kappa = 3$ for linear triangular elements). This shows that the spectral relations between A and $\overline{M}$ and between A and D are of the same quality with respect to (or w.r.t.) $h \mapsto 0$. That is, we have

$$\frac{\mathbf{v}^T A \mathbf{v}}{\mathbf{v}^T \overline{M} \mathbf{v}} \simeq \frac{\mathbf{v}^T A \mathbf{v}}{\mathbf{v}^T D \mathbf{v}},$$

if

$$\frac{\mathbf{v}^T \overline{M} \mathbf{v}}{\mathbf{v}^T D \mathbf{v}} = \mathcal{O}(1) \text{ for } h \mapsto 0.$$

Recall that we proved the estimates

$$\gamma_0\ h^2 \leq \frac{\mathbf{v}^T A \mathbf{v}}{\mathbf{v}^T D \mathbf{v}} \leq \kappa.$$

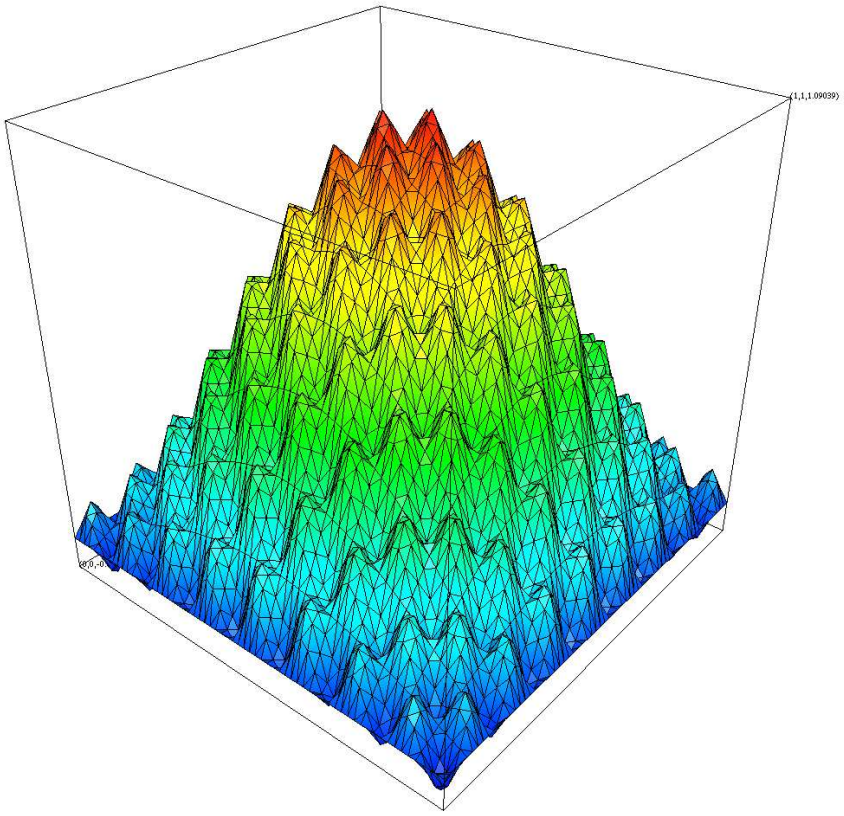

FIGURE 1. *Initial non–smooth function*

Smoothing property of classical iteration methods. It is clear then that the iteration matrix $I-(\omega D)^{-1}A$ will reduce different components of the error differently. More precisely, error components spanned by eigenvector of $D^{-1}A$ corresponding to eigenvalues close to the upper part of the spectrum, i.e., eigenvectors corresponding to eigenvalues that are $\mathcal{O}(1)$ will be reduced with factors uniformly less than one (for $h \mapsto 0$), whereas components of the error that are spanned by eigenvectors corresponding to the lower part of the spectrum, i.e., eigenvalues that are of order $\mathcal{O}(h^2)$, will hardly change. A distinct feature of the finite element stiffness matrices A (scaled symmetrically with their diagonal D) coming from b.v.p. is that their eigenvectors corresponding to the lower part of the spectrum are geometrically smooth and global. Thus the classical iterative methods like weighted Jacobi or Gauss–Seidel damp the geometrically oscillatory components of the error very efficiently. This phenomenon is referred to in the literature as *smoothing*.

To illustrate the smoothing process, we start with a $\mathbf{e}_0$ chosen to be a linear combination of a smooth and a oscillatory component, and then run successively, one, two and three symmetric Gauss–Seidel iterations applied to $A\mathbf{e} = 0$. That is, we compute the iterates $\mathbf{e}_k = (I - \overline{M}^{-1}A)\mathbf{e}_{k-1}$ for $k = 1, 2,\ 3$. The resulting smoothing phenomenon is illustrated in Figs. 1, 2, 3 and 4.

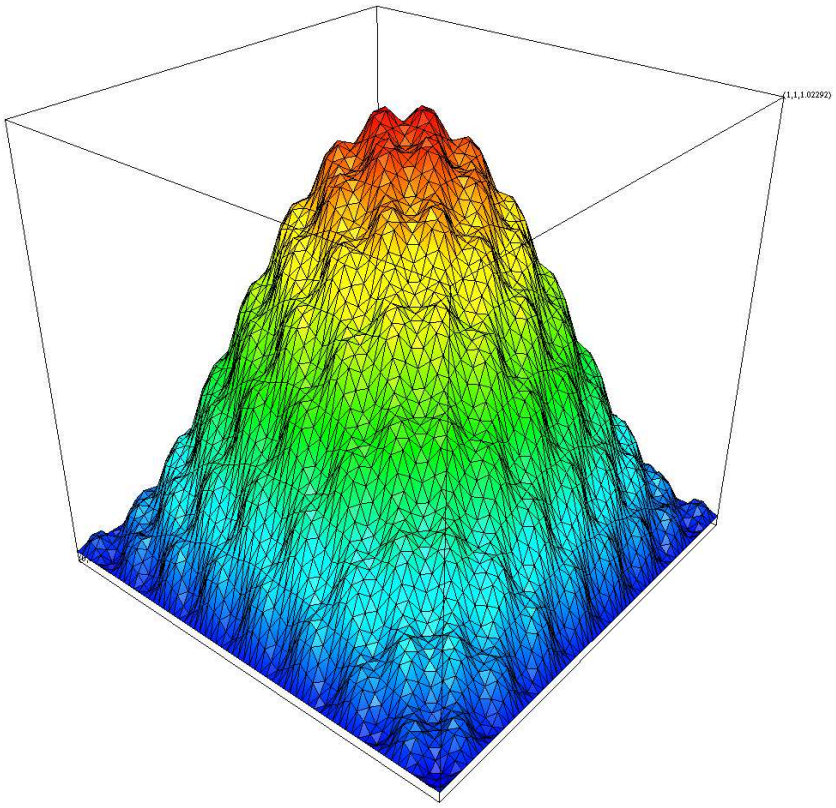

FIGURE 2. *Result after one step of symmetric Gauss–Seidel smoothing*

4. Coarse–grid approximation

Thus a natural idea is after one or few smoothing iterations to approximate the resulting problem on a coarse grid and continue the iteration with a coarse version of the problem. This was the breakthrough observation in the original paper by Fedorenko [**Fe64**], later extended and popularized by Achi Brandt [**AB77**], Hackbusch and others.

The fact that smooth functions can accurately be represented on coarse grids is inherent to any approximation method, in particular, it is inherent to the f.e. method. The latter is illustrated in Figs. 5, 6, 7, and 8.

We summarize the following basic finite element error estimate result (cf., for example, Ciarlet [**Ci02**], Brenner and Scott [**BS02**], Braess [**Br01**])

Since $a(u-u_h,\ \varphi)=0$ for all $\varphi\in V_h$ (a main property of the Galerkin method), we also have the following estimate,

$$|\nabla(u-u_h)|^2 = a(u-u_h,\ u-u_h) = a(u-u_h,\ u-\varphi) \le |\nabla(u-u_h)||\nabla(u-\varphi)|.$$

That is,

$$|\nabla(u-u_h)| = \inf_{\varphi\in V_h} |\nabla(u-\varphi)|.$$

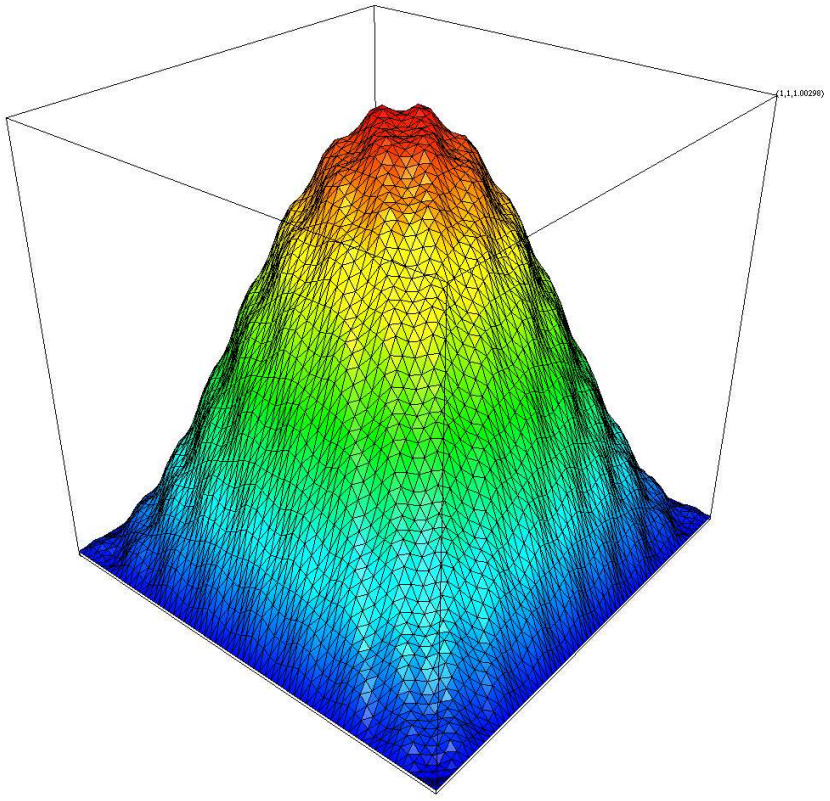

FIGURE 3. *Result after two steps of symmetric Gauss–Seidel smoothing*

Assuming now that u has two derivatives in $L_2(\Omega)$, we immediately get the first order error estimate

$$|\nabla(u - u_h)| \leq Ch\|u\|_2.$$

To be more precise, we first form a nodal interpolant $I_h u = \sum_i u(x_i)\varphi_i$ and then on every triangle τ the following estimate holds

$$\|\nabla(u - I_h u)\|^2 = \sum_{\tau \in \mathcal{T}_h} \int_\tau |\nabla(u - I_h u)|^2 \, dx \leq \sum_{\tau \in \mathcal{T}_h} C_\tau \, h^2 \|u\|^2_{2,\tau} \leq Ch^2\|u\|_2^2.$$

Here, we use the Taylor expansion on every triangle τ and the fact that the triangles are geometrically similar to a fixed number of an initial set of triangles. Hence, C_τ will take be a fixed number of mesh–independent values. The latter estimate shows that for smooth functions u (for example having two derivatives) the finite element approximations on grids $\mathcal{T}_H$ will give approximations u_H such that the error $u - u_H$ in energy norm behaves like $H \, \|u\|_2$. There is one problem with the above argument if we start with a f.e. function u_h and want to measure $u_h - u_H$ in energy norm. This is not immediately possible since the finite element function u_h does not have two derivatives. To overcome this difficulty, we can measure the error in L_2.

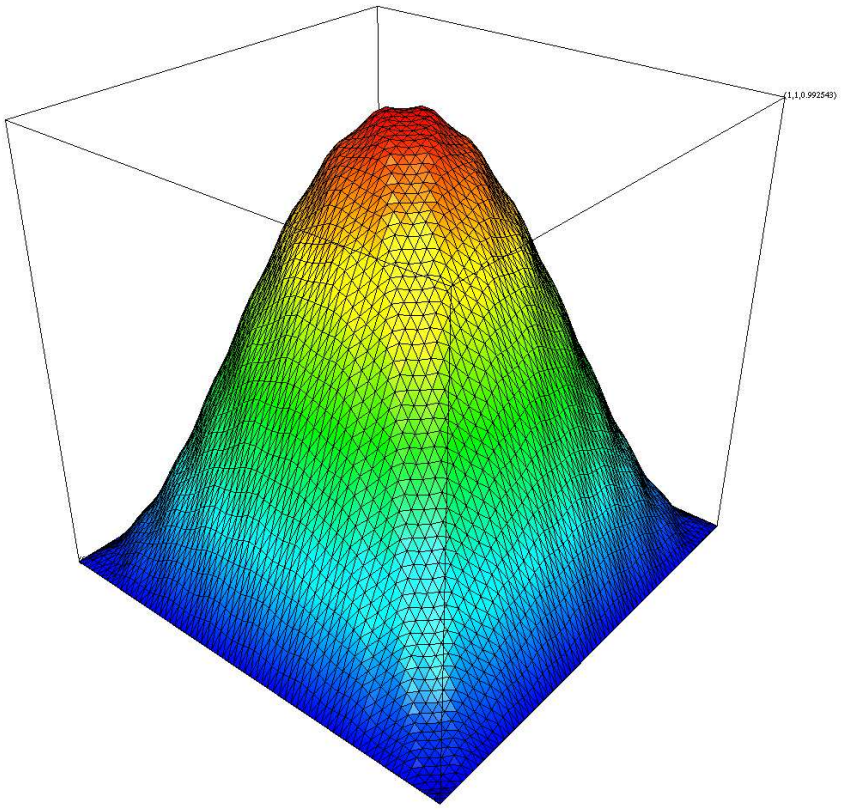

FIGURE 4. *Result after three steps of symmetric Gauss–Seidel smoothing*

L_2–error estimates; Aubin–Nitsche's argument. Consider two finite element spaces V_H and V_h where V_h corresponds to a triangulation $\mathcal{T}_h$ obtained by possibly several steps of refinement from a coarser triangulation $\mathcal{T}_H$. This implies that

$$V_H \subset V_h.$$

Let $u_h \in V_h$ and $u_H \in V_H$ be the Galerkin projection (approximation) of u_h from V_H. This means that

$$a(u_h - u_H,\ \varphi_H) = 0 \text{ for all } \varphi_H \in V_H.$$

Consider the error $e = u_h - u_H \in V_h \subset L_2(\Omega)$ and solve the b.v.p. for the Laplace equation

$$-\Delta w = e(\mathbf{x}) \text{ for } \mathbf{x} \in \Omega,$$

with $w = 0$ on $\partial\Omega$. For convex polygonal domains $\Omega \subset \mathbb{R}^d$, $d = 2,\ 3$, the following regularity result is known

$$\|w\|_2 \leq C\ \|e\|_0.$$

That is, w has derivatives up to second order all in $L_2(\Omega)$ and the above a priori estimate holds. By construction, for the bilinear form $a(.,.)$ coming from the Laplace operator, since $0 = a(e,\ w_H) = a(w_H,\ e)$ for any $w_H \in V_H$, we have

$$\|e\|_0^2 = a(w,\ e) = a(w - w_H,\ e) \leq |w - w_H|_1\ |e|_1 \leq CH\|w\|_2|e|_1 \leq CH\|e\|_0|e|_1.$$

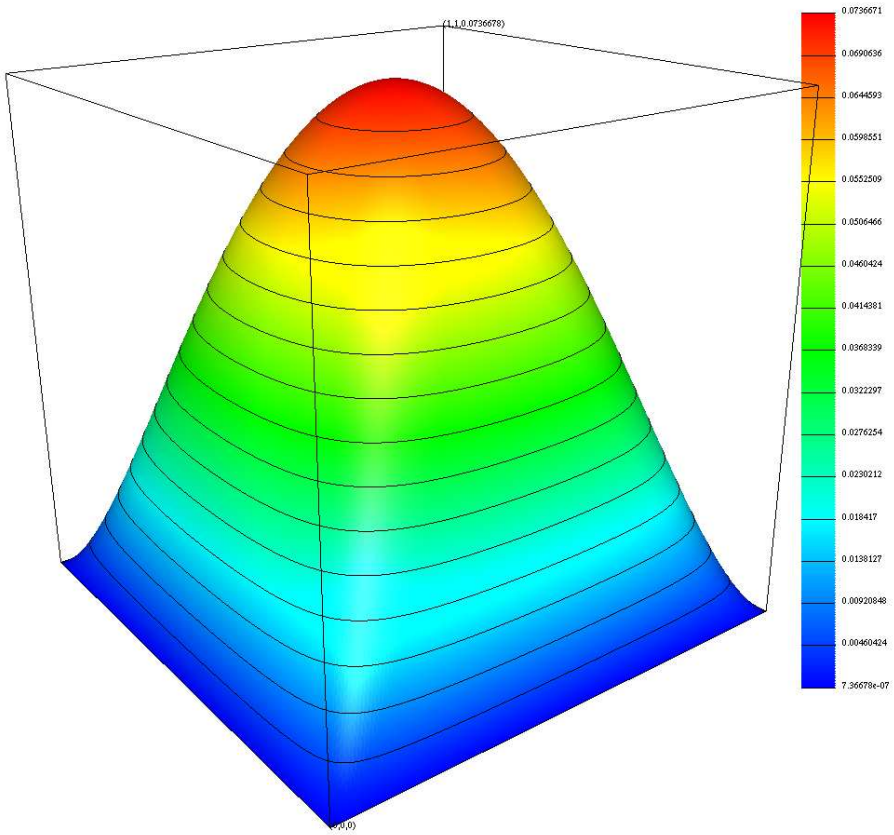

FIGURE 5. *Solution to* $-\Delta u = 1$

In conclusion,

$$\|u_h - u_H\|_0 \leq CH\ \sqrt{a(u_h - u_H,\ u_h - u_H)} \leq CH\ \sqrt{a(u_h,\ u_h)}.$$

Finite element refinement and the interpolation matrix. Consider now two nested finite element spaces $V_H \subset V_h$. Let $V_H = \text{Span}\ (\varphi_{i_c}^{(H)})_{i_c=1}^{n_c}$ and $V_h = \text{Span}\ (\varphi_i^{(h)})_{i=1}^{n}$ with their respective nodal (Lagrangian) bases. Since each $\varphi_{i_c}^{(H)} \in V_H \subset V_h$ we have the expansion

$$\varphi_{i_c}^{(H)} = \sum_{i=1}^{n} \varphi_{i_c}^{(H)}(x_i)\ \varphi_i^{(h)}.$$

Consider the coefficient vector $\boldsymbol{\varphi}_{i_c} = (\varphi_{i_c}^{(H)}(x_i))_{i=1}^{n}$. The matrix $P = (\boldsymbol{\varphi}_{i_c})_{i_c=1}^{n_c}$ is referred to as the *interpolation matrix*. It relates the coefficient vector $\mathbf{v}_c \in \mathbb{R}^{n_c}$ of any function $v_c \in V_H$ expanded in terms of the coarse basis $\{\varphi_{i_c}^{(H)}\}$ to the coefficient vector $P\mathbf{v}_c$ of $v_c \in V_h$ expanded in terms of the fine–grid basis $\{\varphi_i^{(h)}\}$. Since the finite element bases are local, we see that the $n \times n_c$ rectangular matrix P is sparse. The number of non–zero entries of P per column depends on the support of each $\varphi_{i_c}^{(H)}$, namely, depends on the number of fine-grid basis functions $\varphi_i^{(h)}$ that intersect that support. That is, the sparsity pattern of P is controlled by the topology of the triangulations $\mathcal{T}_H$ and $\mathcal{T}_h$.

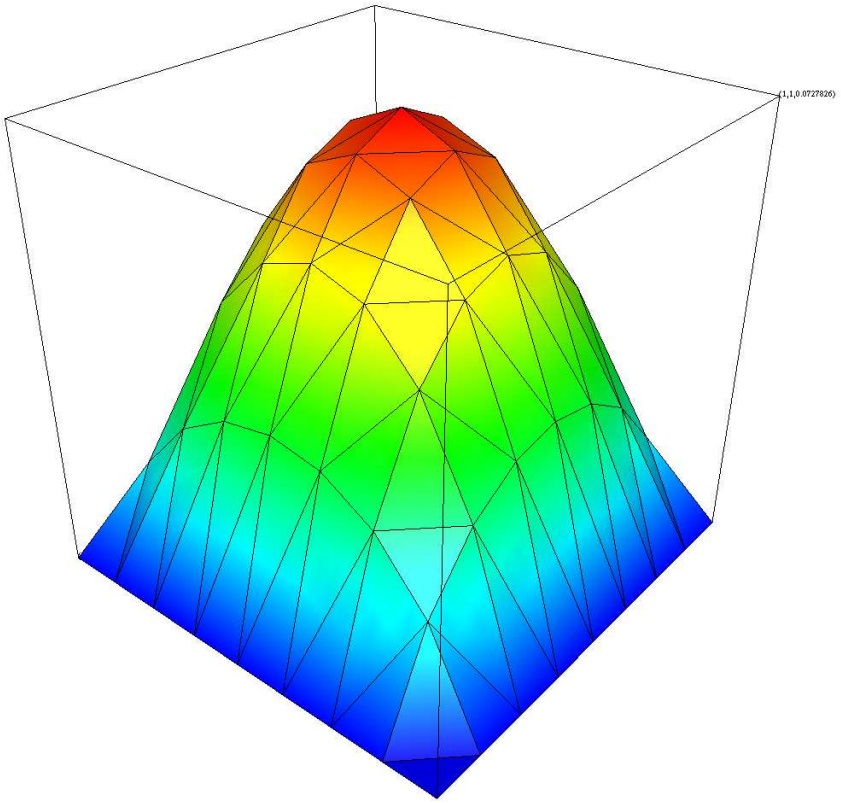

FIGURE 6. *Finite element approximate solution to* $-\Delta u = 1$ *on a coarse mesh*

Matrix–vector form of the L_2-approximation of the Galerkin projection

For a given $u_h \in V_h$ and a subspace V_H of V_h, by definition the Galerkin projection $u_H \in V_H$ of u_h satisfies

$$a(u_h - u_H,\ v_H) = 0 \text{ for all } v_H \in V_H.$$

That is, $u_H \in V_H$ solves

$$a(u_H,\ v_H) = a(u_h,\ v_H) \text{ for all } v_H \in V_H.$$

In terms of coefficient vectors $\mathbf{u}_c$, $\mathbf{u}$, $\mathbf{v}_c$ and $P\mathbf{v}_c$ of u_H, u_h, v_H and v_H as an element of V_h, we have

$$\mathbf{v}_c^T A_c \mathbf{u}_c = a(u_H,\ v_H) = a(u_h,\ v_H) = (P\mathbf{v}_c)^T A\mathbf{u} = \mathbf{v}_c^T P^T A\mathbf{u}.$$

That is, $A_c\mathbf{u}_c = P^T A\mathbf{u}$, or $\mathbf{u}_c = A_c^{-1}P^T A\mathbf{u}$. Hence the coefficient vector of u_H as an element of V_h equals

$$P\mathbf{u}_c = PA_c^{-1}P^T A\mathbf{u}.$$

Another important property is the variational (Galerkin) relation between the coarse matrix A_c and A

$$A_c = P^T AP. \tag{1.17}$$

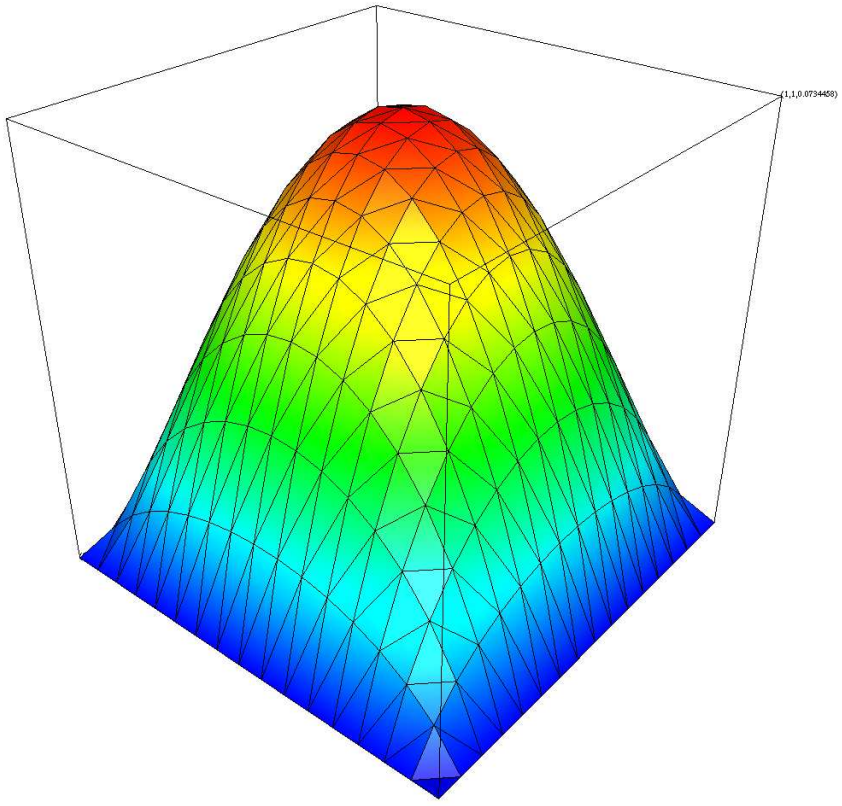

FIGURE 7. *Finite element approximate solution to* $-\Delta u = 1$ *on a refined mesh*

This is seen by construction, since the $(i_c,\ j_c)$ entry of A_c equals $a(\varphi_{j_c}^{(H)},\ \varphi_{i_c}^{(H)}) = \boldsymbol{\varphi}_{i_c}{}^T A \boldsymbol{\varphi}_{j_c} = (P^T A P)_{i_c,\, j_c}$.

In conclusion, we have the following result.

PROPOSITION 4.1. *The Galerkin projection* $u_H \in V_H$ *of* $u_h \in V_h$*, i.e. the coarse finite element function* u_H *that solves*

$$a(u_h - u_H,\ v_H) = 0 \text{ for all } v_H \in V_H,$$

has a fine–grid coefficient vector $\pi_A \mathbf{u} \equiv P A_c^{-1} P^T A \mathbf{u}$ *with* $A_c = P^T A P$.

It is easily checked that $\pi_A^2 = \pi_A$. We have

$$\pi_A^2 = P A_c^{-1} (P^T A P) A_c^{-1} P^T A = P A_c^{-1} A_c A_c^{-1} P^T A = P A_c^{-1} P^T A = \pi_A. \tag{1.18}$$

5. The two–grid algorithm: definition

We conclude this lecture with the

ALGORITHM 5.1 (Two-grid (or TG) algorithm).
Consider the system of equations

$$A\mathbf{x} = \mathbf{b},$$

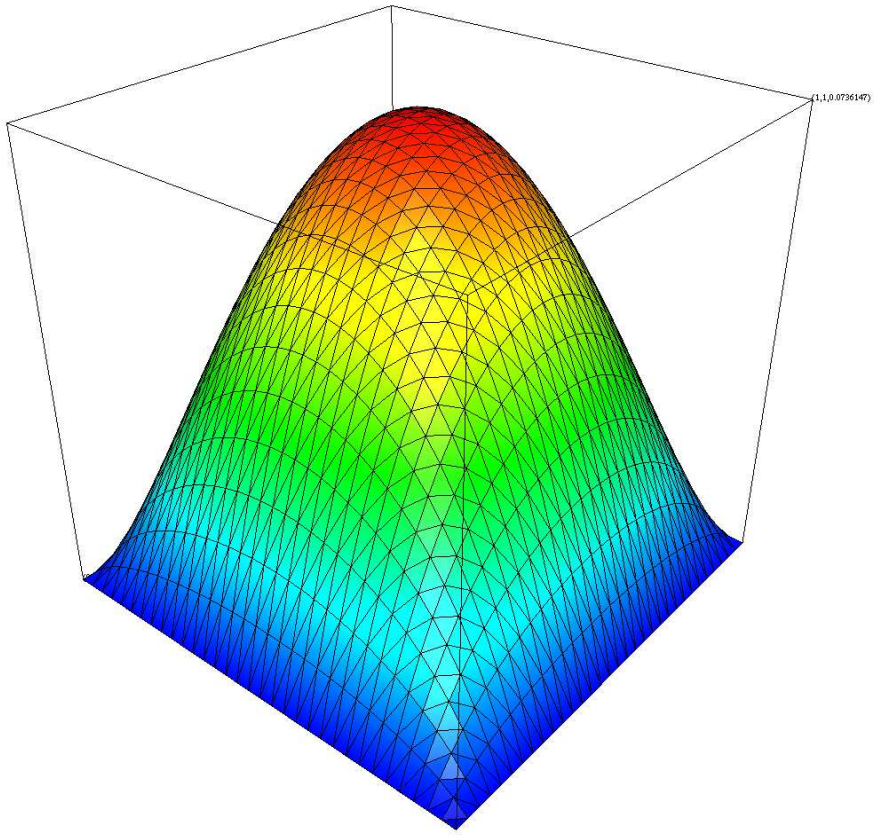

FIGURE 8. *Finite element approximate solution to* $-\Delta u = 1$ *on a more refined mesh*

and let M *be a given smoother,* P *an interpolation matrix, and* $A_c = P^T AP$ *the respective coarse matrix. The (symmetrized) two–grid iteration method computes for any given initial iterate* $\mathbf{x}_0$ *a two–grid iterate* $\mathbf{x}_{TG}$ *in the following steps:*

- *"pre–smoothing step":*
  Compute $\mathbf{y}$ *from*

$$M(\mathbf{y} - \mathbf{x}_0) = \mathbf{b} - A\mathbf{x}_0.$$

- *"coarse–grid correction:*
  Compute $\mathbf{x}_c$ *from*

$$A_c\mathbf{x}_c = P^T(\mathbf{b} - A\mathbf{y}).$$

  The next intermediate iterate is $\mathbf{z} = \mathbf{y} + P\mathbf{x}_c$.
- *"post–smoothing" step:*
  Compute $\mathbf{x}_{TG}$ *from*

$$M^T(\mathbf{x}_{TG} - \mathbf{z}) = \mathbf{b} - A\mathbf{z}.$$

In summary, the TG algorithm involves solutions with M, M^T and A_c, matrix–vector multiplications with sparse matrices A, P^T and P: with A to compute residuals, with P^T to restrict the fine–grid residual and with P to interpolate the coarse–grid correction.

CHAPTER 4

Two-by-two block matrices

This lecture provides some basic facts for two–by-two block matrices, their Schur complements. It also analyzes angles between spaces by introducing an abstract lemma of Kato.

1. Two-by-two block matrices

Let A be a s.p.d. matrix partitioned into a two-by-two blocks $(A_{ij})_{i,j=1}^2$ with square blocks A_{ii}, $i = 1, 2$, which hence, as is easily seen, are s.p.d. as well. Then the following factorization holds

$$A = \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix}.$$

The block $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ is called Schur complement. From the representation $A = L\mathrm{diag}(A_{11},\ S)L^T$ with L being invertible (L is unit triangular), it is clear that S is also s.p.d. The following identity is seen for any vector $\mathbf{v} = (\mathbf{v}_i)_{i=1}^2$,

$$\mathbf{v}^T A\mathbf{v} = (A_{11}\mathbf{v}_1 + A_{12}\mathbf{v}_2)^T A_{11}^{-1} (A_{11}\mathbf{v}_1 + A_{12}\mathbf{v}_2) + \mathbf{v}_2^T S\mathbf{v}_2,$$

which shows the following minimization property of S

$$\mathbf{v}_2^T S\mathbf{v}_2 = \min_{\mathbf{v}_1}\ \mathbf{v}^T A\mathbf{v}.$$

The above minimum is attained for $\mathbf{v}$ in the subspace

$$A_{11}\mathbf{v}_1 + A_{12}\mathbf{v}_2 = 0.$$

Such vector $\mathbf{v}$ is called "minimal energy" extensions of $\mathbf{v}_2$ and it also satisfies the equation

$$A\mathbf{v} = \begin{bmatrix} 0 \\ S\mathbf{v}_2 \end{bmatrix}.$$

The latter formula offers a way to evaluate the actions of S. That is, given $\mathbf{v}_2$, we compute $\mathbf{v}_1$ from $A_{11}\mathbf{v}_1 + A_{12}\mathbf{v}_2 = 0$ and form the product $A\mathbf{v}$. Its second component gives $S\mathbf{v}_2$. Thus, without explicitly forming S its actions can be computed by solving systems with A_{11}. Note that S is in general a dense matrix (even if A is sparse).

S is better conditioned than A. We have the following inequalities valid for the extreme eigenvalues of A and S (which are real and positive):

$$\lambda_{\min}(A) \le \lambda_{\min}(S) \le \lambda_{\max}(S) \le \lambda_{\max}(A). \tag{1.19}$$

From the minimization property of S, we have for any $\mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}$, (using also the trivial inequality $\mathbf{v}_2^T\mathbf{v}_2 \leq \mathbf{v}_1^T\mathbf{v}_1 + \mathbf{v}_2^T\mathbf{v}_2 = \mathbf{v}^T\mathbf{v}$)

$$\frac{\mathbf{v}_2^T S\mathbf{v}_2}{\mathbf{v}_2^T\mathbf{v}_2} = \frac{1}{\mathbf{v}_2^T\mathbf{v}_2} \min_{\mathbf{v}_1} \mathbf{v}^T A\mathbf{v} \geq \min_{\mathbf{v}_1} \frac{\mathbf{v}^T A\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \geq \min_{\mathbf{v}} \frac{\mathbf{v}^T A\mathbf{v}}{\mathbf{v}^T\mathbf{v}} = \lambda_{\min}(A).$$

Hence, $\lambda_{\min}(S) \geq \lambda_{\min}(A)$.

We also have $\mathbf{v}_2^T S\mathbf{v}_2 \leq \mathbf{v}_2^T A_{22}\mathbf{v}_2$, hence

$$\frac{\mathbf{v}_2^T S\mathbf{v}_2}{\mathbf{v}_2^T\mathbf{v}_2} \leq \frac{\mathbf{v}_2^T A_{22}\mathbf{v}_2}{\mathbf{v}_2^T\mathbf{v}_2} = \frac{\begin{bmatrix} 0 \\ \mathbf{v}_2 \end{bmatrix}^T A \begin{bmatrix} 0 \\ \mathbf{v}_2 \end{bmatrix}}{\begin{bmatrix} 0 \\ \mathbf{v}_2 \end{bmatrix}^T \begin{bmatrix} 0 \\ \mathbf{v}_2 \end{bmatrix}} \leq \max_{\mathbf{v}} \frac{\mathbf{v}^T A\mathbf{v}}{\mathbf{v}^T\mathbf{v}} = \lambda_{\max}(A).$$

This shows that $\lambda_{\max}(S) \leq \lambda_{\max}(A)$.

2. Abstract angles between vector spaces

Let A be a $n \times n$ s.p.d. matrix Let J and P be two rectangular matrices with n rows each, such that when put together they form a square invertible matrix $[J,\ P]$. Equivalently, we may say that any vector $\mathbf{v} \in \mathbb{R}^n$ allows for the unique (direct) decomposition

$$\mathbf{v} = J\mathbf{v}_f + P\mathbf{v}_c.$$

Then, the inner product $\mathbf{v}^T A\mathbf{v}$ admits the form

$$\mathbf{v}^T A\mathbf{v} = \overline{\mathbf{v}}^T \overline{A}\overline{\mathbf{v}},$$

where $\overline{\mathbf{v}} = \begin{bmatrix} \mathbf{v}_f \\ \mathbf{v}_c \end{bmatrix}$ and

$$\overline{A} = \begin{bmatrix} \overline{A}_{11} & \overline{A}_{12} \\ \overline{A}_{21} & \overline{A}_{22} \end{bmatrix} = \begin{bmatrix} J^T AJ & J^T AP \\ P^T AJ & P^T AP \end{bmatrix}.$$

A trivial example of J and P is

$$J = \begin{bmatrix} I \\ 0 \end{bmatrix} \text{ and } P = \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

A more interesting example is the so-called "hierarchical" one:

$$J = \begin{bmatrix} I \\ 0 \end{bmatrix} \text{ and } P = \begin{bmatrix} W \\ I \end{bmatrix} \tag{1.20}$$

for a non-zero W.

Since the vector spaces Range(J) and Range(P) have non–trivial angle (any two vectors in this pair of spaces are linearly independent) there is a constant $\gamma \in [0,\ 1)$ (strictly less than one) such that the following strengthened Cauchy-Schwarz inequality holds:

$$\mathbf{v}_f^T J^T AP\mathbf{v}_c \leq \gamma \ \left(\mathbf{v}_f^T J^T AJ\mathbf{v}_f\right)^{\frac{1}{2}} \left(\mathbf{v}_c^T P^T AP\mathbf{v}_c\right)^{\frac{1}{2}}. \tag{1.21}$$

Then, for the Schur complement $\overline{S} = \overline{A}_{22} - \overline{A}_{21}\overline{A}_{11}^{-1}\overline{A}_{12}$ of $\overline{A}$ the following inequality holds

$$(1-\gamma^2)\ \mathbf{v}_c^T\overline{A}_{22}\mathbf{v}_c \le \mathbf{v}_c^T\overline{S}\mathbf{v}_c \le \mathbf{v}_c^T\overline{A}_{22}\mathbf{v}_c.$$

We prove this inequality using the minimization property of the Schur complement $\overline{S}$ and the strengthened Cauchy–Schwarz inequality for the blocks of $\overline{A}$. We have

$$\begin{aligned}
\mathbf{v}_c^T\overline{S}\mathbf{v}_c &= \min_{\mathbf{v}_f} \begin{bmatrix} \mathbf{v}_f \\ \mathbf{v}_c \end{bmatrix}^T \overline{A} \begin{bmatrix} \mathbf{v}_f \\ \mathbf{v}_c \end{bmatrix} \\
&= \min_{\mathbf{v}_f} \left[\mathbf{v}_f^T\overline{A}_{11}\mathbf{v}_f + 2\ \mathbf{v}_f^T\overline{A}_{12}\mathbf{v}_c + \mathbf{v}_c^T\overline{A}_{22}\mathbf{v}_c\right] \\
&\ge \min_{\mathbf{v}_f} \left[\ \mathbf{v}_f^T\overline{A}_{11}\mathbf{v}_f - 2\gamma\ \left(\mathbf{v}_f^T\overline{A}_{11}\mathbf{v}_f\right)^{\frac{1}{2}} \left(\mathbf{v}_c^T\overline{A}_{22}\mathbf{v}_c\right)^{\frac{1}{2}} + \mathbf{v}_c^T\overline{A}_{22}\mathbf{v}_c\right] \\
&= \min_{\mathbf{v}_f} \left[\left(\left(\mathbf{v}_f^T\overline{A}_{11}\mathbf{v}_f\right)^{\frac{1}{2}} - \gamma\ \left(\mathbf{v}_c^T\overline{A}_{22}\mathbf{v}_c\right)^{\frac{1}{2}}\right)^2 + (1-\gamma^2)\mathbf{v}_c^T\overline{A}_{22}\mathbf{v}_c\right] \\
&\ge (1-\gamma^2)\mathbf{v}_c^T\overline{A}_{22}\mathbf{v}_c,
\end{aligned}$$

which is the desired result.

Another result for the special case of hierarchical decomposition (1.20) is that the Schur complements S and $\overline{S}$ of A and $\overline{A}$, respectively, are the same, i.e.,

$$S = A_{22} - A_{21}A_{11}^{-1}A_{12} = \overline{A}_{22} - \overline{A}_{21}\overline{A}_{11}^{-1}\overline{A}_{12} = \overline{S}.$$

We have, with $\mathbf{v}_1 = \overline{\mathbf{v}}_1 + W\mathbf{v}_2$ and $\overline{\mathbf{v}}_2 = \mathbf{v}_2$,

$$\mathbf{v}_2^T S\mathbf{v}_2 = \min_{\mathbf{v}_1} \mathbf{v}^T A\mathbf{v} = \min_{\mathbf{v}_1=\overline{\mathbf{v}}_1+W\mathbf{v}_2} \left(J\overline{\mathbf{v}}_1 + P\overline{\mathbf{v}}_2\right)^T A \left(J\overline{\mathbf{v}}_1 + P\overline{\mathbf{v}}_2\right) = \min_{\overline{\mathbf{v}}_1}\ \overline{\mathbf{v}}^T\overline{A}\overline{\mathbf{v}} = \mathbf{v}_2^T\overline{S}\mathbf{v}_2.$$

That is, $S = \overline{S}$. This is seen from the identity for any $\mathbf{v}_2$ and $\mathbf{w}_2$,

$$0 = (\mathbf{v}_2+\mathbf{w}_2)^T(S-\overline{S})(\mathbf{v}_2+\mathbf{w}_2) = \mathbf{v}_2^T(S-\overline{S})\mathbf{v}_2 + \mathbf{w}_2^T(S-\overline{S})\mathbf{w}_2 + 2\ \mathbf{v}_2^T(S-\overline{S})\mathbf{w}_2 = 2\ \mathbf{v}_2^T(S-\overline{S})\mathbf{w}_2.$$

That is, since $\mathbf{v}_2$ and $\mathbf{w}_2$ are arbitrary, we have $S - \overline{S} = 0$.

3. Kato's lemma

Let π be a projection, i.e., $\pi^2 = \pi$ and $(..,.)$ be an inner product and $\|.\| = \sqrt{(.,.)}$ the associated norm. Kato's lemma relates the cosine of the abstract angle, γ, between the complementary spaces Range(π) and Range($I - \pi$) measured in the inner product $(.,.)$. We assume that these spaces are non–trivial, i.e., that $\pi \neq I$ and $\pi \neq 0$. The following result holds

$$\|\pi\| = \|I - \pi\| = \frac{1}{\sqrt{1-\gamma^2}}.$$

The characterization is seen as follows. For any pair of vectors $\mathbf{v}$, $\mathbf{w}$ and number $t \in \mathbb{R}$ consider the vector $\mathbf{v}_t = \pi\mathbf{v} + t(I-\pi)\mathbf{w}$. We have $\pi\mathbf{v} = \pi\mathbf{v}_t$. Hence

$$\|\pi\mathbf{v}\| = \|\pi\mathbf{v}_t\| \le \|\pi\|\|\mathbf{v}_t\| = \|\pi\|\ \left(\|\pi\mathbf{v}\|^2 + 2t(\pi\mathbf{v},\ (I-\pi)\mathbf{w}) + t^2\ \|(I-\pi)\mathbf{w}\|^2\right)^{\frac{1}{2}}.$$

Therefore, the quadratic form $Q(t) = \left(1 - \frac{1}{\|\pi\|^2}\right)\|\pi\mathbf{v}\|^2 + 2t(\pi\mathbf{v},\ (I-\pi)\mathbf{w}) + t^2\ \|(I-\pi)\mathbf{w}\|^2$ is non-negative. Hence, its discriminant must be non-positive. That is,

$$(\pi\mathbf{v},\ (I-\pi)\mathbf{w})^2 - \left(1 - \frac{1}{\|\pi\|^2}\right)\ \|(I-\pi)\mathbf{w}\|^2\|\pi\mathbf{v}\|^2 \le 0.$$

This shows that the best constant (γ) satisfies the inequality $\gamma^2 \leq 1 - \frac{1}{\|\pi\|^2}$, or equivalently

$$\|\pi\| \geq \frac{1}{\sqrt{1-\gamma^2}}.$$

The fact the we actually have equality is seen by proceeding in a reverse order. From

$$(\pi\mathbf{v},\ (I-\pi)\mathbf{w})^2 - \gamma^2\ \|(I-\pi)\mathbf{w}\|^2\|\pi\mathbf{v}\|^2 \leq 0,$$

it follows that the quadratic form $Q(t) = \gamma^2\ \|\pi\mathbf{v}\|^2 + 2t(\pi\mathbf{v},\ (I-\pi)\mathbf{w}) + t^2\ \|(I-\pi)\mathbf{w}\|^2$ is non-negative. Hence,

$$\|\pi\mathbf{v} + t(I-\pi)\mathbf{w}\|^2 \geq (1-\gamma^2)\ \|\pi\mathbf{v}\|^2.$$

Letting $t = 1$ and $\mathbf{w} = \mathbf{v}$, we get $\|\mathbf{v}\|^2 \geq (1-\gamma^2)\ \|\pi\mathbf{v}\|^2$, that is $\|\pi\mathbf{v}\| \leq \frac{1}{\sqrt{1-\gamma^2}}\ \|\mathbf{v}\|$ which shows that $\frac{1}{\sqrt{1-\gamma^2}}$ is an upper bound for $\|\pi\|$. Thus, we showed $\|\pi\| = \frac{1}{\sqrt{1-\gamma^2}}$. Using the same arguments (replacing π with $I-\pi$), we show that $\|I-\pi\| = \frac{1}{\sqrt{1-\gamma^2}}$.

Part 2

The MG

CHAPTER 5

The TG (two-grid) method

This lecture studies the two-grid (or TG) iteration method. Its relation to a basic two-by-two block factorization preconditioner is described and analyzed.

We also derive one more characteristic identity for the inexact TG operator. The lecture ends up with a number of assumptions on the coarse-grid projection operator π_A combined with the smoother that are useful in the analysis of the method in a multilevel setting.

1. The two–grid algorithm and two–grid operator B_{TG}

The TG iteration matrix. Let $\mathbf{x}$ be the exact solution of $A\mathbf{x} = \mathbf{b}$, $\mathbf{x}_0$ the initial approximation and $\mathbf{x}_{TG}$ the approximation produced by applying one iteration of the TG algorithm described in the previous lecture.

We want to find a representation of the error $\mathbf{x}_{TG} - \mathbf{x}$ in terms of the initial error $\mathbf{x} - \mathbf{x}_0$, i.e., to find a formula for the iteration matrix E_{TG} from the relation

$$\mathbf{x} - \mathbf{x}_{TG} = E_{TG}(\mathbf{x} - \mathbf{x}_0).$$

The following result holds:

PROPOSITION 1.1. *The TG iteration matrix E_{TG} admits the following product form:*

$$E_{TG} = (I - M^{-T}A)(I - \pi_A)(I - M^{-1}A).$$

PROOF. In the TG algorithm, we compute consecutively $\mathbf{y}$, $\mathbf{x}_c$, $\mathbf{z}$ and $\mathbf{x}_{TG}$ in the following steps:

$$\begin{aligned} M(\mathbf{y} - \mathbf{x}_0) &= \mathbf{b} - A\mathbf{x}_0, \\ A_c\mathbf{x}_c &= P^T(\mathbf{b} - A\mathbf{y}), \\ \mathbf{z} &= \mathbf{y} + P\mathbf{x}_c, \\ M^T(\mathbf{x}_{TG} - \mathbf{z}) &= \mathbf{b} - A\mathbf{z}. \end{aligned}$$

Starting from the bottom, we have $\mathbf{x} - \mathbf{x}_{TG} = (I - M^{-T}A)(\mathbf{x} - \mathbf{z})$. Similarly, $\mathbf{x} - \mathbf{y} = (I - M^{-1}A)(\mathbf{x} - \mathbf{x}_0)$. On the other hand, $\mathbf{x} - \mathbf{z} = \mathbf{x} - \mathbf{y} - P\mathbf{x}_c = \mathbf{x} - \mathbf{y} - PA_c^{-1}P^TA(\mathbf{x} - \mathbf{y}) = (I - \pi_A)(I - M^{-1}A)(\mathbf{x} - \mathbf{x}_0)$. Therefore the desired result follows. □

Block–factorization definition of TG.

DEFINITION 1.1 (TG preconditioner). *Let A be a given s.p.d. matrix, P a full–rank rectangular matrix, $A_c = P^TAP$ the s.p.d. coarse matrix, and M, M^T the A-convergent smoothers. The latter is equivalent to the fact that $M + M^T - A$ be s.p.d. Consider also the symmetrized smoother $\overline{M} = M\left(M + M^T - A\right)^{-1} M^T$.*

Define first the block–factored s.p.d. matrix

$$\widehat{B}_{TG} = \begin{bmatrix} I & 0 \\ P^T A M^{-1} & I \end{bmatrix} \begin{bmatrix} \overline{M} & 0 \\ 0 & A_c \end{bmatrix} \begin{bmatrix} I & M^{-T} A P \\ 0 & I \end{bmatrix}.$$

Then, the two–grid (TG) preconditioner, B_{TG} is defined from the formula

$$B_{TG}^{-1} = [I,\ P]\, \widehat{B}_{TG}^{-1} \begin{bmatrix} I \\ P^T \end{bmatrix}.$$

Since

$$\widehat{B}_{TG}^{-1} = \begin{bmatrix} I & -M^{-T} A P \\ 0 & I \end{bmatrix} \begin{bmatrix} \overline{M}^{-1} & 0 \\ 0 & A_c^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -P^T A M^{-1} & I \end{bmatrix},$$

and

$$\begin{aligned} [I,\ P] \begin{bmatrix} I & -M^{-T} A P \\ 0 & I \end{bmatrix} &= [I,\ (I - M^{-T} A) P], \\ \begin{bmatrix} I & 0 \\ -P^T A M^{-1} & I \end{bmatrix} \begin{bmatrix} I \\ P^T \end{bmatrix} &= \begin{bmatrix} I \\ P^T (I - A M^{-1}) \end{bmatrix}, \end{aligned}$$

the following explicit formula is easily seen

$$B_{TG}^{-1} = \overline{M}^{-1} + (I - M^{-T} A) P A_c^{-1} P^T (I - A M^{-1}) = \overline{M}^{-1} + (I - M^{-T} A) \pi_A (I - M^{-1} A) A^{-1}.$$

This shows also the relation

$$I - B_{TG}^{-1} A = I - \overline{M}^{-1} A - (I - M^{-T} A) \pi_A (I - M^{-1} A).$$

Finally, recalling that $I - \overline{M}^{-1} A = (I - M^{-T} A)(I - M^{-1} A)$, the following result can be formulated.

PROPOSITION 1.2. *The TG preconditioner has the explicit form*

$$B_{TG}^{-1} = \overline{M}^{-1} + (I - M^{-T} A) P A_c^{-1} P^T (I - A M^{-1}). \tag{2.1}$$

It is s.p.d. and provides a matrix representation of the TG algorithm since it relates to the TG iteration matrix E_{TG}. More specifically, we have

$$I - B_{TG}^{-1} A = E_{TG} = (I - M^{-T} A)(I - \pi_A)(I - M^{-1} A).$$

Finally, the following spectral inequality holds

$$\mathbf{v}^T A \mathbf{v} \leq \mathbf{v}^T B_{TG} \mathbf{v} \text{ for all } \mathbf{v},$$

which implies that the TG method is A–convergent, i.e. we have

$$0 \leq \mathbf{v}^T A E_{TG} \mathbf{v} \leq \mathbf{v}^T A \mathbf{v} - \mathbf{v}^T A B_{TG}^{-1} A \mathbf{v} \leq \left(1 - \frac{1}{K_{TG}}\right) \mathbf{v}^T A \mathbf{v}. \tag{2.2}$$

Here, K_{TG} is an upper bound of the largest eigenvalue of $A^{-1} B_{TG}$, or equivalently an upper bound in the spectral equivalence estimate

$$\mathbf{v}^T B_{TG} \mathbf{v} \leq K_{TG} \mathbf{v}^T A \mathbf{v} \text{ for all } \mathbf{v}. \tag{2.3}$$

PROOF. We need only show the left hand side of (2.2). For this, it is sufficient to show that $\mathbf{v}^T A\pi_A\mathbf{v} \le \mathbf{v}^T A\mathbf{v}$. Equivalently, we need to show that $\overline{\pi}_A \equiv A^{\frac{1}{2}}\pi_A A^{-\frac{1}{2}} = A^{\frac{1}{2}}PA_c^{-1}P^TA^{\frac{1}{2}}$ has norm one. First, we notice that the symmetric matrix $\overline{\pi}_A$ is also a projection (using the fact that $A_c = P^TAP$). We also have that $\mathbf{v}^T\overline{\pi}_A\mathbf{v} \ge 0$. The desired norm estimate then follows from the identity

$$I = \overline{\pi}_A + (I - \overline{\pi}_A) = \overline{\pi}_A^2 + (I - \overline{\pi}_A)^2.$$

That is,

$$0 \le \mathbf{v}^T\overline{\pi}_A\mathbf{v} = \mathbf{v}^T\overline{\pi}_A^2\mathbf{v} \le \mathbf{v}^T\mathbf{v}.$$

The latter shows that $I - \overline{\pi}_A$ is non-negative matrix, hence

$$A^{\frac{1}{2}}E_{TG}A^{-\frac{1}{2}} = X^T(I - \overline{\pi}_A)X, \quad (\text{with } X = I - A^{\frac{1}{2}}M^{-1}A^{\frac{1}{2}}) \tag{2.4}$$

is also non–negative, which is equivalent to $AE_{TG} = A - AB_{TG}^{-1}A$ being non–negative as well, that is, we have the desired result. □

2. Characterization of K_{TG}

From formula (2.4) and the fact that $(I - \overline{\pi}_A)$ is a symmetric projection, we have

$$\|A^{\frac{1}{2}}E_{TG}A^{-\frac{1}{2}}\| = \|(I - \overline{\pi}_A)X\|^2 = \|X^T(I - \overline{\pi}_A)\|^2.$$

Since

$$(I - \overline{\pi}_A)XX^T(I - \overline{\pi}_A) = (I - \overline{\pi}_A)^2 - (I - \overline{\pi}_A)A^{\frac{1}{2}}\widetilde{M}^{-1}A^{\frac{1}{2}}(I - \overline{\pi}_A),$$

where $\widetilde{M} = M^T(M + M^T - A)^{-1}M$, we have the following formula

$$\|A^{\frac{1}{2}}E_{TG}A^{-\frac{1}{2}}\| = 1 - \frac{1}{K_{TG}},$$

with

$$K_{TG} = \max_{\mathbf{v}=(I-\overline{\pi}_A)\mathbf{w}} \frac{\mathbf{v}^T\mathbf{v}}{\mathbf{v}^TA^{\frac{1}{2}}\widetilde{M}^{-1}A^{\frac{1}{2}}\mathbf{v}}. \tag{2.5}$$

To simplify the above expression introduce a basis in the space $A^{\frac{1}{2}}(I-\pi_A)V$. That is, for some full–rank matrix $\overline{S}$, we have that for any vector $A^{\frac{1}{2}}(I - \pi_A)\mathbf{v}$ there is a unique vector $\mathbf{v}_s$ such that $A^{\frac{1}{2}}(I - \pi_A)\mathbf{v} = \overline{S}\mathbf{v}_s$. We can assume that $\overline{S}^T\overline{S} = I$. Then for any vector $\mathbf{v}$ we can form the orthogonal decomposition (noting that $P^TA(I - \pi_A) = 0$),

$$\mathbf{v} = \overline{S}\mathbf{v}_s + \overline{P}\mathbf{v}_c, \quad \overline{P} = A^{\frac{1}{2}}P.$$

Define $\overline{\pi}_A = A^{\frac{1}{2}}\pi_A A^{-\frac{1}{2}} = A^{\frac{1}{2}}PA_c^{-1}P^TA^{\frac{1}{2}}$. Note first that $(I-\overline{\pi}_A)A^{\frac{1}{2}}P = A^{\frac{1}{2}}(I-\pi_A)P = 0$. We then have

$$\begin{aligned}
(I - \overline{\pi}_A)\mathbf{v} &= (I - \overline{\pi}_A)\left(\overline{S}\mathbf{v}_s + A^{\frac{1}{2}}P\mathbf{v}_c\right) \\
&= (I - \overline{\pi}_A)\overline{S}\mathbf{v}_s \\
&= (I - \overline{\pi}_A)A^{\frac{1}{2}}(I - \pi_A)\mathbf{w} \\
&= A^{\frac{1}{2}}(I - \pi_A)^2\mathbf{w} \\
&= A^{\frac{1}{2}}(I - \pi_A)\mathbf{w} \\
&= \overline{S}\mathbf{v}_s.
\end{aligned}$$

That is,

$$(I-\overline{\pi}_A)\mathbf{v}=\overline{S}\mathbf{v}_s. \tag{2.6}$$

Using the decomposition $\mathbf{v}=\overline{S}\mathbf{v}_s+\overline{P}\mathbf{v}_c$ and the identity (2.6), the formula (2.5) takes the form (since $\overline{S}^T\overline{S}=I$)

$$K_{TG}=\max_{\mathbf{v}_s}\ \frac{\mathbf{v}_s^T\mathbf{v}_s}{\mathbf{v}_s^T\overline{S}^TA^{\frac12}\widetilde{M}^{-1}A^{\frac12}\overline{S}\mathbf{v}_s}. \tag{2.7}$$

Consider now the matrix

$$W=\left[\overline{S},\ \overline{P}\right]^TA^{-\frac12}\widetilde{M}A^{-\frac12}\left[\overline{S},\ \overline{P}\right].$$

Since $\left[\overline{S},\ \overline{P}\right]$ is invertible, we get

$$A^{\frac12}\widetilde{M}^{-1}A^{\frac12}=\left[\overline{S},\ \overline{P}\right]W^{-1}\left[\overline{S},\ \overline{P}\right]^T.$$

Therefore, since $\overline{S}^T\overline{P}=0$, and $\overline{S}^T\overline{S}=I$,

$$\overline{S}^TA^{\frac12}\widetilde{M}^{-1}A^{\frac12}\overline{S}=[I,\ 0]\,W^{-1}\,[I,\ 0].$$

That is, $\overline{S}^TA^{\frac12}\widetilde{M}^{-1}A^{\frac12}\overline{S}$ is the inverse of a Schur complement of W. We write

$$\overline{S}^TA^{\frac12}\widetilde{M}^{-1}A^{\frac12}\overline{S}=(W_{\text{Schur}})^{-1}.$$

The Schur complement W_{Schur} has the following characterization (since W is symmetric positive definite),

$$\mathbf{v}_s^TW_{\text{Schur}}\mathbf{v}_s=\inf_{\mathbf{v}_c}\ \begin{bmatrix}\mathbf{v}_s\\ \mathbf{v}_c\end{bmatrix}^TW\begin{bmatrix}\mathbf{v}_s\\ \mathbf{v}_c\end{bmatrix}=\inf_{\mathbf{v}_c}\ \left(\overline{S}\mathbf{v}_s+\overline{P}\mathbf{v}_c\right)^TA^{-\frac12}\widetilde{M}A^{-\frac12}\left(\overline{S}\mathbf{v}_s+\overline{P}\mathbf{v}_c\right).$$

Finally, from (2.7) based on the above characterization of W_{Schur}, we get

$$\begin{aligned}K_{TG}&=\sup_{\mathbf{v}_s}\ \frac{\mathbf{v}_s^T\mathbf{v}_s}{\mathbf{v}_s^T(W_{\text{Schur}})^{-1}\mathbf{v}_s}\\ &=\sup_{\mathbf{v}_s}\ \frac{\mathbf{v}_s^T(W_{\text{Schur}})\mathbf{v}_s}{\mathbf{v}_s^T\mathbf{v}_s}\\ &=\sup_{\mathbf{v}_s}\inf_{\mathbf{v}_c}\frac{(\overline{S}\mathbf{v}_s+\overline{P}\mathbf{v}_c)^TA^{-\frac12}\widetilde{M}A^{-\frac12}(\overline{S}\mathbf{v}_s+\overline{P}\mathbf{v}_c)}{(A^{-\frac12}\overline{S}\mathbf{v}_s)^TA(A^{-\frac12}\overline{S}\mathbf{v}_s)}\\ &=\sup_{\mathbf{v}_s}\inf_{\mathbf{v}_c}\frac{(A^{-\frac12}\overline{S}\mathbf{v}_s+P\mathbf{v}_c)^T\widetilde{M}(A^{-\frac12}\overline{S}\mathbf{v}_s+P\mathbf{v}_c)}{(A^{-\frac12}\overline{S}\mathbf{v}_s)^TA(A^{-\frac12}\overline{S}\mathbf{v}_s)}.\end{aligned}$$

Noting now that $A^{-\frac12}\overline{S}\mathbf{v}_s=(I-\pi_A)\mathbf{v}$, we end up with the desired formula

$$K_{TG}=\sup_{\mathbf{v}}\frac{\inf_{\mathbf{v}_c}\left((I-\pi_A)\mathbf{v}+P\mathbf{v}_c\right)^T\widetilde{M}\left((I-\pi_A)\mathbf{v}+P\mathbf{v}_c\right)}{\left((I-\pi_A)\mathbf{v}\right)^TA\left((I-\pi_A)\mathbf{v}\right)}. \tag{2.8}$$

Replacing $\pi_A\mathbf{v}-P\mathbf{v}_c=P(A_c^{-1}P^TA\mathbf{v}-\mathbf{v}_c)$ with another $P\mathbf{v}_c$, we end up with the following characterization formula

$$K_{TG}=\sup_{\mathbf{v}}\ \frac{\min_{\mathbf{v}_c}\left(\mathbf{v}-P\mathbf{v}_c\right)^T\widetilde{M}(\mathbf{v}-P\mathbf{v}_c)}{(\mathbf{v}-\pi_A\mathbf{v})^TA(\mathbf{v}-\pi_A\mathbf{v})}=\sup_{\mathbf{v}}\ \frac{(\mathbf{v}-\pi_{\widetilde{M}}\mathbf{v})^T\widetilde{M}(\mathbf{v}-\pi_{\widetilde{M}}\mathbf{v})}{(\mathbf{v}-\pi_A\mathbf{v})^TA(\mathbf{v}-\pi_A\mathbf{v})}. \tag{2.9}$$

Here, $\pi_{\widetilde{M}}=P\widetilde{M}_c^{-1}P^T\widetilde{M}$ where $\widetilde{M}_c=P^T\widetilde{M}P$.

Now, since $(I-\pi_{\widetilde{M}})(I-\pi_A) = I-\pi_{\widetilde{M}}$, $(I-\pi_{\widetilde{M}})^T\widetilde{M}(I-\pi_{\widetilde{M}}) = \widetilde{M}(I-\pi_{\widetilde{M}})$ and $((I-\pi_A)\mathbf{v})^T A(I-\pi_A)\mathbf{v} = \mathbf{v}^T A(I-\pi_A)\mathbf{v} \le \mathbf{v}^T A\mathbf{v}$, the following identities are seen

$$\begin{aligned}
K_{TG} &= \max_{\mathbf{v}=(I-\pi_A)\mathbf{v}} \frac{\mathbf{v}^T\widetilde{M}(I-\pi_{\widetilde{M}})\mathbf{v}}{\mathbf{v}^T A\mathbf{v}} \\
&= \max_{\mathbf{v}=(I-\pi_A)\mathbf{v}} \frac{\left((I-\pi_{\widetilde{M}})\mathbf{v}\right)^T\widetilde{M}(I-\pi_{\widetilde{M}})\mathbf{v}}{\mathbf{v}^T A\mathbf{v}} \\
&= \max_{\mathbf{v}} \frac{\left((I-\pi_{\widetilde{M}})(I-\pi_A)\mathbf{v}\right)^T\widetilde{M}(I-\pi_{\widetilde{M}})(I-\pi_A)\mathbf{v}}{\mathbf{v}^T A(I-\pi_A)\mathbf{v}} \\
&= \max_{\mathbf{v}} \frac{\left((I-\pi_{\widetilde{M}})\mathbf{v}\right)^T\widetilde{M}(I-\pi_{\widetilde{M}})\mathbf{v}}{\mathbf{v}^T A(I-\pi_A)\mathbf{v}} \\
&\ge \max_{\mathbf{v}} \frac{\left((I-\pi_{\widetilde{M}})\mathbf{v}\right)^T\widetilde{M}(I-\pi_{\widetilde{M}})\mathbf{v}}{\mathbf{v}^T A\mathbf{v}} \\
&\ge \max_{\mathbf{v}=(I-\pi_A)\mathbf{v}} \frac{\left((I-\pi_{\widetilde{M}})\mathbf{v}\right)^T\widetilde{M}(I-\pi_{\widetilde{M}})\mathbf{v}}{\mathbf{v}^T A\mathbf{v}} \\
&= K_{TG}.
\end{aligned}$$

That is, the following main result holds.

THEOREM 2.1. *We have that the TG operator B_{TG} and A satisfy the spectral equivalence relations*

$$\mathbf{v}^T A\mathbf{v} \le \mathbf{v}^T B_{TG}\mathbf{v} \le K_{TG}\ \mathbf{v}^T A\mathbf{v},$$

where the best constant K_{TG} is characterized as follows:

$$K_{TG} = \max_{\mathbf{v}} \frac{\mathbf{v}^T\widetilde{M}(I-\pi_{\widetilde{M}})\mathbf{v}}{\mathbf{v}^T A\mathbf{v}}. \tag{2.10}$$

The following corollary is easily seen.

COROLLARY 2.1. *Let the A-convergent smoother M and the s.p.d. matrix D be related so that*

$$c_1\ \mathbf{v}^T D\mathbf{v} \le \mathbf{v}^T\widetilde{M}\mathbf{v} \le c_2\ \mathbf{v}^T D\mathbf{v} \text{ for all } \mathbf{v}. \tag{2.11}$$

Define the D based coarse–grid projection $\pi_D = PD_c^{-1}P^T D$, where $D_c = P^T DP$. Then the following two–sided estimates for K_{TG} hold:

$$c_1\ \max_{\mathbf{v}} \frac{\mathbf{v}^T D(I-\pi_D)\mathbf{v}}{\mathbf{v}^T A\mathbf{v}} \le K_{TG} \le c_2\ \max_{\mathbf{v}} \frac{\mathbf{v}^T D(I-\pi_D)\mathbf{v}}{\mathbf{v}^T A\mathbf{v}}.$$

PROOF. The proof follows from the characterization

$$\mathbf{v}^T\widetilde{M}(I-\pi_{\widetilde{M}})\mathbf{v} = \min_{\mathbf{w}_c}\ \|\mathbf{v}-P\mathbf{w}_c\|^2_{\widetilde{M}},$$

and similarly

$$\mathbf{v}^T D(I-\pi_D)\mathbf{v} = \min_{\mathbf{w}_c}\ \|\mathbf{v}-P\mathbf{w}_c\|^2_D,$$

based on the spectral equivalence relations between $\widetilde{M}$ and D. □

As a an application of the last corollary, we get the following main result.

THEOREM 2.2. *Assume that the TG method based on a smoother M is convergent with a bound K_{TG}. Then the smoother M, or a s.p.d. D, that is spectrally equivalent to the symmetrized smoother $\widetilde{M}$, is efficient for A restricted to a subspace complementary to the coarse space. Equivalently, a necessary condition for the TG convergence is the following "weak approximation property" of the coarse space:*

For any $\mathbf{v}$ there is a coarse–grid interpolant $P\mathbf{v}_c$ such that in the D-norm, where D is spectrally equivalent to the symmetrized smoother $\widetilde{M}$, we have the estimate

$$\|\mathbf{v} - P\mathbf{v}_c\|_D^2 \leq \eta_w \ \mathbf{v}^T A\mathbf{v}.$$

PROOF. The space complementary to the coarse space where the smoother M is efficient, or its symmetrized version $\widetilde{M}$ is efficient, or for that matter, any spectrally equivalent s.p.d. matrix D (as in (2.11)) is efficient, can be chosen as Range $(I - \pi_D)$. In the latter case, from (2.11) we have the spectral equivalence relations

$$\mathbf{v}^T A\mathbf{v} \leq c_2 \ \mathbf{v}^T D\mathbf{v} \text{ for any } \mathbf{v},$$

and from below

$$c_1 \ ((I - \pi_D)\mathbf{v})^T \ D(I - \pi_D)\mathbf{v} \leq K_{TG} \ ((I - \pi_D)\mathbf{v})^T A(I - \pi_D)\mathbf{v}.$$

This shows that D and A are spectrally equivalent on the subspace Range $(I-\pi_D)$ (which is complementary to the coarse space). Also, we have the weak approximation property with $\eta_w = \frac{K_{TG}}{c_1}$, seen from the estimate

$$c_1 \ \min_{\mathbf{v}_c} \ \|\mathbf{v} - P\mathbf{v}_c\|_D^2 = c_1 \ ((I - \pi_D)\mathbf{v})^T \ D(I - \pi_D)\mathbf{v} \leq K_{TG} \ \mathbf{v}^T A\mathbf{v}.$$

□

3. Necessary and sufficient conditions for TG convergence

Here, we summarize the role of the "weak approximation property" corresponding to the smoother M, as a necessary and sufficient condition for TG convergence.

The weak approximation property as a necessary condition. It is immediate to see that the main characterization estimate

$$K_{TG} = \max_{\mathbf{v}} \frac{\min \ \|\mathbf{v} - P\mathbf{v}_c\|_{\widetilde{M}}^2}{\|\mathbf{v}\|_A^2},$$

implies the following "*weak approximation property*"

$$\|\mathbf{v} - P\mathbf{v}_c\|_{\widetilde{M}} \leq \sqrt{K_{TG}} \ \|\mathbf{v}\|_A.$$

In practice, we may replace $\widetilde{M}$ with any spectrally equivalent s.p.d. matrix D, such that

$$c_1 \ \mathbf{v}^T D\mathbf{v} \leq \mathbf{v}^T \widetilde{M}\mathbf{v} \leq c_2 \ \mathbf{v}^T D\mathbf{v}.$$

Then from the inequalities

$$c_1 \ \|\mathbf{v} - P\mathbf{v}_c\|_D^2 \leq \|\mathbf{v} - P\mathbf{v}_c\|_{\widetilde{M}}^2 \leq c_2 \ \|\mathbf{v} - P\mathbf{v}_c\|_D^2, \tag{2.12}$$

it follows, that we equivalently have the following "*weak approximation property*"

$$\|\mathbf{v} - P\mathbf{v}_c\|_D \leq \sqrt{\frac{K_{TG}}{c_1}} \ \|\mathbf{v}\|_A.$$

In some applications, we may choose $D = \|A\|\, I$, then we end up with the more familiar "*weak approximation property*"

$$\|A\|^{\frac{1}{2}}\, \|\mathbf{v} - P\mathbf{v}_c\| \leq \eta_w\, \|\mathbf{v}\|_A,$$

where $\eta_w = \sqrt{K_{TG}/c_1}$.

The weak approximation property as a sufficient condition. Finally, it is clear that we can prove a two–grid convergence estimate if we have a "*weak approximation property*"

$$\|\mathbf{v} - P\mathbf{v}_c\|_D \leq \eta_w\, \|\mathbf{v}\|_A,$$

for a s.p.d. D that is spectrally equivalent to $\widetilde{M}$ as in (2.12). More specifically, the following estimate holds

$$K_{TG} = \sup_{\mathbf{v}} \frac{\min\limits_{\mathbf{v}_c} \|\mathbf{v} - P\mathbf{v}_c\|^2_{\widetilde{M}}}{\mathbf{v}^T A\mathbf{v}} \leq c_2\, \sup_{\mathbf{v}} \frac{\min\limits_{\mathbf{v}_c} \|\mathbf{v} - P\mathbf{v}_c\|^2_D}{\mathbf{v}^T A\mathbf{v}} \leq c_2\eta_w^2.$$

At the end, we recall a result (proven in Lecture # 3), that provides conditions for a s.p.d. matrix D to be spectrally equivalent to the symmetrized smoother $\widetilde{M}$.

LEMMA 3.1. *Let M and the s.p.d. matrix D satisfy the estimates*

$$\mathbf{v}^T(M + M^T - A)\mathbf{v} \geq \delta_0\, \mathbf{v}^T D\mathbf{v} \text{ for all } \mathbf{v}, \tag{2.13}$$

and

$$\mathbf{w}^T M\mathbf{v} \leq \delta_1\, \sqrt{\mathbf{w}^T D\mathbf{w}}\sqrt{\mathbf{v}^T D\mathbf{v}} \text{ for all } \mathbf{v},\ \mathbf{w}. \tag{2.14}$$

Then, for $\widetilde{M} = M^T \left(M + M^T - A\right)^{-1} M$, we have

$$\frac{\delta_0}{4}\, \mathbf{v}^T D\mathbf{v} \leq \mathbf{v}^T \widetilde{M}\mathbf{v} \leq \frac{\delta_1^2}{\delta_0}\, \mathbf{v}^T D\mathbf{v}.$$

4. A main identity for B_{TG}

We showed that B_{TG}^{-1} admits an explicit representation by formula (2.1) in Proposition 1.2. In this section, we will derive an identity characterizing B_{TG}. We consider here a B_{TG} where A_c taking part in its definition is replaced by an inexact solver B_c. We assume that

$$\mathbf{v}_c^T A_c\mathbf{v}_c \leq \mathbf{v}_c^T B_c\mathbf{v}_c \text{ for all } \mathbf{v}_c. \tag{2.15}$$

This inequality implies that the inexact B_{TG} also satisfies the lower bound

$$\mathbf{v}^T B_{TG}\mathbf{v} \geq \mathbf{v}^T A\mathbf{v}.$$

Our goal is the following identity.

THEOREM 4.1. *For any $\mathbf{v} = \mathbf{v}_f + P\mathbf{v}_c$, the following identity holds*

$$\mathbf{v}^T B_{TG}\mathbf{v} = \min_{\mathbf{v}=\mathbf{v}_f+P\mathbf{v}_c} \left(\mathbf{v}_c^T B_c\mathbf{v}_c + (\mathbf{v}_f + M^{-T}AP\mathbf{v}_c)^T \overline{M}(\mathbf{v}_f + M^{-T}AP\mathbf{v}_c)\right).$$

PROOF. From the definition of B_{TG}, we have

$$B_{TG}^{-1} = [I,\ P]\, \widehat{B}_{TG}^{-1} \begin{bmatrix} I \\ P^T \end{bmatrix}.$$

This shows that for $X = B_{TG}^{\frac{1}{2}} [I,\ P]\, \widehat{B}_{TG}^{-\frac{1}{2}}$, we have $\|X\| = \|X^T\| = 1$, which implies the inequality

$$\overline{\mathbf{v}}^T \widehat{B}_{TG}^{-\frac{1}{2}} [I,\ P]^T\, B_{TG}\, [I,\ P]\, \widehat{B}_{TG}^{-\frac{1}{2}} \overline{\mathbf{v}} \leq \overline{\mathbf{v}}^T \overline{\mathbf{v}}.$$

Equivalently,

$$\overline{\mathbf{v}}^T [I,\ P]^T\, B_{TG}\, [I,\ P]\, \overline{\mathbf{v}} \leq \overline{\mathbf{v}}^T \widehat{B}_{TG} \overline{\mathbf{v}},$$

for any $\overline{\mathbf{v}} = \begin{bmatrix} \mathbf{v}_f \\ \mathbf{v}_c \end{bmatrix}$. This shows that for $\mathbf{v} = [I,\ P]\, \overline{\mathbf{v}} = \mathbf{v}_f + P\mathbf{v}_c$, we have

$$\mathbf{v}^T B_{TG} \mathbf{v} \leq \overline{\mathbf{v}}^T \widehat{B}_{TG} \overline{\mathbf{v}} = \mathbf{v}_c^T B_c \mathbf{v}_c + (\mathbf{v}_f + M^{-T} A P \mathbf{v}_c)^T \overline{M} (\mathbf{v}_f + M^{-T} A P \mathbf{v}_c).$$

That is, since the decomposition $\mathbf{v} = \mathbf{v}_f + P\mathbf{v}_c$ is arbitrary, we have

$$\mathbf{v}^T B_{TG} \mathbf{v} \leq \min_{\mathbf{v} = \mathbf{v}_f + P\mathbf{v}_c} \left(\mathbf{v}_c^T B_c \mathbf{v}_c + (\mathbf{v}_f + M^{-T} A P \mathbf{v}_c)^T \overline{M} (\mathbf{v}_f + M^{-T} A P \mathbf{v}_c) \right).$$

The fact that we actually have equality is seen for the choice of $\mathbf{v} = [I,\ P]\overline{\mathbf{v}}$ where $\overline{\mathbf{v}}$ solves the equation

$$\widehat{B}_{TG} \overline{\mathbf{v}} = [I,\ P]^T B_{TG} \mathbf{v}.$$

Indeed, we have then

$$\begin{aligned} \overline{\mathbf{v}}^T \widehat{B}_{TG} \overline{\mathbf{v}} &= \left(\widehat{B}_{TG} \overline{\mathbf{v}} \right)^T \widehat{B}_{TG}^{-1} \widehat{B}_{TG} \overline{\mathbf{v}} \\ &= \mathbf{v}^T B_{TG} [I,\ P] \widehat{B}_{TG}^{-1} [I,\ P]^T B_{TG} \mathbf{v} \\ &= \mathbf{v}^T B_{TG} \mathbf{v}. \end{aligned}$$

□

5. The MG (multigrid) method: definition

The MG method is simply a recursive application of the inexact TG one. Assume, that we have a number of levels $k = 0,\ \ldots,\ \ell$ each coming with its $n_k \times n_k$ s.p.d. matrix A_k, respective smoothers M_k and M_k^T that are A_k–convergent. To be specific, for the time being, assume that $n_0 > n_1 > \cdots > n_\ell$, that is, level 0 is the finest and hence level ℓ is the coarsest. Then, letting P_{k+1}^k be the $n_k \times n_{k+1}$ interpolation matrix from coarse level $k+1$ to the next finer level k, we assume that $A_{k+1} = \left(P_{k+1}^k\right)^T A_k P_{k+1}^k$.

To define the MG preconditioner $B = B_{MG}$, we use induction as follows:

At the coarsest level $k = \ell$, we set $B_k = A_k$. Assuming that at level $k+1$, B_{k+1} has been defined, the kth level one, $B = B_k$, is simply the TG preconditioner with inexact coarse-grid solver $B_c = B_{k+1}$ given by the expression

$$B^{-1} = \overline{M}^{-1} + (I - M^{-T} A) P B_c^{-1} P^T (I - A M^{-1}),$$

where the tools involved in its definition are the respective interpolation matrix $P = P_{k+1}^k$ and smoother $M = M_k$. We recall that $\overline{M} = M \left(M + M^T - A\right)^{-1} M^T$.

Then by definition $B = B_{MG} \equiv B_0$ and it is commonly referred to as the $V(1,1)$–cycle MG operator.

It is clear that B can be implement as in Algorithm 5.1 where the "coarse–grid" correction step uses inexact solve with A_c replaced with B_c involving recursive call to the coarser levels. The MG method will be considered again in a somewhat more general situation (involving more smoothing steps) in Algorithm 2.1.

REMARK 5.1. *In some cases, it is more convenient to use index* 0 *for the coarsest level and* ℓ *for the finest one. Then, the interpolation matrix is denoted by* P_k^{k+1} *and the respective Galerkin relation between the coarse and fine-grid matrices* A_k *and* A_{k+1} *reads* $A_k = (P_k^{k+1})^T A_{k+1} P_k^{k+1}$.

In either case, even when we have many levels, when we consider only two consecutive levels, we omit the fine–grid index and use "c" for the coarse-level index. Also, then P *stands for the interpolation matrix from the coarse level to the given fine–grid level. In particular, we have then the Galerkin relation* $A_c = P^T A P$.

6. Some classical MG convergence results

We recall the symmetrized smoothers

$$\overline{M} = M\left(M + M^T - A\right)^{-1} M^T \text{ and } \widetilde{M} = M^T\left(M + M^T - A\right)^{-1} M.$$

They satisfy the relations

$$I - \overline{M}^{-1}A = (I - M^{-T}A)(I - M^{-1}A) \text{ and } I - \widetilde{M}^{-1}A = (I - M^{-1}A)(I - M^{-T}A).$$

Letting $\overline{E} = I - A^{\frac{1}{2}} M^{-1} A^{\frac{1}{2}}$, we also have

$$A^{\frac{1}{2}}(I - \overline{M}^{-1}A)A^{-\frac{1}{2}} = \overline{E}^T\overline{E} \text{ and } A^{\frac{1}{2}}(I - \widetilde{M}^{-1}A)A^{-\frac{1}{2}} = \overline{E}\overline{E}^T.$$

By definition, we have

$$B^{-1} = \overline{M}^{-1} + (I - M^{-T}A)PB_c^{-1}B^T(I - AM^{-1}).$$

Using the identity $A^{\frac{1}{2}}\overline{M}^{-1}A^{\frac{1}{2}} = I - \overline{E}^T\overline{E}$, we obtain

$$A^{\frac{1}{2}}B^{-1}A^{\frac{1}{2}} = I - \overline{E}^T\overline{E} + \overline{E}^T A^{\frac{1}{2}} P B_c^{-1} P^T A^{\frac{1}{2}} \overline{E}.$$

Assume now that

$$0 \le \mathbf{v}_c^T(B_c - A_c)\mathbf{v}_c \le \eta_c\, \mathbf{v}_c^T A_c \mathbf{v}_c \text{ for all } \mathbf{v}_c.$$

Recalling the projection $\overline{\pi}_A = A^{\frac{1}{2}} P A_c^{-1} P^T A^{\frac{1}{2}}$, we get the following upper bound

$$\begin{aligned}
\tfrac{\mathbf{v}^T B\mathbf{v}}{\mathbf{v}^T A\mathbf{v}} &\le \max_{\mathbf{v}} \tfrac{\mathbf{v}^T A^{-\frac{1}{2}} B A^{-\frac{1}{2}}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \\
&\le \max_{\mathbf{v}} \tfrac{\mathbf{v}^T\left(I - \overline{E}^T\overline{E} + \frac{1}{1+\eta_c}\,\overline{E}^T\overline{\pi}_A\overline{E}\right)^{-1}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \\
&= \max_{\mathbf{v}} \tfrac{\mathbf{v}^T\mathbf{v}}{\mathbf{v}^T\left(I - \overline{E}^T\overline{E} + \frac{1}{1+\eta_c}\,\overline{E}^T\overline{\pi}_A\overline{E}\right)\mathbf{v}} \\
&= (1+\eta_c)\, \max_{\mathbf{v}} \tfrac{\mathbf{v}^T\mathbf{v}}{\mathbf{v}^T\left(\eta_c(I - \overline{E}^T\overline{E}) + I - \overline{E}^T(I - \overline{\pi}_A)\overline{E}\right)\mathbf{v}}.
\end{aligned} \tag{2.16}$$

We make now the following main assumption that relates the smoother M and the coarse–grid projection π_A:

(A) There is a constant $\eta_s > 0$ such that for any vector $\mathbf{v}$, it holds

$$\mathbf{v}^T A(I - M^{-T}A)(I - \pi_A)(I - M^{-1}A)\mathbf{v} \le \eta_s\, \left(\mathbf{v}^T A\mathbf{v} - \mathbf{v}^T A(I - M^{-T}A)(I - M^{-1}A)\mathbf{v}\right).$$

.

Assumption (A) can be rewritten as

$$\mathbf{v}^T\overline{E}^T(I-\overline{\pi}_A)\overline{E}\mathbf{v} \le \eta_s\ \mathbf{v}^T\left(I-\overline{E}^T\overline{E}\right)\mathbf{v}. \tag{2.17}$$

Using this estimate in (2.16), we obtain

$$\frac{1}{1+\eta_c}\ \frac{\mathbf{v}^T B\mathbf{v}}{\mathbf{v}^T A\mathbf{v}} \le \max_{\mathbf{v}}\ \frac{\mathbf{v}^T\mathbf{v}}{\mathbf{v}^T\left(I+(\eta_c-\eta_s)\ (I-\overline{E}^T\overline{E})\right)\mathbf{v}}.$$

Thus if $\eta_c \ge \eta_s$, the coarse-level inequalities

$$0 \le \mathbf{v}_c^T(B_c-A_c)\mathbf{v}_c \le \eta_c\ \mathbf{v}_c^T A_c\mathbf{v}_c, \tag{2.18}$$

imply the same type inequalities on the next finer level

$$0 \le \mathbf{v}^T(B-A)\mathbf{v} \le \eta_c\ \mathbf{v}^T A\mathbf{v}. \tag{2.19}$$

That is, an induction argument over the levels applies, since at the initial (coarsest) level $B_c = A_c$, hence (2.18) holds for any $\eta_c \ge 0$, in particular, it holds for $\eta_c = \eta_s$. Thus, we have the following main V-cycle MG convergence result.

THEOREM 6.1. *The V(1,1)–cycle MG preconditioner B is spectrally equivalent to A with a bound given in* (2.19) *where $\eta_c \ge \eta_s$ if assumption (A) holds.*

6.1. Assumptions that imply main assumption (A). First, we consider an assumption that is equivalent to (A). More specifically, we assume:

(A^*) There is a constant $\delta_s \in (0,1)$ such that

$$\|(I-M^{-T}A)\mathbf{v}\|_A^2 \le \delta_s\|(I-\pi_A)\mathbf{v}\|_A^2 + \|\pi_A\mathbf{v}\|_A^2 \text{ for all } \mathbf{v}.$$

Assumption (A^*) has the following interpretation. The smoother reduces the "oscillatory" component of the error referring to the subspace Range $(I-\pi_A)$, whereas it does not amplify the "smooth" error component referring to the coarse space Range $(\pi_A) =$ Range (P). We show next the following result

PROPOSITION 6.1. *Assumptions (A^*) and (A) are equivalent. with $\delta_s = \frac{\eta_s}{1+\eta_s}$.*

PROOF. Consider Assumption (A) in the form (2.17). By rearranging terms we also have

$$\mathbf{v}^T\overline{E}^T\left(I-\frac{1}{1+\eta_s}\overline{\pi}_A\right)\overline{E}\mathbf{v} \le \frac{\eta_s}{1+\eta_s}\ \mathbf{v}^T\mathbf{v}.$$

Since $\overline{\pi}_A$ and $I-\overline{\pi}_A$ are projections, and due to the same reason $\overline{\pi}_A(I-\overline{\pi}_A)=0$, we have

$$I-\frac{1}{1+\eta_s}\overline{\pi}_A = \delta_s\overline{\pi}_A + (I-\overline{\pi}_A) = \delta_s\overline{\pi}_A^2 + (I-\overline{\pi}_A)^2 = \left(\sqrt{\delta_s}\overline{\pi}_A + (I-\overline{\pi}_A)\right)^2.$$

Thus, we have the norm estimate

$$\|\left(\sqrt{\delta_s}\overline{\pi}_A + (I-\overline{\pi}_A)\right)\overline{E}\mathbf{v}\|^2 \le \frac{\eta_s}{1+\eta_s}\ \mathbf{v}^T\mathbf{v}.$$

The same result holds for the transposed operator, i.e., we have

$$\|\overline{E}^T\left(\sqrt{\delta_s}\overline{\pi}_A + (I-\overline{\pi}_A)\right)\mathbf{v}\|^2 \le \frac{\eta_s}{1+\eta_s}\ \mathbf{v}^T\mathbf{v}.$$

Finally, noticing that

$$\left(\sqrt{\delta_s}\overline{\pi}_A + (I - \overline{\pi}_A)\right)^{-1} = (I - \overline{\pi}_A) + \frac{1}{\sqrt{\delta_s}}\,\overline{\pi}_A,$$

which combined with the preceding estimate gives

$$\begin{aligned}\|\overline{E}^T\mathbf{v}\|^2 &\le \delta_s \,\|\left(\sqrt{\delta_s}\overline{\pi}_A + (I - \overline{\pi}_A)\right)^{-1}\mathbf{v}\|^2\\ &= \delta_s \,\|\left((I - \overline{\pi}_A) + \tfrac{1}{\sqrt{\delta_s}}\,\overline{\pi}_A\right)\mathbf{v}\|^2\\ &= \delta_s\|(I - \overline{\pi}_A)\mathbf{v}\|^2 + \|\overline{\pi}_A\mathbf{v}\|^2.\end{aligned}$$

Letting $\mathbf{v} := A^{\frac{1}{2}}\mathbf{v}$ in the last estimate, assumption (A^*) is obtained. Tracing the above steps backward, it is easily seen that (A^*) implies (A). □

Relation between "strong approximation property" and assumption (A). We formulate next two properties.

(B) "ℓ_2 boundedness of π_A":

$$\|A\|\|(I - \pi_A)\mathbf{v}\| \le \ \eta_b\ \|A\mathbf{v}\|.$$

(C) "Strong approximation property": For every $\mathbf{v}$ there is a coarse interpolant $P\mathbf{v}_c$ such that

$$\|\mathbf{v} - P\mathbf{v}_c\|_A^2 \le \frac{\eta_c}{\|A\|}\ \|A\mathbf{v}\|^2.$$

The following result holds.

PROPOSITION 6.2. *Property (C) implies (B) with $\eta_b = \eta_a$.*

PROOF. The proof is based on the discrete version of Aubin–Nitsche's argument. Consider $\mathbf{e} = (I - \pi_A)\mathbf{v}$ and let $A\mathbf{u} = \mathbf{e}$. Letting $\overline{\eta}_c = \sqrt{\frac{\eta_c}{\|A\|}}$ and using that $\mathbf{e}$ is A-orthogonal to the coarse space, we have for $P\mathbf{u}_c$ the accurate coarse interpolant of $\mathbf{u}$ from (C),

$$\begin{aligned}\|\mathbf{e}\|^2 &= \mathbf{e}^T A\mathbf{u}\\ &= \mathbf{e}^T A(\mathbf{u} - P\mathbf{u}_c)\\ &\le \|\mathbf{e}\|_A\|\mathbf{u} - P\mathbf{u}_c\|_A\\ &\le \overline{\eta}_c\ \|\mathbf{e}\|_A\|A\mathbf{u}\|\\ &= \overline{\eta}_c\ \|\mathbf{e}\|_A\|\mathbf{e}\|.\end{aligned}$$

That is, we have $\|\mathbf{e}\|^2 \le \frac{\eta_c}{\|A\|}\ \|\mathbf{e}\|_A^2 = \frac{\eta_c}{\|A\|}\ \mathbf{e}^T A\mathbf{e} = \frac{\eta_c}{\|A\|}\ \mathbf{e}^T A\mathbf{v} \le \frac{\eta_c}{\|A\|}\ \|\mathbf{e}\|\|A\mathbf{v}\|$ which shows property (B). □

We conclude with the following two results.

PROPOSITION 6.3. *Property (B) implies assumption (A) with $\eta_s = \eta_b\ \frac{\|\widetilde{M}\|}{\|A\|}$.*

PROOF. We have, with $\widetilde{\mathbf{v}} = E\mathbf{v}$, $E = I - M^{-1}A$,

$$\widetilde{\mathbf{v}}^T A(I - \pi_A)\widetilde{\mathbf{v}} \le \|A\widetilde{\mathbf{v}}\|\|(I - \pi_A)\widetilde{\mathbf{v}}\| \le \frac{\eta_b}{\|A\|}\ \|A\widetilde{\mathbf{v}}\|^2.$$

This estimate combined with the bound for $\|A\widetilde{\mathbf{v}}\|^2$ which we derive below imply the desired result.

Using the identity $A^{\frac{1}{2}}\widetilde{M}^{-1}A^{\frac{1}{2}} = I - \overline{EE}^T$, we also have

$$\begin{aligned}
\|A\widetilde{\mathbf{v}}\|^2 &\le \|\widetilde{M}\|\ \mathbf{v}^T E^T A\widetilde{M}^{-1}AE\mathbf{v} \\
&= \|\widetilde{M}\|\ (A^{\frac{1}{2}}\mathbf{v})^T A^{-\frac{1}{2}}E^T A^{\frac{1}{2}}(I - \overline{EE}^T)A^{\frac{1}{2}}EA^{-\frac{1}{2}}(A^{\frac{1}{2}}\mathbf{v}) \\
&= \|\widetilde{M}\|\ (A^{\frac{1}{2}}\mathbf{v})^T \overline{E}^T(I - \overline{EE}^T)\overline{E}(A^{\frac{1}{2}}\mathbf{v}) \\
&= \|\widetilde{M}\|\ (A^{\frac{1}{2}}\mathbf{v})^T \left(\overline{E}^T\overline{E} - (\overline{E}^T\overline{E})^2\right)(A^{\frac{1}{2}}\mathbf{v}) \\
&\le \|\widetilde{M}\|\ (A^{\frac{1}{2}}\mathbf{v})^T \left(I - \overline{E}^T\overline{E}\right)(A^{\frac{1}{2}}\mathbf{v}) \\
&= \|\widetilde{M}\|\ \mathbf{v}^T \left(A - A(I - M^{-T}A)(I - M^{-1}A)\right)\mathbf{v}.
\end{aligned}$$

We used the elementary inequality $t - t^2 \le 1 - t$ for the symmetric matrix $\overline{E}^T\overline{E}$ which has eigenvalues between zero and one. □

LEMMA 6.1. *If the smoother M is efficient on the A–orthogonal complement of the coarse space, i.e., on the subspace Range $(I - \pi_A)$ in the sense that*

$$\mathbf{v}_s^T\overline{M}\mathbf{v}_s \le \eta_s\ \mathbf{v}_s^T A\mathbf{v}_s \text{ for all } \mathbf{v}_s = (I - \pi_A)\mathbf{v}. \tag{2.20}$$

Then, assumption (A) holds (with M replaced with M^T).

If the smoother M is s.p.d. and properly scaled such that

$$\mathbf{v}^T A\mathbf{v} \le \mathbf{v}^T M\mathbf{v} \text{ for all } \mathbf{v},$$

then assuming that (2.20) *holds with $\overline{M}$ replaced with M, assumption (A) also holds.*

PROOF. The result follows from the Cauchy–Schwarz inequality

$$\mathbf{w}^T A(I - \pi_A)\mathbf{w} \le \|A\mathbf{w}\|_{\overline{M}^{-1}}\|(I - \pi_A)\mathbf{w}\|_{\overline{M}} \le \sqrt{\eta_s}\ \|A\mathbf{w}\|_{\overline{M}^{-1}}\|(I - \pi_A)\mathbf{w}\|_A$$

That is,

$$\mathbf{w}^T A(I - \pi_A)\mathbf{w} \le \eta_s\ \mathbf{w}^T A\overline{M}^{-1}A\mathbf{w}. \tag{2.21}$$

Using the latter inequality for $\mathbf{w} = (I - M^{-T}A)\mathbf{v}$, noticing that $A\overline{M}^{-1}A = A - A(I - M^{-T}A)(I - M^{-1}A)$, we see that the r.h.s. of the latter inequality takes the form (letting $\overline{E} = I - A^{\frac{1}{2}}M^{-1}A^{\frac{1}{2}}$ and using the inequality $t - t^2 \le 1 - t$ for $\overline{EE}^T$)

$$\mathbf{w}^T A\overline{M}^{-1}A\mathbf{w} = \left(A^{\frac{1}{2}}\mathbf{v}\right)^T \overline{E}(I - \overline{E}^T\overline{E})\overline{E}^T(A^{\frac{1}{2}}\mathbf{v}) \le (A^{\frac{1}{2}}\mathbf{v})^T(I - \overline{EE}^T)(A^{\frac{1}{2}}\mathbf{v}),$$

which is the r.h.s. of (A) (with M is replaced with M^T). Also, the left–hand side of (2.21) is the left–hand-side of (A) (with M replaced with M^T).

The second statement of the Lemma is proven noticing that in the case of M being s.p.d. and scaled so that $\mathbf{v}^T A\mathbf{v} \le \mathbf{v}^T M\mathbf{v}$, we have that

$$\frac{1}{2}\ \mathbf{v}^T M\mathbf{v} \le \mathbf{v}^T\overline{M}\mathbf{v} \le \mathbf{v}^T M\mathbf{v}.$$

That is, efficiency of M in the subspace Range$(I - \pi_A)$ implies efficiency of $\overline{M}$ in the same subspace with the same constant η_s. □

CHAPTER 6

The MG: a recursive application of inexact TG

This lecture studies the multigrid (or MG) iteration method as a recursive application of inexact TG method. We also study the effect of more smoothing steps on the MG convergence factor. The lecture ends with some stable multilevel decompositions of finite element spaces obtained by successive steps of mesh refinement. The latter function decomposition provides stable decomposition of the corresponding coefficient vector spaces and hence offer tools to prove MG convergence without assuming "strong approximation property". The lecture ends with an example of stable decomposition for a non–convex domain (hence no full regularity result is available).

1. Composite iterations and the respective iteration matrix

Given an s.p.d. matrix A, let M_0 provide an A-convergent iteration. The latter is equivalent to the fact of $M_0^T + M_0 - A$ being s.p.d.

Given an integer m and let $m = 2\nu + \theta$ where $\theta = 0$ or 1. For a given initial iterate $\mathbf{x}_0$ for solving the s.p.d. problem $A\mathbf{x} = \mathbf{b}$ perform the following iteration steps for $k = 1, \ldots, \nu$

$$\begin{aligned} M_0(\mathbf{x}_{2k-1} - \mathbf{x}_{2k-2}) &= \mathbf{b} - A\mathbf{x}_{2k-2}, \\ M_0^T(\mathbf{x}_{2k} - \mathbf{x}_{2k-1}) &= \mathbf{b} - A\mathbf{x}_{2k-1}. \end{aligned}$$

If $\theta = 0$, let $\mathbf{x}_m = \mathbf{x}_{2\nu}$; otherwise (if $\theta = 1$) perform one more step

$$M_0(\mathbf{x}_m - \mathbf{x}_{2\nu}) = \mathbf{b} - A\mathbf{x}_{2\nu}.$$

The iteration matrix E of the above composite process takes the product form

$$E = (I - M_0^{-1}A)^\theta \left((I - M_0^{-T}A)(I - M_0^{-1}A)\right)^\nu = (I - M_0^{-1}A)^\theta (I - \overline{M}^{-1}A)^\nu.$$

Recall that $\overline{M} = M(M^T + M - A)^{-1}M^T$.

Based on E, we can define implicitly the matrix M from the equation

$$I - M^{-1}A = E.$$

That is, $M^{-1} = (I - E)A^{-1}$. Since $\|A^{\frac{1}{2}}EA^{-\frac{1}{2}}\| < 1$ (M_0 is A–convergent), it is clear that M is well-defined (i.e., $I - E$ or equivalently $I - A^{\frac{1}{2}}EA^{-\frac{1}{2}}$ is invertible).

Introduce the scaled iteration matrices $\overline{E}_0 = I - A^{\frac{1}{2}}M_0^{-1}A^{\frac{1}{2}}$, $\overline{E} = A^{\frac{1}{2}}EA^{-\frac{1}{2}}$. The following relation holds

$$\overline{E} = \overline{E}_0^\theta \left(\overline{E}_0^T \overline{E}_0\right)^\nu.$$

Hence,

$$\overline{E}^T\overline{E} = \left(\overline{E}_0^T \overline{E}_0\right)^m.$$

2. Multigrid V-cycle algorithm with more smoothing steps

We now define the MG algorithm as inexact TG algorithm recursively calling the coarse-level MG operator B_c defined by induction on the previous coarse levels. At the initial (coarsest) level $B_c = A_c$, i.e., we use exact solve there.

Assuming that at some coarse level B_c has been defined at the next fine level, we define B using the following inexact TG algorithm.

ALGORITHM 2.1 (Inexact TG algorithm with several smoothing steps). *Let A be s.p.d., $P : \mathbb{R}^{n_c} \mapsto \mathbb{R}^n$ be the interpolation matrix, and let M_0 be an A-convergent smoother. Finally let B_c be an s.p.d. approximation to the exact coarse-grid matrix $A_c = P^T AP$.*

For a given initial iterate $\mathbf{x}_0$, to define the next TG iterate $\mathbf{x}_{TG}$, we perform the following steps:

- *Perform $m = 2\nu + \theta$ ($\theta = 0$ or 1) pre–smoothing iterations with the composite smoother M defined implicitly from the relation $I - M^{-1}A = (I - M_0^{-1}A)^\theta \left(E_0^T E_0\right)^\nu$, i.e., compute $\mathbf{y}$ from*

$$M(\mathbf{y} - \mathbf{x}_0) = \mathbf{b} - A\mathbf{x}_0.$$

- *Solve the inexact coarse problem*

$$B_c \mathbf{x}_c = P^T(\mathbf{b} - A\mathbf{y}).$$

- *interpolate: $\mathbf{z} = \mathbf{y} + P\mathbf{x}_c$.*
- *Perform m post–smoothing composite iterations in reverse order, i.e., compute $\mathbf{x}_{TG}$ from the equations*

$$M^T(\mathbf{x}_{TG} - \mathbf{z}) = \mathbf{b} - A\mathbf{z}.$$

The above algorithm defines, at a given level of a hierarchy of grids, the actions of B^{-1} assuming the actions of B_c^{-1} are available. Applying recursion over the levels, a multigrid method is defined by initially letting $B_c = A_c$ and then B is defined on the basis of B_c and at the next level setting $B_c := B$ the next level B is defined as above.

Assuming now assumption (B) (which holds if the strong approximation property (C) holds). I.e., using the Cauchy–Schwarz inequality and assumption (B), we have

$$\widetilde{\mathbf{v}}^T A(I - \pi_A)\widetilde{\mathbf{v}} \le \|A\widetilde{\mathbf{v}}\|\|(I - \pi_A)\widetilde{\mathbf{v}}\| \le \frac{\eta_b}{\|A\|} \|A\widetilde{\mathbf{v}}\|^2. \tag{2.22}$$

We will use this estimate for $\widetilde{\mathbf{v}} = E\mathbf{v}$. Our goal is to prove estimate as in assumption (A), which as we know, implies uniform MG convergence i.e., a uniformly bounded ρ away from unity, or equivalently a uniformly bounded spectral equivalence constant $K \le 1+\eta_s$. More specifically, the following result holds.

THEOREM 2.1 (Braess and Hackbusch). *The following spectral equivalence relations hold for the MG V-cycle with m-step composite smoother if the strong approximation property (C) holds if $m = 2\nu + 1$*

$$\mathbf{v}^T A\mathbf{v} \le \mathbf{v}^T B\mathbf{v} \le \left(1 + \frac{1}{m}\, \eta_b\, \frac{\|\widetilde{M_0}\|}{\|A\|}\right)\, \mathbf{v}^T A\mathbf{v}.$$

For even m, we have

$$\mathbf{v}^T A\mathbf{v} \le \mathbf{v}^T B\mathbf{v} \le \left(1 + \frac{1}{m}\,\eta_b\,\frac{\|\overline{M}_0\|}{\|A\|}\right)\,\mathbf{v}^T A\mathbf{v}.$$

PROOF. Assume that $m = 2\nu + 1$. The case of even m is proved similarly. Using the identity $A^{\frac12}\widetilde{M}_0^{-1}A^{\frac12} = I - \overline{E}_0\overline{E}_0^T$, we have

$$\begin{aligned}
\|A\widetilde{\mathbf{v}}\|^2 &\le \|\widetilde{M}_0\|\ \mathbf{v}^T E^T A\widetilde{M}_0^{-1}AE\mathbf{v}\\
&= \|\widetilde{M}_0\|\ (A^{\frac12}\mathbf{v})^T A^{-\frac12}E^T A^{\frac12}(I - \overline{E}_0\overline{E}_0^T)A^{\frac12}EA^{-\frac12}(A^{\frac12}\mathbf{v})\\
&= \|\widetilde{M}_0\|\ (A^{\frac12}\mathbf{v})^T\overline{E}^T(I - \overline{E}_0\overline{E}_0^T)\overline{E}(A^{\frac12}\mathbf{v})\\
&= \|\widetilde{M}_0\|\ (A^{\frac12}\mathbf{v})^T\left(\overline{E}^T\overline{E} - \overline{E}^T\overline{E}_0\overline{E}_0^T\overline{E}\right)(A^{\frac12}\mathbf{v})\\
&= \|\widetilde{M}_0\|\ (A^{\frac12}\mathbf{v})^T\left(\overline{E}_0^T\overline{E}_0\right)^m\left(I - \overline{E}_0^T\overline{E}_0\right)(A^{\frac12}\mathbf{v})\\
&\le \tfrac{1}{m}\ \|\widetilde{M}_0\|\ (A^{\frac12}\mathbf{v})^T\left(I - \left(\overline{E}_0^T\overline{E}_0\right)^m\right)(A^{\frac12}\mathbf{v})\\
&= \tfrac{1}{m}\ \|\widetilde{M}_0\|\ (A^{\frac12}\mathbf{v})^T\left(I - \overline{E}^T\overline{E}\right)(A^{\frac12}\mathbf{v})\\
&= \tfrac{1}{m}\ \|\widetilde{M}_0\|\ \mathbf{v}^T\left(A - A(I - M^{-T}A)(I - M^{-1}A)\right)\mathbf{v}.
\end{aligned} \tag{2.23}$$

We used the elementary inequality for $t \in [0,1]$, $t^m \le t^k$ for $k \le m$, i.e., $mt^m \le \sum_{k=0}^{m-1} t^k$, hence

$$t^m(1-t) \le (1-t)\,\frac{1}{m}\sum_{k=0}^{m-1} t^k = \frac{1}{m}(1 - t^m).$$

We applied this inequality to the symmetric matrix $\overline{E}_0^T\overline{E}_0$ which has eigenvalues between zero and one. Combining the latter estimate (2.23) with (2.22) the desired property (A) follows with $\eta_s = \frac{1}{m}\,\eta_b\,\frac{\|\widetilde{M}_0\|}{\|A\|}$.

□

3. MG analysis without the strong approximation property (C)

Here, we use indices k, $0 \le k \le \ell$ to denote the level index of the grids generated by successive steps of refinement, $\mathcal{T}_k$ with respective meshsize $h_k = 2^{-k}h_0$, $k \ge 0$. The corresponding finite element spaces are V_k. For $k = \ell$, we use the notation $V = V_h = V_\ell$, $h = h_\ell$, which is the space of our main interest. We also have A_k and G_k the kth level stiffness and mass matrices, respectively. They are $n_k \times n_k$ s.p.d. sparse matrices. We recall the spectral relations between A_k, the diagonal D_k of A_k, G_k and the standard vector–inner product in $\mathbb{R}^{n_k}$:

$$\mathbf{v}_k^T A_k\mathbf{v}_k \le \kappa\ \mathbf{v}_k^T D_k\mathbf{v}_k \simeq h_k^{-2}\ \mathbf{v}_k^T G_k\mathbf{v}_k \simeq h_k^{d-2}\ \mathbf{v}_k^T\mathbf{v}_k. \tag{2.24}$$

The constant κ depends on the type of elements we use; for the case of piecewise linear functions $\kappa = 3$. The number $d = 2$ or 3 stands for the dimension of the domain $\Omega \subset \mathbb{R}^d$, where the b.v.p. is posed.

3.1. The "XZ"-identity. We now formulate a main identity that characterizes the V–cycle MG operator $B = B_\ell$. This identity is referred to as the "XZ"-identity due to a result of Xu and Zikatanov in [**XZ02**] (see also [**Va08**]).

THEOREM 3.1. *Given a sequence of s.p.d. matrices A_k, and Let M_k be the kth level smoother convergent in A_k-norm, and let $P_{k-1} : \mathbb{R}^{n_{k-1}} \mapsto \mathbb{R}^{n_k}$ the corresponding interpolation matrix form coarse level $k-1$ to the next finer level k. We assume the Galerkin relation $A_{k-1} = P_{k-1}^t A_k P_{k-1}$. Consider the respective V(1,1)-cycle MG operator $B = B_{MG}$ defined based on the specified smoothers and interpolation matrices.*

For any fine–grid vector $\mathbf{v} = \mathbf{v}_\ell$, the XZ-identity in a matrix-vector form reads:

$$\begin{aligned} \mathbf{v}^T B \mathbf{v} \;=\; & \min_{(\mathbf{v}_k = \mathbf{v}_k^f + P_k \mathbf{v}_{k-1})_{k=1}^{\ell}} \Big[\mathbf{v}_0^T A_0 \mathbf{v}_0 \\ & + \sum_{k=1}^{\ell} \left(\mathbf{v}_k^f + M_k^{-T} A_k P_{k-1} \mathbf{v}_{k-1} \right)^T \overline{M}_k \left(\mathbf{v}_k^f + M_k^{-T} A_k P_{k-1} \mathbf{v}_{k-1} \right) \Big]. \end{aligned} \tag{2.25}$$

PROOF. The proof follows as a recursive application of the TG identity found in Theorem 4.1. □

Using the triangle inequality, it is clear that in order to bound $\mathbf{v}^T B \mathbf{v}$ in terms of $\mathbf{v}^T A \mathbf{v}$, it is sufficient to bound the following two sums (i)', (ii)' for some particular decomposition of $\mathbf{v}$ involving the components $\mathbf{v}_k$ and $\mathbf{v}_k^f = \mathbf{v}_k - P_{k-1}\mathbf{v}_{k-1}$:

(i)'

$$\sum_k (\mathbf{v}_k^f)^T \overline{M}_k \mathbf{v}_k^f,$$

and

(ii)'

$$\mathbf{v}_0^T A_0 \mathbf{v}_0 + \sum_k (\mathbf{v}_{k-1})^T P_{k-1}^T A_k \left(M_k^T + M_k - A_k \right)^{-1} A_k P_{k-1} \mathbf{v}_{k-1}.$$

We proved for M_k being the forward Gauss–Seidel smoother that

$$\mathbf{v}_k^T \overline{M}_k \mathbf{v}_k \simeq \mathbf{v}_k^T D_k \mathbf{v}_k.$$

For the same smoother, we have $M_k^T + M_k - A_k = D_k$. Finally, recalling (2.24), i.e., that $D_k \simeq h_k^{-2} G_k$, it is clear that to estimate (i)' and (ii)' for smoothers M_k equivalent to the Gauss–Seidel, it is equivalent to bound the sums

(i)

$$\sum_k h_k^{-2} \; (\mathbf{v}_k^f)^T G_k \mathbf{v}_k^f,$$

and

(ii)

$$\mathbf{v}_0^T A_0 \mathbf{v}_0 + \sum_k h_k^2 \; (\mathbf{v}_{k-1})^T P_{k-1}^T A_k G_k^{-1} A_k P_{k-1} \mathbf{v}_{k-1}$$

Rewriting sums (i)-(ii) using finite element functions. Let $v \in V_h$ and $v_k, v_k^f \in V_k$ correspond to the coefficient vectors $\mathbf{v}$ and $\mathbf{v}_k,\ \mathbf{v}_k^f$, respectively. Assume that $v = \sum_k v_k^f$ where $v_k^f = v_k - v_{k-1},\ k \geq 1$ and $v_0^f = v_0$ We recall that the finite element spaces are nested, i.e., $V_{k-1} \subset V_k$. Define the vector $\boldsymbol{\psi}_k = G_k^{-1} A_k P_{k-1} \mathbf{v}_{k-1}$ and let $\psi_k \in V_k$ be the finite element function corresponding to the coefficient vector $\boldsymbol{\psi}_k$. Then, by definition

$$\boldsymbol{\psi}_k^T A_k P_{k-1} \mathbf{v}_{k-1} = a(v_{k-1},\ \psi_k).$$

We also have $v_{k-1} = \sum_{j=0}^{k-1} v_j^f$. Hence,

$$\boldsymbol{\psi}_k^T A_k P_{k-1} \mathbf{v}_{k-1} = a(v_{k-1},\ \psi_k) = \sum_{j=0}^{k-1} a(v_j^f,\ \psi_k).$$

We make now the following assumptions:

(S) "stable decomposition": The decomposition of $\mathbf{v}$, based on the components $\mathbf{v}_k^f$ is such that for $v_k = \sum_{j=0}^{k} v_j^f$ we have

$$\sum_k h_k^{-2} \|v_k^f\|_0^2 \leq C_S\ a(v,v).$$

(I) "strengthened inverse inequality:" for $j \leq k$ and any $\psi_j \in V_j$ and $\psi_k \in V_k$, it holds

$$a(\psi_k,\ \psi_j) \leq C_I\ h_k^{-\frac{1}{2}} h_j^{-\frac{1}{2}}\ \|\psi_k\|_0 |\psi_j|_1.$$

The following main result then holds:

THEOREM 3.2. *Under the assumptions (S) and (I), the sums (i) and (ii) are bounded in terms of* $\mathbf{v}^T A \mathbf{v}$ *uniformly with respect to the number of levels* ℓ *and the fine–grid mesh size* $h \mapsto 0$. *Equivalently, the* V*–cycle MG operator* B *is spectrally equivalent to the fine–grid stiffness matrix* A.

PROOF. The sum (i) is actually assumption (S). To bound sum (ii), we use the finite element representation

$$\sum_k h_k^2\ (G_k^{-1} A_k P_{k-1} \mathbf{v}_{k-1})^T A_k P_{k-1} \mathbf{v}_{k-1} = \sum_k h_k^2\ \sum_{j \leq k} a(\psi_k,\ v_j^f).$$

The sum (ii) also equals to

$$\sum_k h_k^2\ (G_k^{-1} A_k P_{k-1} \mathbf{v}_{k-1})^T G_k \left(G_k^{-1} A_k P_{k-1} \mathbf{v}_{k-1}\right) = \sum_k h_k^2 (\boldsymbol{\psi}_k)^T G_k \boldsymbol{\psi}_k = \sum_k h_k^2 \|\psi_k\|_0^2.$$

Using assumption (I), we have

$$\begin{aligned}\sum_k h_k^2\|\psi_k\|_0^2 &= \sum_k h_k^2 \sum_{j\le k} a(\psi_k,\ v_j^f)\\ &\le C_I \sum_k h_k^2 \sum_{j\le k} h_k^{-\frac12} h_j^{-\frac12} |v_j^f|_1 \|\psi_k\|_0\\ &= C_I \sum_k h_k \|\psi_k\|_0 \sum_{j\le k} \sqrt{\tfrac{h_k}{h_j}}\ |v_j^f|_1\\ &\le C_I \left(\sum_k \sum_{j\le k} h_k^2\|\psi_k\|_0^2 \left(\tfrac{1}{\sqrt2}\right)^{k-j}\right)^{\frac12} \left(\sum_k \sum_{j\le k} \left(\tfrac{1}{\sqrt2}\right)^{k-j}\ |v_j^f|_1^2\right)^{\frac12}\\ &\le \tfrac{\sqrt2}{\sqrt2-1}\ C_I\ \left(\sum_k h_k^2\|\psi_k\|_0^2\right)^{\frac12} \left(\sum_j |v_j^f|_1^2\right)^{\frac12}.\end{aligned}$$

We used $\sqrt{h_k/h_j} = \left(\frac{1}{\sqrt2}\right)^{k-j}$. Since $|v_j^f|_1 \le C h_j^{-1}\|v_j^f\|_0$ (based on (2.24)), we have using assumption (S),

$$\sum_j |v_j^f|_1^2 \le C\ \sum_j h_j^{-2}|v_j^f|_1^2 \le C\ a(v,\ v).$$

That is, we showed

$$\sum_k h_k^2\|\psi_k\|_0^2 \le C\ \left(\sum_k h_k^2\|\psi_k\|_0^2\right)^{\frac12} |v|_1.$$

This shows the desired mesh-independent bound of sum (ii) in terms of $|v|_1^2 = a(v,\ v) = \mathbf{v}^T A\mathbf{v}$. □

4. Verification of assumption (I)

Consider two nested finite element spaces $V_H \subset V_h$ corresponding to respective coarse triangulation $\mathcal{T}_H$ and a refined one $\mathcal{T}_h$. Let T be a coarse mesh element (triangle). For functions $\psi_H \in V_H$ and $\psi_h \in V_h$, noticing that the gradient of ψ_H is constant on T (since ψ_H is linear on T) the integration by parts formula gives

$$\int_T \nabla\psi_H \cdot \nabla\psi_h\ d\mathbf{x} = -\int_T \operatorname{div}(\nabla\psi_H)\psi_h\ d\mathbf{x} + \int_{\partial T} (\nabla\psi_H)\cdot\mathbf{n}\ \psi_h\ d\sigma = \int_{\partial T} (\nabla\psi_H)\cdot\mathbf{n}\ \psi_h\ d\sigma.$$

Hence by Cauchy–Schwarz inequality we get

$$\int_T \nabla\psi_H \cdot \nabla\psi_h\ d\mathbf{x}\ \ \le \left(\int_{\partial T} (\nabla\psi_H\cdot\mathbf{n})^2\ d\boldsymbol{\sigma}\right)^{\frac12} \left(\int_{\partial T}\ \psi_h^2\ d\boldsymbol{\sigma}\right)^{\frac12}.$$

Use now the following inverse inequality valid for any f.e. function ψ_h; namely, the equivalence between discrete ℓ_2–norm and the respective integral L_2-norm of f.e. functions,

$$\int_{\partial T} \psi_h^2\ d\boldsymbol{\sigma} \simeq h^{d-1} \sum_{\mathbf{x}_i\in\partial T} \psi^2(\mathbf{x}_i) \le h^{d-1} \sum_{\mathbf{x}_i\in T} \psi^2(\mathbf{x}_i) \simeq h^{-1} \int_T \psi_h^2\ d\mathbf{x}.$$

Also, since $\nabla\psi_H$ is a constant vector on T, we have

$$\int_{\partial T}(\nabla\psi_H\cdot\mathbf{n})^2\,d\boldsymbol{\sigma}\simeq\frac{|\partial T|}{|T|}\int_T|\nabla\psi_H|^2\,d\mathbf{x}\simeq H^{-1}\int_T|\nabla\psi_H|^2\,d\mathbf{x}.$$

Therefore,

$$\int_T\nabla\psi_H\cdot\nabla\psi_h\,d\mathbf{x}\le C_I\,H^{-\frac12}\left(\int_T|\nabla\psi_H|^2\,d\mathbf{x}\right)^{\frac12}\,h^{-\frac12}\left(\int_T\psi_h^2\,d\mathbf{x}\right)^{\frac12}.$$

Using summation over $T\in\mathcal{T}_H$, $a(\psi_H,\ \psi_h)=\sum\limits_T\int\limits_T\nabla\psi_H\cdot\nabla\psi_h\,d\mathbf{x}$, based on the last inequality, we have then

$$a(\psi_H,\ \psi_h)\le C_IH^{-\frac12}h^{-\frac12}\sum_T\left(\int_T|\nabla\psi_H|^2\,d\mathbf{x}\right)^{\frac12}\left(\int_T\psi_h^2\,d\mathbf{x}\right)^{\frac12}.$$

The desired strengthened inverse inequality (I) follows then applying the Cauchy–Schwarz inequality

$$\begin{aligned}\sum_T\left(\int_T|\nabla\psi_H|^2\,d\mathbf{x}\right)^{\frac12}\left(\int_T\psi_h^2\,d\mathbf{x}\right)^{\frac12}&\le\left(\sum_T|\nabla\psi_H|^2\,d\mathbf{x}\right)^{\frac12}\left(\sum_T\int_T\psi_h^2\,d\mathbf{x}\right)^{\frac12}\\&=\|\nabla\psi_H\|_0\|\psi_h\|_0.\end{aligned}$$

5. Verification of assumption (S)

Consider the Galerkin (also called elliptic) projections $\pi_k:\ H_0^1\mapsto V_k$ defined as the solution of the Galerkin finite element problem: For any $u\in H_0^1$ solve for $\pi_ku\in V_k$ the Galerkin f.e. problem

$$a(\pi_ku,\ \varphi)=a(u,\ \varphi)\text{ for all }\varphi\in V_k.$$

Since the spaces are nested, i.e., $V_k\subset V_{k+1}$, it is easily seen that

$$\pi_k\pi_{k+1}=\pi_k.$$

Assume now full regularity of the b.v.p.. Then as we know (by Aubin–Nitsche's argument) the following L_2–error estimate holds

$$\|u-\pi_ku\|_0\le Ch_k\,|u-\pi_ku|_1. \tag{2.26}$$

Using this estimate for $u=\pi_{k+1}u$, based on the fact that $\pi_k\pi_{k+1}=\pi_k$, we obtain

$$\|\pi_ku-\pi_{k+1}u\|_0\le Ch_k\,|(\pi_k-\pi_{k+1})u|_1. \tag{2.27}$$

The decomposition of our interest is based on the components $v_k^f=(\pi_k-\pi_{k-1})v$. That is, we have

$$v=\sum_kv_k^f=\sum_k(\pi_k-\pi_{k-1})v.$$

To verify (S), we need to bound the sum

$$\sum_kh_k^{-2}\|v_k^f\|_0^2$$

in terms of $|v|_1^2 = a(v,\ v) = \mathbf{v}^T A\mathbf{v}$. For this, we use the $a(.,.)$-orthogonality of the components v_k^f and v_j^f for $j \neq k$. Indeed, assuming $j < k$, we have that $v_k^f = (I - \pi_{k-1})(\pi_k v)$ is $a(.,.)$-orthogonal to V_{k-1} which contains V_j (for $j \leq k-1$). This shows that v_k^f is $a(.,.)$-orthogonal to $v_j^f \in V_j$. As a corollary, we obtain the identity

$$|v|_1^2 = \sum_k |(\pi_k - \pi_{k-1})v|_1^2.$$

Then, using the L_2–error estimate (2.27), combined with the orthogonality for the components, we obtain

$$\sum_k h_k^{-2} \|(\pi_k - \pi_{k-1})v\|_0^2 \leq C \sum_k |(\pi_k - \pi_{k-1})v|_1^2 = C|v|_1^2 = Ca(v,v).$$

6. Lions' example

We recall that to prove (2.26), we assumed full regularity of the b.v.p. To avoid this assumption, other projections that are accurate in L_2 and stable in H_0^1 are needed. It appears that the L_2 projections $Q_k :\ L_2 \mapsto V_k$ satisfy that property (see later Section 1). Alternatively, we may want to partition the domain Ω into overlapping set of $m_0 \geq 1$ convex polygons. Then if any $H_0^1(\Omega)$ function can be decomposed into H^1 components supported into one of the convex polygonal subdomains, a stable decomposition of each component would imply a stable decomposition of the original $H_0^1(\Omega)$ function.

Explicit construction of continuous H_0^1–stable decomposition with components supported in convex polygons was shown in Lions [**Li87**] for a model L–shaped domain Ω with $m_0 = 2$. We present this example next.

EXAMPLE 6.1. *Given the L–shaped domain Ω shown in Figure 1. Consider the following cut–off function*

$$\chi = \begin{cases} 1, & x \leq 0, \\ 1 - \frac{bx}{ay}, & (x,y) \in T = \{b \geq y \geq \frac{b}{a}x,\ 0 \leq x \leq a\}, \\ 0, & y \leq \frac{b}{a}x,\ x \in [0,a]. \end{cases}$$

Its gradient is non–zero only on T and it equals

$$\nabla\chi = \frac{b}{a} \begin{bmatrix} -\frac{1}{y} \\ \frac{x}{y^2} \end{bmatrix}.$$

On T, we have $\frac{x^2}{y^2} \leq \frac{a^2}{b^2}$ and $\frac{1}{x^2+y^2} \geq \frac{1}{\frac{a^2}{b^2}y^2+y^2}$. This shows that

$$|\nabla\chi|^2 = \frac{b^2}{a^2}\frac{1}{y^2}\left[1 + \frac{x^2}{y^2}\right] \leq \frac{b^2}{a^2}\frac{1}{x^2+y^2}\left[1 + \frac{a^2}{b^2}\right]\left[1 + \frac{a^2}{b^2}\right].$$

The decomposition of our main interest reads

$$v = \chi v + (1-\chi)v.$$

Note that $v_1 = \chi v$ is supported in the convex domain (rectangle) $\Omega_1 = (-c,a) \times (0,b)$ and $v_2 = (1-\chi)v$ is supported in the convex domain (rectangle) $\Omega_2 = (0,a) \times (-a,b)$.

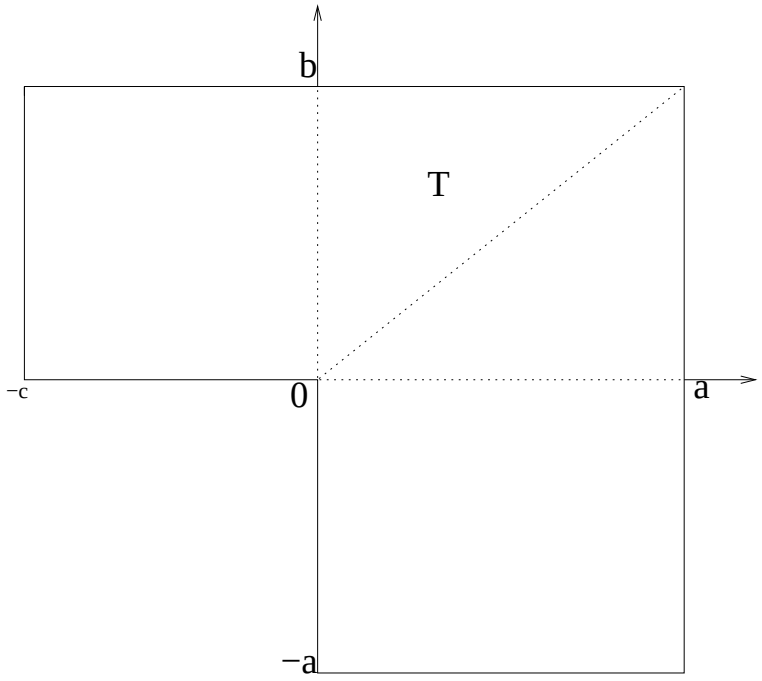

FIGURE 1. *L–shaped domain Ω partitioned into two overlapping rectangles $\Omega_1 = (-c, a) \times (0,\ b)$ and $\Omega_2 = (0,\ a) \times (-a, b)$.*

To show the desired H_0^1–stability, we have to estimate $|v_1|_1$ in terms of $|v|_1$. We have

$$\begin{aligned}\int\limits_{\Omega} |\nabla v_1|^2\, d\mathbf{x} &\le 2 \int\limits_{\Omega} v^2 |\nabla \chi|^2\, d\mathbf{x} + 2 \int\limits_{\Omega} \chi^2 |\nabla v|^2\, d\mathbf{x} \\ &\le 2 \int\limits_{\Omega} \chi^2 |\nabla v|^2\, d\mathbf{x} + C \int\limits_{\Omega} \tfrac{v^2(\mathbf{x})}{dist^2\ (\mathbf{x},\, \partial\Omega)}\, d\mathbf{x}.\end{aligned}$$

The stability follows due to a classical inequality

$$\int\limits_{\Omega} \frac{v^2(\mathbf{x})}{dist^2\ (\mathbf{x},\ \partial\Omega)}\, d\mathbf{x} \le C\ |v|_1^2,$$

valid for $H_0^1(\Omega)$–functions.

CHAPTER 7

Additive MG and MG as block–Gauss-Seidel on an extended system

This lecture studies the additive MG (or BPX) method and its relation to the more traditional (multiplicative) MG by viewing both as block Gauss–Seidel and Jacobi methods on an extended semi-definite system.

1. The additive MG or BPX method

One way to define the MG preconditioner B is based on the following block–matrix factorization:

We first introduce

$$\widehat{B} = \begin{bmatrix} I & 0 \\ P^T A M^{-1} & I \end{bmatrix} \begin{bmatrix} \overline{M} & 0 \\ 0 & B_c \end{bmatrix} \begin{bmatrix} I & M^{-T} A P \\ 0 & I \end{bmatrix},$$

and then

$$B^{-1} = [I,\ P]\, \widehat{B}^{-1} \begin{bmatrix} I \\ P^T \end{bmatrix}.$$

In the "additive" MG we ignore the unit triangular factors in the matrix $\widehat{B}$, i.e., we consider instead

$$\widehat{B}_{add} = \begin{bmatrix} \overline{M} & 0 \\ 0 & B_c \end{bmatrix},$$

and then define as before

$$B_{add}^{-1} = [I,\ P]\, \widehat{B}_{add}^{-1} \begin{bmatrix} I \\ P^T \end{bmatrix},$$

or more explicitly, we have

$$B_{add}^{-1} = \overline{M}^{-1} + P B_c^{-1} P^T.$$

It is also clear that we do not have to use a composite smoother $\overline{M}$ (coming from both M and M^T), instead a single s.p.d. one, Λ, suffices. I.e., we have then

$$B_{add}^{-1} = \Lambda^{-1} + P B_c^{-1} P^T.$$

The following algorithm can be used to evaluate $B_{add}^{-1} = B_\ell^{-1}$ in the case of $\ell \geq 1$ levels. For this purpose, introduce a hierarchy of $n_k \times n_k$ s.p.d. matrices A_k, s.p.d. smoothers Λ_k, and for $k = 1, \ldots, \ell$ the interpolation matrices $P_{k-1} : \mathbb{R}^{n_{k-1}} \mapsto \mathbb{R}^{n_k}$ such that $A_{k-1} = P_{k-1}^T A_k P_{k-1}$. Here, $n_{k-1} < n_k$ and we also let $\Lambda_0 = A_0$ and $P_\ell = I$.

ALGORITHM 1.1 (The multilevel additive MG (BPX) algorithm). *To compute $B_{add}^{-1}\mathbf{b} = B_{BPX}^{-1}\mathbf{b}$ for a given $\mathbf{b}$, we compute $\mathbf{r}_k$ and $\mathbf{x}_k$, for $k = 0, \ldots, \ell$ and let $B_{add}^{-1}\mathbf{b} = \mathbf{x}_\ell$, in the following steps:*

(1) *Let* $\mathbf{r}_0 = \mathbf{b}$ *and for* $k = \ell$ *down to* 1 *compute*

$$\mathbf{r}_{k-1} = P_{k-1}^T \mathbf{r}_k.$$

(2) *Compute* $\mathbf{x}_0 = A_0^{-1}\mathbf{r}_0$ *and for* $k = 1, \ldots, \ell$, *compute*

$$\mathbf{x}_k = \Lambda_k^{-1}\mathbf{r}_k + P_{k-1}\mathbf{x}_{k-1}.$$

(3) *The output is* $B_{add}^{-1}\mathbf{b} = \mathbf{x}_\ell$.

From step (2) above, it is clear that $B_k^{-1}\mathbf{r}_k = \mathbf{x}_k$ satisfies the relation $B^{-1}\mathbf{r}_k = \Lambda_k^{-1}\mathbf{r}_k + P_{k-1}B_{k-1}^{-1}\mathbf{r}_{k-1}$. Using now step (1), i.e., $\mathbf{r}_{k-1} = P_{k-1}^T\mathbf{r}_k$, we get $B^{-1}\mathbf{r}_k = \left(\Lambda_k^{-1} + P_{k-1}B_{k-1}^{-1}P_{k-1}^T\right)\mathbf{r}_k$, that is, the following recurrence holds, starting with $B_0 = A_0$,

$$B_k^{-1} = \Lambda_k^{-1} + P_{k-1}B_{k-1}^{-1}P_{k-1}^T.$$

which shows that Algorithm 1.1 does implement the multilevel additive preconditioner.

It also gives us the following more explicit definition of the method.

DEFINITION 1.1 (Additive MG (BPX) preconditioner). *Introduce a hierarchy of* $n_k \times n_k$ *s.p.d. matrices* A_k, *s.p.d. smoothers* Λ_k, *and for* $k = 1, \ldots, \ell$ *and interpolation matrices* $P_{k-1} : \mathbb{R}^{n_{k-1}} \mapsto \mathbb{R}^{n_k}$ *such that* $A_{k-1} = P_{k-1}^T A_k P_{k-1}$ *and* $n_{k-1} < n_k$. *Let also* $\Lambda_0 = A_0$ *and* $P_\ell = I$.

The multilevel additive V-cycle preconditioner $B_{add} = B_\ell$, *also referred to as the BPX preconditioner, admits the following explicit form*

$$B_{add}^{-1} = \sum_{j=0}^{\ell} (P_\ell \ \ldots \ P_j)\Lambda_j^{-1}(P_j^T \ \ldots \ P_\ell^T). \tag{2.28}$$

1.1. Additive MG: convergence properties. Similarly to the traditional MG, the following main result holds. sometimes referred to as "Lions′ Lemma".

THEOREM 1.1. *Consider for any* $\mathbf{v}$ *decompositions of the form:*

(o) $\mathbf{v}_\ell = \mathbf{v}$,

(i) *for* $k = \ell, \ldots, 1$ *let* $\mathbf{v}_k = [I, \ P_{k-1}] \begin{bmatrix} \mathbf{v}_k^f \\ \mathbf{v}_{k-1} \end{bmatrix}$.

Then for the k*th level additive MG operator* B_k, *based on s.p.d. smoothers* Λ_k *for* A_k *(for example,* $\Lambda_k = M_k\left(M_k^T + M_k - A_k\right)^{-1} M_k^T$ *or* $\Lambda_k = D_k$*–the diagonal of* A_k*) the following identity holds: for any* $k \geq 0$ *and* $k \geq s$,

$$\mathbf{v}_k^T B_k \mathbf{v}_k \ = \inf_{(\mathbf{v}_j = \mathbf{v}_j^f + P_{j-1}\mathbf{v}_{j-1})_{j=s+1}^k} \left[\mathbf{v}_s^T B_s \mathbf{v}_s + \sum_{j=s+1}^{k} (\mathbf{v}_j^f)^T \Lambda_j \mathbf{v}_j^f\right].$$

Note that at the coarsest level $s = 0$, *we typically set* $B_0 = A_0$.

PROOF. We have to note that since the additive MG is also defined via a relation $B_k^{-1} = [I, \ P_{k-1}]\,\widehat{B}_k^{-1}\,[I, \ P_{k-1}]^T$ the same proof as for the standard MG applies in this case. That is, we use the fact that $\|X\| = \|X^T\| = 1$, for $X = \widehat{B}_k^{-\frac{1}{2}}\,[I, \ P_{k-1}]^T\,B_k^{\frac{1}{2}}$. This shows that for any decomposition $\mathbf{v}_k = \mathbf{v}_k^f + P_{k-1}\mathbf{v}_{k-1} = [I, \ P_{k-1}]\,\widehat{\mathbf{v}}_k$, $\widehat{\mathbf{v}}_k = \begin{bmatrix} \mathbf{v}_k^f \\ \mathbf{v}_{k-1} \end{bmatrix}$,

$$\mathbf{v}_k^T B_k \mathbf{v}_k = \widehat{\mathbf{v}}_k^T\,[I, \ P_{k-1}]^T\,B_k\,[I, \ P_{k-1}]\,\widehat{\mathbf{v}}_k \leq \widehat{\mathbf{v}}_k^T \widehat{B}_k \widehat{\mathbf{v}}_k.$$

That is,

$$\mathbf{v}_k^T B_k \mathbf{v}_k \leq (\mathbf{v}_k^f)^T \Lambda_k \mathbf{v}_k^f + \mathbf{v}_{k-1}^T B_{k-1} \mathbf{v}_{k-1} \leq \mathbf{v}_s^T B_s \mathbf{v}_s + \sum_{j=s+1}^{k} (\mathbf{v}_j^f)^T \Lambda_j \mathbf{v}_j^f.$$

The rest of the proof is identical to the one of the MG (cf., Theorem 3.1 which is based on Theorem 4.1.)

□

Our main goal is to prove the spectral equivalence relations

$$\mathbf{v}^T A \mathbf{v} \simeq \mathbf{v}^T B_{add} \mathbf{v}.$$

For the estimate from above, we need to show that for any decomposition $\mathbf{v}_k = \mathbf{v}_k^f + P_{k-1}\mathbf{v}_{k-1}$ the following inequalities hold

$$\mathbf{v}^T A \mathbf{v} \leq C \left(\mathbf{v}_0^T A_0 \mathbf{v}_0 + \sum_{j=1}^{k} (\mathbf{v}_j^f)^T \Lambda_j \mathbf{v}_j^f \right).$$

Then, the desired upper bound would follow by taking minimum over all possible decompositions. To prove the above estimate, we will use the "strengthened inverse inequality"for any pair of functions $v_l^f \in V_l$ and $v_j^f \in V_j$

$$a(v_l^f,\ v_j^f) \leq C_I\ h_l^{-\frac{1}{2}} h_j^{-\frac{1}{2}} \|v_j^f\|_0\ |v_l^f|_1 \text{ if } j \leq l.$$

In terms of finite element functions, we have the decomposition $v = \sum_j v_j^f,\ v_0^f = v_0$, $v_k^f \in V_k$ and $v_k = \sum_{j=0}^{k} v_j^f \in V_k$. We have ($h_l = h_j 2^{j-l}$),

$$\begin{aligned}
\mathbf{v}^T A \mathbf{v} = a(v,\ v) &= a(\sum_j v_j^f,\ \sum_l v_l^f) \\
&= \sum_l a(v_l^f,\ v_l^f) + 2 \sum_{j<l} a(v_l^f,\ v_j^f) \\
&\leq C_I^2 \sum_l h_l^{-2} \|v_l^f\|_0^2 + 2C_I \sum_{j<l} h_l^{-\frac{1}{2}} h_j^{-\frac{1}{2}} \|v_l^f\|_0 |v_j^f|_1 \\
&= C_I^2 \sum_l h_l^{-2} \|v_l^f\|_0^2 + 2C_I \sum_{j<l} h_l^{\frac{1}{2}} h_j^{-\frac{1}{2}} |v_j^f|_1\ \left(h_l^{-1}\ \|v_l^f\|_0 \right) \\
&= C_I^2 \sum_l h_l^{-2} \|v_l^f\|_0^2 + 2C_I \sum_{j<l} \left(\tfrac{1}{\sqrt{2}} \right)^{l-j}\ |v_j^f|_1\ \left(h_l^{-1}\ \|v_l^f\|_0 \right) \\
&\leq C_I^2 \sum_l h_l^{-2} \|v_l^f\|_0^2 + 2C_I \tfrac{\sqrt{2}}{\sqrt{2}-1}\ \left(\sum_j |v_j^f|_1^2 \right)^{\frac{1}{2}} \left(\sum_l h_l^{-2} \|v_l^f\|_0^2 \right)^{\frac{1}{2}} \\
&\leq C_I^2 \left(1 + \tfrac{2\sqrt{2}}{\sqrt{2}-1} \right)\ \sum_l h_l^{-2} \|v_l^f\|_0^2.
\end{aligned}$$

Finally, using the fact that $\Lambda_j \simeq D_j$ (the diagonal of A_j) and that $D_j \simeq h_j^{-2} G_j$ (scaled mass matrix), we see that $\sum_j h_j^{-2} \|v_j^f\|_0^2 \simeq \sum_j h_j^{-2} (\mathbf{v}_j^f)^T G_j \mathbf{v}_j^f \simeq \sum_j (\mathbf{v}_j^f)^T \Lambda_j \mathbf{v}_j^f$, which implies the desired upper bound $\mathbf{v}^T A \mathbf{v} \leq C\ \mathbf{v}^T B_{add} \mathbf{v}$. For the bound in the other direction,

we need to prove that for some particular decomposition $v = \sum_j v_j^f$, the following bound holds

$$\begin{aligned}\mathbf{v}^T B_{add}\mathbf{v} &\leq \sum_j (\mathbf{v}_j^f)^T \Lambda_j \mathbf{v}_j^f \simeq \sum_j (\mathbf{v}_j^f)^T D_j \mathbf{v}_j^f \simeq \sum_j h_j^{-2} (\mathbf{v}_j^f)^T G_j \mathbf{v}_j^f \\ &= \sum_j h_j^{-2} \|v_j^f\|_0^2 \leq C\ a(v,\ v).\end{aligned}$$

The latter inequality, we have verified for the finite element projections $v_j^f = (\pi_j - \pi_{j-1})v$ and convex domain and we commented out how to handle the non-convex domain cases.

Change of notation

As already mentioned (cf. Remark 5.1) in some cases it is more convenient to label the level indices so that $k < l$ refer to fine (k) and coarse (l), respectively. In particular, level 0 stands for the finest level whereas level ℓ is the coarsest one. This convention agrees better with commonly accepted linear algebra (matrix) notation that we often use. With this convention, P_k refers to the interpolation from coarse level $k+1$ to fine level k. The respective vector spaces are $\mathbf{V}_{k+1} = \mathbb{R}^{n_{k+1}}$-coarse and $\mathbf{V}_k = \mathbb{R}^{n_k}$-fine, i.e., we then have $n_{k+1} < n_k$.

2. MG as product iteration method

Introduce next the composite interpolation matrices $\overline{P}_k = P_0 \dots P_{k-1}$ from kth level coarse vector space $\mathbf{V}_k$ all the way up to the finest level vector space $\mathbf{V} = \mathbf{V}_0$. The following result will allow us to view the symmetric $V(1,1)$-cycle MG as a product iterative method performed on the finest level. The iterations exploit corrections from the subspaces $\overline{P}_k \mathbf{V}_k$ of the original vector space $\mathbf{V} = \mathbf{V}_0$. Such methods are sometimes called "subspace correction" methods.

We recall the recursive two-level definition of B_k,

$$B_k^{-1} = \overline{M}_k^{-1} + (I - M_k^{-T} A_k) P_k B_{k+1}^{-1} P_k^T (I - A_k M_k^{-1}). \tag{2.29}$$

PROPOSITION 2.1. *The following recursive relation between the subspace iteration matrices $I - \overline{P}_k B_k^{-1} \overline{P}_k^T A$ and $I - \overline{P}_{k+1} B_{k+1}^{-1} \overline{P}_{k+1}^T A$ holds,*

$$I - \overline{P}_k B_k^{-1} \overline{P}_k^T A = (I - \overline{P}_k M_k^{-T} \overline{P}_k^T A)(I - \overline{P}_{k+1} B_{k+1}^{-1} \overline{P}_{k+1}^T A)(I - \overline{P}_k M_k^{-1} \overline{P}_k^T A).$$

PROOF. We have, from the definition 2.29

$$\overline{P}_k B_k^{-1} \overline{P}_k^T = \overline{P}_k \overline{M}_k^{-1} \overline{P}_k^T + \overline{P}_k (I - M_k^{-T} A_k) P_k B_{k+1}^{-1} P_k^T (I - A_k M_k^{-1}) \overline{P}_k^T.$$

Now use the fact that $A_k = \overline{P}_k^T A \overline{P}_k$ and $\overline{P}_{k+1} = \overline{P}_k P_k$ to arrive at the expression,

$$\overline{P}_k B_k^{-1} \overline{P}_k^T = \overline{P}_k \overline{M}_k^{-1} \overline{P}_k^T + (I - \overline{P}_k M_k^{-T} \overline{P}_k^T A) \overline{P}_{k+1} B_{k+1}^{-1} \overline{P}_{k+1}^T (I - A \overline{P}_k M_k^{-1} \overline{P}_k^T).$$

Then forming $I - \overline{P}_k B_k^{-1} \overline{P}_k^T A$ gives,

$$I - \overline{P}_k B_k^{-1} \overline{P}_k^T A = I - \overline{P}_k \overline{M}_k^{-1} \overline{P}_k^T A - (I - \overline{P}_k M_k^{-T} \overline{P}_k^T A) \overline{P}_{k+1} B_{k+1}^{-1} \overline{P}_{k+1}^T A (I - \overline{P}_k M_k^{-1} \overline{P}_k^T A).$$

It remains to notice that $\overline{M}_k^{-1} = M_k^{-1} + M_k^{-T} - M_k^{-T} A_k M_k^{-1} = M_k^{-1} + M_k^{-T} - M_k^{-T} \overline{P}_k^T A \overline{P}_k M_k^{-1}$ implies

$$I - \overline{P}_k \overline{M}_k^{-1} \overline{P}_k^T A = (I - \overline{P}_k M_k^{-T} \overline{P}_k^T A)(I - \overline{P}_k M_k^{-1} \overline{P}_k^T A),$$

which combined with the previous identity gives the desired result.

□

MG as block Gauss–Seidel. Based on the above product form of the MG V-cycle iteration matrix, the following interpretation of the V-cycle MG is seen.

In the downward cycle, we compute corrections $\overline{P}_k \mathbf{x}_k^f$ from the coarse subspaces Range $(\overline{P}_k)$, $k = 0, \ldots, \ell$ by solving the following systems

$$M_k \mathbf{x}_k^f = \overline{P}_k^T \left(\mathbf{b} - A \sum_{j=0}^{k-1} \overline{P}_j \mathbf{x}_j^f \right) \tag{2.30}$$

The current approximation is $\sum_{j=0}^{k} \overline{P}_j \mathbf{x}_j^f$. At the coarsest level we solve for a correction $\mathbf{x}_\ell^f = \mathbf{x}_\ell$ the system

$$A_\ell \mathbf{x}_\ell = \overline{P}_\ell^T \left(\mathbf{b} - A \sum_{j=0}^{\ell-1} \overline{P}_j \mathbf{x}_j^f \right).$$

We let $\mathbf{y}_\ell^f = \mathbf{x}_\ell$. On the way back, at level $k < \ell$, we compute an update $\mathbf{y}_k^f$ to $\mathbf{x}_k^f$, by solving for the correction $\mathbf{y}_k^f - \mathbf{x}_k^f$, the equation

$$M_k^T (\mathbf{y}_k^f - \mathbf{x}_k^f) = \overline{P}_k^T \left(\mathbf{b} - A \sum_{j=k+1}^{\ell} \overline{P}_j \mathbf{y}_j^f - A \sum_{j=0}^{k} \overline{P}_j \mathbf{x}_j^f \right). \tag{2.31}$$

That is, after step k on the way back, the current approximation is

$$\sum_{j=k}^{\ell} \overline{P}_j \mathbf{y}_j^f + \sum_{j=0}^{k-1} \overline{P}_j \mathbf{x}_j^f.$$

Using the fact that $\mathbf{x}_k^f$ solves the equation (2.30) and $A_k = \overline{P}_k^T A \overline{P}_k$, the system for $\mathbf{y}_k^f$ can be rewritten as

$$M_k^T \mathbf{y}_k^f = (M_k^T + M_k - A_k) \mathbf{x}_k^f - \overline{P}_k^T A \sum_{j=k+1}^{\ell} \overline{P}_j \mathbf{y}_j^f.$$

The final MG V-cycle approximation is

$$\sum_{j=0}^{\ell} \overline{P}_j \mathbf{y}_j^f.$$

In conclusion, introducing the blocks $T_{kj} = \overline{P}_k^T A \overline{P}_j$, and the block–lower triangular matrix

$$L_B = \begin{bmatrix} M_0 & 0 & \ldots & 0 \\ T_{10} & M_1 & \ldots & 0 \\ \vdots & \ddots & \ddots & 0 \\ T_{\ell,0} & \ldots & T_{\ell,\ell-1} & M_\ell \end{bmatrix},$$

the inverse of the V-cycle preconditioner B_{MG} can be represented by the following block-matrix formula

$$B_{MG}^{-1} = \left[\overline{P}_0,\ \ldots,\ \overline{P}_\ell\right] L_B^{-T} \left(\operatorname{diag}\left(M_k^T + M_k - A_k\right)_{k=0}^{\ell}\right) L_B^{-1} \left[\overline{P}_0,\ \ldots,\ \overline{P}_\ell\right]^T. \tag{2.32}$$

This formula gives the following alternative representation of the XZ-identity (Theorem 3.1).

THEOREM 2.1. *The following main identity holds for the V-cycle MG operator B_{MG}:*

$$\mathbf{v}^T B_{MG}\mathbf{v} = \min_{\mathbf{v}=\sum_{j=0}^{\ell}\overline{P}_j\mathbf{v}_j^f} \begin{bmatrix} \mathbf{v}_0^f \\ \vdots \\ \mathbf{v}_\ell^f \end{bmatrix}^T L_B \left(\mathit{diag}\left(M_k^T + M_k - A_k\right)\right)^{-1} L_B^T \begin{bmatrix} \mathbf{v}_0^f \\ \vdots \\ \mathbf{v}_\ell^f \end{bmatrix}.$$

We note that the block-factored matrix $L_B \left(\operatorname{diag}\left(M_k^T + M_k - A_k\right)\right)^{-1} L_B^T$ is the inexact symmetric block Gauss–Seidel preconditioner for the block–matrix T. Indeed, decompose $T = D_T + L_T + L_T^T$ where L_T is the strictly block–lower triangular part of T and $D_T = \operatorname{diag}\left(A_k\right)_{k=0}^{\ell}$ is the block-diagonal part of T. Finally, let $M = \operatorname{diag}\left(M_k\right)_{k=0}^{\ell}$. It is clear then that $L_B = M + L_T$ and hence

$$L_B \left(\operatorname{diag}\left(M_k^T + M_k - A_k\right)\right)^{-1} L_B^T = L_B \left(L_B^T + L_B - T\right)^{-1} L_B^T,$$

which shows the desired result.

To solve a given system $A\mathbf{x} = \mathbf{b}$ we can proceed by first transforming it (cf. [**Gr94**]) based on the fact that any $\mathbf{x}$ allows for a (non-unique) decomposition $\mathbf{x} = \sum_{k=0}^{\ell} \overline{P}_k \mathbf{x}_k^f$ and then after forming $\overline{P}_k^T A\mathbf{x} = \sum_{l=0}^{\ell} \overline{P}_k^T A \overline{P}_l \mathbf{x}_l^f = \overline{P}_k^T \mathbf{b}$ to end up with the following consistent extended system

$$T \begin{bmatrix} \mathbf{x}_0^f \\ \vdots \\ \mathbf{x}_\ell^f \end{bmatrix} = \begin{bmatrix} \overline{P}_0^T \mathbf{b} \\ \vdots \\ \overline{P}_\ell^T \mathbf{b} \end{bmatrix}.$$

Note that the matrix of this system $T = (T_{kj})$, $T_{kj} = \overline{P}_k^T A \overline{P}_j$, is symmetric and only positive semi-definite. The latter consistent semi-definite system is solved then by the CG method using either the (inexact) symmetric Gauss–Seidel matrix

$$L_B \left(\operatorname{diag}\left(M_k^T + M_k - A_k\right)_{k=0}^{\ell}\right)^{-1} L_B^T,$$

or the (inexact) block–Jacobi one

$$\operatorname{diag}\ \left(M_k (M_k^T + M_k - A_k)^{-1} M_k^T\right)_{k=0}^{\ell} = \operatorname{diag}\ \left(\overline{M}_k\right)_{k=0}^{\ell} \simeq \operatorname{diag}\ \left(\Lambda_k\right)_{k=0}^{\ell},$$

as preconditioner. The original solution is recovered then as

$$\mathbf{x} = \left[\overline{P}_0,\ \ldots,\ \overline{P}_\ell\right] \begin{bmatrix} \mathbf{x}_0^f \\ \vdots \\ \mathbf{x}_\ell^f \end{bmatrix} = \sum_{k=0}^{\ell} \overline{P}_k \mathbf{x}_k^f.$$

CHAPTER 8

MG complexity and analysis of variable-step (nonlinear) AMLI-cycle MG

This lecture studies the complexity of various multigrid (or MG) iteration methods (V-cycle, W-cycle, or more general AMLI-cycle). Then, we analyze the AMLI-cycle MG method - both the stationary and conjugate gradient (or CG) based one. For this, we introduce a variable–step CG method and prove some convergence rate estimates.

1. Arithmetic complexity of MG cycles

Consider a hierarchy of meshes $\mathcal{T}_k$ obtained by successive steps of uniform refinement of an initial coarse triangulation $\mathcal{T}_H$; $\mathcal{T}_0 = \mathcal{T}_H$ and $\mathcal{T}_k$ is obtained by refining $\mathcal{T}_{k-1}$. The corresponding meshsizes are related $h_k = \frac{1}{2}\, h_{k-1}$ and the size n_k of the nodesets $\mathcal{N}_k$ (vertices of the triangles in 2D) are of order $2^{dk} n_0$, where $d = 2$ or 3 is the dimension of the computational domain (polygon or polytope) $\Omega \subset \mathbb{R}^d$. The corresponding finite element spaces V_k are nested, i.e., $V_{k-1} \subset V_k$ and there is an interpolation mapping $P_{k-1} : \mathbb{R}^{n_{k-1}} \mapsto \mathbb{R}^{n_k}$ that relates the corresponding coefficient vectors $\mathbf{v}_{k-1}$ of a function $v_{k-1} \in V_{k-1}$ to $P\mathbf{v}_{k-1}$ viewed as an element of V_k (since $V_{k-1} \subset V_k$). Also, the respective stiffness matrices are variationally related, i.e., $A_{k-1} = P_{k-1}^T A_k P_{k-1}$.

Assume that one smoothing iteration with M_k and M_k^T costs $\mathcal{O}(n_k)$ operations. This is the case if M_k comes from the sparse stiffness matrix A_k (that has $\mathcal{O}(n_k)$ non-zero entries), for example if M_k is the forward Gauss–Seidel iteration matrix of the scaled Jacobi (ωD_k, D_k being the diagonal of A_k and ω suitable weight).

In a typical inexact TG algorithm, we perform

(1) three residual computations, $\mathbf{b} - A\mathbf{x}_0$, $\mathbf{b} - A\mathbf{y}$ and $\mathbf{b} - A\mathbf{z}$ which is an order $\mathcal{O}(n_k)$ operations;
(2) one solve with the coarse–grid operator B_c, the cost denoted by w_{k-1} operations;
(3) one restriction based on the action of P^T and one interpolation of the form $\mathbf{z} = \mathbf{y} + P\mathbf{x}_c$, both requiring $O(n_k)$ operations.

Thus the following recursive relation is immediately seen:

$$w_k = w_{k-1} + C n_k.$$

Thus

$$w_{V-cycle} \equiv w_\ell = w_0 + C \sum_k n_k = w_0 + C \sum_k 2^{dk} = w_0 + C\, n_\ell \sum_{k=1}^{\ell} 2^{d(k-\ell)} \le w_0 + C\, n_\ell.$$

2. W–cycle and more general AMLI, or polynomially-based, MG–cycles

Assume that we have defined at a given coarse "c"-level an s.p.d. approximation B_c to A_c, such that

$$\mathbf{v}_c^T A_c \mathbf{v}_c \le \mathbf{v}_c^T B_c \mathbf{v}_c \text{ for all } \mathbf{v}_c.$$

In general, we may not have the actions of B_c on vectors available, what is important, we assume that the actions B_c^{-1} on vectors are readily available. As we saw earlier these actions for B_c (and B) being the V-cycle MG operator are computable by the recursive inexact TG algorithm.

Having actions of B_c^{-1} on vectors available, we may define a more accurate approximation $B_c^{(\nu)}$ to A_c by the following inner iterative method: For any given vector $\mathbf{b}_c$, the more accurate approximation to the solution of $A\mathbf{x}_c = \mathbf{b}_c$, than $B_c^{-1}\mathbf{b}_c$ equals the νth iterate $\mathbf{x}_c^{(\nu)}$ of the inner iterative method:

Let $\mathbf{x}_c^{(0)} = 0$. For $s = 1, \ \ldots, \ \nu$, we compute

$$B_c(\mathbf{x}_c^{(s)} - \mathbf{x}_c^{(s-1)}) = \mathbf{b}_c - A_c\mathbf{x}_c^{(s-1)}. \tag{2.33}$$

This shows that with $E_c = I - B_c^{-1}A_c$,

$$B_c^{(\nu)^{-1}}\mathbf{b}_c \equiv \mathbf{x}_c^{(\nu)} = B_c^{-1}\mathbf{b}_c + (I - B_c^{-1}A_c)\mathbf{x}_c^{(\nu-1)} = (I + E_c + E_c^2 + \cdots + E_c^{\nu-1})B_c^{-1}\mathbf{b}_c + E_c^{\nu}\mathbf{x}_c^{(0)}.$$

That is, for $\mathbf{x}_c^{(0)} = 0$, we have

$$B_c^{(\nu)^{-1}}\mathbf{b}_c = (I - E_c^{\nu})(I - E_c)^{-1}B_c^{-1}\mathbf{b}_c = (I - E_c^{\nu})A_c^{-1}\mathbf{b}_c.$$

W-cycle and AMLI-cycle. Thus, introducing the polynomial $p_\nu(t) = (1-t)^\nu$, we have the following equivalent definition

$$B_c^{(\nu)^{-1}} = \left[I - p_\nu(B_c^{-1}A_c)\right] A_c^{-1}. \tag{2.34}$$

The latter definition can be used for more general polynomials p_ν as long as $p_\nu(0) = 1$ and $|p_\nu(t)| < 1$ on an interval containing the spectrum of $B_c^{-1}A_c$. Typically, we choose $p_\nu(t)$ to be non-negative on the spectrum of $B_c^{-1}A_c$, or simply being nonnegative. For example, if we choose $\nu = 2$ and

$$p_\nu(t) = (1-t)^2 \ge 0,$$

the resulting MG–cycle is referred to as the so–called W-cycle. This means that we use two recursive stationary inner iterations (as in (2.33)).

We estimate next the complexity of the following generalized cycle MG algorithm. Given an approximation B_c to A_c for an integer $\nu \ge 1$, $\nu = \nu_c$ (i.e., it may depend on the level index), we define the more accurate approximation $B_c^{(\nu)}$ to A_c and use it in the inexact TG–algorithm.

More specifically we consider:

ALGORITHM 2.1 (MG algorithm with arbitrary number of inner iterations). *Given B_c - an s.p.d. approximation to A_c, for an integer $\nu = \nu_c \ge 1$ and a suitable polynomial p_ν of degree ν such that $p_\nu(0) = 1$, we define $B_c^{(\nu)^{-1}}$ as in* (2.34).

Then one iteration for solving $A\mathbf{x} = \mathbf{b}$ for a given $\mathbf{x}_0$ computes $\mathbf{x}_{MG} = \mathbf{x}_0 + B^{-1}(\mathbf{b} - A\mathbf{x}_0)$ in the following steps:

(i) *"pre-smoothing iteration:"*

$$M(\mathbf{y}-\mathbf{x}_0)=\mathbf{b}-A\mathbf{x}_0.$$

(ii) *"inexact coarse–grid correction" using polynomial-type inner iterations, i.e, compute* $\mathbf{x}_c$ *from*

$$\mathbf{x}_c=B_c^{(\nu)^{-1}}P^T(\mathbf{b}-A\mathbf{y}).$$

(iii) *"interpolate:"*

$$\mathbf{z}=\mathbf{y}+P\mathbf{x}_c.$$

(iv) *"post–smoothing iteration:"*

$$M^T(\mathbf{x}_{MG}-\mathbf{z})=\mathbf{b}-A\mathbf{z}.$$

We have $\mathcal{O}(n_k)$ operations in total for the smoothing steps (i) and (iv), for computing the residuals in (i), (ii), and (iv), to implement the interpolation step (iii) as well as the restriction of the residual in (ii). The cost of the inner iterations used to implement the inverse action of $B_c^{(\nu)}$ as implemented in (2.33) is readily estimated as

$$\nu_c(w_c+\mathcal{O}(n_c)),$$

where the cost $\mathcal{O}(n_c)$ stands for computing the coarse-level residuals in (2.33).

Thus the following recursion holds

$$w_k=\nu_{k-1}(w_{k-1}+\mathcal{O}(n_{k-1}))+O(n_k). \tag{2.35}$$

Let us now assume the following behavior of ν_k. Given an integer parameter $k_0\geq 1$ and another fixed integer $\nu\geq 1$, assume that ν_k takes one of the following two values

$$\nu_k=\begin{cases}\nu, & \text{if } k=sk_0,\\ 1, & \text{otherwise.}\end{cases} \tag{2.36}$$

The above cycling strategy is sometimes referred to as the AMLI-cycle ("Algebraic Multi-Level Iteration" cycle) originally used in combination with some optimal (Chebyshev) polynomials to define p_ν in (2.34).

In the AMLI-cycle, the general work estimate recursion (2.35) simplifies to

$$w_{(s+1)k_0}=\nu\ w_{sk_0}+\nu\mathcal{O}(n_{sk_0})+\mathcal{O}(n_{sk_0+1}+\cdots+n_{sk_0+k_0}).$$

Applying recursion, we have

$$\begin{aligned}
w_{(s+1)k_0} &= \mathcal{O}(n_{(s+1)k_0})+\nu\mathcal{O}(n_{sk_0})+\nu\ w_{sk_0}\\
&= \mathcal{O}(n_{(s+1)k_0})+\nu\mathcal{O}(n_{sk_0})\\
&\quad +\nu\ \mathcal{O}(n_{sk_0})+\nu^2\ w_{(s-1)k_0}\\
&= \mathcal{O}(n_{(s+1)k_0})+\nu\mathcal{O}(n_{sk_0})\\
&\quad +\nu\mathcal{O}(n_{sk_0})+\nu^2\mathcal{O}(n_{(s-1)k_0})\\
&\quad +\nu^2\mathcal{O}(n_{(s-1)k_0})+\nu^3\mathcal{O}(n_{(s-2)k_0})\\
&\quad \vdots\\
&\quad +\nu^{s-1}\mathcal{O}(2n_{k_0})+\nu^s w_{k_0}\\
&= \nu^s w_{k_0}+\mathcal{O}(\sum_{j=2}^{s+1}\nu^{s+1-j}\ n_{jk_0}).
\end{aligned}$$

Since $w_{k_0} = \mathcal{O}(n_{k_0})$, and $\frac{n_{jk_0}}{n_{(s+1)k_0}} = 2^{jk_0 d} 2^{-(s+1)k_0 d} = \frac{1}{2^{d(s+1-j)k_0}}$, we end up with the final work estimate

$$w_{(s+1)k_0} = \mathcal{O}(\sum_{j=1}^{s+1} \nu^{s+1-j}\ n_{jk_0}) = \mathcal{O}(n_{(s+1)k_0})\ \sum_{j=0}^{s} \left(\frac{\nu}{2^{dk_0}}\right)^j.$$

The latter sum of the geometric progression is $\mathcal{O}(1)$ if

$$\nu < 2^{dk_0}. \tag{2.37}$$

In that case we have that $w_k = \mathcal{O}(n_k)$, i.e., the resulting multilevel cycle leads to a MG method of optimal cost.

The W–cycle corresponds to $\nu = 2$ and $k_0 = 1$. It is clear that it has always an optimal complexity for $d \geq 2$ (since $2 < 2^d$).

In some applications, we may have to choose large ν (i.e., sufficiently many inner iterations) to improve the quality of the cycle. To control the complexity of the resulting MG method then, we have to skip k_0 levels (and use only simple V-cycle recursion there) where k_0 satisfies the inequality (2.37).

3. Analysis of the AMLI-cycle

We consider the simple choice of polynomial $p_\nu(t) = (1-t)^\nu$. Then if B_c is s.p.d. such that $\mathbf{v}_c^T A_c \mathbf{v}_c \leq \mathbf{v}_c^T B_c \mathbf{v}_c$, it follows that the modified one $B_c^{(\nu)}$ also satisfies the same inequality

$$\mathbf{v}_c^T {B_c^{(\nu)}}^{-1} \mathbf{v}_c = \mathbf{v}_c^T \left(I - (I - B_c^{-1} A_c)^\nu\right) A_c^{-1} \mathbf{v}_c \leq \mathbf{v}_c^T A_c^{-1} \mathbf{v}_c.$$

For the upper bound, we have

$$\mathbf{v}_c^T B_c^{(\nu)} \mathbf{v}_c \leq \max_{t \in [\frac{1}{1+\eta_c},\, 1]}\ \frac{1}{1 - p_\nu(t)}\ \mathbf{v}_c^T A_c \mathbf{v}_c,$$

where η_c satisfies the inequality

$$\mathbf{v}_c^T B_c \mathbf{v}_c \leq (1 + \eta_c)\ \mathbf{v}_c^T A_c \mathbf{v}_c.$$

For the particular polynomial $p_\nu = (1-t)^\nu$, we have

$$\max_{t \in [\frac{1}{1+\eta_c},\, 1]}\ \frac{1}{1 - p_\nu(t)} = \frac{1}{1 - (1 - \frac{1}{1+\eta_c})^\nu} = \frac{1 + \eta_c}{1 + q_c + q_c^2 + \cdots + q_c^{\nu-1}}, \quad q_c = \frac{\eta_c}{1 + \eta_c}.$$

Let us now use the XZ-identity for the V-cycle between levels sk_0 and $m \leq (s+1)k_0$ with inexact solve at the coarse level sk_0 using the modified $B_c^{(\nu)} = B_{sk_0}^{(\nu)}$ based on $B_c = B_{sk_0}$. We have

$$\begin{aligned} \mathbf{v}^T B \mathbf{v} \ = \min_{(\mathbf{v}_k = \mathbf{v}_k^f + P_{k-1}\mathbf{v}_{k-1})_{k=sk_0+1}^m} & \Big(\mathbf{v}_{sk_0}^T B_{sk_0}^{(\nu)} \mathbf{v}_{sk_0} \\ & + \sum_{j=sk_0+1}^{m} \left(\mathbf{v}_j^f + M_j^{-T} A_j P_{j-1} \mathbf{v}_{j-1}\right)^T \overline{M}_j \left(\mathbf{v}_j^f + M_j^{-T} A_j P_{j-1} \mathbf{v}_{j-1}\right) \Big) \end{aligned}$$

Using the estimate between $B_c^{(\nu)}$ and A_c in the latter identity, we end up

$$\begin{aligned}\mathbf{v}^T B\mathbf{v} &\leq \tfrac{1+\eta_c}{1+q_c+q_c^2+\cdots+q_c^{\nu-1}} \min_{(\mathbf{v}_k=\mathbf{v}_k^f+P_{k-1}\mathbf{v}_{k-1})_{k=sk_0+1}^m} \Big(\mathbf{v}_{sk_0}^T A_{sk_0}\mathbf{v}_{sk_0} \\ &\quad + \sum_{j=sk_0+1}^{m} \left(\mathbf{v}_j^f + M_j^{-T}A_jP_{j-1}\mathbf{v}_{j-1}\right)^T \overline{M}_j \left(\mathbf{v}_j^f + M_j^{-T}A_jP_{j-1}\mathbf{v}_{j-1}\right)\Big) \\ &= \tfrac{1+\eta_c}{1+q_c+q_c^2+\cdots+q_c^{\nu-1}}\ \mathbf{v}^T B^{(sk_0)\mapsto m}\mathbf{v}.\end{aligned}$$

In the last line above we used the V–cycle MG operator that uses exact solve at its coarsest level sk_0. Thus, the inexact V–cycle one, B, is bounded by the exact V–cycle operator multiplied by the factor $\frac{1+\eta_c}{1+q_c+q_c^2+\cdots+q_c^{\nu-1}}$ resulting from the error $\eta_c \geq 0$ that we commit at level sk_0.

Assume now that the V–cycle operators $B^{(jk_0)\mapsto(j+1)k_0}$ of level length k_0, i.e., between any pair of levels jk_0 and $(j+1)k_0$ and exact solve at their respective coarse level jk_0 can be bounded in terms of $A^{(j+1)k_0}$ uniformly with respect to j by a constant κ_{k_0}. That is this constant may depend on k_0 but is independent of j. Hence

$$\mathbf{v}^T A\mathbf{v} \leq \mathbf{v}^T B^{(jk_0)\mapsto(j+1)k_0}\mathbf{v} \leq \kappa_{k_0}\ \mathbf{v}^T A\mathbf{v}, \quad A = A^{((j+1)k_0)}.$$

In what follows, we want to choose ν sufficiently large so that

$$\frac{1+\eta_c}{1+q_c+q_c^2+\cdots+q_c^{\nu-1}}\ \kappa_{k_0} \leq 1+\eta_c.$$

It is clear that if $\nu > \kappa_{k_0}$, we can find a $q_c \in (0,1)$ such that the above inequality holds.

In the case of W-cycle, we are in the situation of $k_0 = 1$ and $\kappa_{k_0} = K_{TG} = \frac{1}{1-\rho_{TG}}$, where ρ_{TG} is the convergence factor of any exact TG method. Thus the condition for uniform W-cycle convergence factor is

$$\frac{1}{1-\varrho_{TG}} = K_{TG} < 2.$$

That is, if at all levels the exact TG method has a convergence factor

$$\varrho_{TG} < \frac{1}{2},$$

then the W-cycle is also uniformly convergent with a factor

$$\varrho_{W-cycle} = 1 - \frac{1}{K_{W-cycle}} \leq 1 - \frac{1}{1+\eta_c} = \frac{\eta_c}{1+\eta_c} = q_c.$$

The constant $q_c \in (0,1)$ solves the equation $\frac{K_{TG}}{1+q_c} = 1$, or $q_c = K_{TG} - 1 = \frac{1}{1-\varrho_{TG}} - 1 = \frac{\varrho_{TG}}{1-\varrho_{TG}}$, i.e., we have

$$\varrho_{W-cycle} \leq \frac{\varrho_{TG}}{1-\varrho_{TG}} < 1.$$

4. Using nonlinear approximate coarse-grid operators

In this section, we assume that the operators B_c^{-1} are approximated by some nonlinear mappings $B_c[\cdot]$ that satisfy the deviation estimate

$$\|B_c^{-1}\mathbf{v}_c - B_c[\mathbf{v}_c]\|_{B_c} \leq \delta_c\ \|\mathbf{v}_c\|_{B_c^{-1}}.$$

Using the nonlinear mapping $B_c[.]$ in a (conjugate–gradient like) iterative procedure, we can define an approximation $B_c^{(\nu)}[\cdot]$ to A_c^{-1} that is again a nonlinear mapping. By increasing $\nu \geq 1$, the number of these iterations, we get better approximation that satisfies the estimate

$$\|A_c^{-1}\mathbf{v}_c - B_c^{(\nu)}[\mathbf{v}_c]\|_{A_c} \leq \overline{\delta}_c^{\nu} \, \|\mathbf{v}_c\|_{A_c^{-1}}. \tag{2.38}$$

where

$$\overline{\delta}_c = \sqrt{1 - \frac{1-\delta_c^2}{\kappa_c}}. \tag{2.39}$$

Here, κ_c is the condition number of $B_c^{-1}A_c$.

A non–linear TG operator with inexact coarse-grid solve is defined as follows

$$B[\mathbf{v}] = \overline{M}^{-1}\mathbf{v} + (I - M^{-T}A)PB_c^{(\nu)}\left[P^T(I - AM^{-1})\mathbf{v}\right].$$

We can also define the companion linear TG operator using B_c^{-1} as inexact coarse–grid solver,

$$B^{-1}\mathbf{v} = \overline{M}^{-1}\mathbf{v} + (I - M^{-T}A)PB_c^{-1}P^T(I - AM^{-1})\mathbf{v}.$$

We have certain monotonicity property. More specifically, consider for any $\mathbf{v}$ the coarse vector $\overline{\mathbf{v}}_c = P^T(I - AM^{-1})\mathbf{v}$. Then

$$\|B[\mathbf{v}] - B^{-1}\mathbf{v}\|_B \leq \|B_c[\overline{\mathbf{v}}_c] - B_c^{-1}\overline{\mathbf{v}}_c\|_{B_c}.$$

From the definitions of $\overline{\mathbf{v}}_c$ and B^{-1}, we have

$$\begin{aligned}
\|\overline{\mathbf{v}}_c\|^2_{B_c^{-1}} &= \|B_c^{-\frac{1}{2}}\overline{\mathbf{v}}_c\|^2 \\
&= \mathbf{v}^T(I - M^{-T}A)PB_c^{-1}P^T(I - AM^{-1})\mathbf{v} \\
&\leq \mathbf{v}^T\left(\overline{M}^{-1} + (I - M^{-T}A)PB_c^{-1}P^T(I - AM^{-1})\right)\mathbf{v} \\
&= \|\mathbf{v}\|^2_{B^{-1}}.
\end{aligned}$$

This inequality also shows that $\|B_c^{-\frac{1}{2}}P^T(I - AM^{-1})B^{\frac{1}{2}}\| \leq 1$. The desired result is seen from the inequalities

$$\begin{aligned}
\|B^{-1}\mathbf{v} - B[\mathbf{v}]\|_B &= \|B^{\frac{1}{2}}(I - M^{-T}A)P\left(B_c^{(\nu)}[\overline{\mathbf{v}}_c] - B_c^{-1}\overline{\mathbf{v}}_c\right)\| \\
&\leq \|B^{\frac{1}{2}}(I - M^{-T}A)PB_c^{-\frac{1}{2}}\|\|B_c^{(\nu)}[\overline{\mathbf{v}}_c] - B_c^{-1}\overline{\mathbf{v}}_c\|_{B_c} \\
&= \|B_c^{-\frac{1}{2}}P^T(I - AM^{-1})B^{\frac{1}{2}}\|\|B_c^{(\nu)}[\overline{\mathbf{v}}_c] - B_c^{-1}\overline{\mathbf{v}}_c\|_{B_c} \\
&\leq \|B_c^{(\nu)}[\overline{\mathbf{v}}_c] - B_c^{-1}\overline{\mathbf{v}}_c\|_{B_c}.
\end{aligned}$$

Now, let us consider the following nonlinear AMLI-type MG cycle. For a given integer parameter $k_0 \geq 1$ and a fixed number if inner iterations $\nu \geq 1$ at every level k of multiplicity k_0, i.e., $k = sk_0$ for some $s \geq 1$, we run ν inner iterations that from a nonlinear mapping $B_k[.]$ that approximates the linear V–cycle mapping (its inverse) B_k^{-1} (with exact solve at level $k - k_0 = (s-1)k_0$, we define an iterated one $B_k^{(\nu)}[.]$ that approximates A_k^{-1} with certain accuracy. Our goal is to estimate the quality of the nonlinear mapping $B_{(s+1)k_0}[.]$ and its iterated version $B^{(nu)}_{(s+1)k_0}[\cdot]$ as an approximation to $A^{-1}_{(s+1)k_0}$.

Using the monotonicity result, we get the inequalities

$$\|B^{-1}_{(s+1)k_0}\mathbf{v} - B_{(s+1)k_0}[\mathbf{v}]\|_{B_{(s+1)k_0}} \leq \|B^{-1}_{sk_0}\overline{\mathbf{v}}_c - B^{(\nu)}_{sk_0}[\overline{\mathbf{v}}_c]\|_{B_{sk_0}},$$

for a vector $\overline{\mathbf{v}}_c$ such that

$$\|\overline{\mathbf{v}}_c\|_{B^{-1}_{sk_0}} \leq \|\mathbf{v}\|_{B^{-1}_{(s+1)k_0}}.$$

Since at level sk_0, $B^{-1}_{(s+1)k_0}$ uses exact solve, i.e., we have $B_{sk_0} = A_{sk_0}$, hence

$$\|B^{-1}_{(s+1)k_0}\mathbf{v} - B_{(s+1)k_0}[\mathbf{v}]\|_{B_{(s+1)k_0}} \leq \|A^{-1}_{sk_0}\overline{\mathbf{v}}_c - B^{(\nu)}_{sk_0}[\overline{\mathbf{v}}_c]\|_{A_{sk_0}}.$$

Assume now by induction, that

$$\|A^{-1}_{sk_0}\overline{\mathbf{v}}_c - B^{(\nu)}_{sk_0}[\overline{\mathbf{v}}_c]\|_{A_{sk_0}} \leq \delta \, \|\overline{\mathbf{v}}_c\|_{A^{-1}_{sk_0}}.$$

Due to the monotonicity, we have then for $k \leq (s+1)k_0$,

$$\|B^{-1}_k\mathbf{v} - B_k[\mathbf{v}]\|_{B_k} \leq \|A^{-1}_{sk_0}\overline{\mathbf{v}}_c - B^{(\nu)}_{sk_0}[\overline{\mathbf{v}}_c]\|_{A_{sk_0}} \leq \delta \, \|\overline{\mathbf{v}}_c\|_{A^{-1}_{sk_0}} \leq \delta \, \|\mathbf{v}\|_{B^{-1}_k}.$$

Assume also that the k_0th length V–cycle MG operator B_k has a relative condition number with respect to A_k bounded by κ_{k_0}. Then applying the estimate for the iterated nonlinear mapping, we obtain the estimate

$$\|A^{-1}_k\mathbf{v} - B^{(\nu)}_k[\mathbf{v}]\|_{A_k} \leq \overline{\delta}^{\nu}\|\mathbf{v}\|_{A^{-1}_k}.$$

where

$$\overline{\delta} \leq \sqrt{1 - \frac{1-\delta}{\kappa_{k_0}}}.$$

To complete the induction argument, we have to show that we can choose ν sufficiently large for a fixed k_0 such that the inequality

$$\left(\sqrt{1 - \frac{1-\delta^2}{\kappa_{k_0}}}\right)^{\nu} \leq \delta,$$

has a solution $\delta \in (0,1)$. Equivalently, letting $t = \delta^{\frac{2}{\nu}} \in (0,1)$, we need to solve the inequality

$$1 - t \leq \frac{1-t^{\nu}}{\kappa_{k_0}}.$$

That is,

$$\kappa_{k_0} \leq 1 + t + \cdots + t^{\nu-1}.$$

This is solvable, if $\nu > \kappa_{k_0}$ (noting that $\kappa_{k_0} \geq 1$) since then the function $f(t) = \kappa_{k_0} - (1 + t + \cdots + t^{\nu-1})$ changes sign in the interval $[0,1]$.

The following result similar to the (linear) W-cycle holds.

COROLLARY 4.1. *If the two–grid method at all levels (with exact solve at its coarse level) is uniformly convergent so that $\varrho_{TG} < \frac{1}{2}$, i.e., $K_{TG} = \frac{1}{1-\varrho_{TG}} < 2$, then the nonlinear W–cycle (or nonlinear AMLI-cycle with $\nu = 2$ preconditioned CG-based recursive calls at all levels) is uniformly convergent with a factor*

$$\delta \leq \kappa_{TG} - 1 = \frac{\varrho_{TG}}{1 - \varrho_{TG}} < 1.$$

5. Steepest descent algorithm with nonlinear preconditioner

A nonlinear coercive mapping. Let B be a s.p.d. mapping with relative condition number κ with respect to a given s.p.d. matrix A, i.e., if we have

$$\gamma_1 \mathbf{v}^T A\mathbf{v} \le \mathbf{v}^T B\mathbf{v} \le \gamma_2\, \mathbf{v}^T A\mathbf{v} \text{ for all } \mathbf{v}, \tag{2.40}$$

then we can choose $\kappa = \frac{\gamma_2}{\gamma_1}$ as an estimate of the condition number of $B^{-1}A$.

We assume now, that by some algorithm we can approximate B^{-1} with a computable nonlinear mapping $B[.]$ such that the following estimate holds:

$$\|B^{-1}\mathbf{v} - B[\mathbf{v}]\|_B \le \delta\, \|\mathbf{v}\|_{B^{-1}}.$$

Here δ is a tolerance between zero and one. The above estimate is equivalent to

$$\|\mathbf{v}\|^2_{B^{-1}} - 2\mathbf{v}^T B[\mathbf{v}] + \|B[\mathbf{v}]\|^2_B \le \delta^2\, \|\mathbf{v}\|^2_{B^{-1}}.$$

This inequality implies the coercivity estimate

$$\mathbf{v}^T B[\mathbf{v}] \ge \frac{1}{2}\left((1-\delta^2)\, \|\mathbf{v}\|^2_{B^{-1}} + \|B[\mathbf{v}]\|^2_B\right) \ge \sqrt{1-\delta^2}\, \|\mathbf{v}\|_{B^{-1}} \|B[\mathbf{v}]\|_B.$$

Using then the relations (2.40) we arrive at the modified coercivity estimate

$$\mathbf{v}^T B[\mathbf{v}] \ge \sqrt{\frac{1-\delta^2}{\kappa}}\, \|\mathbf{v}\|_{A^{-1}} \|B[\mathbf{v}]\|_A. \tag{2.41}$$

A nonlinearly preconditioned steepest descent algorithm. Now consider the following iteration method that minimizes the A^{-1}-norm of the residual $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha AB[\mathbf{r}_k]$ along the search direction $\mathbf{d}_k = B[\mathbf{r}_k]$. That is, starting with some iteration $\mathbf{x}_0$ for solving $A\mathbf{x} = \mathbf{b}$, for $k \ge 0$ we compute $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$ and form $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha B[\mathbf{r}_k]$ where α is chosen so that

$$\|\mathbf{r}_{k+1}\|_{A^{-1}} = \|\mathbf{r}_k - \alpha AB[\mathbf{r}_k]\|_{A^{-1}} \mapsto \min.$$

This minimization problem gives the following formula

$$\alpha = \frac{\mathbf{r}_k^T B[\mathbf{r}_k]}{\|B[\mathbf{r}_k]\|^2_A}.$$

With this choice of α the following relation is seen

$$\|\mathbf{r}_{k+1}\|^2_{A^{-1}} = \|\mathbf{r}_k\|^2_{A^{-1}} - \frac{(\mathbf{r}_k^T B[\mathbf{r}_k])^2}{\|B[\mathbf{r}_k]\|^2_A}.$$

Using the coercivity estimate (2.41) the following convergence rate estimate is immediately seen

$$\|\mathbf{r}_{k+1}\|^2_{A^{-1}} \le \left(1 - \frac{1-\delta^2}{\kappa}\right)\, \|\mathbf{r}_k\|^2_{A^{-1}}.$$

This leads to the kind of estimates (2.38)–(2.39) that we used previously in Section 4. Indeed, letting $\mathbf{b} = \mathbf{v}$ and $\mathbf{x}_0 = 0$ defining $B^{(\nu)}[\mathbf{v}] = \mathbf{x}_\nu$, we arrive at the estimate

$$\|A^{-1}\mathbf{v} - B^{(\nu)}[\mathbf{v}]\|_A \le \left(1 - \frac{1-\delta^2}{\kappa}\right)^{\frac{\nu}{2}}\, \|\mathbf{v}\|_{A^{-1}}$$

A nonlinearly preconditioned CG algorithm. At the end we remark that in practice we use the potentially more accurate than the above steepest descent algorithm, namely, the conjugate gradient method with possibly nonlinear preconditioner $B[.]$. It can be summarized as follows:

ALGORITHM 5.1 (Variable–step (Flexible) Preconditioned CG Algorithm). *Given the system $A\mathbf{x} = \mathbf{b}$ with a s.p.d. matrix A and let $B[.]$ be a nonlinear mapping that approximates the inverse of a linear preconditioner B, also a s.p.d. matrix.*

The algorithm below uses a sequence of integers $\{m_k\}_{k\geq 0}$, $0 \leq m_k \leq m_{k-1}+1 \leq k-1$ for $k \geq 1$ ($m_0 = m_1 = 0$). A typical choice is $m_k = 0$.

For a given initial iterate $\mathbf{x}_0$, for $k \geq 0$ the method computes $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$, $\widetilde{\mathbf{r}}_k = B[\mathbf{r}_k]$ and respective search vectors $\{\mathbf{d}_j\}_{j\geq 0}$.

More specifically, the algorithm consists of the following steps:

(1) *Letting $\mathbf{x}_0 = 0$, hence $\mathbf{r}_0 = \mathbf{b}$, and $\widetilde{\mathbf{r}}_0 = B[\mathbf{r}_0]$. We let $\mathbf{d}_0 = \widetilde{\mathbf{r}}_0$. The first iterate then equals*

$$\mathbf{x}_1 = \frac{\mathbf{d}_0^T \widetilde{\mathbf{r}}_0}{\mathbf{d}_0^T A \mathbf{d}_0}\, \mathbf{d}_0.$$

The corresponding residual is $\mathbf{r}_1 = \mathbf{r}_0 - \frac{\mathbf{d}_0^T \widetilde{\mathbf{r}}_0}{\mathbf{d}_0^T A \mathbf{d}_0}\, A\mathbf{d}_0$.

(2) *For $k \geq 1$, compute $\widetilde{\mathbf{r}}_k = B[\mathbf{r}_k]$ and based on the most recent $m_k + 1$ search vectors $\{\mathbf{d}_j\}_{j=k-1-m_k}^{k-1}$, the next search vector is computed as follows:*

$$\mathbf{d}_k = \widetilde{\mathbf{r}}_k - \sum_{j=k-1-m_k}^{k-1} \frac{\widetilde{\mathbf{r}}_k^T A \mathbf{d}_j}{\mathbf{d}_j^T A \mathbf{d}_j}\, \mathbf{d}_j,$$

(3) *The new iterate is*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\widetilde{\mathbf{r}}_k^T \mathbf{d}_k}{\mathbf{d}_k^T A \mathbf{d}_k}\, \mathbf{d}_k,$$

and

(4) *the corresponding residual is*

$$\mathbf{r}_{k+1} = \mathbf{b} - A\mathbf{x}_{k+1} = \mathbf{r}_k - \frac{\widetilde{\mathbf{r}}_k^T \mathbf{d}_k}{\mathbf{d}_k^T A \mathbf{d}_k}\, A\mathbf{d}_k.$$

We remark that it can be shown that $\mathbf{d}_k^T \widetilde{\mathbf{r}}_k = \mathbf{r}_k^T \widetilde{\mathbf{r}}_k$ and that the above algorithm computes at every step $k+1$ an iterate so that its residual is minimized in A^{-1}-norm along the most recent $m_k + 2$ search directions $\{A\mathbf{d}_j\}_{j=k-1-m_k}^{k}$. Since they span the preconditioned residual $A\widetilde{\mathbf{r}}_k$, the method is at least as accurate as the preconditioned steepest descent method that we described earlier.

CHAPTER 9

Smoothing rates of iterative methods and the *cascadic* MG

This lecture introduces and studies an optimal Chebyshev-like polynomial. Then, the so-called "cascadic" MG is introduced and analyzed based on properties of this polynomial.

1. An optimal Chebyshev–like polynomial

Consider the Chebyshev polynomials $T_k(t)$ defined by recursion as follows, $T_0 = 1$, $T_1(t) = t$ and for $k \geq 1$, $T_{k+1}(t) = 2tT_k(t) - T_{k-1}(t)$. Letting $t = \cos\alpha \in [-1, 1]$, we have the explicit representation $T_k(t) = \cos k\alpha$, which is seen from the trigonometric identity $\cos(k+1)\alpha + \cos(k-1)\alpha = 2\cos\alpha\cos k\alpha$.

We now prove some properties of T_k that will be needed in the analysis of two MG methods later on.

PROPOSITION 1.1. *We have the expansion $T_{2k+1}(t) = c_{2k+1}t + tQ_k(t^2)$, $c_{2k+1} = (-1)^k(2k+1)$, for $k \geq 0$, where Q_k is a polynomial of degree k such that $Q(0) = 0$. Similarly, $T_{2k}(t) = (-1)^k + P_k(t^2)$, where P_k is a polynomial of degree k such that $P_k(0) = 0$.*

PROOF. We have $T_1 = t$, $T_2 = 2tT_1 - T_0 = 2t^2 - 1$, and $T_3 = 2tT_2 - T_1 = 2t(2t^2-1) - t = 4t^3 - 3t$. That is, assume by induction that for $k \geq 1$, $T_{2k-1}(t) = c_{2k-1}t + tQ_{k-1}(t^2)$ and $T_{2k}(t) = (-1)^k + P_k(t^2)$ for some polynomials Q_{k-1} and P_k of respective degrees $k-1$ and k, and such that $Q_{k-1}(0) = 0$ and $P_k(0) = 0$. Then, from $T_{2k+1} = 2tT_{2k} - T_{2k-1}$, we get

$$\begin{aligned} T_{2k+1} &= 2t((-1)^k + P_k(t^2)) - (-1)^{k-1}(2k-1)t - tQ_{k-1}(t^2) \\ &= (-1)^k(2k+1)t + t\left(2P_k(t^2) - Q_{k-1}(t^2)\right). \end{aligned}$$

That is, the induction assumption for T_{2k+1} is confirmed with $Q_k(t) = 2P_k - Q_{k-1}$, and hence, $Q_k(0) = 0$. Similarly, for T_{2k+2}, we have

$$\begin{aligned} T_{2k+2} &= -T_{2k} + 2tT_{2k+1} \\ &= -(-1)^k - P_k(t^2) + 2t\left((-1)^k(2k+1)t + tQ_k(t^2)\right) \\ &= (-1)^{k+1} + \left(2(-1)^k(2k+1)t^2 + 2t^2Q_k(t^2) - P_k(t^2)\right). \end{aligned}$$

The latter confirms the induction assumption for T_{2k+2} with $P_{k+1}(t^2) = 2(-1)^k(2k+1)t^2 + 2t^2Q_k(t^2) - P_k(t^2)$ and hence $P_{k+1}(0) = 0$. □

PROPOSITION 1.2. *The following estimate holds for any $t \in [0, 1]$,*

$$|T_{2k+1}(t)| \leq (2k+1)t.$$

PROOF. Note that for $t = \cos\alpha \in [-1, 1]$, $|T_k(t)| = |\cos k\alpha| \leq 1$. Therefore, assuming by induction that $|T_{2k-1}(t)| \leq (2k-1)t$ for $t \in [0, 1]$, we have

$$|T_{2k+1}(t)| = |2tT_{2k}(t) - T_{2k-1}(t)| \leq 2t + (2k-1)t = (2k+1)t,$$

which confirms the induction assumption.

□

PROPOSITION 1.3. *For a given $b > 0$, consider for $t \in [0,\ b]$ the function*

$$\varphi_\nu(t) = (-1)^\nu \frac{1}{2\nu+1}\ \frac{\sqrt{b}}{\sqrt{t}}\ T_{2\nu+1}\left(\frac{\sqrt{t}}{\sqrt{b}}\right). \tag{2.42}$$

We have that $\varphi_\nu(t)$ is a polynomial of degree ν such that $\varphi_\nu(0) = 1$, that is, $\varphi_\nu(t) = 1 - tq_{\nu-1}(t)$ for some polynomial $q_{\nu-1}(t)$ of degree $\nu - 1$.

PROOF. For $\nu = 0$, $\varphi_\nu = 1$. Consider the case $\nu \geq 1$. Due to Proposition 1.1, we have with $\lambda = \sqrt{\frac{t}{b}} \in [0, 1]$, that $\varphi_\nu(t) = \frac{1}{c_{2\nu+1}} \frac{1}{\lambda}\ \lambda\left(c_{2\nu+1} + Q_\nu(\lambda^2)\right) = 1 - \lambda^2 q_{\nu-1}(\lambda^2)$, since $Q_\nu(0) = 0$ hence $\frac{1}{c_{2\nu+1}} Q_\nu(\lambda^2) = -\lambda^2 q_{\nu-1}(\lambda^2)$ for some polynomial $q_{\nu-1}(\lambda)$ of degree $\nu - 1$. That is, we showed that $\varphi_\nu(t)$ as defined in (2.42) is a polynomial of degree ν such that $\varphi_\nu(0) = 1$.

□

PROPOSITION 1.4. *The polynomial φ_ν defined in* (2.42) *has the following optimality property:*

$$\min_{p_\nu:\ p_\nu(0)=1}\ \max_{t\in[0,\ b]}\ |\sqrt{t}\ p_\nu(t)| = \max_{t\in[0,\ b]}\ |\sqrt{t}\ \varphi_\nu(t)| = \frac{\sqrt{b}}{2\nu+1}. \tag{2.43}$$

We have $\varphi_\nu(0) = 1$ and also

$$\max_{t\in[0,\ b]}\ |\varphi_\nu(t)| = 1. \tag{2.44}$$

PROOF. The first fact follows from the optimality property of the Chebyshev polynomials, since letting $\lambda = \sqrt{\frac{t}{b}} \in [0, 1]$ $\sqrt{t}\varphi_\nu(t)$ equals $T_{2\nu+1}(\lambda)$ times a constant.
The fact that $|\varphi_\nu(t)| \leq 1$ follows from Proposition 1.2.

□

Here are some particular cases of the polynomials φ_ν.
Using the definition of the Chebyshev polynomials, $T_0 = 1$, $T_1 = t$, $T_{k+1} = 2tT_k - T_{k-1}$, for $k \geq 1$, we get $T_2 = 2t^2 - 1$ and hence

$$T_3(t) = 4t^3 - 3t.$$

Thus,

$$\varphi_1(t) = -\frac{1}{3}\sqrt{b}(4\frac{t}{b^{\frac{3}{2}}} - \frac{3}{\sqrt{b}}) = 1 - \frac{4}{3}\ \frac{t}{b}.$$

This in particular shows that

$$\sup_{t\in(0,\ b]}\ \frac{|1-\varphi_1(t)|}{\sqrt{t}} = \frac{4}{3}\ \frac{1}{\sqrt{b}}.$$

The next polynomial is based on $T_5 = 2tT_4 - T_3 = 2t(2tT_3 - T_2) - T_3 = (4t^2 - 1)(4t^3 - 3t) - 4t^3 + 2t = 16t^5 - 20t^3 + 5t$. Therefore,

$$\varphi_2(t) = \frac{1}{5}\sqrt{\frac{b}{t}}\left(16\sqrt{t}t^2\frac{1}{b^{\frac{5}{2}}} - 20\sqrt{t}t\frac{1}{b^{\frac{3}{2}}} + 5\sqrt{t}\frac{1}{\sqrt{b}}\right).$$

This shows,

$$\varphi_2(t) = \frac{16}{5}\frac{t^2}{b^2} - 4\frac{t}{b} + 1.$$

We also have,

$$\sup_{t\in(0,\,b]} \frac{1-\varphi_2(t)}{\sqrt{t}} = \frac{4}{\sqrt{b}} \sup_{x\in(0,1]} (x - \frac{4}{5}x^3) = \frac{4}{3}\sqrt{\frac{5}{3}}\frac{1}{\sqrt{b}}.$$

In general, it is clear that the following result holds.

PROPOSITION 1.5. *There is a constant C_ν independent of b such that the following estimate holds,*

$$\sup_{t\in(0,\,b]} \frac{|1-\varphi_\nu(t)|}{\sqrt{t}} \le C_\nu \frac{1}{b^{\frac{1}{2}}}. \tag{2.45}$$

PROOF. We have, $1-\varphi_\nu(t) = tq_{\nu-1}(t)$, that is, $\frac{1-\varphi_\nu}{\sqrt{t}} = \sqrt{t}\; q_{\nu-1}(t)$ and therefore the quotient in question is bounded for $t \in (0, b]$. More specifically, the following dependence on b is seen:

$$\sup_{t\in(0,\,b]} \frac{|1-\varphi_\nu(t)|}{\sqrt{t}} = \frac{1}{b^{\frac{1}{2}}} \sup_{\lambda\in(0,\,1]} \frac{\left|1 - \frac{(-1)^\nu}{2\nu+1}\frac{T_{2\nu+1}(\sqrt{\lambda})}{\sqrt{\lambda}}\right|}{\sqrt{\lambda}}.$$

Clearly, the constant $C_\nu = \sup\limits_{\lambda\in(0,\,1]} \frac{\left|1 - \frac{(-1)^\nu}{2\nu+1}\frac{T_{2\nu+1}(\sqrt{\lambda})}{\sqrt{\lambda}}\right|}{\sqrt{\lambda}}$ is independent of b.

□

1.1. Application to smoothing rate estimates of the preconditioned CG method. Consider $A\mathbf{x} = \mathbf{b}$ where A is a s.p.d. matrix. Let also Λ be a s.p.d. preconditioner to A. The kth iterate $\mathbf{x}_k$, $k \ge 1$, of the preconditioned conjugate gradient (or PCG) method is characterized as certain best polynomial approximation to $\mathbf{x} = A^{-1}\mathbf{b}$. Introducing $\overline{A} = \Lambda^{-\frac{1}{2}} A \Lambda^{-\frac{1}{2}}$ and $\overline{\mathbf{x}} = \Lambda^{\frac{1}{2}}\mathbf{x}$ and $\overline{\mathbf{b}} = \Lambda^{-\frac{1}{2}}\mathbf{b}$, the standard convergence estimate of the CG method reads,

$$\|\mathbf{x}-\mathbf{x}_k\|_A = \|\overline{\mathbf{x}}-\overline{\mathbf{x}}_k\|_{\overline{A}} = \min_{p_k:\, p_k(0)=1} \|p_k(\overline{A})(\overline{\mathbf{x}}-\overline{\mathbf{x}}_k)\|_{\overline{A}} = \min_{p_k:\, p_k(0)=1} \|p_k(\overline{A})\Lambda^{\frac{1}{2}}(\mathbf{x}-\mathbf{x}_0)\|_{\overline{A}}.$$

In other words, we have

$$\|\mathbf{x}-\mathbf{x}_k\|_A = \min_{p_k:\, p_k(0)=1} \|\overline{A}^{\frac{1}{2}} p_k(\overline{A})\Lambda^{\frac{1}{2}}(\mathbf{x}-\mathbf{x}_0)\|.$$

Since the eigenvalues of the s,p.d. matrix $\overline{A}$ vary in the interval $(0,\ \|\overline{A}\|)$, we can use the polynomial in (2.42) with $b = \|\overline{A}\|$. This will give us the following estimate for the PCG method

$$\|\mathbf{x}-\mathbf{x}_k\|_A = \min_{p_k:\, p_k(0)=1} \|\overline{A}^{\frac{1}{2}} p_k(\overline{A})\Lambda^{\frac{1}{2}}(\mathbf{x}-\mathbf{x}_0)\| \le \max_{t\in(0,\,\|\overline{A}\|]} |\sqrt{t}\varphi_k(t)|\ \|\mathbf{x}-\mathbf{x}_0\|_\Lambda.$$

That is, we have the following, sometimes referred to as "smoothing rate" estimate of the PCG method:

$$\|\mathbf{x}-\mathbf{x}_k\|_A \le \frac{\|\overline{A}\|^{\frac{1}{2}}}{2k+1}\ \|\mathbf{x}-\mathbf{x}_0\|_\Lambda. \tag{2.46}$$

1.2. Smoothing rate estimates for stationary iterative methods. Estimates similar to the PCG smoothing rate estimate (2.46) can be derived for stationary iterative methods. Let M be a matrix that provides A-convergent iterations for $A\mathbf{x} = \mathbf{b}$. Equivalently, let $M + M^T - A$ be s.p.d. Define $\overline{L} = M \left(M + M^T - A\right)^{-\frac{1}{2}}$ and let $\overline{A} = \overline{L}^{-1} A \overline{L}^{-T}$ and $\overline{\mathbf{b}} = \overline{L}^{-1}\mathbf{b}$. Consider the iteration process

$$(\overline{L}^T \mathbf{x}_k) = (\overline{L}^T \mathbf{x}_{k-1}) + (\overline{\mathbf{b}} - \overline{A}(\overline{L}^T \mathbf{x}_{k-1}).$$

Its computationally feasible equivalent version reads

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \overline{L}^{-T}\overline{L}^{-1}\left(\mathbf{b} - A\mathbf{x}_{k-1}\right) = \mathbf{x}_{k-1} + \overline{M}^{-1}\left(\mathbf{b} - A\mathbf{x}_{k-1}\right).$$

Another, more familiar form of the above iteration reads

$$\begin{aligned} \mathbf{x}_{k-\frac{1}{2}} &= \mathbf{x}_{k-1} + M^{-1}(\mathbf{b} - A\mathbf{x}_k) \\ \mathbf{x}_k &= \mathbf{x}_{k-\frac{1}{2}} + M^{-T}(\mathbf{b} - A\mathbf{x}_{k-\frac{1}{2}}). \end{aligned}$$

Introducing $\overline{E} = I - \overline{A}$, since its eigenvalues are between zero and one, we have the following convergence estimate,

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}_k\|_A &= \|\overline{L}^T(\mathbf{x} - \mathbf{x}_k)\|_{\overline{A}} \\ &= \|\overline{E}^k \overline{L}^T(\mathbf{x} - \mathbf{x}_0)\|_{\overline{A}} \\ &= \|\overline{A}^{\frac{1}{2}}(I - \overline{A})^k\ \overline{L}^T(\mathbf{x} - \mathbf{x}_0)\| \\ &\leq \max_{t\in[0,\,1]}\ \sqrt{t}(1-t)^k\ \|\overline{L}^T(\mathbf{x} - \mathbf{x}_0)\| \\ &= \tfrac{1}{\sqrt{2k+1}}\left(1 - \tfrac{1}{2k+1}\right)^k\ \|\mathbf{x} - \mathbf{x}_0\|_{\overline{M}}. \end{aligned}$$

2. Cascadic Multigrid

We first describe the following two-grid algorithm.

ALGORITHM 2.1 (Two-grid cascadic algorithm). *Consider $A\mathbf{x} = \mathbf{b}$. Let $P : \mathbb{R}^{n_c} \mapsto \mathbb{R}^n$ be the interpolation matrix, $A_c = P^T A P$ the coarse matrix and Λ an s.p.d. preconditioner to A (such as symmetric Gauss–Seidel, $\overline{M}$ or the Jacobi one F). The two-grid cascadic algorithm computes an approximation $\mathbf{x}_{CTG}$ to the exact solution $\mathbf{x} = A^{-1}\mathbf{b}$ in the following steps:*

(i) *Solve the coarse-grid problem $A_c\mathbf{x}_c = P^T\mathbf{b}$.*
(ii) *Interpolate and compute the residual $\mathbf{r} = \mathbf{b} - AP\mathbf{x}_c = (I - \pi_A)^T\mathbf{b}$, where $\pi_A = PA_c^{-1}P^TA$ is the coarse-grid projection.*
(iii) *Apply $m \geq 1$ iterations by the preconditioned using conjugate gradient method (PCG) based on Λ to $A\mathbf{v} = \mathbf{r}$ with initial iterate $\mathbf{v}_0 = 0$. Let $\mathbf{v}_m$ be the resulting mth iterate.*
(iv) *Compute the cascadic TG approximation $\mathbf{x}_{CTG} = \mathbf{v}_m + P\mathbf{x}_c$.*

Using the smoothing property (2.46) of the PCG method, we have the following estimate

$$\|\mathbf{v} - \mathbf{v}_m\|_A \leq \frac{\|\overline{A}\|^{\frac{1}{2}}}{2m+1}\ \|\mathbf{v}\|_\Lambda$$

where $\mathbf{v}:\ A\mathbf{v} = \mathbf{r} = (I-\pi_A)^T\mathbf{b}$, that is $\mathbf{v} = A^{-1}(I-\pi_A)^T\mathbf{b} = (I-\pi_A)A^{-1}\mathbf{b} = (I-\pi_A)\mathbf{x}$. Therefore, the following estimate holds

$$\|\mathbf{v}-\mathbf{v}_m\|_A \leq \frac{\|\overline{A}\|^{\frac{1}{2}}}{2m+1}\ \|\Lambda\|^{\frac{1}{2}}\ \|(I-\pi_A)\mathbf{x}\|.$$

Assume now property (B), i.e., the ℓ_2–boundedness of the projection π_A

$$\|A\|\|(I-\pi_A)\mathbf{v}\|^2 \leq \eta_b\ \|\mathbf{v}\|_A^2.$$

The latter holds if the "strong approximation property" $\|\mathbf{v}-P\mathbf{v}_c\|_A^2 \leq \frac{\eta_a}{\|A\|}\ \|A\mathbf{v}\|^2$ holds which is the case for finite element matrices coming from boundary value problems for Laplace operator posed on convex polygonal domain.

Under the assumption of ℓ_2–boundedness of the projection π_A, we arrive at the estimate

$$\|\mathbf{v}-\mathbf{v}_m\|_A \leq \frac{\|\overline{A}\|}{2m+1}\left(\frac{\|\Lambda\|\overline{\eta}_b}{\|A\|}\right)^{\frac{1}{2}}\ \|(I-\pi_A)\mathbf{x}\|_A.$$

The final estimate reads,

$$\begin{aligned}\|\mathbf{x}-\mathbf{x}_{CTG}\|_A &= \|(\mathbf{x}-P\mathbf{x}_c)-\mathbf{v}_m\|_A\\ &= \|\mathbf{v}-\mathbf{v}_m\|_A\\ &\leq \tfrac{\sqrt{\overline{\eta}_b}}{2m+1}\ \|(I-\pi_A)\mathbf{x}\|_A.\end{aligned}$$

where

$$\overline{\eta}_b = \eta_b\ \|\Lambda^{-\frac{1}{2}}A\Lambda^{-\frac{1}{2}}\|\ \frac{\|\Lambda\|}{\|A\|}.$$

The multilevel version of the method takes the following form.

ALGORITHM 2.2 (Multilevel Cascadic MG). *Let $\overline{P}_k:\ \mathbb{R}^{n_k}\mapsto\mathbb{R}^{n_\ell}$ be the composite interpolants, i.e., $\overline{P}_k = P_\ell\ \ldots\ P_k$. The coarse matrices are $A_k = \overline{P}_k^TA\overline{P}_k = P_k^TA_{k+1}P_k$ and $A = A_\ell$ is the fine-grid matrix, whereas A_0 is the coarse matrix at the initial level $k=0$. Let Λ_k be s.p.d. preconditioner for A_k such as symmetric Gauss–Seidel, or Jacobi matrix coming from A_k. The multilevel cascadic MG algorithm computes $\mathbf{x}_{CMG}$ as an approximation to the exact solution of $A\mathbf{x}=\mathbf{b}$ for a given r.h.s. $\mathbf{b}$ in the following steps:*

(i) *Let $\mathbf{r}_\ell = \mathbf{b}$ and for $k=\ell,\ \ldots\ 1$ compute $\mathbf{b}_{k-1} = P_{k-1}^T\mathbf{b}_k$.*
(ii) *Solve $A_0\mathbf{x}_0 = \mathbf{b}_0$,*
(iii) *For $k=1,\ \ldots,\ \ell$ with initial iterate $\mathbf{x}_k^{(0)} = P_{k-1}\mathbf{x}_{k-1}$ perform $m=m_k$ PCG iterations for computing an approximate solution to the kth level coarse system $A_k\widehat{\mathbf{x}}_k = \mathbf{b}_k$. Then let $\mathbf{x}_k = \mathbf{x}_k^{(m)}$.*
(iv) *The cascadic multigrid (CMG) approximation is $\mathbf{x}_{CMG} = \mathbf{x}_\ell$.*

In what follows we need to estimate the difference $P_{k-1}\widehat{\mathbf{x}}_{k-1} - \widehat{\mathbf{x}}_k$. Since $A_{k-1}\widehat{\mathbf{x}}_{k-1} = \mathbf{b}_{k-1} = P_{k-1}^T\mathbf{b}_k = P_{k-1}^TA_k\widehat{\mathbf{x}}_k$, we obtain

$$P_{k-1}\widehat{\mathbf{x}}_{k-1} - \widehat{\mathbf{x}}_k = (P_{k-1}A_{k-1}^{-1}P_{k-1}^TA_k - I)\widehat{\mathbf{x}}_k = -(I-\pi_{k-1}^k)\widehat{\mathbf{x}}_k.$$

Above, π_{k-1}^k is the two-level A_k–based projection. We assume the uniform in k, ℓ_2-boundedness of π_{k-1}^k (cf. Assumption (B))

$$\|A_k\|\|(I-\pi_{k-1}^k)\mathbf{v}\|^2 \leq \eta_b\ \|\mathbf{v}\|_{A_k}^2.$$

Using the above estimate for $\mathbf{v} = (I - \pi_{k-1}^k)\widehat{\mathbf{x}}_k$, we obtain the desired auxiliary estimate

$$\|P_{k-1}\widehat{\mathbf{x}}_{k-1} - \widehat{\mathbf{x}}_k\| \leq \frac{\sqrt{\eta_b}}{\|A_k\|^{\frac{1}{2}}}\, \|P_{k-1}\widehat{\mathbf{x}}_{k-1} - \widehat{\mathbf{x}}_k\|_{A_k}. \tag{2.47}$$

Using the best polynomial approximation property of the PCG method and the special polynomial $p_m(t) = \varphi_m(t)$ for the interval $(0,\ \|\overline{A}_k\|]$, assuming the uniform in $k \geq 0$ bound

$$\eta_b\, \|\overline{A}_k\|\, \frac{\|\Lambda_k\|}{\|A_k\|} \leq \overline{\eta}_b,$$

the following estimate for the CMG approximation is obtained (based also on (2.47)),

$$\begin{aligned}
\|\mathbf{x}_k^{(m)} - \widehat{\mathbf{x}}_k\|_{A_k} &= \min_{p_m:\, p_m(0)=1} \|p_m(\overline{A}_k)\Lambda_k^{\frac{1}{2}}(P_{k-1}\mathbf{x}_{k-1} - \widehat{\mathbf{x}}_k)\|_{\overline{A}_k} \\
&\leq \tfrac{\sqrt{\overline{\eta}_b}}{2m+1}\, \|P_{k-1}\widehat{\mathbf{x}}_{k-1} - \widehat{\mathbf{x}}_k\|_{A_k} + \|\Lambda_k^{\frac{1}{2}}(P_{k-1}\widehat{\mathbf{x}}_{k-1} - P_{k-1}\mathbf{x}_{k-1})\|_{\overline{A}_k} \\
&= \tfrac{\sqrt{\overline{\eta}_b}}{2m+1}\, \|\overline{P}_k(P_{k-1}\widehat{\mathbf{x}}_{k-1} - \widehat{\mathbf{x}}_k)\|_A + \|P_{k-1}(\widehat{\mathbf{x}}_{k-1} - \mathbf{x}_{k-1})\|_{A_k} \\
&= \tfrac{\sqrt{\overline{\eta}_b}}{2m+1}\, \|\overline{P}_k(P_{k-1}\widehat{\mathbf{x}}_{k-1} - \widehat{\mathbf{x}}_k)\|_A + \|\mathbf{x}_{k-1} - \widehat{\mathbf{x}}_{k-1}\|_{A_{k-1}} \\
&= \tfrac{\sqrt{\overline{\eta}_b}}{2m+1}\, \|\overline{P}_{k-1}\widehat{\mathbf{x}}_{k-1} - \overline{P}_k\widehat{\mathbf{x}}_k\|_A + \|\widehat{\mathbf{x}}_{k-1} - \mathbf{x}_{k-1}\|_{A_{k-1}}.
\end{aligned} \tag{2.48}$$

Introduce the projections $\pi_k = \overline{P}_k A_k^{-1}\overline{P}_k^T A$. Then

$$\pi_k\mathbf{x} = \pi_k A^{-1}\mathbf{b} = \overline{P}_k A_k^{-1}\overline{P}_k^T\mathbf{b} = \overline{P}_k A_k^{-1}\mathbf{b}_k = \overline{P}_k\widehat{\mathbf{x}}_k.$$

Therefore the preceding estimate (2.48) reads

$$\|\mathbf{x}_k - \widehat{\mathbf{x}}_k\|_{A_k} = \|\mathbf{x}_k^{(m)} - \widehat{\mathbf{x}}_k\|_{A_k} \leq \frac{\sqrt{\overline{\eta}_b}}{2m+1}\, \|(\pi_k - \pi_{k-1})\mathbf{x}\|_A + \|\mathbf{x}_{k-1} - \widehat{\mathbf{x}}_{k-1}\|_{A_{k-1}}.$$

Now, using recursion with $m = m_k$ and the fact that $\mathbf{x}_0 = \widehat{\mathbf{x}}_0$, we end up with the estimate

$$\begin{aligned}
\|\mathbf{x} - \mathbf{x}_{CMG}\|_A &\leq \sqrt{\overline{\eta}_b}\, \sum_{k=1}^{\ell} \tfrac{1}{2m_k+1}\, \|(\pi_k - \pi_{k-1})\mathbf{x}\|_A \\
&\leq \sqrt{\overline{\eta}_b}\, \left(\sum_{k=1}^{\ell} \tfrac{1}{(2m_k+1)^2}\right)^{\frac{1}{2}} \left(\sum_k \|(\pi_k - \pi_{k-1})\mathbf{x}\|_A^2\right)^{\frac{1}{2}}.
\end{aligned}$$

As a corollary, if we use m_k smoothing PCG iterations at level k that satisfy the geometric rule

$$2m_k + 1 = \mu^{\ell-k}\, (2m+1),$$

for a given $m \geq 1$ and a $\mu \geq 2$, we end up with the following final estimate

$$\|\mathbf{x} - \mathbf{x}_{CMG}\|_A \leq \frac{\mu}{\sqrt{\mu^2-1}} \frac{\sqrt{\overline{\eta}_b}}{2m+1}\, \left(\sum_{k=1}^{\ell} \|(\pi_k - \pi_{k-1})\mathbf{x}\|_A^2\right)^{\frac{1}{2}} = \frac{\mu}{\sqrt{\mu^2-1}}\, \frac{\sqrt{\overline{\eta}_b}}{2m+1}\, \|\mathbf{x}\|_A.$$

If we run stationary iterations to define $\mathbf{x}_k = \mathbf{x}_k^{(m)}$ the starting estimate reads

$$\begin{aligned}
\|\mathbf{x}_k^{(m)} - \widehat{\mathbf{x}}_k\|_{A_k} &= \|\overline{E}_k^m \Lambda_k^{\frac{1}{2}}(P_{k-1}\mathbf{x}_{k-1} - \widehat{\mathbf{x}}_k)\|_{\overline{A}_k} \\
&\le \|\overline{E}_k^m \Lambda_k^{\frac{1}{2}}(P_{k-1}\widehat{\mathbf{x}}_{k-1} - \widehat{\mathbf{x}}_k)\|_{\overline{A}_k} \\
&\quad + \|\overline{E}_k^m \Lambda_k^{\frac{1}{2}}(P_{k-1}\widehat{\mathbf{x}}_{k-1} - P_{k-1}\mathbf{x}_{k-1})\|_{\overline{A}_k} \\
&\le \|\overline{E}_k^m \Lambda_k^{\frac{1}{2}}(P_{k-1}\widehat{\mathbf{x}}_{k-1} - \widehat{\mathbf{x}}_k)\|_{\overline{A}_k} \\
&\quad + \|\Lambda_k^{\frac{1}{2}}(P_{k-1}\widehat{\mathbf{x}}_{k-1} - P_{k-1}\mathbf{x}_{k-1})\|_{\overline{A}_k} \\
&\le \tfrac{1}{\sqrt{2m+1}} \, \|\Lambda_k^{\frac{1}{2}}(P_{k-1}\widehat{\mathbf{x}}_{k-1} - \widehat{\mathbf{x}}_k)\| \\
&\quad + \|\widehat{\mathbf{x}}_{k-1} - \mathbf{x}_{k-1}\|_{A_{k-1}} \\
&\le \tfrac{\sqrt{\overline{\eta}_b}}{\sqrt{2m+1}} \, \|(\pi_k - \pi_{k-1})\mathbf{x}\|_A + \|\mathbf{x}_{k-1} - \widehat{\mathbf{x}}_{k-1}\|_{A_{k-1}}.
\end{aligned}$$

The final estimate translates to

$$\|\mathbf{x} - \mathbf{x}_{CMG}\|_A \le \frac{\sqrt{\overline{\eta}_b \mu/(\mu-1)}}{\sqrt{2m+1}} \left(\sum_{k=1}^{\ell} \|(\pi_k - \pi_{k-1})\mathbf{x}\|_A^2 \right)^{\frac{1}{2}} = \frac{\sqrt{\overline{\eta}_b \mu/(\mu-1)}}{\sqrt{2m+1}} \, \|\mathbf{x}\|_A.$$

Here we have assumed that the smoothing iterations m_k vary according to the rule

$$2m_k + 1 = \mu^{\ell-k}(2m+1).$$

Complexity of CMG. Assuming that $n_k \simeq 2^d n_{k-1}$ ($d = 2$ or 3) and $m_k \simeq \mu^{\ell-k}\, m$, the complexity of the CMG is readily estimated to be of order

$$\sum_k n_k m_k \simeq n_\ell m \, \sum_k \left(\frac{\mu}{2^d}\right)^{\ell-k} \simeq \mathcal{O}(n_\ell)$$

if $2 \le \mu < 2^d$, which we can always satisfy if $d \ge 2$.

Part 3

Algebraic MG: main principles and algorithms for finite element problems

CHAPTER 10

Algebraic MG: coarse degrees of freedom and interpolation matrices

This lecture introduces the main principles of algebraic multigrid method (or AMG). In particular, we provide arguments for construction of interpolation matrices, both the selection of their domain of definition (or the selection of coarse-grid dofs) and the computation of the actual interpolation weights (or entries of P). We describe a classical choice of the coarse-dofs as subset of fine-grid dofs, the notion of compatible relaxation and the energy boundedness of the of a hierarchical coarse-grid projection and their role of proving weak approximation property and hence TG convergence. We also describe a spectral way of selecting coarse dofs and present two important examples, one coming from finite element discretization and another one handling fairly general matrices utilizing a least–squares approach and leading to a "strong approximation property" hence providing TG convergence that improves with increasing the number of smoothing steps.

1. Algebraic MG (or AMG) as an "*inverse problem*"

Consider a linear system of equations $A\mathbf{x} = \mathbf{b}$ with a sparse s.p.d. matrix which, we may not have knowledge about its origin. Since the MG methods have proven optimal convergence properties, we may want to utilize the MG principle to design preconditioners for A. For example, to design a two–grid preconditioner (and then continue by recursion), we need to construct an interpolation matrix P. This involves, in particular, the selection of the domain of definition of P, which is referred to as "*coarse degrees of freedom*" (or coarse dofs). Then, we need to build the actual mapping P. In matrix notation, this means that we need to select the number of columns of P, the sparsity pattern of P, i.e., the number of nonzero entries of for each row of P, their positions and finally the actual entries of P. Using geometric MG language, we need to choose for each fine-grid dof (row of P) a coarse-grid neighborhood (i.e., the column indices of P corresponding to the non-zero entries in that fine-grid row) and the actual interpolation weights (the non-zero entries of P at the corresponding positions).

Once, a P has been constructed, the coarse–grid matrix A_c is typically defined variationally, i.e., $A_c = P^T A P$. Since, we want to apply the same construction recursively, we want A_c to have similar properties as A (however being with smaller size). At the minimum, we want that A_c be sparse. This imposes the requirement on P to be sparse as well. That is, each fine–grid dof should interpolate from a bounded number of coarse dofs.

We know, that a two-grid (and MG for that matter) to be successful, a balance between the smoother M and the coarse-space Range (P) must be established, in the sense, that the coarse space should ensure a "*weak approximation property*":

For each fine-grid vector $\mathbf{v}$ there is a coarse–grid interpolant $P\mathbf{v}_c$ such that

$$\|\mathbf{v} - P\mathbf{v}_c\|_{\widetilde{M}} \leq \eta_w \, \|\mathbf{v}\|_A, \tag{3.1}$$

for a constant η_w independent of $\mathbf{v}$. Here, $\widetilde{M} = M^T(M+M^T-A)^{-1}M$ is the symmetrized smoother corresponding to an A-convergent iteration matrix M^T.

Traditionally, to define a TG method, i.e., to construct a P, the smoother M is pre–selected to provide a convergent iterative method (in A-norm), in the simplest cases like weighted Jacobi, Gauss–Seidel, incomplete factorization matrices (in the M–matrix case), overlapping Schwarz methods (block Gauss–Seidel with small overlapping blocks), etc.

Then, given M (and A), P is constructed so that the target is to ensure the weak approximation estimate (3.1). It is clear that this procedure is an "*inverse problem*" and as any inverse problem it is "*ill–posed*". The latter means that there is not a unique solution to this task. Part of the problem is that many coarse spaces (or equivalently, many interpolation matrices) can lead to equally good (or bad) TG methods. The least rigorous part in the construction of P is the choice of the coarse degrees of freedom. At any rate, all resulting procedures that lead to a P to be used in a two-grid iteration process are commonly refereed to as "algebraic" two–grid or algebraic MG, and abbreviated as AMG. Originally, the AMG concept was proposed by Achi Brandt, Steve McCormick and John Ruge in 1982.

Choosing coarse dofs to be subset of fine-grid dofs. A typical case (as originally proposed) is to have a mapping represented by a rectangular $n_c \times n$ matrix R which represents the embedding $\mathcal{N}_c = R\mathcal{N} \subset \mathcal{N}$. Under proper ordering (the coarse dofs ordered last) the matrix R admits the following block form

$$R = [0, \; I] \, .$$

Then a natural assumption is to have the interpolation matrix P being identity at the coarse dofs, i.e.,

$$P = \begin{bmatrix} W \\ I \end{bmatrix} \begin{matrix} \} \, \mathcal{N}_f \equiv \mathcal{N} \setminus \mathcal{N}_c \\ \} \, \mathcal{N}_c \end{matrix} .$$

It is clear then that $RP = I$ and hence the mapping $Q = PR$ being a projection. Indeed $Q^2 = P(RP)R = PR = Q$.

Weak approximation property and compatible relaxation. We recall that if D is an s.p.d. matrix, spectrally equivalent to the symmetrized smoother $\widetilde{M}$, then a *necessary* condition for the TG convergence is the *weak approximation property*

$$\|(I - \pi_D)\mathbf{v}\|_D \leq \eta_w \, \|\mathbf{v}\|_A.$$

Note that the projection $Q = PR$ can be seen as an analog of the "best" one, $\pi_D = PR_*$, where $R_* = (P^TDP)^{-1}P^TD$. We note, that R_* is not sparse in general. A "good" choice of the coarse dofs would correspond to a fast decay of the inverse of P^TDP, so that it can be approximated by a sparse matrix, hence the choice PR for proving a weak approximation property of the form

$$\|(I - PR)\mathbf{v}\|_D \leq \eta_w \, \|\mathbf{v}\|_A, \tag{3.2}$$

would be justified (in the simplest cased when D is diagonal). We note that this is only a sufficient condition for TG convergence, since it implies $\|(I-\pi_D)\mathbf{v}\|_D = \min_{\mathbf{w}_c}\|\mathbf{v} - P\mathbf{w}_c\|_D \le \|\mathbf{v}-PR\mathbf{v}\|_D \le \eta_w \|\mathbf{v}\|_A$.

Assume (3.2) for a s.p.d. D that is spectrally equivalent to $\widetilde{M}$, i.e.,

$$c_1\mathbf{v}^T D\mathbf{v} \le \mathbf{v}^T\widetilde{M}\mathbf{v} \le c_2\ \mathbf{v}^T D\mathbf{v}. \tag{3.3}$$

First, we show that D and A are spectrally equivalent, when restricted to the set $\mathcal{N}_f = \mathcal{N}\setminus\mathcal{N}_c$, the hierarchical complement of the set of coarse dofs $\mathcal{N}_c$. This is trivially seen by choosing $\mathbf{v} = \begin{bmatrix}\mathbf{v}_f\\ 0\end{bmatrix}$. We have then $PR\mathbf{v}=0$, and hence (3.2) takes the form

$$\mathbf{v}_f^T D_{ff}\mathbf{v}_f \le \eta_w^2 \mathbf{v}_f^T A_{ff}\mathbf{v}_f. \tag{3.4}$$

where D_{ff} comes from the block partitioning of $D = \begin{bmatrix} D_{ff} & D_{fc}\\ D_{cf} & D_{cc}\end{bmatrix}\begin{matrix}\}\,\mathcal{N}_f\\ \}\,\mathcal{N}_c\end{matrix}$ – the same as for A. Using finally the fact that $\widetilde{M}$ comes from an A–convergent smoother M^T for A, we have $\mathbf{v}^T A\mathbf{v} \le \mathbf{v}^T\widetilde{M}\mathbf{v}$, which together with (3.3) implies

$$\mathbf{v}^T A\mathbf{v} \le c_2\mathbf{v}^T D\mathbf{v}. \tag{3.5}$$

Using (3.5) for $\mathbf{v} = \begin{bmatrix}\mathbf{v}_f\\ 0\end{bmatrix}$ gives the desired upper bound

$$\mathbf{v}_f^T A_{ff}\mathbf{v}_f \le c_2\mathbf{v}_f^T D_{ff}\mathbf{v}_f. \tag{3.6}$$

That is, both (3.4) and (3.6) represent the fact that $D_{ff}^{-1}A_{ff}$ is well–conditioned, and is sometimes referred to as "*compatible relaxation*" (or CR), a concept introduced by Achi Brandt in 2000.

Energy boundedness of the projection PR. The weak approximation property (3.2) and the fact that D is spectrally equivalent to $\widetilde{M}$ imply that $I-PR$ is bounded in energy.

To show this result we note first, that since $\widetilde{M}$ comes from an A-convergent smoother M^T which together with the spectral equivalence of D and $\widetilde{M}$ show estimate (3.5). Using the latter estimate combined with the weak approximation property (3.2) gives

$$\|(I-PR)\mathbf{v}\|_A \le \|(I-PR)\mathbf{v}\|_{\widetilde{M}} \le \sqrt{c_2}\ \|(I-PR)\mathbf{v}\|_D \le \sqrt{c_2}\ \eta_w\|\mathbf{v}\|_A.$$

That is, the desired result follows (using also Kato's lemma)

$$\|PR\|_A = \|I-PR\|_A \le \sqrt{c_2}\ \eta_w. \tag{3.7}$$

Energy boundedness of PR and CR imply TG convergence. We show next that the energy boundedness of PR and good CR bound κ_{ff} defined below (see also (3.6))

$$c_2 \ge \lambda_{max}(D_{ff}^{-1}A_{ff}) \ge \lambda_{\min}\,(D_{ff}^{-1}A_{ff}) \ge \kappa_{ff}, \tag{3.8}$$

imply weak approximation property of the form (3.2) with

$$\eta_w \le \frac{1}{\sqrt{\kappa_{ff}}}\,\|PR\|_A,$$

and hence a TG convergence holds with $\varrho_{TG} = 1 - \frac{1}{K_{TG}}$ where

$$K_{TG} \leq c_2\, \eta_w^2 \leq c_2 \lambda_{\min}^{-1}(D_{ff}^{-1} A_{ff})\, \|PR\|_A^2 \leq \frac{c_2}{\kappa_{ff}}\, \|PR\|_A^2. \tag{3.9}$$

Since $(I - PR)\mathbf{v} = \begin{bmatrix} \overline{\mathbf{v}}_f \\ 0 \end{bmatrix}$ first using the spectral equivalence between D_{ff} and A_{ff}, and then the norm bound of $I - PR$ (which is the same as for PR), we have

$$\|(I - PR)\mathbf{v}\|_D \leq \lambda_{\max}^{\frac{1}{2}}(D_{ff} A_{ff}^{-1})\, \|(I - PR)\mathbf{v}\|_A \leq \lambda_{\min}^{-\frac{1}{2}}(D_{ff}^{-1} A_{ff})\, \|PR\|_A\, \|\mathbf{v}\|_A.$$

That is letting $\eta_w = \lambda_{\min}^{-\frac{1}{2}}(D_{ff}^{-1} A_{ff})\, \|PR\|_A \leq \frac{\|PR\|_A}{\sqrt{\kappa_{ff}}}$ the desired estimate (3.2) holds.

2. Heuristic algorithms for coarse-grid selection

Assume that an initial set of coarse dofs $\mathcal{N}_c$ has been selected. To test if the respective lower CR bound κ_{ff} is acceptable, we run the "source" iteration: For any $i_f \in \mathcal{N}_f \equiv \mathcal{N} \setminus \mathcal{N}_c$, perform PCG iteration to solve the problem

$$A_{ff}\mathbf{x}_f = \mathbf{e}_{i_f},$$

where the vector $\mathbf{e}_{i_f}$ has a single non–zero entry at position $i_f \in \mathcal{N}_f$. For a preconditioner we use D_{ff} coming from a matrix D that has sparse inverse. Since the iterates have the form $p(D_{ff}^{-1} A_{ff})\mathbf{e}_{i_f}$ for a polynomial p, it is clear that they will have non-zero entries only in a neighborhood around the fine-dof i_f. It is clear also that if $D_{ff}^{-1} A_{ff}$ is well-conditioned then these iterates will have fast decay away from position i_f. If this is not the case, the dof i_f is a candidate to be added to the set of coarse dofs $\mathcal{N}_c$. Among all such candidates we select a subset (according to some criterion) and augment $\mathcal{N}_c$ with it, and then repeat the process until we are satisfied with the resulting decay.

Research tasks. Design a criterion for measuring decay rates and a criterion for selecting coarse dofs from a candidate set. Use so–called maximal independent set algorithms.

3. Algorithms for computing P

We proved that a sufficient condition for TG convergence is to have PR bounded in energy. Assuming that we have selected the coarse dofs set $\mathcal{N}_c$ and the sparsity pattern of the columns of P, i.e., the fine-dof neighbors that a given coarse dof interpolates to, we need to compute the actual entries corresponding to the selected sparsity pattern. In other words if $P = \big[\boldsymbol{\psi}_1,\ \boldsymbol{\psi}_2,\ \ldots,\ \boldsymbol{\psi}_{n_c}\big]$, and the columns $\boldsymbol{\psi}_{i_c}$ of P have prescribed support sets Ω_{i_c} of fine–grid dofs. Let $I_{i_c} = \begin{bmatrix} 0 \\ I \\ 0 \end{bmatrix} \Big\} \Omega_{i_c}$ be extension of vectors $\overline{\mathbf{v}}_{i_c}$ defined on Ω_{i_c} by zero outside Ω_{i_c} to vectors $\mathbf{v}_{i_c}$ defined on $\mathcal{N}$.

From the weak approximation property (with $D = \|A\|\, I$)

$$\|A\|^{\frac{1}{2}}\, \|(I - PR)\mathbf{v}\| \leq \eta_w\, \|\mathbf{v}\|_A,$$

it is clear that if $\mathbf{v}$ is a near-null vector of A, i.e., $\|\mathbf{v}\|_A \approx 0$, then $PR\mathbf{v} \approx \mathbf{v}$. Thus, a heuristic approach to construct P is to have $PR\mathbf{1} = \mathbf{1}$ for any near–null vector $\mathbf{1}$ of A.

Such vectors are sometimes referred to as "*algebraically smooth vectors*". Since we also want PR be bounded in energy, a sufficient condition for this is to minimize the following quadratic functional

$$\sum_{i_c} \boldsymbol{\psi}_{i_c}^T A \boldsymbol{\psi}_{i_c} \mapsto \min,$$

subject to $\sum_{i_c} \boldsymbol{\psi}_{i_c} = \mathbf{1}$. Note that the above quadratic expression is the trace of the matrix $P^T A P$, and its square root defines a matrix norm that we use as a more computationally feasible approximation to the desired A–norm of PR.

Since $\boldsymbol{\psi}_{i_c} = I_{\Omega_{i_c}} \overline{\psi}_{i_c}$, the above constrained minimization problem can be rewritten as follows. Introduce the local matrices $A_{\Omega_{i_c}} = I_{\Omega_{i_c}}^T A I_{\Omega_{i_c}}$. Then

$$\sum_{i_c} \boldsymbol{\psi}_{i_c}^T A \boldsymbol{\psi}_{i_c} = \sum_{i_c} \overline{\boldsymbol{\psi}}_{i_c}^T A_{\Omega_{i_c}} \overline{\boldsymbol{\psi}}_{i_c} \mapsto \min .$$

Forming the Lagrangian

$$\mathcal{L}((\underline{\boldsymbol{\psi}}_{i_c}),\ \boldsymbol{\lambda}) = \frac{1}{2} \sum_{i_c} \overline{\boldsymbol{\psi}}_{i_c}^T A_{\Omega_{i_c}} \overline{\boldsymbol{\psi}}_{i_c} - \boldsymbol{\lambda}^T \left(\mathbf{1} - \sum_{i_c} I_{\Omega_{i_c}} \overline{\boldsymbol{\psi}}_{i_c} \right) \mapsto \min,$$

and minimizing it leads to the following saddle–point system

$$\begin{array}{ll} A_{\Omega_{i_c}} \overline{\boldsymbol{\psi}}_{i_c} + I_{\Omega_{i_c}}^T \boldsymbol{\lambda} & = 0, \text{ for } i_c = 1, \ \ldots, \ n_c, \\ \sum\limits_{i_c} I_{\Omega_{i_c}} \overline{\boldsymbol{\psi}}_{i_c} & = \mathbf{1}. \end{array}$$

To solve the above saddle–point problem we introduce the local matrices $T_{i_c} = I_{\Omega_{i_c}} A_{\Omega_{i_c}}^{-1} I_{\Omega_{i_c}}^T$ and let $T = \sum_{i_c} T_{i_c}$. We have then

$$\overline{\boldsymbol{\psi}}_{i_c} = -A_{\Omega_{i_c}}^{-1} I_{\Omega_{i_c}}^T \boldsymbol{\lambda},$$

which used in the second equation above gives

$$\mathbf{1} = -\sum_{i_c} I_{\Omega_{i_c}} A_{\Omega_{i_c}}^{-1} I_{\Omega_{i_c}}^T \boldsymbol{\lambda} = -T \boldsymbol{\lambda}.$$

Hence $\boldsymbol{\lambda} = -T^{-1}\mathbf{1}$ and therefore

$$\boldsymbol{\psi}_{i_c} = I_{\Omega_{i_c}} \overline{\boldsymbol{\psi}}_{i_c} = T_{i_c} T^{-1} \mathbf{1}.$$

It is clear that $\sum_{i_c} \boldsymbol{\psi}_{i_c} = \mathbf{1}$.

We note that to compute the vector $T^{-1}\mathbf{1}$ in practice, we can use the PCG method with preconditioner the diagonal matrix $\Lambda = \sum_{i_c} I_{\Omega_{i_c}} \Lambda_{i_c} I_{\Omega_{i_c}}^T$ coming from the diagonals Λ_{i_c} of T_{i_c}. The local matrices $A_{\Omega_{i_c}}^{-1}$ are explicitly computed, hence T_{i_c} are explicitly available.

Finally, we comment on the choice of the vector $\mathbf{1}$. As mentioned above, it corresponds to an approximation to the minimal eigenvector of $D^{-1}A$, or more generally to a vector $\mathbf{1}$ corresponding to $\lambda_{\min}(\widetilde{M}^{-1}A)$. For matrices coming from finite element discretization

of elliptic PDEs (like Laplacian), a common choice is the constant vector $\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$, or a vector obtained by applying $m \geq 1$ times the iteration matrix $(I - \widetilde{M}^{-1}A)$ to it, i.e., $(I - \widetilde{M}^{-1}A)^m \mathbf{1}$. Since $A\mathbf{1}$ is zero except near the boundary of the domain, the smoothed version of $\mathbf{1}$ differs from it, only within a strip near the domain boundary (assuming diagonal or sparse $\widetilde{M}^{-1}$).

For more general applications, the vector $\mathbf{1}$ is more difficult to compute and in general we may need more than one vector to design a successful AMG method.

4. Spectral choice of coarse dofs

In some applications, the s.p.d. matrix A defines a quadratic form $\mathbf{v}^T A\mathbf{v}$ which can be assembled from local quadratic forms $\mathbf{v}_\tau^T A_\tau \mathbf{v}_\tau$. More specifically, let $\{\tau\}_{\tau\in\mathcal{T}}$ be an overlapping partition of relatively small sets τ that cover the set $\mathcal{N}$ of fine-grid dofs. The matrices A_τ act on vectors $\mathbf{v}_\tau$ defined on the local sets τ. Introduce the extension matrices I_τ. Then any fine–grid vector $\mathbf{v}$ restricted to τ can be represented as $\mathbf{v}_\tau = \mathbf{v}|_\tau = I_\tau^T \mathbf{v}$. We assume

$$\mathbf{v}^T A\mathbf{v} = \sum_\tau (I_\tau^T \mathbf{v})^T A_\tau I_\tau^T \mathbf{v} = \sum_\tau \mathbf{v}_\tau^T A_\tau \mathbf{v}_\tau.$$

We also assume that the local matrices A_τ are symmetric positive semi–definite.

Solve now the local eigenvalue problems

$$A_\tau \mathbf{q}_k = \lambda_k \, \mathbf{q}_k, \; k = 1, \; \ldots, \; n_\tau.$$

Note that the eigenvalues λ_k are non-negative.

For a given tolerance $\theta \in (0, 1)$, we choose $n_\tau^c \leq n_\tau$ such that for $k > n_\tau^c$, we have $\lambda_k > \theta \, \lambda_{\max} = \lambda_{n_\tau}$. We define then the local interpolation matrices $P_\tau = \left[\mathbf{q}_1, \; \ldots, \; \mathbf{q}_{n_\tau^c}\right]$. To define a global interpolation matrix, we need some diagonal matrices W_τ to be used as weights in what follows. The diagonal matrices W_τ have non-negative entries and are such that

$$\sum_\tau I_\tau W_\tau I_\tau^T = I.$$

The latter property is called "*partition of unity*" property.

To define the global interpolation matrix, we first introduce the sets τ_c consisting of the indices $1, \; \ldots, \; n_{\tau_c}$, corresponding to the eigenvalues λ_k for $k \leq n_\tau^c$. The set $\mathcal{N}_c$, is the union of all τ_c with their entries renumbered with global indices from one to $n_c = \sum_\tau n_\tau^c$. Let I_{τ_c} be the mapping that implements the local-to-global numbering of the coarse dofs in each τ_c.

The global interpolation matrix P takes then the form:

$$P = \sum_\tau \; I_\tau W_\tau P_\tau I_{\tau_c}^T.$$

Since each coarse vector $\mathbf{v}_c$ has block components $\mathbf{v}_{\tau_c}^c$, where $\mathbf{v}_{\tau_c}^c = I_{\tau_c}^T \mathbf{v}_c$, the actions of P are computed as follows

$$P\mathbf{v}_c = \sum_\tau I_\tau W_\tau P_\tau \mathbf{v}_{\tau_c}^c.$$

If the tolerance θ is properly chosen, we can ensure the following local estimates: for any given $\mathbf{v}$ and its restriction to τ, $\mathbf{v}_\tau$, there is a coarse vector $\mathbf{v}_{\tau_c}^c$ such that

$$\|A_\tau\| \|\mathbf{v}_\tau - P_\tau\, \mathbf{v}_{\tau_c}^c\|^2 \leq \delta\, \mathbf{v}_\tau^T\, A_\tau\, \mathbf{v}_\tau. \tag{3.10}$$

We assume that $\|A_\tau\|$ are uniformly bounded from below by $\eta\|A\|$ for a constant η, i.e.,

$$\|A_\tau\| \geq \eta\, \|A\|. \tag{3.11}$$

We remark that $\|A_\tau\| \leq \|A\|$. The latter is seen from the inequality $\mathbf{v}_\tau^T A_\tau \mathbf{v}_\tau \leq \mathbf{v}^T A \mathbf{v} \leq \|A\|\|\mathbf{v}\|^2$. Choosing $\mathbf{v} = 0$ outside the set τ which gives $\mathbf{v}_\tau^T A_\tau \mathbf{v}_\tau \leq \|A\|\|\mathbf{v}_\tau\|^2$, that is $\|A_\tau\| \leq \|A\|$.

We have the following main result.

THEOREM 4.1. *The local estimates* (3.10)–(3.11) *imply the following global weak approximation property*

$$\|A\|\|\mathbf{v} - P\mathbf{v}_c\|^2 \leq \frac{\delta}{\eta}\, \|\mathbf{v}\|_A^2.$$

PROOF. Given $\mathbf{v}$ and its restrictions $\mathbf{v}_\tau$ to the sets τ. Let $\mathbf{v}_c = (\mathbf{v}_{\tau_c}^c)$ be the local coarse components for which the estimates (3.10) hold.

We have the identity $\mathbf{v} = \sum_\tau I_\tau W_\tau \mathbf{v}_\tau$ and $P\mathbf{v}_c = \sum_\tau I_\tau W_\tau \mathbf{v}_{\tau_c}^c$. Hence $\mathbf{v} - P\mathbf{v}_c = \sum_\tau I_\tau W_\tau (\mathbf{v}_\tau - P_\tau \mathbf{v}_{\tau_c}^c)$. Therefore

$$\begin{aligned}
\|\mathbf{v} - P\mathbf{v}_c\|^2 &= (\mathbf{v} - P\mathbf{v}_c)^T \left(\sum_\tau I_\tau W_\tau (\mathbf{v}_\tau - P_\tau \mathbf{v}_{\tau_c}^c) \right) \\
&= \sum_\tau \left(W_\tau^{\frac{1}{2}} I_\tau^T (\mathbf{v} - P\mathbf{v}_c) \right)^T W_\tau^{\frac{1}{2}} \left(\mathbf{v}_\tau - P_\tau \mathbf{v}_{\tau_c}^c \right) \\
&\leq \left(\sum_\tau (\mathbf{v} - P\mathbf{v}_c)^T I_\tau W_\tau I_\tau^T (\mathbf{v} - P\mathbf{v}_c) \right)^{\frac{1}{2}} \\
&\qquad \times \left(\sum_\tau (\mathbf{v}_\tau - P_\tau \mathbf{v}_{\tau_c}^c)^T W_\tau (\mathbf{v}_\tau - P_\tau \mathbf{v}_{\tau_c}^c) \right)^{\frac{1}{2}} \\
&= \|\mathbf{v} - P\mathbf{v}_c\| \left(\sum_\tau (\mathbf{v}_\tau - P_\tau \mathbf{v}_{\tau_c}^c)^T W_\tau (\mathbf{v}_\tau - P_\tau \mathbf{v}_{\tau_c}^c) \right)^{\frac{1}{2}} \\
&\leq \|\mathbf{v} - P\mathbf{v}_c\| \left(\sum_\tau (\mathbf{v}_\tau - P_\tau \mathbf{v}_{\tau_c}^c)^T (\mathbf{v}_\tau - P_\tau \mathbf{v}_{\tau_c}^c) \right)^{\frac{1}{2}} \\
&= \|\mathbf{v} - P\mathbf{v}_c\| \left(\sum_\tau \|\mathbf{v}_\tau - P_\tau \mathbf{v}_{\tau_c}^c)\|^2 \right)^{\frac{1}{2}}.
\end{aligned}$$

Therefore

$$\|\mathbf{v} - P\mathbf{v}_c\|^2 \leq \sum_\tau \|\mathbf{v}_\tau - P_\tau \mathbf{v}_{\tau_c}^c)\|^2. \tag{3.12}$$

This shows that

$$\begin{aligned}\|\mathbf{v}-P\mathbf{v}_c\|^2 &\leq \sum_{\tau}\|\mathbf{v}_\tau - P_\tau \mathbf{v}^c_{\tau_c})\|^2\\ &\leq \sum_{\tau}\tfrac{\delta}{\|A_\tau\|}\mathbf{v}_\tau^T A_\tau \mathbf{v}_\tau\\ &\leq \tfrac{\delta}{\eta\|A\|}\sum_{\tau}\mathbf{v}_\tau^T A_\tau \mathbf{v}_\tau\\ &= \tfrac{\delta}{\eta\|A\|}\mathbf{v}^T A\mathbf{v},\end{aligned}$$

which is the desired result. □

5. Examples

Finite element matrices. A natural example of matrices A assembled from local positive semi-definite matrices A_τ comes from finite element discretization of elliptic PDEs (such as Laplace equation). The sets τ, when we apply the method recursively, can be agglomerates T of fine–grid (one level finer) elements τ (connected unions of fine–grid elements). The agglomerates T are viewed as sets in terms of one level higher fine–grid dofs. The sets, where the local eigenproblems are defined are unions of agglomerates, denoted by Ω. We remark that each agglomerate T in general belongs to a number of such local subdomains Ω. The local matrices A_Ω are assembled from the fine–grid element matrices A_τ for $\tau \subset \Omega$. Typically, such subdomain Ω is the union of all agglomerates T that share a common fine–grid dof (here the agglomerates are viewed as sets on the initial fine–grid). The coarse level element matrices are then defined as the symmetric positive semi–definite matrices $PI_TA_TI_T^TP$, where I_T stands for extension by zero outside the agglomerate T (viewed as a set of one-level higher fine-grid dofs). The local matrix A_T is assembled from the fine-grid element matrices A_τ, $\tau \subset T$. Once having coarse-level element matrices the method can be recursively applied. It requires agglomeration procedure that generates the next level agglomerates and respective local subdomains where the associated eigenproblems are posed.

5.1. The window-based spectral AMG method. A purely algebraic way of constructing local quadratic forms is based on the following least–squares approach.

Let $\{w\}$ provide an overlapping partition of the set $\mathcal{N}$ of fine–degrees of freedom. From our given $n \times n$ sparse matrix A, extract its rows that correspond to the index set w and form a rectangular matrix A_w of size $|w| \times n$ ($|w|$ stands for the number of entries in w). Let $\{Q_w\}$ provide a partition of unity, i.e., $\sum_w I_w Q_w I_w^T = I$, where as before I_w stands for extension of vectors $\mathbf{v}_w$ defined on w to vector $I_w\mathbf{v}_w = \begin{bmatrix} 0 \\ \mathbf{v}_w \\ 0 \end{bmatrix}$ where the zero entries corresponds to indices in $\mathcal{N} \setminus w$. Since $A_w = I_w^T A$, the following identity is easily seen

$$\begin{aligned}\sum_w (A_w\mathbf{v})^T Q_w A_w\mathbf{v} &= \sum_w (I_w^T A\mathbf{v})^T Q_w I_w^T(A\mathbf{v})\\ &= \sum_w (A\mathbf{v})^T I_w Q_w I_w^T (A\mathbf{v})\\ &= (A\mathbf{v})^T\left(\sum_w I_w Q_w I_w^T\right)(A\mathbf{v})\\ &= \mathbf{v}^T A^T A\mathbf{v}.\end{aligned}$$

Hence, the local matrices $A_w^T Q_w A_w$, or in fact some semi-definite Schur complements S_w of them, can be used to solve local eigenproblems associated with the sets w, and thus ensure local approximation properties.

The observation that $A_w^T Q_w A_w$ are local, is seen from the assumption on sparsity of A, i.e., that each row of A has bounded number of nonzero entries. The actual local matrices that will be used to compute the eigenvectors are defined as the Schur complements

$$\mathbf{v}_w^T S_w \mathbf{v}_w = \min_{\mathbf{v}_\chi} \begin{bmatrix} \mathbf{v}_w \\ \mathbf{v}_\chi \end{bmatrix}^T A_w^T Q_w A_w \begin{bmatrix} \mathbf{v}_w \\ \mathbf{v}_\chi \end{bmatrix}.$$

We recall again that the block $\mathbf{v}_\chi = \mathbf{v}|_{\chi=\mathcal{N}\setminus w}$ enters the above quadratic minimization problem with a small (bounded) number of its entries, corresponding to the non-zero entries $a_{i,j}$ of A_w ($i \in w$ and j in $\chi = \mathcal{N} \setminus w$).

To analyze the method we first notice that based on the definition of S_w, we have $\mathbf{v}_w^T S_w \mathbf{v}_w \leq \mathbf{v}^T A_w^T Q_w A_w \mathbf{v} \leq \sum_w \mathbf{v}^T A_w^T Q_w A_w \mathbf{v} = \mathbf{v}^T A^T A \mathbf{v} \leq \|A\|^2 \|\mathbf{v}\|^2$. Hence for $\mathbf{v}$ vanishing outside the set w, we have $\mathbf{v}_w^T S_w \mathbf{v}_w \leq \|A\|^2 \|\mathbf{v}_w\|^2$. That is,

$$\|S_w\| \leq \|A\|^2.$$

Based on S_w, we can construct the local P_w that ensures the local weak approximation properties

$$\|S_w\| \|\mathbf{v}_w - P_w \mathbf{v}_{w_c}^c\|^2 \leq \delta \mathbf{v}_w^T S_w \mathbf{v}_w.$$

In the same way (as in the case of local matrices A_τ before) we define a global P (using another partition of unity matrix set $\{W_w\}$),

$$P = \sum_w I_w W_w P_w I_{w_c}^T,$$

for which we can prove the estimate (see (3.12))

$$\|\mathbf{v} - P\mathbf{v}_c\|^2 \leq \sum_w \|\mathbf{v}_w - P_w \mathbf{v}_{w_c}^c\|^2.$$

Assuming again the quasiuniformity of the windows, i.e., the estimate

$$\|S_w\| \geq \eta \|A\|^2,$$

for a constant $\eta \in (0, 1]$ independent of w, we end up with the final estimates

$$\begin{aligned} \|\mathbf{v} - P\mathbf{v}_c\|^2 &\leq \sum_w \|\mathbf{v}_w - P_w \mathbf{v}_{w_c}^c\|^2 \\ &\leq \sum_w \tfrac{\delta}{\|S_w\|} \mathbf{v}_w^T S_w \mathbf{v}_w \\ &\leq \tfrac{\delta}{\eta \|A\|^2} \sum_w \mathbf{v}^T A_w Q_w A_w \mathbf{v} \\ &= \tfrac{\delta}{\eta \|A\|^2} \|A\mathbf{v}\|^2. \end{aligned}$$

That is, we proved the following "*strong approximation property*":

$$\|A\| \|\mathbf{v} - P\mathbf{v}_c\|^2 \leq \frac{\delta}{\eta \|A\|} \|A\mathbf{v}\|^2. \tag{3.13}$$

We remark at the end, that we did not use symmetry, nor positive definiteness of A.

CHAPTER 11

Adaptive AMG and Smoothed Aggregation (SA) AMG

This lecture introduces the concept of adaptive AMG methods and motivates the need for constructing of interpolation mappings that fit (approximately or exactly) a set of "algebraically smooth" vectors. We study several approaches of such interpolation rules. The lecture ends up with a formulation and multilevel analysis of the smoothed aggregation (or SA) algebraic multigrid method.

1. The concept of adaptive AMG

The standard smoothing (or relaxation) iterations, possibly combined with a coarse-grid correction based on a projection $\pi = \pi_A := P(P^TAP)^{-1}P^TA$, can be formulated as

$$\mathbf{x} := (I - M^{-T}A)(I - \pi)(I - M^{-1}A)\mathbf{x}. \tag{3.14}$$

If $P = 0$ (or not defined) we set $\pi = 0$. By monitoring the norm of two consecutive iterates (viewed as errors for solving the trivial equation $A\mathbf{x} = 0$),we can get an indication about the quality of the respective (TG) iteration method. Since inverting the coarse–grid matrix $A_c = P^TAP$ can be expensive, we approximate the projection π by either using $\pi_{\overline{M}} = P\left(P^T\overline{M}P\right)^{-1}P^T\overline{M}$ (when $\overline{M}$ is sparse), or we can construct an initial (tentative and possibly not very efficient) V–cycle operator B_c and replace $I-\pi$ with $I-PB_c^{-1}P^TA$. To begin with, when we have not constructed even a single interpolation matrix P (hence $\pi = 0$), we simply run the relaxation process (letting $\pi = 0$ above).

At any rate, at some point after $m \geq 1$ iterations, we may encounter very slow convergence, which means that

$$\mathbf{x}^TA\mathbf{x} \simeq \mathbf{x}_{\text{old}}^TA\mathbf{x}_{\text{old}}. \tag{3.15}$$

In other words, we have that $\mathbf{x}_m := \mathbf{x}$ is a good approximation to the minimal eigenvector $\mathbf{q}$ of the generalized eigenproblem

$$A\mathbf{q} = \lambda_{\min}\overline{M}\mathbf{q},$$

(assuming $\pi = 0$). Indeed, (3.15) implies that $\|(I - \overline{M}^{-1}A)\mathbf{x}\|_A \simeq \|\mathbf{x}\|_A$, that is $\frac{\mathbf{x}^TA\mathbf{x}}{\mathbf{x}^T\overline{M}\mathbf{x}} \approx \lambda_{\min} \simeq 0$.

Our goal would therefore be to incorporate this "algebraically smooth" vector $\mathbf{1} := \mathbf{q} \approx \mathbf{x}_m$ into the "to be constructed" coarse space. Equivalently, we want to construct a P such that

$$\mathbf{1} \in \text{Range}\,(P).$$

Assume that, we have constructed a P, and by recursion we have constructed an initial coarse V-cycle operator B_c. Then, we repeat the above procedure, where now we run the

modified (inexact) two–grid iteration starting with a random initial iterate $\mathbf{x} = \mathbf{x}_0$:

$$\mathbf{x} := (I - M^{-T}A)(I - PB_c^{-1}P^TA)(I - M^{-1}A)\mathbf{x}.$$

In general, by testing the current method available we eventually end up with a component $\mathbf{x}$ that the current level V–cycle cannot handle; that is, the A–norms of two successive iterates $\mathbf{x}$ and $\mathbf{x}_{\rm new}$ are not too different, i.e.,

$$\mathbf{x}_{\rm new}^T A\mathbf{x}_{\rm new} \simeq \mathbf{x}^T A\mathbf{x}.$$

The reasons for this to happen are, either,

- the current coarse space cannot well approximate $\mathbf{e} = \mathbf{x} - P\mathbf{x}_c = \begin{bmatrix} \mathbf{e}_f \\ 0 \end{bmatrix}$, and/or
- B_c cannot successfully damp the coarse interpolant $\mathbf{x}_c$ of $\mathbf{x}$.

A possible remedy to the above is to improve the coarse space and/or coarse solvers B_c^{-1} by augmenting the current interpolation matrix $P = \begin{bmatrix} W \\ I \end{bmatrix}$ by adding few more columns (or one block–column) $P_{\rm new}$, i.e., to construct

$$\overline{P} = \begin{bmatrix} W & \overline{P}_{\rm new} \\ I & 0 \end{bmatrix}.$$

The new columns of $\overline{P}$ are based on additional coarse dofs $\mathcal{N}_{c,\,\rm new} \subset \mathcal{N} \setminus \mathcal{N}_c$. The latter can be chosen in the same way as for P noting that the interpolation error $\mathbf{e} = \mathbf{x} - P\mathbf{x}_c$ vanishes at the current coarse dofs set $\mathcal{N}_c$, i.e., $\mathbf{e}|_{\mathcal{N}_c} = 0$.

In conclusion, we see that we need to be able to construct interpolation matrices P that fit (interpolate exactly or approximately) several "algebraically smooth" vectors $\mathbf{1}_1, \ldots, \mathbf{1}_m$, for any given (small) number $m \geq 1$.

2. Algorithms to fit several vectors

If the set of vectors $\mathbf{1}_k$ restricted to small neighborhood sets of indices $\{\mathcal{A}\}$ provide some reasonable approximation properties, i.e.,

$$\|\mathbf{v}_{\mathcal{A}} - \sum_k \alpha_k \ \mathbf{1}_k|_{\mathcal{A}}\,\|^2 \leq \eta_{\mathcal{A}}\ \mathbf{v}_{\mathcal{A}}^T A_{\mathcal{A}}\mathbf{v}_{\mathcal{A}}.$$

then, we may construct a tentative $\overline{P}$ by simply putting together the pieces of the vectors $\mathbf{1}_k$, $\mathbf{1}_k|_{\mathcal{A}}$, using nonnegative diagonal PU (partition of unity) matrices, similarly to the spectral AMG methods. Here, we assume that there is a set of local matrices $\{A_{\mathcal{A}}\}$ that provide a sense of "local" energy. In other words, we assume that the global quadratic form associated with the original $n \times n$ s.p.d. matrix A can be split into a sum of local quadratic forms associated with the local symmetric positive semi–definite matrices $A_{\mathcal{A}}$ where $\{\mathcal{A}\}$ provides a (overlapping or non-overlapping) partition of the index set $1, 2, \ldots, n$.

2.1. Interpolation by constrained energy minimization. To fix the ideas, we assume here a finite element setting. In particular, we assume that the sets $\mathcal{A}$ are covered exactly by fine-grid elements $\tau \in \mathcal{T}_h$ and use the notation T instead of $\mathcal{A}$. In addition to the given s.p.d. matrix A, we assume access to the local mass matrices G_T (assembled, for every T, from the fine–grid element mass matrices G_τ for $\tau \subset T$).

Given the set of vectors $\mathbf{1}_k,\ k = 1,\ \ldots,\ m$, to each T, we associate m basis functions $\varphi_T^{(k)}$ supported in a neighborhood Ω_T of T that is contained in the union $\cup T^{'}$ of all neighbors $T^{'}$ to T. The function $\varphi_T^{(k)}$, for a fixed k, solves the following local constrained minimization problem

$$a(\varphi_T^{(k)},\ \varphi_T^{(k)}) = \boldsymbol{\varphi}_T^{(k)^T} A \boldsymbol{\varphi}_T^{(k)} \mapsto \min,$$

subject to the prescribed integral moments

$$\mathbf{1}_l^T G_{T^{'}} I_{T^{'}}^T \boldsymbol{\varphi}_T^{(k)} = \delta_{T,\,T^{'}}\ \mathbf{1}_l^T G_T \mathbf{1}_k,\ \text{for all } T^{'} \cap \Omega_T \neq \emptyset \text{ and } l = 1,\ \ldots,\ m.$$

Here, $\delta_{T,\,T^{'}} = 0$ for $T \neq T^{'}$ and $\delta_{T,\,T} = 1$. We also use the notation I_X for zero extension of vectors defined on X to vectors of full size. Since $\boldsymbol{\varphi}_T^{(k)}$ is supported in Ω_T, we have $\boldsymbol{\varphi}_T^{(k)} = I_{\Omega_T} \underline{\boldsymbol{\varphi}}_T^{(k)}$, where now the vector $\underline{\boldsymbol{\varphi}}_T^{(k)}$ is defined only on Ω_T.

The coefficient vectors $\boldsymbol{\varphi}_T^{(k)}$ for $k = 1,\ \ldots,\ m$ running over all $T \in \mathcal{T}_H$ provide the columns of the desired interpolation matrix P. It is clear that by construction

$$\sum_T \boldsymbol{\varphi}_T^{(k)} - \mathbf{1}_k$$

is $G_{T^{'}}$–orthogonal to all $\mathbf{1}_l$ when restricted to any fixed $T^{'}$. Hence, in a weak sense

$$\sum_T \boldsymbol{\varphi}_T^{(k)} \approx \mathbf{1}_k.$$

That is, the resulting P approximately fits all given vectors $\mathbf{1}_k$.

The accuracy can be improved, by performing some iterations used to minimize the difference

$$\| \sum_T \boldsymbol{\varphi}_T^{(k)} - \mathbf{1}_k \|_A^2$$

subject to the above integral moments constraints. One possible algorithm is as follows. Given current approximations $\boldsymbol{\varphi}_T^{(k)},\ T \in \mathcal{T}_H$ for a fixed k that satisfy the respective integral constraints.

Then a new set is obtained by updating each $\boldsymbol{\varphi}_T^{(k)}$ running over all T and solving the local constrained minimization problem for $\mathbf{g}_T$ supported in Ω_T. More specifically, we solve

$$\|\mathbf{g}_T + \sum_{T^{'}} \boldsymbol{\varphi}_{T^{'}}^{(k)} - \mathbf{1}_k\|_A^2 \mapsto \min$$

subject to the constraints

$$\mathbf{1}_l^T G_{T^{'}} I_{T^{'}}^T \mathbf{g}_T = 0 \text{ for all } T^{'} :\ T^{'} \cap \Omega_T \neq \emptyset \text{ and } l = 1,\ \ldots,\ m. \tag{3.16}$$

Equivalently, introducing the error $\mathbf{e} = \mathbf{1}_k - \sum_{T'} \boldsymbol{\varphi}_{T'}^{(k)}$, we solve

$$J(\mathbf{g}_T) \equiv \frac{1}{2}\, \mathbf{g}_T^T A \mathbf{g}_T - \mathbf{e}^T A \mathbf{g}_T \mapsto \min$$

subject to (3.16). This leads to a small (local) saddle–point system for the non-zero entries of $\mathbf{g}_T$ (which is supported in Ω_T) and the respective Lagrange multiplier. More specifically, we have for $\mathbf{g}_T = I_{\Omega_T}\, \underline{\mathbf{g}}_T$ and a Lagrange multiplier $\boldsymbol{\lambda}$ of size the number of neighbors of T (including T) times m, that both solve the saddle–point system (letting $A_{\Omega_T} = I_{\Omega_T}^T A I_{\Omega_T}$)

$$\begin{bmatrix} A_{\Omega_T} & [\ldots,\ I_{\Omega_T}^T I_{T'} G_{T'} \mathbf{1}_l,\ \ldots] \\ \begin{bmatrix} \vdots \\ \mathbf{1}_l^T G_{T'} I_{T'}^T I_T \\ \vdots \end{bmatrix} & 0 \end{bmatrix} \begin{bmatrix} \underline{\mathbf{g}}_T \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} I_{\Omega_T}^T A \mathbf{e} \\ 0 \end{bmatrix}.$$

After $\mathbf{g}_T$ is being computed, we update the current $\boldsymbol{\varphi}_T^{(k)} = I_{\Omega_T} \underline{\boldsymbol{\varphi}}_T^{(k)}$,

$$\underline{\boldsymbol{\varphi}}_T^{(k)} := \underline{\boldsymbol{\varphi}}_T^{(k)} + \underline{\mathbf{g}}_T,$$

and move onto the next set T.

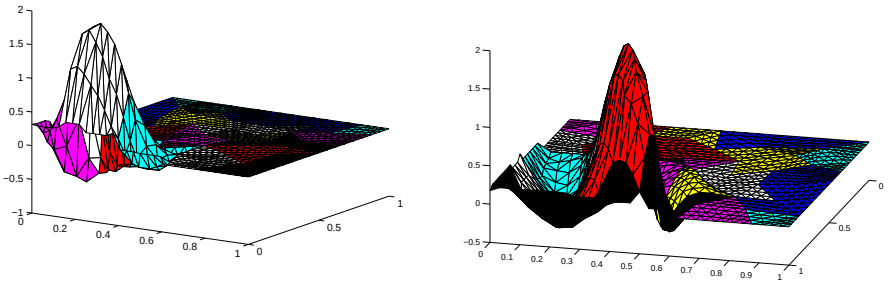

FIGURE 1. Typical coarse basis functions based on fitting one (constant) function.

Examples of fitting one (constant) function and matrix A coming from Laplace operator are seen on Fig. 1. Fitting several ($m = 4$) functions $v_k = \sin(\Pi k_x x)\sin(\Pi k_y y)$, $k = (k_x, k_y)$, $k_x,\ k_y = 1,\ \ldots,\ \sqrt{m}$, and matrix A coming from Laplace operator, is illustrated on Fig. 2.

2.2. Smoothed Aggregation (SA) AMG. If the partitioning $\{\mathcal{A}\}$ is non-overlapping, the respective sets $\mathcal{A}$ are referred to as aggregates. The simple block–diagonal $\overline{P}$ then may not as good as an interpolation matrix for use in a multilevel cycle. The respective coarse vector spaces can be viewed as "piecewise" constant, which in terms of functions, are discontinuous. Thus, we may need to "smooth" out the block–diagonal (tentative) $\overline{P}$. This leads to a method proposed by P. Vaněk (1992), [**VanSA**], known as the "smoothed aggregation" AMG or SA AMG.

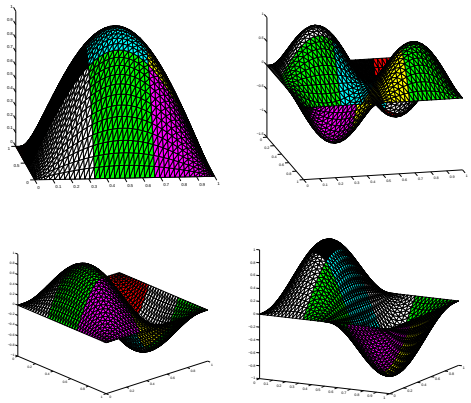

FIGURE 2. $\sum_T \Phi_T^{(k)}$ based on fitting four sin functions v_k on a 3×3 coarse mesh ($H = 1/3$); $h = 1/36$.

Construction of locally supported basis by SA. To illustrate the method we assume in the present section that A is a given symmetric positive semi–definite matrix and let $\mathbf{1}$ be a given null-vector of A, i.e., $A\mathbf{1} = 0$. The method will be applied to a matrix A_0 that coincides with A (after certain boundary conditions are imposed).

For a given integer $\nu \geq 1$ partition the set of degrees of freedom of A, i.e., the fine grid into nonoverlapping sets $\mathcal{A}_i$ such that $\mathcal{A}_i$ contains an index i with the following property. Namely, for any integer $s \leq \nu$, the entries of $(A^s)_{ij}$ away from i are zero. More specifically, we assume,

$$(A^s)_{ij} = 0, \text{ for all indices } j \text{ outside } \mathcal{A}_i. \tag{3.17}$$

Let $\mathbf{1}_i = \mathbf{1}|_{\mathcal{A}_i}$ and be zero outside $\mathcal{A}_i$. It is clear that

$$\sum_i \mathbf{1}_i = \mathbf{1}. \tag{3.18}$$

For a given diagonal matrix D (to be specified later on), let $\overline{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$.

Let φ_ν be a given polynomial (to be specified later on) of degree $\nu \geq 1$ such that $\varphi_\nu(0) = 1$. Hence $\varphi_\nu(t) = 1 - tq_{\nu-1}(t)$ for another polynomial $q_{\nu-1}$.

In what follows we use the notation $\mathbf{v}(x_i)$ to denote the ith entry of a vector $\mathbf{v}$. This is motivated by the fact that very often in practice $\mathbf{v}$ are coefficient vectors of functions v when expanded in terms of a given Lagrangian finite element basis.

Define now,

$$\boldsymbol{\psi}_i = (I - D^{-1} A q_{\nu-1}(D^{-1}A))\mathbf{1}_i. \tag{3.19}$$

We have,

$$\sum_i \boldsymbol{\psi}_i = (I - D^{-1} A q_{\nu-1} D^{-1} A) \sum_i \mathbf{1}_i = (I - D^{-1} A q_{\nu-1}(D^{-1}A))\mathbf{1} = \mathbf{1}, \tag{3.20}$$

since $A\mathbf{1} = 0$. Also $(A^s \mathbf{1}_j)_i = 0$ implies that $((D^{-1}A)^s \mathbf{1}_j)_i = 0$ since D is diagonal and hence A^s and $(D^{-1}A)^s$ have the same sparsity pattern. Thus, $\mathbf{1}(x_i) = \sum_j (\boldsymbol{\psi}_j)(x_i) =$

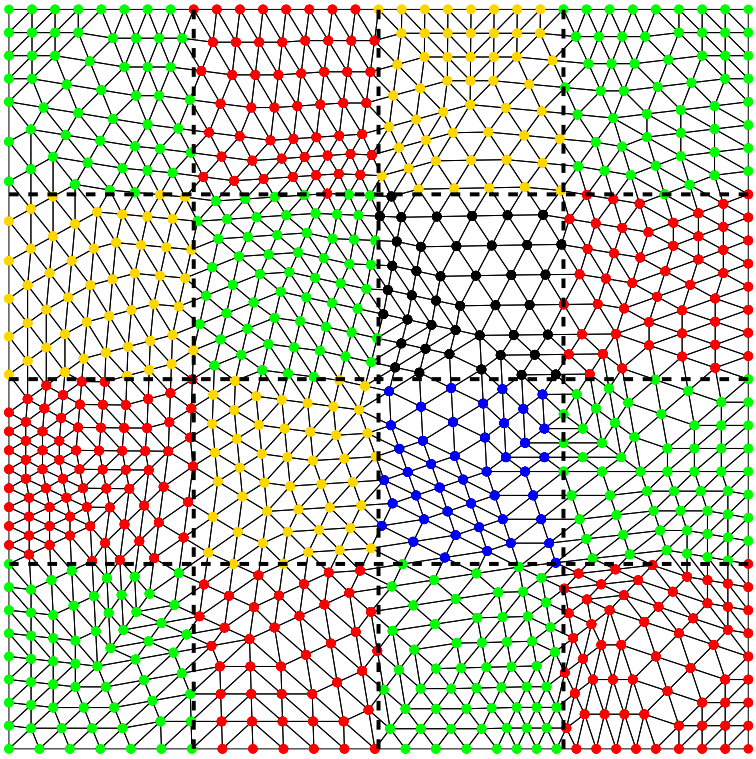

FIGURE 3. Formation of aggregates to guarantee sparsity of all coarse-level operators.

$(\mathbf{1}_i)(x_i) - (D^{-1}Aq_{\nu-1}(D^{-1}A)\mathbf{1}_i)(x_i)$, for all $s \leq \nu$, and $j \neq i$. The latter implies

$$(D^{-1}Aq_{k-1}(D^{-1}A)\mathbf{1}_i)(x_i) = 0.$$

Therefore

$$\boldsymbol{\psi}_i(x_i) = \mathbf{1}(x_i).$$

The vectors $\boldsymbol{\psi}_i$ will form our coarse basis. Note that they have local support and form a partition of unity (in the sense of identity (3.20)) and they also provide a Lagrangian basis.

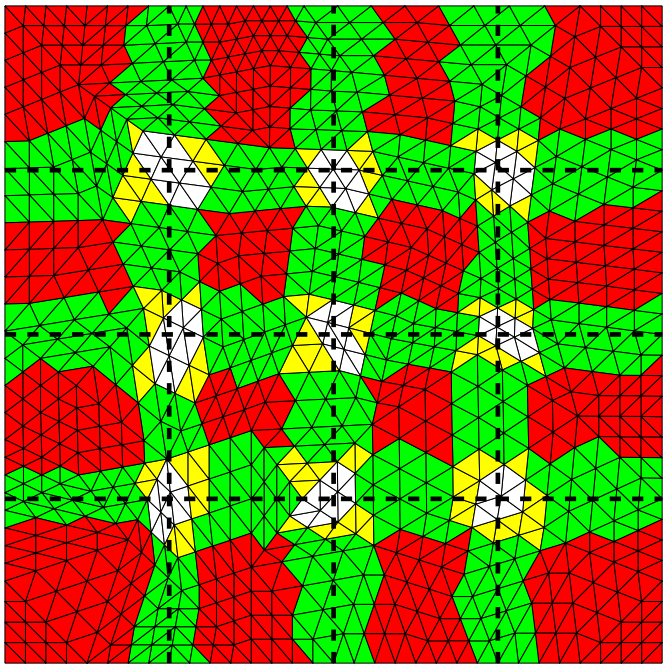

FIGURE 4. The overlap of the extended aggregates obtained by applying two actions of A illustrating the sparsity of the resulting SA coarse-level operator. Darker color corresponds to elements that belong to fewer extended aggregates.

We comment next on one way of constructing aggregates that leads to coarse matrices with controlled sparsity pattern. Namely, assume we are given a quasi-uniform mesh $\mathcal{T}_h$ that triangulates our polygonal (or polyhedral) domain Ω. Choose a parameter H and generate a uniform mesh $\mathcal{T}_H$ with boxes of size $H \times H$ ($\times H$ in 3D). Consider only those boxes that provide covering of Ω. Each box Ω_{ij}(or Ω_{ijk} in 3D) intersects part of the mesh $\mathcal{T}_h$. In this way we construct aggregates $\mathcal{A}_{ij}$ (or $\mathcal{A}_{ijk}$) each containing all fine–grid vertices that are within a particular box (with some arbitration of nodes on box boundaries if any). The only requirement is that the resulting aggregates have large enough interior which can be ensured if H is large enough and $\mathcal{T}_h$ is fine enough. Fig. 3 illustrates this geometric way to generate aggregates with guaranteed diameter bound, whereas Fig. 4 illustrates the overlap of the support of the polynomially smoothed basis functions defined as in (3.20) for $\nu = 2$. It is clear that after one level of smoothed aggregation, the resulting coarse matrix will have sparsity pattern corresponding to a finite difference matrix on uniform grid (9-point stencil in 2D and 27-point stencil in 3D).

Finally, we note that the above properties hold for any null–vector $\mathbf{1}$ of A. We note that A may have several null-vectors such as in the case of matrices coming from linear elasticity (the respective null-vectors or functions are called "rigid body modes").

To continue the process by recursion define $\mathbf{1}_c = [1, \ldots, 1]^T \in \mathbb{R}^{n_c}$. We have, $A_c\mathbf{1}_c = P^T A \sum_i \boldsymbol{\psi}_i = P^T A\mathbf{1} = 0$. Here $P = \left[\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{n_c}\right]$ is the interpolation matrix. Due to the Lagrangian property of the basis $\{\boldsymbol{\psi}_i\}$, i.e., $\boldsymbol{\psi}_i(x_j) = \delta_{ij}$ it follows that P has a full column rank.

We then generate coarse aggregates with corresponding polynomial property (3.17). Note that we have the flexibility to change ν, i.e., to have $\nu = \nu_k$ depending on the level number.

Assume that we have generated $\ell \geq 1$ levels and at every level k we have constructed the respective interpolation matrices P_k. Then after a proper choice of smoothers M_k we end up with a symmetric $V(1,1)$–cycle smoothed aggregation AMG. Our goal is to analyze the method, by only assuming that the vector $\mathbf{1}$ ensures a multilevel approximation property formulated later on (see (3.23)).

The fact that $\mathbf{1}$ is a (near)–null–vector of A is not needed. That is why in practice the resulting coarse bases are not necessarily Lagrangian. Nevertheless convergence is guaranteed as we can see next.

3. A general setting for the SA method

In this section we select the parameters of the smoothed aggregation method.

To simplify the analysis we assume that $\nu \geq 1$ is independent of k. We assume that we are given a set of block–diagonal matrices $\overline{I}_{k-1} : \mathbb{R}^{n_k} \mapsto \mathbb{R}^{n_{k-1}}$. We assume that $\overline{I}_{k-1}$ has the following block–diagonal form,

$$\overline{I}_{k-1} = \begin{bmatrix} \mathbf{1}_1 & 0 & 0 & \ldots & 0 \\ 0 & \mathbf{1}_2 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & \mathbf{1}_{n_k-1} & 0 \\ 0 & 0 & \ldots & 0 & \mathbf{1}_{n_k} \end{bmatrix} \begin{matrix} \} \, \mathcal{A}_1 \\ \} \, \mathcal{A}_2 \\ \} \, \vdots \\ \} \, \mathcal{A}_{n_k-1} \\ \} \, \mathcal{A}_{n_k} \end{matrix}$$

where, for $k > 1$, $\mathbf{1}_i = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$. Note that the vector $\mathbf{1}_i \in \mathbb{R}^{|\mathcal{A}_i|}$, has as many entries of ones, as the size of the fine–grid set (called aggregate) $\mathcal{A}_i$ they interpolate to. We stress upon the fact that, the SA method will be well–defined as soon as the first "piecewise–constant" interpolant $\overline{I}_0$ is specified. We outlined earlier a choice of $\overline{I}_0$ based on a nullvector of A. We can of course select other initial coarse level interpolants, that for example fit several a priori given vectors.

Let $\overline{I}_{k-1}$ be the piecewise constant interpolant from level k to level $k-1$ and let $I_{k-1} = \overline{I}_0 \ldots \overline{I}_{k-1}$ be the composite one. We define $D_k = I_{k-1}^T I_{k-1}$. Denote then $\overline{A}_{k-1} = D_{k-1}^{-\frac{1}{2}} A_{k-1} D_{k-1}^{-\frac{1}{2}}$. Then, the interpolation matrix P_{k-1} is constructed as before, on the basis of A_{k-1}, D_{k-1} and the norm of $\overline{A}_{k-1}$ for our fixed ν. More specifically, we have

$$P_{k-1} = S_{k-1}\overline{I}_{k-1},$$

where

$$S_{k-1} = \varphi_\nu \left(D_{k-1}^{-1} A_{k-1} \right),$$

and $\varphi_\nu(t) = (-1)^\nu \frac{1}{2\nu+1} \frac{\sqrt{b}}{\sqrt{t}}\ T_{2\nu+1}\left(\frac{\sqrt{t}}{\sqrt{b}}\right)$ for $b = b_{k-1} \geq \|\overline{A}_{k-1}\|$. We will show later on (in Lemma 3.1) that $b = b_{k-1} \leq \frac{\|A\|}{(2\nu+1)^{2(k-1)}}$.

The smoother M_k is chosen such that

$$M_k \simeq \|\overline{A}_k\|\ D_k.$$

More specifically, we assume that, M_k is s.p.d. and spectrally equivalent to the diagonal matrix $\|\overline{A}_k\|\ D_k$, and scaled so that,

$$\mathbf{v}^T A_k \mathbf{v} \leq \|\overline{A}_k\| \mathbf{v}^T D_k \mathbf{v} \leq \mathbf{v}^T M_k \mathbf{v}. \tag{3.21}$$

Based on the above choice of P_k, A_k and M_k, for $0 \leq k \leq \ell$, starting with $B_\ell = A_\ell$, for $k = \ell - 1,\ \ldots,\ 1,\ 0$, we recursively define a V–cycle preconditioner B_k to A_k in the following standard way,

$$I - B_k^{-1} A_k = (I - M_k^{-T} A_k)\left(I - P_k B_{k+1}^{-1} P_k^T A_k\right)(I - M_k^{-1} A_k).$$

Letting $B = B_0$, we are concerned in what follows with the (upper) bound K_* in the estimate

$$\mathbf{v}^T A \mathbf{v} \leq \mathbf{v}^T B \mathbf{v} \leq K_\star\ \mathbf{v}^T A \mathbf{v}. \tag{3.22}$$

3.1. The result of Vaněk, Mandel and Brezina. We present here perhaps the only known multilevel convergence result for algebraic multigrid; namely, the suboptimal convergence of the smoothed aggregation (or SA) AMG. The original proof is found in [**SA**] and targeted matrices A coming from second order elliptic PDEs (scalar, like Laplace equation, or systems, such as elasticity).

One of the main assumption in the analysis is a "weak approximation property" of certain coarse spaces of piecewise constant vectors. Namely, that a f.e. function v can be approximated by a piecewise constant interpolant I_H in L_2. The latter is defined based on sets $\mathcal{A}_i$ (the union of fine–grid elements that cover our aggregates, which we later do not distinguish, i.e., treat as the same sets of degrees of freedom) with diameter $\mathcal{O}(H)$. On each set $\mathcal{A}_i$, $I_H v$ is constant, for example equal to an average value of v over $\mathcal{A}_i$, i.e., $I_H v = \frac{1}{|\mathcal{A}_i|} \int\limits_{\mathcal{A}_i} v\ dx$. Then, if A comes from a Laplace–like discrete problem, the following is a standard estimate in L_2 in terms of the energy norm $\|.\|_A$,

$$\|v - I_H v\|_0 \leq c_a H\ \|v\|_A.$$

Rewriting this in terms of vectors leads to the following one

$$h^{\frac{d}{2}}\ \|\mathbf{v} - \underline{I_H v}\| \leq c_a H\ \|\mathbf{v}\|_A,$$

where $d = 2$ or $d = 3$ is the dimension of the domain where the corresponding PDE (Laplacian–like) is posed. Since then $\|A\| \simeq h^{d-2}$, we arrive at

$$\|\mathbf{v} - \underline{I_H v}\| \leq c_a \frac{H}{h}\ \frac{1}{\|A\|^{\frac{1}{2}}}\ \|\mathbf{v}\|_A.$$

In the application of the SA we will have $\frac{H}{h} \simeq (2\nu+1)^{k+1}$, where $\nu \geq 1$ will be the polynomial degree of a polynomial used to smooth out the piecewise constant interpolants, that we start with. Also, $k = 0, 1, \ldots,\ \ell$ stands for the coarsening level. We summarize this estimate as our main assumption. Given the nonoverlapping sets $\mathcal{A}_i^{(k)}$ (aggregates)

at coarsening level $k \geq 0$ viewed as sets of fine–grid dofs. Let $\overline{Q}_k$ be the block–diagonal ℓ_2–projections that for every vector $\mathbf{v}$ restricted to an aggregate $\mathcal{A}_i^{(k)}$ assigns a scalar value $\overline{\mathbf{v}}_i$, the average of $\mathbf{v}|_{\mathcal{A}_i}$ over $\mathcal{A}_i$. Finally, let I_k interpolate them back all–the way up to the finest level as constants over $\mathcal{A}_i^{(k)}$ (equal to the average value $\overline{\mathbf{v}}_i$). Finally, assume that the diameter of $\mathcal{A}_i^{(k)}$ is of order $(2\nu+1)^{k+1}h$ where h is the finest mesh size. Then, the following approximation property is our main assumption:

$$\|\mathbf{v} - I_k\overline{Q}_k\mathbf{v}\| \leq c_a \, \frac{(2\nu+1)^{k+1}}{\|A\|^{\frac{1}{2}}} \, \|\mathbf{v}\|_A. \tag{3.23}$$

The latter assumption is certainly true if the matrix A comes from elliptic PDEs discretized on a uniformly refined mesh, and the corresponding aggregates at every level k are constructed based on the uniform hierarchy of the geometric meshes. In the applications when we have access to the fine–grid matrix only (and possibly to the fine–grid mesh) when constructing the hierarchy of aggregates we have to follow the rule that their graph diameter grows like $(2\nu+1)^{k+1}$. A typical choice in practice is $\nu = 1$.

An optimal Chebyshev like polynomial. We first revisit a Chebyshev–like polynomial introduced earlier (in a previous lecture).

Consider the Chebyshev polynomials $T_k(t)$ defined by recursion as follows, $T_0 = 1$, $T_1(t) = t$ and for $k \geq 1$, $T_{k+1}(t) = 2tT_k(t) - T_{k-1}(t)$. Letting $t = \cos\alpha \in [-1, 1]$, we have the explicit representation $T_k(t) = \cos k\alpha$, which is seen from the trigonometric identity $\cos(k+1)\alpha + \cos(k-1)\alpha = 2\cos\alpha\cos k\alpha$.

PROPOSITION 3.1. *For a given $b > 0$, the function defined for $t \in [0,\ b]$*

$$\varphi_\nu(t) = (-1)^\nu \frac{1}{2\nu+1} \, \frac{\sqrt{b}}{\sqrt{t}} \, T_{2\nu+1}\left(\frac{\sqrt{t}}{\sqrt{b}}\right), \tag{3.24}$$

is a polynomial of degree ν such that $\varphi_\nu(0) = 1$, that is, $\varphi_\nu(t) = 1 - tq_{\nu-1}(t)$ for some polynomial $q_{\nu-1}(t)$ of degree $\nu - 1$.

PROPOSITION 3.2. *The polynomial φ_ν defined in* (3.24) *has the following optimality property:*

$$\min_{p_\nu:\, p_\nu(0)=1} \max_{t\in[0,\, b]} |\sqrt{t}\, p_\nu(t)| = \max_{t\in[0,\, b]} |\sqrt{t}\, \varphi_\nu(t)| = \frac{\sqrt{b}}{2\nu+1}. \tag{3.25}$$

Also, $\varphi_\nu(0) = 1$ and

$$\max_{t\in[0,\, b]} |\varphi_\nu(t)| = 1. \tag{3.26}$$

Here are some particular cases of the polynomials φ_ν.

Using the definition of the Chebyshev polynomials, $T_0 = 1$, $T_1 = t$, $T_{k+1} = 2tT_k - T_{k-1}$, for $k \geq 1$, we get $T_2 = 2t^2 - 1$ and hence

$$T_3(t) = 4t^3 - 3t.$$

Thus,

$$\varphi_1(t) = -\frac{1}{3}\sqrt{b}(4\frac{t}{b^{\frac{3}{2}}} - \frac{3}{\sqrt{b}}) = 1 - \frac{4}{3}\,\frac{t}{b}.$$

This in particular shows that

$$\sup_{t\in(0,\,b]}\frac{|1-\varphi_1(t)|}{\sqrt{t}}=\frac{4}{3}\frac{1}{\sqrt{b}}.$$

The next polynomial is based on $T_5=2tT_4-T_3=2t(2tT_3-T_2)-T_3=(4t^2-1)(4t^3-3t)-4t^3+2t=16t^5-20t^3+5t$. Therefore,

$$\varphi_2(t)=\frac{1}{5}\sqrt{\frac{b}{t}}\left(16\sqrt{t}t^2\frac{1}{b^{\frac{5}{2}}}-20\sqrt{t}t\frac{1}{b^{\frac{3}{2}}}+5\sqrt{t}\frac{1}{\sqrt{b}}\right).$$

This shows,

$$\varphi_2(t)=\frac{16}{5}\frac{t^2}{b^2}-4\frac{t}{b}+1.$$

We also have,

$$\sup_{t\in(0,\,b]}\frac{1-\varphi_2(t)}{\sqrt{t}}=\frac{4}{\sqrt{b}}\sup_{x\in(0,1]}(x-\frac{4}{5}x^3)=\frac{4}{3}\sqrt{\frac{5}{3}}\frac{1}{\sqrt{b}}.$$

In general, it is clear that the following result holds.

PROPOSITION 3.3. *There is a constant C_ν independent of b such that the following estimate holds,*

$$\sup_{t\in(0,\,b]}\frac{|1-\varphi_\nu(t)|}{\sqrt{t}}\le C_\nu\frac{1}{b^{\frac{1}{2}}}. \tag{3.27}$$

PROOF. We have, $1-\varphi_\nu(t)=tq_{\nu-1}(t)$, that is, $\frac{1-\varphi_\nu}{\sqrt{t}}=\sqrt{t}\;q_{\nu-1}(t)$ and therefore the quotient in question is bounded for $t\in(0,b]$. More specifically, the following dependence on b is seen:

$$\sup_{t\in(0,\,b]}\frac{|1-\varphi_\nu(t)|}{\sqrt{t}}=\frac{1}{b^{\frac{1}{2}}}\sup_{\lambda\in(0,\,1]}\frac{\left|1-\frac{(-1)^\nu}{2\nu+1}\frac{T_{2\nu+1}(\sqrt{\lambda})}{\sqrt{\lambda}}\right|}{\sqrt{\lambda}}.$$

Clearly, the constant $C_\nu=\sup_{\lambda\in(0,\,1]}\frac{\left|1-\frac{(-1)^\nu}{2\nu+1}\frac{T_{2\nu+1}(\sqrt{\lambda})}{\sqrt{\lambda}}\right|}{\sqrt{\lambda}}$ is independent of b.

□

Preliminary estimates. Our second main assumption is that we can construct at every level $k\ge 1$ aggregates with the polynomial property (3.17). The latter is needed to keep the sparsity pattern of the resulting coarse matrices under control. We also assume that the the size of the composite aggregates coming from level k onto the finest level satisfy the estimate

$$\max_{i\in\mathcal{N}_k}|\mathrm{diam}(\mathcal{A}_i)|\le(2\nu+1)^k h.$$

As already mentioned above the above assumption is easily met in practice for meshes that are obtained by uniform refinement. For more general unstructured finite element meshes the above assumption is only a practical rule to construct the coarse level aggregates.

The analysis in what follows closely follows [**SA**].

LEMMA 3.1. *The following main estimate holds true:*

$$\|\overline{A}_k\| \leq \frac{\|A\|}{(2\nu+1)^{2k}}.$$

PROOF. Recall that $D_{k+1} = I_k^T I_k$. Then, with $S_k = I - D_k^{-1} A_k q_\nu(D_k^{-1} A_k)$, using the fact that $P_k = S_k \overline{I}_k$ and $D_{k+1} = \overline{I}_k^T D_k \overline{I}_k$, we have

$$\begin{aligned}
\|D_{k+1}^{-\frac{1}{2}} A_{k+1} D_{k+1}^{-\frac{1}{2}}\| &= \sup_{\mathbf{v}} \frac{\mathbf{v}^T A_{k+1}\mathbf{v}}{\mathbf{v}^T D_{k+1}\mathbf{v}} \\
&= \sup_{\mathbf{v}} \frac{\mathbf{v}^T \overline{I}_k^T S_k^T A_k S_k \overline{I}_k \mathbf{v}}{(\overline{I}_k\mathbf{v})^T D_k (\overline{I}_k\mathbf{v})} \\
&\leq \sup_{\mathbf{v}} \frac{\mathbf{v}^T S_k^T A_k S_k \mathbf{v}}{\mathbf{v}^T D_k \mathbf{v}}.
\end{aligned}$$

Therefore, based on property (3.25) of φ_ν, we get,

$$\begin{aligned}
\mathbf{v}^T D_k^{-\frac{1}{2}} S_k^T A_k S_k D_k^{-\frac{1}{2}} \mathbf{v} &\leq \sup_{t \in \left[0,\, \|D_k^{-\frac{1}{2}} A_k D_k^{-\frac{1}{2}}\|\right]} t(1 - t q_{\nu-1}(t))^2 \, \|\mathbf{v}\|^2 \\
&\leq \tfrac{\|D_k^{-\frac{1}{2}} A_k D_k^{-\frac{1}{2}}\|}{(2\nu+1)^2} \, \|\mathbf{v}\|^2.
\end{aligned}$$

That is, by recursion (with $D_0 = I$, $A_0 = A$), we end up with the estimate

$$\|D_{k+1}^{-\frac{1}{2}} A_{k+1} D_{k+1}^{-\frac{1}{2}}\| \leq \frac{\|A\|}{(2\nu+1)^{2(k+1)}}.$$

Thus the proof is complete. □

We will be using the main result regarding the relative spectral condition number of the ℓth level V–cycle preconditioner B with respect to A, which we restate here.

Given smoothers M_j and interpolation matrices P_j and respective coarse matrices relates as $A_{j+1} = P_j^T A_j P_j$. Each smoother M_j is such that $M_j^T + M_j - A_j$ is s.p.d.. Then, the following main identity holds:

$$\begin{aligned}
\mathbf{v}^T A \mathbf{v} \leq \mathbf{v}^T B \mathbf{v} \;&= \inf_{(\mathbf{v}_k)} \Bigg[\mathbf{v}_\ell^T A_\ell \mathbf{v}_\ell + \sum_{j<\ell} \left(M_j^T \mathbf{v}_j^f + A_j P_j \mathbf{v}_{j+1} \right)^T \\
&\qquad \left(M_j^T + M_j - A_j \right)^{-1} \left(M_j^T \mathbf{v}_j^f + A_j P_j \mathbf{v}_{j+1} \right) \Bigg].
\end{aligned} \tag{3.28}$$

The inf here is taken over the components $(\mathbf{v}_k)$ of all possible decompositions of $\mathbf{v}$:

(i) starting with $\mathbf{v}_0 = \mathbf{v}$, and
(ii) for $k \geq 0$, $\mathbf{v}_k = \mathbf{v}_k^f + P_k \mathbf{v}_{k+1}$.

Introduce now the following averaging operators,

$$\overline{Q}_{k-1} = (I_{k-1}^T I_{k-1})^{-1} I_{k-1}^T \; : \; \mathbb{R}^{n_0} \mapsto \mathbb{R}^{n_k}. \tag{3.29}$$

Note that $I_{k-1}\overline{Q}_{k-1}$ are ℓ_2–orthogonal projections.

We will be interested in a particular recursive decomposition for any given fine–grid vector $\mathbf{v}$. Based on the characterization identity (3.28) utilizing an energy stable particular decomposition of the fine–gird vectors, we can get an upper bound of K_*,

which is our goal. Introduce $\overline{Q}_{-1} = I$, and let for $k \geq 0$, $\mathbf{v}_k = \overline{Q}_{k-1}\mathbf{v} \in \mathbb{R}^{n_k}$. We have the two–level decomposition

$$\mathbf{v}_k = \left(\overline{Q}_{k-1}\mathbf{v} - P_k\overline{Q}_k\mathbf{v}\right) + P_k\overline{Q}_k\mathbf{v} = \mathbf{v}_k^f + P_k\mathbf{v}_{k+1}.$$

In order to bound the relative condition number of the V–cycle preconditioner B with respect to A, (due to estimate (3.28)), based on our choice of the smoother as in (3.21), it is sufficient to bound the expressions (i) and (ii) below:

$$\text{(i)} \qquad \sum_{k<\ell}(\mathbf{v}_k^f)^T M_k \mathbf{v}_k^f = \sum_{k<\ell}\left(\overline{Q}_{k-1}\mathbf{v} - P_k\overline{Q}_k\mathbf{v}\right)^T M_k \left(\overline{Q}_{k-1}\mathbf{v} - P_k\overline{Q}_k\mathbf{v}\right)$$

and

$$\text{(ii)} \qquad \sum_{k\leq\ell}\mathbf{v}_k^T A_k \mathbf{v}_k = \sum_{k<\ell}\mathbf{v}^T\overline{Q}_{k-1}^T A_k \overline{Q}_{k-1}\mathbf{v},$$

both in terms of $\mathbf{v}^T A\mathbf{v}$.

Estimating the first sum (i). Recall that $P_k = S_k\overline{I}_k$, $S_k = I - D_k^{-1}A_k q_{\nu-1}(D_k^{-1}A_k)$, $I_k = \overline{I}_0\overline{I}_1\ldots\overline{I}_k$ and $D_k = (I_{k-1})^T I_{k-1}$. Note that (see (3.26)) $\|D_k^{\frac{1}{2}} S_k D_k^{-\frac{1}{2}}\| = \sup_{t\in[0,\,\|\overline{A}_k\|]} |\varphi_\nu(t)| = 1$. We start with the inequality,

$$\begin{aligned}
\|\left(\overline{Q}_{k-1} - P_k\overline{Q}_k\right)\mathbf{v}\|_{D_k} &= \|\left(\overline{Q}_{k-1} - S_k\overline{I}_k\overline{Q}_k\right)\mathbf{v}\|_{D_k} \\
&= \|D_k^{\frac{1}{2}} S_k\left(\overline{Q}_{k-1} - \overline{I}_k\overline{Q}_k\right)\mathbf{v} + (I - S_k)\overline{Q}_{k-1}\mathbf{v}\| \\
&\leq \|D_k^{\frac{1}{2}} S_k\left(\overline{Q}_{k-1} - \overline{I}_k\overline{Q}_k\right)\mathbf{v}\| + \|D_k^{\frac{1}{2}}(I - S_k)\overline{Q}_{k-1}\mathbf{v}\| \\
&\leq \|D_k^{\frac{1}{2}} S_k D_k^{-\frac{1}{2}}\|\|D_k^{\frac{1}{2}}\left(\overline{Q}_{k-1} - \overline{I}_k\overline{Q}_k\right)\mathbf{v}\| \\
&\qquad + \|(I - D_k^{\frac{1}{2}} S_k D_k^{-\frac{1}{2}})D_k^{\frac{1}{2}}\overline{Q}_{k-1}\mathbf{v}\| \\
&\leq \|D_k^{\frac{1}{2}}\left(\overline{Q}_{k-1} - \overline{I}_k\overline{Q}_k\right)\mathbf{v}\| + \|(I - D_k^{\frac{1}{2}} S_k D_k^{-\frac{1}{2}})D_k^{\frac{1}{2}}\overline{Q}_{k-1}\mathbf{v}\| \\
&= \|I_{k-1}\left(\overline{Q}_{k-1} - \overline{I}_k\overline{Q}_k\right)\mathbf{v}\| + \|(I - D_k^{\frac{1}{2}} S_k D_k^{-\frac{1}{2}})D_k^{\frac{1}{2}}\overline{Q}_{k-1}\mathbf{v}\|.
\end{aligned}$$

Let $(0,\ b]$ be the interval that contains the eigenvalues of $\overline{A}_k = D_k^{-\frac{1}{2}} A_k D_k^{-\frac{1}{2}}$ which is used to construct the optimal polynomial $\varphi_\nu(t) = 1 - tq_{\nu-1}(t)$, i.e., $b \geq \|\overline{A}_k\|$. Notice that

$$I - D_k^{\frac{1}{2}} S_k D_k^{-\frac{1}{2}} = I - \varphi_\nu(\overline{A}_k) = \overline{A}_k^{-\frac{1}{2}}\left(I - \varphi_\nu(\overline{A}_k)\right)\overline{A}_k^{\frac{1}{2}}.$$

Based on estimate (3.27) we then get,

$$\begin{aligned}
\|(I - D_k^{\frac{1}{2}} S_k D_k^{-\frac{1}{2}})D_k^{\frac{1}{2}}\overline{Q}_{k-1}\mathbf{v}\| &\leq \max_{t\in(0,\,b]} \tfrac{1-\varphi_\nu(t)}{\sqrt{t}}\ \|\overline{A}_k^{\frac{1}{2}} D_k^{\frac{1}{2}}\overline{Q}_{k-1}\mathbf{v}\| \\
&\leq C_\nu \tfrac{1}{\sqrt{b}}\ \|\overline{A}_k^{\frac{1}{2}} D_k^{\frac{1}{2}}\overline{Q}_{k-1}\mathbf{v}\| \\
&\leq C_\nu \tfrac{1}{\|\overline{A}_k\|^{\frac{1}{2}}}\ \|\overline{Q}_{k-1}\mathbf{v}\|_{A_k}.
\end{aligned}$$

Thus, we arrived at the estimate

$$\|(\overline{Q}_{k-1} - P_k\overline{Q}_k)\mathbf{v}\|_{D_k} \leq \|\left(I_{k-1}\overline{Q}_{k-1} - I_k\overline{Q}_k\right)\mathbf{v}\| + \frac{C_\nu}{\|\overline{A}_k\|^{\frac{1}{2}}}\ \|\overline{Q}_{k-1}\mathbf{v}\|_{A_k}. \tag{3.30}$$

The final bound on sum (i) will be derived after an estimate of the terms in sum (ii) is obtained.

Estimating the second sum (ii). We bound next $\|\overline{Q}_k\mathbf{v}\|_{A_{k+1}}$.

Since $\|A_k^{\frac{1}{2}}D_k^{-1}A_k^{\frac{1}{2}}\| = \|\overline{A}_k\|$, we have $\|A_k^{\frac{1}{2}}S_kA_k^{-\frac{1}{2}}\| = \|\varphi_\nu(A_k^{\frac{1}{2}}D_k^{-1}A_k^{\frac{1}{2}})\| \leq 1$ and similarly $\|D_k^{\frac{1}{2}}S_kD_k^{-\frac{1}{2}}\| = \|\varphi_\nu(D_k^{-\frac{1}{2}}A_kD_k^{-\frac{1}{2}})\| = \|\varphi_\nu(\overline{A}_k)\| \leq 1$. The first estimate shows that

$$\mathbf{w}^TS_k^TA_kS_k\mathbf{w} \leq \mathbf{w}^TA_k\mathbf{w}.$$

Then, based on Lemma 3.1, we obtain

$$\begin{aligned}
\|\overline{Q}_k\mathbf{v}\|_{A_{k+1}} &= \|P_k\overline{Q}_k\mathbf{v}\|_{A_k}\\
&= \|S_k\overline{I}_k\overline{Q}_k\mathbf{v}\|_{A_k}\\
&\leq \|S_k\left(\overline{I}_k\overline{Q}_k - \overline{Q}_{k-1}\right)\mathbf{v}\|_{A_k} + \|S_k\overline{Q}_{k-1}\mathbf{v}\|_{A_k}\\
&\leq \|S_k\left(\overline{I}_k\overline{Q}_k - \overline{Q}_{k-1}\right)\mathbf{v}\|_{A_k} + \|\overline{Q}_{k-1}\mathbf{v}\|_{A_k}\\
&\leq \|\left(A_k^{\frac{1}{2}}D_k^{-\frac{1}{2}}\right)\left(D_k^{\frac{1}{2}}S_kD_k^{-\frac{1}{2}}\right)D_k^{\frac{1}{2}}\left(\overline{I}_k\overline{Q}_k - \overline{Q}_{k-1}\right)\mathbf{v}\| + \|\overline{Q}_{k-1}\mathbf{v}\|_{A_k} \qquad (3.31)\\
&\leq \|\overline{A}_k\|^{\frac{1}{2}}\,\|D_k^{\frac{1}{2}}\left(\overline{I}_k\overline{Q}_k - \overline{Q}_{k-1}\right)\mathbf{v}\| + \|\overline{Q}_{k-1}\mathbf{v}\|_{A_k}\\
&\leq \tfrac{\|A\|^{\frac{1}{2}}}{(2\nu+1)^k}\,\|I_{k-1}\left(\overline{I}_k\overline{Q}_k - \overline{Q}_{k-1}\right)\mathbf{v}\| + \|\overline{Q}_{k-1}\mathbf{v}\|_{A_k}\\
&\leq \tfrac{\|A\|^{\frac{1}{2}}}{(2\nu+1)^k}\|I_{k-1}\left(\overline{I}_k\overline{Q}_k - \overline{Q}_{k-1}\right)\mathbf{v}\| + \|\overline{Q}_{k-1}\mathbf{v}\|_{A_k}.
\end{aligned}$$

We have,

$$\|\mathbf{v} - I_k\overline{Q}_k\mathbf{v}\|^2 = \|\left(I_{k-1}\overline{Q}_{k-1} - I_k\overline{Q}_k\right)\mathbf{v}\|^2 + \|\mathbf{v} - I_{k-1}\overline{Q}_{k-1}\mathbf{v}\|^2,$$

since $I_{k-1}^TI_{k-1}\overline{Q}_{k-1} = I_{k-1}^T$ and $I_k = I_{k-1}\overline{I}_k$, which imply

$$\left(\mathbf{v} - I_{k-1}\overline{Q}_{k-1}\mathbf{v}\right)^T\left(I_{k-1}\overline{Q}_{k-1} - I_k\overline{Q}_k\right)\mathbf{v} = \left(\mathbf{v} - I_{k-1}\overline{Q}_{k-1}\mathbf{v}\right)^TI_{k-1}(\star) = 0,$$

Therefore,

$$\|\left(I_{k-1}\overline{Q}_{k-1} - I_k\overline{Q}_k\right)\mathbf{v}\| \leq \|\mathbf{v} - I_k\overline{Q}_k\mathbf{v}\|.$$

That is, if we bound $\|\mathbf{v} - I_k\overline{Q}_k\mathbf{v}\|$ the result will follow.

Use now the main estimate (3.23) which was our main assumption. It reads,

$$\|\mathbf{v} - I_k\overline{Q}_k\mathbf{v}\|^2 \leq \sigma_a^2\frac{(2\nu+1)^{2(k+1)}}{\|A\|}\mathbf{v}^TA\mathbf{v}.$$

Then,

$$\|\left(I_{k-1}\overline{Q}_{k-1} - I_k\overline{Q}_k\right)\mathbf{v}\| \leq \sigma_a\frac{(2\nu+1)^{k+1}}{\|A\|^{\frac{1}{2}}}\|\mathbf{v}\|_A. \qquad (3.32)$$

Substituting the latter estimate in (3.31), leads to the following main recursive estimate,

$$\|\overline{Q}_k\mathbf{v}\|_{A_{k+1}} \leq \|\overline{Q}_{k-1}\mathbf{v}\|_{A_k} + \sigma_a\frac{(2\nu+1)^k}{\|A\|^{\frac{1}{2}}}\,\frac{\|A\|^{\frac{1}{2}}}{(2\nu+1)^k}\,\|\mathbf{v}\|_A$$

That is, we proved the following main estimate,

$$\|\overline{Q}_k\mathbf{v}\|_{A_{k+1}} \leq \|\overline{Q}_{k-1}\mathbf{v}\|_{A_k} + \Delta\,\|\mathbf{v}\|_A \leq (1+\sigma_ak)\,\|\mathbf{v}\|_A. \qquad (3.33)$$

Thus the second sum is bounded as follows

$$\sum_{l\leq\ell}\mathbf{v}_k^TA_k\mathbf{v}_k = \sum_{k\leq\ell}\|\overline{Q}_{k-1}\mathbf{v}\|_{A_k}^2 \leq C\ell^3\,\mathbf{v}^TA\mathbf{v}. \qquad (3.34)$$

Completing the bound of the first sum (i). We showed, see estimate (3.30)

$$\| \left(\overline{Q}_{k-1} - P_k\overline{Q}_k\right) \mathbf{v}\|_{D_k} \leq \| \left(I_{k-1}\overline{Q}_{k-1} - I_k\overline{Q}_k\right) \mathbf{v}\| + \frac{C_\nu}{\|\overline{A}_k\|^{\frac{1}{2}}} \|\overline{Q}_{k-1}\mathbf{v}\|.$$

This estimate, together with (3.32) and (3.33), imply

$$\| \left(\overline{Q}_{k-1} - P_k\overline{Q}_k\right) \mathbf{v}\|_{D_k} \leq \sigma_a \frac{(2\nu+1)^k}{\|A\|^{\frac{1}{2}}}\|\mathbf{v}\|_A + \frac{C_\nu}{\|\overline{A}_k\|^{\frac{1}{2}}} \left(1+\sigma_a k\right) \|\mathbf{v}\|_A.$$

We need to bound $\|\overline{A}_k\|^{\frac{1}{2}}\| \left(\overline{Q}_{k-1} - P_k\overline{Q}_k\right) \mathbf{v}\|_{D_k}$. (Recall that $M_k \simeq \|\overline{A}_k\|\ D_k$.) This implies

$$\begin{aligned} \|\overline{A}_k\|^{\frac{1}{2}}\| \left(\overline{Q}_{k-1} - P_k\overline{Q}_k\right) \mathbf{v}\|_{D_k} &\leq \|\overline{A}_k\|^{\frac{1}{2}} \left(\sigma_a \frac{(2\nu+1)^k}{\|A\|^{\frac{1}{2}}} \right. \\ &\qquad \left. + \frac{C_\nu}{\|\overline{A}_k\|^{\frac{1}{2}}} \left(1+\sigma_a k\right) \right) \|\mathbf{v}\|_A \\ &\leq \frac{\|A\|^{\frac{1}{2}}}{(2\nu+1)^k} \sigma_a \frac{(2\nu+1)^k}{\|A\|^{\frac{1}{2}}} \|\mathbf{v}\|_A \\ &\qquad + C_\nu(1+\sigma_a k) \|\mathbf{v}\|_A \\ &= \left[\sigma_a + C_\nu \left(1+\sigma_a k\right)\right] \|\mathbf{v}\|_A. \end{aligned} \tag{3.35}$$

Final estimates. In conclusion, we are ready to complete the proof of the following main result (given for $\nu = 1$ in [**SA**]).

THEOREM 3.1. *Under the following assumptions:*

- *the approximation property* (3.23) *of the piecewise constant interpolants I_k (from coarse level $k+1$ all the way up to finest level 0) holds. This is the case if the kth level composite aggregates have diameter that grows not faster than $(2\nu+1)^k h$ (where h is the finest meshsize).*
- *the choice of smoother is $M_k \simeq \|\overline{A}_k\|\ D_k$, where $D_k = I_{k-1}^T I_{k-1}$ and $\overline{A}_k = D_k^{-\frac{1}{2}} A_k D_k^{-\frac{1}{2}}$;*
- *the choice* (3.24) *of the polynomials φ_ν with $b \geq \|\overline{A}_k\|$ at every level k used in the construction of the smoothed interpolation matrices $P_k = \varphi_\nu(D_k^{-1}A_k)\overline{I}_k$, where $\overline{I}_k$ is the piecewise constant interpolant from coarse level $k+1$ to the next fine level k;*

the resulting $V(1,1)$–cycle MG preconditioner B is nearly spectrally equivalent to A with $K_* \leq C\ell^3$, where K_* is the constant in (3.22).

PROOF. It remains to use the estimates (3.35) and (3.34), for the particular decomposition $\mathbf{v}_k = (\overline{Q}_{k-1} - P_k\overline{Q}_k)\mathbf{v} + P_k\overline{Q}_k\mathbf{v}$ and $\mathbf{v}_{k+1} = \overline{Q}_k\mathbf{v}$. We have, see identity (3.28),

$$\begin{aligned}
\mathbf{v}^T B\mathbf{v} &\leq \left[\|P_\ell \overline{Q}_\ell \mathbf{v}\|^2_{A_\ell} + 2\sum_k \|\overline{A}_k\| \| \left(\overline{Q}_{k-1} - P_k\overline{Q}_k\right)\mathbf{v}\|^2_{D_k} + 2\sum_k \|P_k\overline{Q}_k\mathbf{v}\|^2_{A_k} \right] \\
&\leq \left(2\sum_{k\leq \ell} \|\overline{Q}_{k-1}\mathbf{v}\|^2_{A_k} + 2\sum_{k<\ell} \left(\sigma_a + C_\nu(1+k\sigma_a)\right)^2 \|\mathbf{v}\|^2_A \right) \\
&\leq C\left[\ell^3 + \sum_{k<\ell} k^2 \right] \|\mathbf{v}\|^2_A \\
&\leq C\ell^3 \|\mathbf{v}\|^2_A.
\end{aligned}$$

□

CHAPTER 12

Appendix: H^1_0-norm characterization

Here we provide some auxiliary results on boundedness and approximation properties of finite element quasi-interpolants and respective L_2 projections.

1. A H^1-bounded approximation operator

Let V_h be a given finite element space spanned by the Lagrangian basis $\{\varphi_i^{(h)}\}_{\mathbf{x}_i\in\mathcal{N}_h}$. Define the linear operator

$$\widetilde{Q}_h v = \sum_{\mathbf{x}_i\in\mathcal{N}_h} \frac{(v,\ \varphi_i^{(h)})}{(1,\ \varphi_i^{(h)})}\ \varphi_i^{(h)}.$$

For any element τ from the triangulation $\mathcal{T}_h$ consider its neighborhood Ω_τ of immediate element neighbors. It is clear then that the diameter of Ω_τ is of order $\mathcal{O}(h)$. Since the sets Ω_τ have bounded overlap the sum of integrals $\sum\limits_{\tau\in\mathcal{T}_h}\int\limits_{\Omega_\tau}\psi^2(\mathbf{x})\,d\mathbf{x}$ is bounded by a constant times the integral $\int\limits_{\Omega}\psi^2(\mathbf{x})\,d\mathbf{x}$ for any function $\psi\in L_2(\Omega)$.

Due to the local support of the basis functions $\varphi_i^{(h)}$, the following local estimate is immediate (using Cauchy–Schwarz inequality and $\int\varphi_i^{(h)}\simeq h^d$, $\int(\varphi_i^{(h)})^2\simeq h^d$)

$$\begin{aligned}\int\limits_\tau(\widetilde{Q}_h v)^2\,d\mathbf{x} &= \int\limits_\tau\left(\sum_{\mathbf{x}_i\in\tau}\tfrac{(v,\,\varphi_i^{(h)})}{(1,\,\varphi_i^{(h)})}\,\varphi_i^{(h)}\right)^2 d\mathbf{x}\\ &\le C\sum_{\mathbf{x}_i\in\tau}\int\limits_{\Omega_\tau} v^2\,d\mathbf{x}\frac{\left(\int\left(\varphi_i^{(h)}\right)^2 d\mathbf{x}\right)^2}{\left(\int\limits_{\Omega_\tau}\varphi_i^{(h)}\,d\mathbf{x}\right)^2}\\ &\le C\int\limits_{\Omega_\tau} v^2(\mathbf{x})\,d\mathbf{x}.\end{aligned}$$

Therefore, we also have

$$\int\limits_\tau(v-\widetilde{Q}_h v)^2\,d\mathbf{x}\le 2\int\limits_\tau v^2\,d\mathbf{x}+2\int\limits_\tau\left(\widetilde{Q}_h v\right)^2\,d\mathbf{x}\le C\int\limits_{\Omega_\tau}v^2(\mathbf{x})\,d\mathbf{x}.$$

Applying the same inequality for $v:=v-c$ on Ω_τ, since then $\left((I-\widetilde{Q}_h)c\right)\Big|_\tau=0$ for any constant function c on Ω_τ, we also get

$$\int\limits_\tau(v-\widetilde{Q}_h v)^2\,d\mathbf{x}\le C\int\limits_{\Omega_\tau}(v-c)^2\,d\mathbf{x}.$$

Using this inequality for c being the average value of v over Ω_τ, and applying the Poincaré inequality, we arrive at the following local approximation estimate

$$\int_\tau (v - \widetilde{Q}_h v)^2 \, d\mathbf{x} \leq C \, \mathrm{diam}^2(\Omega_\tau) \int_{\Omega_\tau} |\nabla v|^2 \, d\mathbf{x} \leq C \, h^2 \int_{\Omega_\tau} |\nabla v|^2 \, d\mathbf{x}.$$

The desired global L_2–approximation property follows after summation over $\tau \in \mathcal{T}_h$ and using the bounded overlap of the neighborhood sets $\{\Omega_\tau\}$, i.e., we have

$$\|v - \widetilde{Q}_h v\|_0^2 \leq Ch^2 \, |v|_1^2 = Ch^2 \, a(v, \, v) \tag{3.36}$$

The next property is to show that $\widetilde{Q}_h v$ is bounded in $|\cdot|_1$. Introducing the weighted average values $\overline{v}_i = \frac{(v, \, \varphi_i^{(h)})}{(1, \, \varphi_i^{(h)})}$, we have

$$|\widetilde{Q}_h v|_1^2 = \sum_\tau \int_\tau \left| \sum_{\mathbf{x}_i \in \tau} \overline{v}_i \nabla \varphi_i^{(h)} \right|^2 \, d\mathbf{x}.$$

Now use the fact that on τ, $\sum\limits_{\mathbf{x}_i \in \tau} \varphi_i^{(h)} = 1$, hence $\sum\limits_{\mathbf{x}_i \in \tau} \nabla \varphi_i^{(h)} = 0$. That is, $\nabla \varphi_{i_0}^{(h)} = - \sum\limits_{\mathbf{x}_i \in \tau \setminus \{\mathbf{x}_{i_0}\}} \nabla \varphi_i^{(h)}$, which implies

$$|\widetilde{Q}_h v|_1^2 = \sum_\tau \int_\tau \left| \sum_{\mathbf{x}_i \in \tau} (\overline{v}_i - \overline{v}_{i_0}) \nabla \varphi_i^{(h)} \right|^2 \, d\mathbf{x}.$$

Applying Cauchy-Schwarz inequality, we then obtain

$$\begin{aligned} |\widetilde{Q}_h v|_1^2 &\leq C \sum_\tau \sum_{\mathbf{x}_i \in \tau} (\overline{v}_i - \overline{v}_{i_0})^2 \int_\tau \left| \nabla \varphi_i^{(h)} \right|^2 \, d\mathbf{x} \\ &\leq C \sum_\tau \sum_{\mathbf{x}_i \in \tau} (\overline{v}_i - \overline{v}_{i_0})^2 h^{d-2}. \end{aligned} \tag{3.37}$$

In what follows, we need the following estimate bounding the deviation of the weighted averages $\overline{v}_i$ from the simple averages $\overline{v}_\tau = \frac{1}{|\Omega_\tau|} \int\limits_{\Omega_\tau} v \, d\mathbf{x}$. We have, for any $\mathbf{x}_i \in \tau$, based on the Cauchy-Schwarz and Poincaré inequalities, and the fact that $\frac{\|\varphi_i\|_0}{(1, \, \varphi_i)} \simeq h^{-\frac{d}{2}}$,

$$\overline{v}_i - \overline{v}_\tau = \frac{(v - \overline{v}_\tau, \, \varphi_i)}{(1, \, \varphi_i)} \leq Ch \, \left(\int_{\Omega_\tau} |\nabla v|^2 \, d\mathbf{x} \right)^{\frac{1}{2}} \, \frac{\|\varphi_i\|_0}{(1, \, \varphi_i)} \leq Ch^{1-\frac{d}{2}} \, \left(\int_{\Omega_\tau} |\nabla v|^2 \, d\mathbf{x} \right)^{\frac{1}{2}}.$$

This shows then for any $\mathbf{x}_i \in \tau$, that the difference of the weighted average values $\overline{v}_i - \overline{v}_{i_0}$, $\mathbf{x}_i$, $\mathbf{x}_{i_0} \in \tau$, can be bounded by the seminorm $|.|_{1, \, \Omega_\tau}$ of v over the element neighborhood Ω_τ. That is, we have the local estimates

$$(\overline{v}_i - \overline{v}_{i_0})^2 \leq 2 \, (\overline{v}_i - \overline{v}_\tau)^2 + 2 \, (\overline{v}_{i_0} - \overline{v}_\tau)^2 \leq C \, h^{2-d} \int_{\Omega_\tau} |\nabla v|^2 \, d\mathbf{x}.$$

The later estimates, after summation over $\tau \in \mathcal{T}_h$, used in (3.37) based on the bounded overlap of the local subdomains Ω_τ gives the desired energy bound

$$|\widetilde{Q}_h v|_1^2 \leq C \sum_\tau \int_{\Omega_\tau} |\nabla v|^2 \, d\mathbf{x} \leq C \, |v|_1^2 = C \, a(v, \, v). \tag{3.38}$$

COROLLARY 1.1. *The L_2–projection operator $Q_h : \; L_2(\Omega) \mapsto V_h$ is bounded in H^1 and has 1st order approximation in L_2 for functions in $H^1(\Omega)$.*

PROOF. The L_2–approximation is seen from the minimization property of the L_2–projection, i.e.,

$$\|v - Q_h v\|_0 = \inf_{\varphi \in V_h} \|v - \varphi\|_0 \leq \|v - \widetilde{Q}_h v\|_0,$$

and estimate (3.36). The H^1–boundedness follows from the triangle inequality

$$|Q_h v|_1 \leq |Q_h v - \widetilde{Q}_h v|_1 + |\widetilde{Q}_h v|_1,$$

the inverse inequality $|\psi_h|_1 \leq C_I h^{-1} \, \|\psi_h\|_0$ used for $\psi_h = Q_h v - \widetilde{Q}_h v \in V_h$, the proven L_2–approximation properties of Q_h and $\widetilde{Q}_h$ all used after using the triangle inequality

$$\|Q_h v - \widetilde{Q}_h v\|_0 \leq \|v - Q_h v\|_0 + \|v - \widetilde{Q}_h v\|_0 \leq 2 \, \|v - \widetilde{Q}_h v\|_0 \leq C \, h \, |v|_1,$$

and the H^1–boundedness (3.38) of $\widetilde{Q}_h$. In conclusion, we have the estimate

$$|Q_h v|_1 \leq C \, |v|_1.$$

□

REMARK 1.1. *In the analysis of MG, we can use multilevel decompositions based on the operators $\widetilde{Q}_k = \widetilde{Q}_{h_k}$ by letting $v_k^f = (\widetilde{Q}_k - \widetilde{Q}_{k-1})v$ for $k \geq 1$, and $v_0^f = v_0 = \widetilde{Q}_0 v$. Alternatively, we may use decompositions based on the L_2–projections Q_k. Then, we need to verify assumption (S) for the decomposition $v = \sum_k v_k^f$ based either on $\widetilde{Q}_k$ or Q_k. It is true (see Section 2) that*

$$\sum_k h_k^{-2} \|v_k^f\|_0^2 \simeq |v|_1^2 = C \, a(v, \, v).$$

An application of $\widetilde{Q}_h$ to Schwarz methods. Assume that a given computational domain (polygon or polytope) $\Omega \subset \mathbb{R}^d$ is covered by a set of overlapping subdomains Ω_i, $i = 1, \, 2, \, \ldots, \, m$ with bounded diameter of order $\mathcal{O}(H)$. Also, let $\{\theta_i\}$ be a partition of unity of smooth functions θ_i that are supported in $\overline{\Omega}_i$ such that

$$\nabla \theta_i \leq C H^{-1}.$$

Let $\mathcal{T}_h$ be a given triangulation of Ω such that each Ω_i is completely covered by elements from $\mathcal{T}_h$. We assume that $h \leq H$ but no restriction on the size of H/h is assumed. In practice, the domains Ω_i can be constructed as unions of elements T from a coarse triangulation $\mathcal{T}_H$ and then $\mathcal{T}_h$ is obtained by several steps of refinement of $\mathcal{T}_H$. For a partition of unity functions θ_i, we can simply use the basis of a H^1-conforming finite element space V_H associated with $\mathcal{T}_H$.

Given a H^1-conforming finite element space V_h associated with $\mathcal{T}_h$ and let $\{\varphi_i\}_{\mathbf{x}_i \in \mathcal{N}_h}$ be its nodal basis. We are interested, for a given $v \in V_h$, in the following local components

$$\psi_i \equiv \widetilde{Q}_h(\chi_i v) = \sum_{\mathbf{x}_j \in \mathcal{N}_h \cap \overline{\Omega}_i} \frac{(\chi_i v,\ \varphi_j)}{(1,\ \varphi_j)}\ \varphi_j \in V_h.$$

It is clear that ψ_i are supported in $\widehat{\Omega}_i$ which is the union of Ω_i and the neighboring elements $\tau \in \mathcal{T}_h$. The components ψ_i also satisfy

$$\sum_i \psi_i = \widetilde{Q}_h \left(\sum_i \chi_i v \right) = \widetilde{Q}_h v.$$

The difference $v - \widetilde{Q}_h v$ can be decomposed as $\sum\limits_i \epsilon_i$, where each

$$\epsilon_i = \sum_{\mathbf{x}_j \in \mathcal{N}_h \cap \Omega_i} w_{j,i} \left(v(\mathbf{x}_j) - \frac{(v,\ \varphi_j)}{(1,\ \varphi_j)} \right)\ \varphi_j,$$

is supported in $\overline{\Omega}_i$. The weights $w_{j,i}$ are between zero and one and reflect the fact that each node $\mathbf{x}_j$ can belong to several subdomains Ω_i (due to their overlap).

Our goal is to bound the local components $\epsilon_i + \psi_i$ in L_2 and H^1. For any i, each individual term in ϵ_i,

$$\delta_j \equiv \left(v(\mathbf{x}_j) - \frac{(v,\ \varphi_j)}{(1,\ \varphi_j)} \right)\ \varphi_j$$

is bounded in L_2 by $Ch\|\nabla v\|_{0,\Omega(\mathbf{x}_j)}$, where $\Omega(\mathbf{x}_j)$ is the union of all elements τ that share node $\mathbf{x}_j$. To prove this, we first notice that

$$\|v(\mathbf{x}_j)\varphi_j\|_0^2 = v^2(\mathbf{x}_j)\ \|\varphi_j\|_0^2 \le Ch^d\ v^2(\mathbf{x}_j) \le C\|v\|_{0,\,\Omega(\mathbf{x}_j)}^2,$$

and

$$\| \left(\frac{(v,\ \varphi_j)}{(1,\ \varphi_j)} \right)\ \varphi_j\|_0^2 \le \|v\|_{0,\,\Omega(\mathbf{x}_j)}^2\ \frac{\|\varphi_j\|_0^4}{(1,\ \varphi_j)^2} \le C\ \|v\|_{0,\,\Omega(\mathbf{x}_j)}^2.$$

That is, $\|\delta_j\|_0 \le C\ \|v\|_{0,\,\Omega(\mathbf{x}_j)}$. Using this result for $v := v - \text{const}$, noticing that δ_j does not change then, letting const $= \overline{v}_j$, the average of v over $\Omega(\mathbf{x}_j)$, we obtain $\|\delta_j\|_0 \le C\ \|v - \overline{v}_j\|_{0,\,\Omega(\mathbf{x}_j)}$. The estimate $\|\delta_j\|_0 \le Ch\ \|\nabla v\|_{0,\,\Omega(\mathbf{x}_j)}$ follows then by Poincaré inequality.

To bound the local terms ϵ_i, we use Cauchy–Schwarz inequality and the last estimate. We have

$$\|\epsilon_i\|_0^2 \le C \sum_{\mathbf{x}_j \in \mathcal{N}_h \cap \Omega_i} \|\delta_j\|_0^2 \le Ch^2 \sum_{\mathbf{x}_j \in \mathcal{N}_h \cap \Omega_i} \|\nabla v\|_{0,\,\Omega(\mathbf{x}_j)}^2 \le Ch^2\ \|\nabla v\|_{0,\,\widehat{\Omega}_i}^2.$$

The bound in H^1 follows by an inverse inequality used for $\epsilon_i \in V_h$ and the last L_2–estimate, i.e., we have

$$\|\nabla \epsilon_i\|_0^2 \le C\ h^{-2}\ \|\epsilon_i\|_0^2 \le C\ \|\nabla v\|_{0,\,\widehat{\Omega}_i}^2.$$

In conclusion, by summing up the last two estimates, using the bounded overlap of the subdomains Ω_i (and hence of $\widehat{\Omega}_i$), we have

$$h^{-2} \sum_i \|\epsilon_i\|_0^2 + \sum_i \|\nabla \epsilon_i\|_0^2 \leq C \, \|\nabla v\|_0^2. \tag{3.39}$$

We bound now the local functions $\psi_i = \widetilde{Q}_h \, (\chi_i v) \in V_h$. Since $\widetilde{Q}_h$ is bounded in both L_2 and H^1, it is sufficient to bound the functions $\chi_i v$ instead. The L_2–bound is trivial since χ_i is between zero and one. For H^1, using the product rule for derivatives and the assumed bound $|\nabla \chi_i| \leq C H^{-1}$, we have

$$C^{-1} \, \|\nabla \psi_i\|_0^2 \leq \|\nabla(\chi_i v)\|_0^2 \leq 2\|v \nabla \chi_i\|_0^2 + 2\|\chi_i \nabla v\|_0^2 \leq C H^{-2} \|v\|^2_{0,\,\Omega_i} + 2\|\nabla v\|^2_{0,\,\Omega_i}.$$

The final estimate follows by summation over i, using again the bounded overlap of the Schwarz subdomains Ω_i, i.e., we have

$$\sum_i \|\nabla \psi_i\|_0^2 \leq C H^{-2} \, \|v\|_0^2 + C \, \|\nabla v\|_0^2. \tag{3.40}$$

The following result is the essence of the analysis of the so-called overlapping Schwarz methods that exploit subdomain solvers (locally in each subdomain Ω_i). Also, to achieve optimal condition number, the Schwarz methods utilize in addition a coarse–grid solver based on a coarse subspace V_H associated with a coarse mesh $\mathcal{T}_H$ of size H that is comparable to the characteristic diameter of the subdomains Ω_i. We recall, that the subdomains Ω_i are assumed to have overlap of size comparable to their diameter. The latter property is needed to show that a partition of unity functions χ_i, with controlled bound $\mathcal{O}(H^{-1})$ on their gradient, is possible to construct.

THEOREM 1.1. *Assume that for any $v \in V_h$ there is a coarse-grid function $v_H \in V_H$ such that*

$$\|v - v_H\|_0 \leq C \, H \, \|\nabla v\|_0 \text{ and } \|\nabla v_H\|_0 \leq C \, \|\nabla v\|_0.$$

For example, we can choose $v_H = \widetilde{Q}_H v$. Consider the local components $v_i = \epsilon_i + \psi_i$ supported in $\widehat{\Omega}_i$ constructed for the function $v - v_H \in V_h$. Then, for the decomposition

$$v = v_H + \sum_i v_i,$$

the following stability estimate holds

$$\|\nabla v_H\|_0^2 + \sum_i \|\nabla v_i\|_0^2 \leq C \, \|\nabla v\|_0^2.$$

2. H^1_0–norm characterization

In this section we present in a constructive way a $H^1_0(\Omega)$–norm characterization. First the result is proven for a convex polygonal domain Ω. Consider

$$-\Delta u = f(x), \; x \in \Omega,$$

subject to $u = 0$ on $\partial\Omega$. Since Ω is convex the following full regularity estimate holds

$$\|u\|_2 \leq C \, \|f\|_0.$$

We assume now that Ω is triangulated on a sequence of uniformly refined triangulations with characteristic mesh size $h_k = h_0 2^{-k}$, $k \geq 0$, and it is well–known that the respective finite element spaces of piecewise linear functions $V_k = V_{h_k}$ satisfy $\overline{\cup V_k} = H_0^1(\Omega)$. Define the L_2–projections $Q_k :\ L_2(\Omega) \mapsto V_k$. Then, we can prove the following main result

$$\sum_k h_k^{-2} \|(Q_k - Q_{k-1})v\|_0^2 \simeq \|v\|_1^2. \tag{3.41}$$

More generally, we have the following main characterization of $H_0^1(\Omega)$, a result originally proven by Oswald [**Os94**],

$$\|v\|_1^2 \simeq \inf_{v=\sum\limits_k v_k,\ v_k \in V_k} \sum_k h_k^{-2} \|v_k\|_0^2. \tag{3.42}$$

To this end let us define the elliptic–projections $\pi_k :\ H_0^1(\Omega) : \mapsto V_k$ in the standard way

$$(\nabla \pi_k v,\ \varphi) = (\nabla v,\ \nabla \varphi),\ \text{for all } \varphi \in V_k.$$

Based on the optimal L_2–error estimate $\|v - \pi_{k-1} v\|_0 \leq C h_k \|v\|_1$, for $v := (\pi_k - \pi_{k-1})v$, and using the H_0^1–orthogonality of the projections, we have

$$\sum_k h_k^{-2} \|(\pi_k - \pi_{k-1})v\|_0^2 \leq C \sum_k \|(\pi_k - \pi_{k-1})v\|_1^2 = C \sum_k \left(\|\pi_k v\|_1^2 - \|\pi_{k-1} v\|_1^2 \right) = \|v\|_1^2.$$

Finally, from the following chain of inequalities, using the fact that Q_k are L_2–symmetric and that $(Q_k - Q_{k-1})^2 = Q_k - Q_{k-1}$, and the optimal L_2–error estimate, we have

$$\begin{aligned}
\sum_k h_k^{-2} \|(Q_k - Q_{k-1})v\|_0^2 &= \sum_k h_k^{-2} ((Q_k - Q_{k-1})v,\ v) \\
&= \sum_k h_k^{-2} ((Q_k - Q_{k-1})v,\ \sum_{j \geq k} (\pi_j - \pi_{j-1})v) \\
&\leq \sum_k h_k^{-2} \|(Q_k - Q_{k-1})v\|_0 \sum_{j \geq k} \|(\pi_j - \pi_{j-1})v\|_0 \\
&\leq C \sum_k \sum_{j \geq k} \frac{1}{2^{j-k}} \left(h_k^{-1} \|(Q_k - Q_{k-1})v\|_0 \right) \|(\pi_j - \pi_{j-1})v\|_1 \\
&\leq C \left(\sum_k \sum_{j \geq k} \frac{1}{2^{j-k}} \ h_k^{-2} \|(Q_k - Q_{k-1})v\|_0^2 \right)^{\frac{1}{2}} \\
&\quad \times \left(\sum_j \sum_{k \leq j} \frac{1}{2^{j-k}} \ \|(\pi_j - \pi_{j-1})v\|_1^2 \right)^{\frac{1}{2}} \\
&\leq C \left(\sum_k h_k^{-2} \|(Q_k - Q_{k-1})v\|_0^2 \right)^{\frac{1}{2}} \|v\|_1.
\end{aligned}$$

The latter shows the first desired result (3.41). Applying exactly the same argument as above to any decomposition $v = \sum\limits_j v_j,\ v_j \in V_j$ (formally replacing $(\pi_j - \pi_{j-1})v$ with

$v_j \in V_j$) we show the inequality

$$\sum_k h_k^{-2}\|(Q_k - Q_{k-1})v\|_0^2 \le \sum_k \sum_{j\ge k} \frac{1}{2^{j-k}} \left(h_k^{-1}\|(Q_k - Q_{k-1})v\|_0\right)\left(h_j^{-1}\|v_j\|_0\right).$$

The latter implies,

$$\sum_k h_k^{-2}\|(Q_k - Q_{k-1})v\|_0^2 \le \mathcal{C} \inf_{v=\sum\limits_j v_j,\ v_j\in V_j} \sum_j h_j^{-2}\|v_j\|_0^2.$$

That is, the decomposition $v = \sum\limits_j (Q_j - Q_{j-1})v$ is quasi–optimal. This (together with (3.43) below) shows the well–known norm characterization (3.42) of $H^1_0(\Omega)$.

For a more general domain Ω we assume that it can be split into overlapping convex subdomains Ω_m, $m = 1, \ldots, m_0$ for a fixed number m_0. We also assume that there is a partition of unity of smooth functions θ_m such that $0 \le \theta_m \le 1$, $\sum\limits_m \theta_m = 1$ and θ_m is supported in $\overline{\Omega}_m$. Then, since $v = \sum\limits_m \theta_m v$ and $\|v\theta_m\|_1^2 \le C\|v\nabla\theta_m\|^2_{0,\,\Omega_m} + C\|v\|^2_{1,\,\Omega_m}$, if we can choose θ_m such that $v\nabla\theta_m \in H^1_0(\Omega_m)$ with H^1–norm bounded in terms of $\|v\|_1$, then the decomposition $v = \sum\limits_m \theta_m v$ will be stable in $H^1_0(\Omega)$ and the functions $\theta_m v$ have the proven $H^1_0(\Omega_m)$ norm characterization (since Ω_m are convex). Such a result has been shown in Lions [**Li87**] for a L–shaped domain Ω with $m_0 = 2$. Thus we can find a stable decomposition for each v_m (supported in $\overline{\Omega}_m$) and thus a stable decomposition of the finite sum $v = \sum\limits_m v_m$ is constructed which proves (3.42) in one of the directions.

For any decomposition $v = \sum\limits_j v_j$, $v_j \in V_j$, and for a fixed $\alpha \in (0, \frac{1}{2})$, using the inequality $(p,q) \le \|p\|_\alpha \|q\|_{-\alpha}$ and appropriate inverse inequalities, we have

$$\begin{aligned}
\|v\|_1^2 &= \sum_k (\nabla(\pi_k - \pi_{k-1})v,\ \sum_{j\ge k} \nabla v_j) \\
&\le \sum_k \sum_{j\ge k} \|(\pi_k - \pi_{k-1})v\|_{1+\alpha}\ \|v_j\|_{1-\alpha} \\
&\le C \sum_k \sum_{j\ge k} h_k^{-\alpha}\|(\pi_k - \pi_{k-1})v\|_1 h_j^{-1+\alpha}\ \|v_j\|_0 \\
&= C \sum_k \sum_{j\ge k} \left(\frac{1}{2^\alpha}\right)^{j-k} \|(\pi_k - \pi_{k-1})v\|_1 \left(h_j^{-1}\ \|v_j\|_0\right) \\
&\le C \left(\sum_k \sum_{j\ge k} \left(\frac{1}{2^\alpha}\right)^{j-k} \|(\pi_k - \pi_{k-1})v\|_1^2\right)^{\frac{1}{2}} \\
&\quad \times \left(\sum_j \sum_{k\le j} \left(\frac{1}{2^\alpha}\right)^{j-k} h_j^{-2}\|v_j\|_0^2\right)^{\frac{1}{2}} \\
&\le C/(1 - 2^{-\alpha})\ \|v\|_1 \left(\sum_j h_j^{-2}\|v_j\|_0^2\right)^{\frac{1}{2}}.
\end{aligned} \tag{3.43}$$

This shows the remaining fact that $\|v\|_1^2$ is bounded in terms of the r.h.s. of (3.42).

Bibliography

[Br01] D. Braess, "*Finite Elements, Theory, fast solvers and applications in solid mechanics*", 2nd edition, Cambridge University Press, Cambridge, 2001.

[BH83] D. Braess and W. Hackbusch, "A new convergence proof of the multigrid method including the V-cycle," SIAM Journal on Numerical Analysis **20**(1983), pp. 967-975.

[Br93] J. H. Bramble, "*Multigrid Methods*", Pitman Research Notes in Mathematics Series, No. **294**, Longman Scientific and Technical, John Wiley & Sons Inc, New York, 1993.

[AB77] A. Brandt, *Multilevel adaptive solutions to boundary-value problems*, Mathematics of Computation **31**(1977), pp. 333–390.

[BS02] S. C. Brenner and L. R. Scott, *"The Mathematical Theory of Finite Element Methods"*, 2nd edition, Springer, New York, 2002.

[Ci02] P. G. Ciarlet, *"The Finite Element Method for Elliptic Problems"*, Classics in Applied Mathematics 40, SIAM, Philadelphia, 2002.

[Fe64] R. P. Fedorenko, *The speed of convergence of one iterative process*, USSR Comput. Math. Math. Phys. **4**(1964), pp. 227–235.

[Gr94] M. Griebel, "*Multilevel algorithms considered as iterative methods on semidefinite systems,*" SIAM J. Scientific Computing **15**(1994), pp. 547–565.

[Ha82] W. Hackbusch, "Multi-grid convergence theory," In: "Multi-Grid Methods" (W. Hackbusch and U. Trottenberg, eds.) Springer Lecture Notes in Mathematics **960**(1982), pp. 177-219.

[Li87] P.-L. Lions, "*On the Schwarz alternating method. I*", in: R. Glowinski, G. H. Golub, G. A. Meurant, and J. Periaux, eds., *1st International Symposium on Domain Decomposition Methods for PDEs, held in Paris, France, January 7–9, 1987*. SIAM, Philadelphia, PA, 1988, pp. 1–42.

[Os94] Peter Oswald, "Multilevel Finite Element Approximation.*Theory and Applications.*" B.G. Teubner Stuttgart, 1994.

[XZ02] 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**(2002), pp. 573-597.

[VanSA] P. Vaněk, "Acceleration of convergence of a two-level algorithm by smoothing transfer operator," Applications of Mathematics **37**(1992), pp. 265–274.

[SA] P. Vaněk, M. Brezina, and J. Mandel, "*Convergence of algebraic multigrid based on smoothed aggregation,*" Numerische Mathematik **88**(2001), pp. 559–579.

[Va08] Panayot S. Vassilevski, "Multilevel Block Factorization Preconditioners, Matrix-based Analysis and Algorithms for Solving Finite Element Equations," Springer, New York, 2008. 514 p.