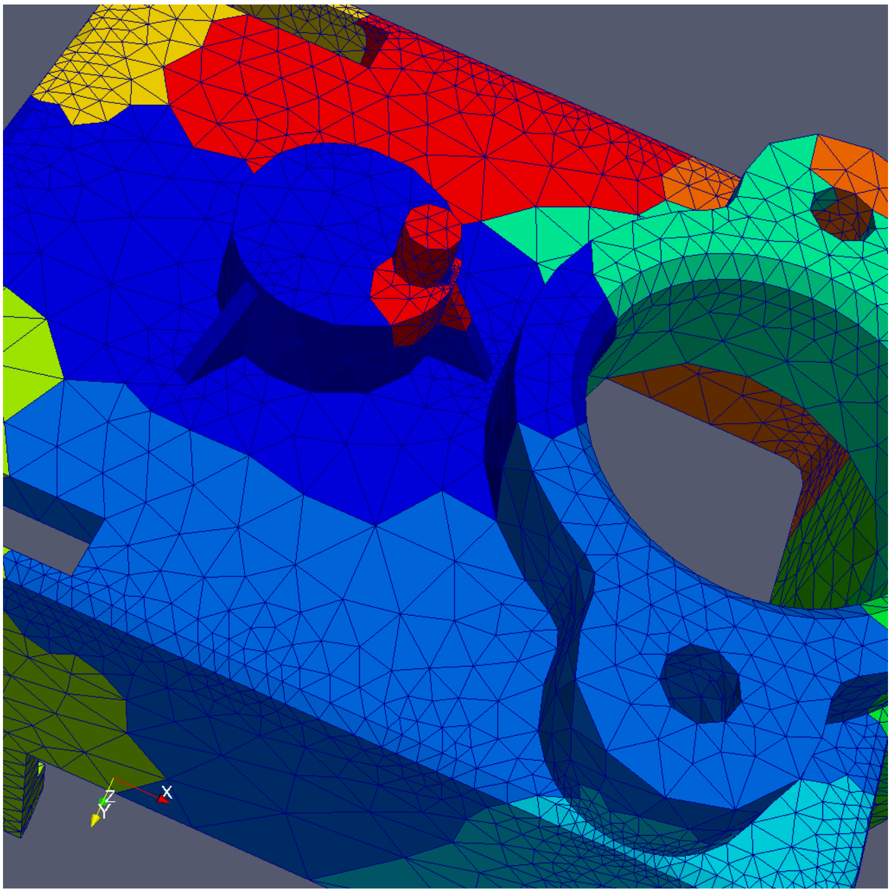
Lecture Notes on Parallel Solution of Large Sparse Linear Systems

Peter Bastian
Interdisziplinäres Zentrum für Wissenschaftliches Rechnen
Universität Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg
`Peter.Bastian@iwr.uni-heidelberg.de`



July 12, 2015

Acknowledgements

These lecture notes grew out of a lecture I have been giving more or less regularly since 1999. The current english version would not exist without the excellent work of Ansgar Burchardt and Judith Stein who typed the manuscript in the winter semester 2012 and summer semester 2015, respectively. I thank them very sincerely for their work. All remaining errors are of course my own.

Peter Bastian Heidelberg, July 2015

Contents

Chapter 1

Recapitulation of the Finite Element Method

In this chapter we want to give a short summary about the Finite Element Method, a numerical technique for finding approximate solutions to boundary value problems for partial differential equations. Introductions to the finite element method can be found in Eriksson et al. [1996]; Braess [2003]; Ciarlet [2002]; Ern and Guermond [2004]; Brenner and Scott [1994]; Rannacher [2006]; Bastian [2014].

Elliptic Model Problem: "Strong Formulation"

Now we consider linear elliptic problems that are commonly found in mechanical and physical partial differential equation models. The aim is to introduce the notion of a weak formulation that gives access to existence and uniqueness results for the solutions and that is well suited for the numerical approximation of such problems.

In the theory of partial differential equations, elliptic operators are differential operators that generalize the Laplace operator. An elliptic differential equation of second order has the form

$$\begin{aligned} -\nabla \cdot (K(x)\nabla u(x)) + c(x)u(x) &= f(x) && x \in \Omega \subset \mathbb{R}^n \\ u(x) &= g(x) && x \in \Gamma_D \subseteq \partial\Omega \\ -K(x)\nabla u(x) \cdot n(x) &= j(x) && x \in \Gamma_N = \partial\Omega \backslash \Gamma_D \end{aligned} \tag{1.1}$$

with the coefficient functions K and c.

We assume Ω to be open, connected and bounded. An important assumption on the coefficient K is that for all $\xi \in \mathbb{R}^n$ we have

$$k_0 \|\xi\|^2 \leq \xi^T K(x) \xi \qquad \forall x \in \Omega$$

which is called uniform ellipticity and that

$$\xi^T K(x) \xi \leq K_0 \|\xi\|^2 \qquad \forall x \in \Omega$$

which is boundedness. Furthermore $K(x)$ is assumed to be symmetric and $c(x) \geq 0$.

Regarding the Problem (1.1) we can investigate the following questions:

- For the problem to be well-posed we have to guarantee that
  - the solution exists,
  - it is unique
  - and stable: $\|u\| \leq c\,(\underbrace{\|f\| + \|g\| + \|j\|}_{data})$.
- For a numerical solution producing an approximation u_h on would like to guarantee an priori error estimate of the form

  $$\|u - u_h\| \leq ch^k \|u\|$$

  where h is a mesh size parameter.
- Guaranteed error control of the numerical solution requires an posteriori error estimate of the form

  $$\|u - u_h\| \leq \eta(u_h)$$

  with an η that is effectively computable.

Note that $\|\cdot\|$ means a "generic" norm in these lecture notes. More over, the strong formulation (1.1) requires very restrictive demands placed on the data (f, g, j) to answer these questions. For this reason we consider the weak/variational formulation.

1.1 The Variational Formulation of Elliptic Partial Differential Equations

We describe the general abstract framework for elliptic problems with homogeneous Dirichlet data, $\partial\Omega = \Gamma_D$ and $g = 0$. To get the variational form we multiply the equation by a "test function" $v(x)$ and do integration by parts:

$$\begin{aligned}\int\limits_{\Omega} [-\nabla \cdot (K\nabla u) + cu]v\,dx &= \int\limits_{\Omega} (K\nabla u) \cdot \nabla v + c\,uv\,dx + \int\limits_{\partial\Omega} (K\nabla u) \cdot \nu v\,ds \\ &= \int\limits_{\Omega} (K\nabla u) \cdot \nabla v + c\,uv\,dx \quad (v = 0 \text{ on } \partial\Omega) \\ &=: a(u, v).\end{aligned}$$

This relation holds true for all test functions $v(x) \in \mathcal{C}^1(\Omega) \cap \mathcal{C}^0(\overline{\Omega})$. The idea is now to reverse the argument and to *define* the function u by requiring

$$a(u, v) = l(v) := \int\limits_{\Omega} fv\,dx$$

for “sufficiently many” test functions v.

Put in an abstract way, the problem reads as follows. Given suitable function spaces U and V (see below) define the function u by the variational formulation:

$$\text{Find } u \in U : \quad a(u,v) = l(v) \qquad \forall v \in V. \tag{1.2}$$

Here, $a(\cdot,\cdot) \in \mathcal{L}(U \times V, \mathbb{R})$ is a so-called bilinear form and $l(\cdot) \in \mathcal{L}(V, \mathbb{R})$ is a linear functional.

Remark 1.1. $\mathcal{L}(U \times V, \mathbb{R})$ is the space of continuous bilinear forms and $\mathcal{L}(V, \mathbb{R})$, is the space of continuous bilinear functionals. $\mathcal{L}(V, \mathbb{R})$ is also abbreviated by V' and is called the dual space of V. □

The following two theorems ensure the existence, uniqueness and stability of the solution given by (1.2).

Theorem 1.2 (Banach-Nečas-Babuška)**.** Let U and V be Banach spaces (complete, linear, normed spaces), let V be reflexive and $a \in \mathcal{L}(U \times V, \mathbb{R})$, $l \in \mathcal{L}(V, \mathbb{R})$. Then (1.2) is well-posed if and only if

$$\exists \alpha > 0 : \inf_{u \in U} \sup_{v \in V} \frac{a(u,v)}{\|u\|_U \|v\|_V} \geq \alpha, \tag{1.3}$$

$$\forall v \in V : (\forall u \in U : a(u,v) = 0) \Rightarrow (v = 0). \tag{1.4}$$

Furthermore, the following stability estimate holds:

$$\|u\|_U \leq \frac{1}{\alpha} \|l\|_{V'}.$$ □

Additional Comments

- The dual space V' is equipped with the norm

$$\|l\|_{V'} = \sup_{\substack{w \in V \\ w \neq 0}} \frac{l(w)}{\|w\|_V}.$$

- $a(u,\cdot) \in V'$ for given $u \in U$.
- The linear operator $A : U \to V'$ is defined by $Au := a(u,\cdot)$.
- (1.2) $\Leftrightarrow$ $Au = l$. In that sense eqref1.2 is a linear equation in function spaces.
- (1.3) $\Leftrightarrow$ A is injective.
- (1.4) $\Leftrightarrow$ A is surjective.
- $f \in L^2(\Omega)$ implies that $l(v) = \int_\Omega f v \, dx = (f,v)_{L^2(\Omega)} \in V'$.

Theorem 1.3 (Lax-Milgram)**.** Let V be a Hilbert space, $a \in \mathcal{L}(V \times V, \mathbb{R})$, $(U = V!)$, and $l \in V'$, i.e. $a(\cdot,\cdot)$ is a continuous bilinear form and $l(\cdot)$ a continuous functional. If the bilinear form $a(\cdot,\cdot)$ is coercive (also called V-elliptic), i.e.

$$\exists \alpha > 0, \forall u \in V : \quad a(u,u) \geq \alpha \|u\|_V^2,$$

then there exists a unique solution to model problem (1.2) and the following stability estimate holds

$$\|u\|_V \leq \frac{1}{\alpha}\|l\|_{V'}. \qquad \square$$

Remark 1.4.

- The Lax-Milgram theorem is proved with the help of the Riesz representation theorem (which requires V to be a Hilbert space) and the Banach fixed-point theorem.
- One can show that Lax-Milgram theorem 1.3 implies Banach-Nečas-Babuška theorem 1.2, but not vice versa.
- Note that we do <u>not</u> assume $a(\cdot,\cdot)$ to be symmetric in order to proof well-posedness.
- For our model problem Lax-Milgram theorem is sufficient. Banach-Nečas-Babuška theorem 1.2 is needed in more complex situations. It is used to proof well-posedness to parabolic equations or even more complex systems of partial differential equations (e.g. Stokes equations).

Sobolev Spaces

In order to prove the well-posedness with the help of Lax-Milgram theorem, we have to find an appropriate Hilbert space. Such spaces are given by so-called Sobolev spaces that consist of weakly differentiable functions.

Definition 1.5 ($L^2(\Omega)$)**.** Sobolev spaces are based on the space of functions which are square integrable in the sense of Lebesgue, i.e.

$$L^2(\Omega) = \left\{ u \ : \ \int\limits_\Omega u^2(x)\,dx < \infty \right\}.$$

Functions in $L^2(\Omega)$ are equipped with the scalar product and norm

$$(u,v)_{0,\Omega} = \int\limits_\Omega uv\,dx, \qquad \|u\|_{0,\Omega} = \sqrt{(u,u)_{0,\Omega}}. \qquad \square$$

L^2 functions are not differentiable in the classical sense and one needs an alternative notion of differentiability. The idea is to use integration by parts to transfer derivatives to a function that is differentiable in the classical sense.

Definition 1.6 (Weak Derivative)**.** Let $\alpha \in \mathbb{N}_0^d$ be a multi-index, that is

$$\alpha := (\alpha_1, ..., \alpha_d) \quad \text{and} \quad |\alpha|_1 := \sum_{i=1}^{d} \alpha_i.$$

Considering a function $u \in L^2(\Omega)$, we say that u is called weakly differentiable, if a function $g \in L^2(\Omega)$ exists, so that for all test functions $\phi \in C_0^\infty(\Omega)$ the following condition holds

$$\int\limits_{\Omega} g(x)\phi(x)\,dx = (-1)^{|\alpha|_1} \int\limits_{\Omega} u(x) \tfrac{\partial^{|\alpha|}}{\partial x^\alpha} \phi(x)\,dx.$$

Such a function g is called the α-th weak derivative of u in the $L^2(\Omega)$ sense and we define $\partial^\alpha u := \frac{\partial^{|\alpha|}}{\partial x^\alpha} u := g$. Here the multi-index notation

$$\frac{\partial^{|\alpha|_1}}{\partial x^\alpha} u(x) = \frac{\partial^{|\alpha|_1}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} u(x)$$

has been used. □

Definition 1.7 (Sobolev space $H^k(\Omega)$)**.** The Hilbert space of all elements $u \in L^2(\Omega)$ with square integrable weak derivatives $\partial^\alpha u \in L^2(\Omega)$ for all α with $|\alpha|_1 \leq k$ is called Sobolev space of order k and will be denoted by $H^k(\Omega)$, i.e.

$$H^k(\Omega) := \{u \in L^2(\Omega) : \; \partial^\alpha u \in L^2(\Omega) \;\; \forall 0 \leq |\alpha|_1 \leq k\}.$$

The Sobolev space $H^k(\Omega)$ is equipped with the inner product

$$(u, v)_{k,\Omega} := \sum_{0 \leq |\alpha|_1 \leq k} \int\limits_{\Omega} (\partial^\alpha u)\,(\partial^\alpha v)\,dx$$

and the induced norm

$$\|u\|_{k,\Omega} := \sqrt{(u, u)_{k,\Omega}}.$$ □

Definition 1.8. The space of all linear continuous functionals $u^* : H^k(\Omega) \to \mathbb{R}$ is denoted by

$$H^{-k}(\Omega) := \mathcal{L}(H^k(\Omega), \mathbb{R}) = (H^k(\Omega))'$$

and is also called the dual space of $H^k(\Omega)$. □

According to the Riesz representation theorem any continuous linear functional $l \in H^{-k}(\Omega)$ can be represented by an element $u_l \in H^k(\Omega)$ via

$$l(v) := (u_l, v)_{k,\Omega}. \tag{1.5}$$

Since we consider Dirichlet boundary conditions in this lecture the following subspaces of Sobolev spaces will be of importance.

Definition 1.9 (Sobolev space $H_0^k(\Omega)$)**.** The Sobolev space of all functions vanishing in a weak sense on the boundary of Ω is given by

$$H_0^k(\Omega) := \{u \in H^k(\Omega) : \, u|_{\partial\Omega} = 0 \text{ “almost everywhere”}\}. \qquad \square$$

Remark 1.10 (Subset relations)**.** By Definition 1.7 the identity $H^0(\Omega) = L^2(\Omega)$ follows. Moreover, we have the following relations

$$\begin{array}{ccccccccccc} \dots & \supset & H^{-1}(\Omega) & \supset & L^2(\Omega) & \supset & H^1(\Omega) & \supset & H^2(\Omega) & \supset & \dots \\ & & & & & & \cup & & \cup & & \\ & & & & & & H_0^1(\Omega) & \supset & H_0^2(\Omega) & \supset & \dots \end{array}$$

where the dual space $L^2(\Omega)'$ has been identified with the space $L^2(\Omega)$ itself. Regarding equation 1.5 the dual space of a Sobolev space is even bigger than the space itself. □

Remark 1.11 (Construction of Sobolev spaces)**.** An alternative way to define Sobolev spaces is to think of them as the completion of a certain function space with respect to a certain norm. These spaces are often labeled as $W^k(\Omega)$.
It can be shown that $W^k(\Omega) = H^k(\Omega)$ holds.

- For $k \geq 0$ the Sobolev space $H^k(\Omega)$ is given as the completion of $\mathcal{C}^k(\overline{\Omega})$ with respect to $\|\cdot\|_{k,\Omega}$.
- For $k > 0$ the Sobolev space $H_0^k(\Omega)$ is given as the completion of $\mathcal{C}_0^\infty(\Omega)$ with respect to $\|\cdot\|_{k,\Omega}$. □

A relation between classical function spaces and Sobolev spaces is given by the following

Proposition 1.12 (Sobolev embedding theorem)**.** For dimension n, $k \in \mathbb{N}_0$ and $k - \frac{n}{2} > m$ there exists a continuous embedding

$$W^{k,p}(\Omega) \hookrightarrow \mathcal{C}^m(\Omega) \subset \mathcal{C}(\Omega). \qquad \square$$

Application of Lax-Milgram-Theorem 1.3

Now, we want to apply Lax-Milgram Theorem to our model problem in order to proof the well-posedness of the problem. To do so, we have to determine an appropriate Hilbert space V and show that the bilinear form $a(\cdot,\cdot)$ is coercive and continuous with respect to the norm of the Hilbert space. Moreover, continuity of the linear functional l is required which we will presuppose in the considered examples and can be easily achieved since $f \in L^2(\Omega)$ already implies $l(v) = \int_\Omega fv\,dx \in V'$. The following examples differ only in the given boundary conditions.

Example: Homogeneous Dirichlet boundary conditions Let us considering Problem 1.1 with $\Gamma_D = \partial\Omega$, $g = 0$, so called homogenous Dirichlet boundary conditions.

We take the Hilbert space $V = H_0^1(\Omega)$ equipped with the inner product $(\cdot,\cdot)_{1,\Omega}$. In order to prove continuity and coercivity of the bilinear form with respect to V, we need Friedrich's inequality, which can be proved by the fundamental theorem of calculus and the Cauchy-Schwarz inequality.

Theorem 1.13 (Friedrich's inequality)**.** For every function $v \in H_0^1(\Omega)$

$$\|v\|_{0,\Omega} \le s_\Omega\,|v|_{1,\Omega} = s_\Omega\|\nabla v\|_{0,\Omega}$$

holds with the diameter $s_\Omega = \operatorname{diam}(\Omega)$ of the domain Ω and the semi-norm

$$|v|_{k,\Omega} = \Bigg(\sum_{|\alpha|=k}\int\limits_\Omega (\partial^\alpha v)^2\,dx\Bigg)^{\frac{1}{2}} \quad \forall v \in H_0^1(\Omega). \qquad \square$$

Using Friedrich's inequality one can show that $|.|_{1,\Omega}$ is a norm on V and this norm is equivalent to $\|.\|_{1,\Omega}$.

Example: Pure Neumann boundary conditions Now we consider the problem with pure Neumann boundary conditions, i.e. $\Gamma_D = \emptyset$ and $\Gamma_N = \partial\Omega$.

Here we use the Sobolev space $V = \{v \in H^1(\Omega) : \int_\Omega v\,dx = 0\}$ with inner product $(\cdot,\cdot)_{1,\Omega}$ to guarantee the well-posedness of the regarded problem. Note that this space does not explicitely include a boundary condition as it has been in the previous case. Instead we expect all functions to have a mean value equal to zero in order to assure the uniqueness of the solution. For the proof of coercivity and continuity we need:

Theorem 1.14 (Poincaré's inequality)**.** There exist positive constants c_1, c_1 such that

$$\|v\|_{0,\Omega}^2 \le c_1|v|_{1,\Omega}^2 + c_2\Bigg(\int\limits_\Omega v\,dx\Bigg)^2 \quad \forall v \in H^1(\Omega). \qquad \square$$

Theorem 1.15 (Trace Theorem)**.** Assume Ω is bounded and has Lipschitz boundary. Then there exists a bounded linear operator $\gamma : H^1(\Omega) \to L^2(\partial\Omega)$ such that

$$\|\gamma v\|_{0,\partial\Omega} \leq c\|v\|_{1,\Omega} \quad \forall v \in H^1(\Omega).$$

In the original version the existing operator is even stronger: $\gamma : H^1(\Omega) \to H^{\frac{1}{2}}(\partial\Omega)$, but the above formulation is sufficient for our purposes. □

Example: Inhomogeneous Dirichlet boundary conditions As in the first example, we assume $\Gamma_D = \partial\Omega$ but with the difference that now we have $g \neq 0$. In this case we decompose our solution into a homogeneous $u_0 \in H_0^1(\Omega)$ and non-homogeneous part $u_g \in H^1(\Omega)$, i.e.

$$u = u_0 + u_g$$

and we further assume the inhomogeneous part to be an extension of the boundary values $\gamma u_g = g$ with the operator $\gamma : H^1(\Omega) \to H^{\frac{1}{2}}(\Omega)$ from the trace theorem. Note that this requires $g \in H^{\frac{1}{2}}(\Omega)$.

With the help of this decomposition we can treat the problem similar to the homogeneous Dirichlet example:

$$\text{Find } u_0 \in H_0^1(\Omega) : \quad a(u_0, v) = l(v) - a(u_g, v) \quad \forall v \in H_0^1(\Omega).$$

Mixed boundary conditions Regarding mixed boundary conditions $\Gamma_D \subset \partial\Omega, \Gamma_D \neq \emptyset$ we can use the Hilbert space

$$V = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D \text{ “almost everywhere”}\}$$

in order to prove well-posedness. The proof of coercivity then requires a variant of Friedrich's inequality.

1.2 Conforming Finite Element Method

Definition 1.16 (Conformity)**.** Let V be an adapted Sobolev space to the variational problem (1.2) and V_h be the finite-dimensional Finite Element ansatz space. Then the discretization V_h is called “conforming”, if

$$V_h \subset V$$

or else it is called “non-conforming”. □

An important characterization of finite-dimensional subspaces of Sobolev spaces can be deduced from the following theorem.

Theorem 1.17. Let Ω be a bounded domain, $\{\omega_1, \dots, \omega_N\}$ a partitioning of Ω into a finite number of subdomains and V_h a space of functions such that for $v \in V_h$ we have $v|_{\omega_i} \in \mathcal{C}^\infty$. Then $V_h \subset H^k(\Omega)$, $k \geq 1$, if and only if $V_h \subset \mathcal{C}^{k-1}(\overline{\Omega})$. □

In our applications we need $k = 1$. From the theorem we conclude that a piecewise infinitely differentiable function, e.g. a piecewise polynomial, is in H^1 if and only if the function is globally continuous. The conforming finite element method comprises a specific way to construct the finite-dimensional space V_h using piecewise polynomial functions that are globally continuous.

The Lax-Milgram theorem immediately establishes the solution of the variational problem

$$\text{Find } u_h \in V_h\text{:} \quad a(u_h, v) = l(v) \quad \forall v \in V_h \tag{1.6}$$

in the subspace V_h.

Any finite dimensional vector space is spanned by a set of basis functions

$$V_h = \operatorname{span}\{\varphi_1^h, ..., \varphi_{N_h}^h\}.$$

Using the basis, for every $u_h \in V_h$ we have the representation

$$u_h = \sum_{j=1}^{N_h} z_j^h \varphi_j^h.$$

Inserting the basis representation into the weak discrete problem (1.6) results in a linear system of equations:

$$\begin{aligned}
&\text{Find } u_h \in V_h\text{:} && a(u_h, v) = l(v) && \forall v \in V_h \\
&\Leftrightarrow && a\left(\sum_{j=1}^{N_h} z_j^h \varphi_j^h, \varphi_i^h\right) = l(\varphi_i^h) && i = 1, \dots, N_h \\
&\Leftrightarrow && \sum_{j=1}^{N_h} z_j^h a(\varphi_j^h, \varphi_i^h) = l(\varphi_i^h) && i = 1, \dots, N_h \\
&\Leftrightarrow && A^h z^h = b^h.
\end{aligned}$$

with the unknown vector $z^h \in \mathbb{R}^{N_h}$, the stiffness matrix $A^h \in \mathbb{R}^{N_h \times N_h}$ and the load vector $b^h \in \mathbb{R}^{N_h}$, which are defined by

$$(A^h)_{ij} := a(\varphi_j^h, \varphi_i^h), \quad (b^h)_i := l(\varphi_i^h).$$

The matrix A^h is sparse because of the small overlap of the basis functions and its elements can be computed by an element-wise evaluation of the integral.

It can be shown that A^h is symmetric and positive definite, if the bilinearform $a(\cdot, \cdot)$ is symmetric and coercive.

Finite Element Mesh

An important prerequisite for the practical construction of the space V_h and its basis is the partitioning of the domain Ω. This partitioning is called a mesh or grid in finite element terminology consists of so called elements or cells:

$$\mathcal{T}_h = \{t_1, ..., t_m\}.$$

Each element t_i is an open, bounded and connected subset of $\mathbb{R}^n$. The partitioning property is expressed by

$$\bigcup_{i=1}^{m} \overline{t_i} = \overline{\Omega}, \qquad\qquad t_i \cap t_j = \emptyset \qquad \forall i, j \in \{1, ..., m\}, i \neq j.$$

$h_t = \operatorname{diam}(t)$ is the diameter of an element and

$$h := \max_{t \in \mathcal{T}_h} h_t$$

denotes the mesh size.

In order to speak of convergence of the finite element approximation we actually need a sequence of meshes with $h \to 0$.

The individual elements t_i of the mesh typically have simple shape and in order to simplify the calculations t_i is given by a transformation from a reference element. In figure 1.1 shows different types of reference elements $\hat{t}$ in different space dimensions that are used in practice: the simplex and the cube family.

Proposition 1.18 (Reference transformation)**.** Every element $t_i \subset \mathbb{R}^n, i \in \mathcal{T}_h$ can be obtained from the reference element $\hat{t} \subset \mathbb{R}^n$ by using an invertible affine-linear transformation (shifting, rotation, scaling...)

$$\mu_i : \hat{S}_n \text{ or } \hat{Q}_n \to \bar{t}_i, \quad t_i = \mu_i(\hat{t}) = B_i\hat{t} + z_i,$$

with $B_i \in \mathbb{R}^{d\times d}$, $\det B_i > 0$ and $z_i \in \mathbb{R}^d$. □

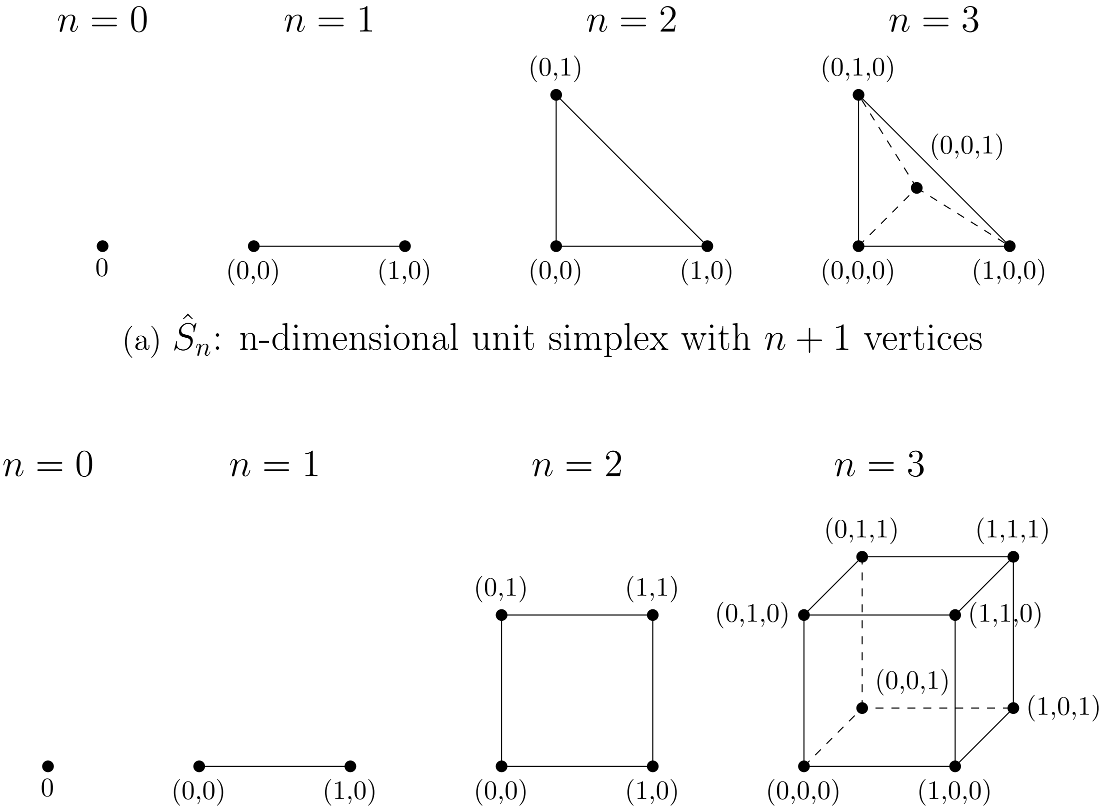
As a consequence we have

Corollary 1.19. $\overline{\Omega}$ is a polyhedral domain (polygon in two space dimensions)!

In general, nonlinear transformations μ can also be considered which then allows one to handle domains with curved boundaries but this will not be considered in this lecture.

It turns out that the mesh $\mathcal{T}_h$ has to satisfy the following additional properties:

1. **Regularity of structure:** Two cells have at most one vertex or one edge (or one face in 3D) in common (no “hanging nodes”).



(a) $\hat{S}_n$: n-dimensional unit simplex with $n+1$ vertices

(b) $\hat{Q}_n$: n-dimensional unit cube with 2^n vertices

Figure 1.1: Examples for reference elements on simplices and cubes

2. **Regularity of form:** For every cell it holds

$$\exists c_1 > 0 : \quad h_t \leq c_1 \rho_t$$

with the apothem ρ_t and the circumscribed radius h_t.

3. **Regularity of size:** Every cell is of the same size.

$$\exists c_2 > 0 : \quad \max_{t \in \mathcal{T}_h} h_t \leq c_2 \min_{t \in \mathcal{T}_h} h_t.$$

Remark 1.20. This only make sense if we have a sequence of grids $\mathcal{T}_{h,\nu}$ and $\nu \in \mathbb{N}$ such that $h_\nu \to 0$ and all constants c_i, $i = \{1, 2\}$, are the same for every ν. □

Finite Element Spaces

Using the mesh we now we can construct Finite Element ansatz spaces and deal with questions about the practical realization of the method. Ω is a polygon domain with the decomposition $\mathcal{T}_h$ in triangles or rectangles (triangular pyramid or hexahedron in 3-D) and all the properties given above are satisfied.

Generally we define the following multivariate polynomial spaces of degree k or smaller:

$$\mathbb{P}^n_k := \{u \in C^\infty(\mathbb{R}^n) : u(x) = \sum_{0 \le |\alpha|_1 \le k} c_\alpha x^\alpha\},$$

$$\mathbb{Q}^n_k := \{u \in C^\infty(\mathbb{R}^n) : u(x) = \sum_{0 \le |\alpha|_\infty \le k} c_\alpha x^\alpha\}$$

with $|\alpha|_1 = \alpha_1 + ... + \alpha_n$, $|\alpha|_\infty = \max_{i=1,...,n} \alpha_i$ and $x^\alpha = x_1^{\alpha_1} \cdot \ldots \cdot x^{\alpha_n}$. In $\mathbb{R}^2$ this looks like

$$\mathbb{P}^2_k := \{u \in \mathcal{C}^\infty(\mathbb{R}^2) : u(x) = \sum_{0 \le i+j \le k} c_{ij} x_1^i x_2^j\},$$

$$\mathbb{Q}^2_k := \{u \in \mathcal{C}^\infty(\mathbb{R}^2) : u(x) = \sum_{0 \le i,j \le k} c_{ij} x_1^i x_2^j\}$$

With that we may define the following function spaces:

$$\mathcal{P}^n_k(\mathcal{T}_h) := \{u \in \mathcal{C}^0(\overline{\Omega}) : \forall t \in \mathcal{T}_h : u\big|_{\bar{t}} \in \mathbb{P}^n_k\},$$

$$\mathcal{Q}^n_k(\mathcal{T}_h) := \{u \in \mathcal{C}^0(\overline{\Omega}) : \forall t \in \mathcal{T}_h : u\big|_{\bar{t}} = \hat{u}_t \circ \mu_t^{-1}, \hat{u}_t \in \mathbb{Q}^n_k\}.$$

It can be checked that this definition is in fact proper, i.e. the requirement of global continuity does not contradict the polynomial form within each element.

The next step is to construct a basis for the finite element space. In particular, for the finite element spaces considered here, a so-called Lagrange basis can be found which is characterized by the property

$$\varphi^h_i(s_j) = \delta_{ij}, \quad j = 1, ..., N_h,$$

for certain points s_j. In the lowest order case $k = 1$ the points s_j are the vertices of the mesh $\mathcal{T}_h$.

Approximation Properties of FE spaces

Definition 1.21 (Lagrange-Interpolation)**.** Given a Lagrange basis we can define the Lagrange interpolation operator acting on continuous functions:

$$\mathcal{I} : \mathcal{C}^0(\Omega) \to \mathcal{P}^n_k(\mathcal{T}_h), \quad \mathcal{I}v = \sum_{i=1}^{N_h} v(s_i)\varphi^h_i. \qquad \square$$

Remark 1.22. Note that $Iv_h \equiv v_h$ for every $v_h \in \mathcal{P}^n_k(\mathcal{T}_h)$. □

Remark 1.23. In order to define Lagrange interpolation for Sobolev functions we need $k > \frac{n}{2}$ for $H^k(\Omega) \subset \mathcal{C}^0(\Omega)$. Then the Sobolev embedding theorem ensures that functions are continuous and pointwise evaluation is well-defined. This means for $n = 1$ that $k \geq 1$ and for $n = 2, 3$ that $k \geq 2$. □

A cornerstone of the finite element a-priori error estimate is the following approximation property of finite element functions:

Proposition 1.24. For $k \in \mathbb{N}$, $k > \min(1, n/2)$ and Lagrange interpolation $\mathcal{I} : \mathcal{H}^k(\Omega) \to \mathcal{P}^n_{k-1}(\mathcal{T}_h)$ (note the polynomial degree is $k - 1$!) and $m \in \{0, 1\}$ we have the estimate

$$\|u - \mathcal{I}u\|_{m,\Omega} \leq ch^{k-m}|u|_{k,\Omega}$$

with a constant $c = c(n, k, \hat{t}, \mathcal{T}_h)$. In particular, the constant depends on the size of the angles of the triangulation. □

As an example consider $n = 2$ and $k > 1$ (required to make Lagrange interpolation well-defined), i.e. the smallest k is 2 and the corresponding polynomial degree is 1 (piecewise linear functions). Then we have $\|u - \mathcal{I}u\|_{1,\Omega} \leq ch|u|_{2,\Omega}$. However, the Lax-Milgram theorem establishes only a solution in H^1. Thus one has to *assume* that a solution with "additional regularity" exists.

Regularity Assumptions

We now discuss briefly under which assumptions solution in higher-order Sobelev spaces actually do exist.

Example 1.25. For domains with smooth boundary or convex polygonal domain Ω it has been proved that $u \in H^2(\Omega)$.

Example 1.26. Ω has a $\mathcal{C}^s$ boundary $\partial\Omega$ (s times continuously differentiable parameterized), then one can show $u \in H^s(\Omega)$.

The regularity of solutions of problem (1.2) can be also "very low". For this discussion fractional order Sobolev spaces are required, i.e. $H^s(\Omega)$ with $s \in \mathbb{R}$.

Example 1.27. Consider the problem $-\nabla\cdot(K(x)\nabla u) = f$ (in weak form) where the coefficient $K(x) > 0$ is discontinuous and has the following "checkerboard" form:

K_1	K_2
K_2	K_1

Then one can show that for $0 < K_1 \leq K_2$ the solution satisfies $u \in H^{1+\alpha}$ with $\alpha = \frac{2}{\pi} \arctan\left(\frac{2\sqrt{K_1 K_2}}{K_2 - K_1}\right) \approx \frac{4}{\pi}\sqrt{\frac{K_1}{K_2}}$ which approaches zero for $K_1 \ll K_2$. Correspondingly, the convergence of the finite element method is h^α which is extremely slow and is observed in practice. □

A-priori Error Estimates

We start with a very important property of the finite element solution.

Proposition 1.28 (Galerkin orthogonality)**.** Suppose $u \in V$ solves (1.2) and u_h solves (1.6), i.e.

$$a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h. \tag{1.7}$$

Then it follows that the error $e = u - u_h$ satisfies

$$a(e, v_h) = 0 \quad v_h \in V_h.$$

Proof. Since $V_h \subset V$, we can use v_h as the test function in the original equation

$$a(u, v_h) = l(v_h) \quad \forall v_h \in V_h.$$

Subtracting from this equation (1.7), we get the Galerkin orthogonality relation for the error $u - u_h$:

$$a(u - u_h, v_h) = a(u, v_h) - a(u_h, v_h) = l(v_h) - l(v_h) = 0 \quad \forall v_h \in V_h. \qquad \square$$

If $a(\cdot,\cdot)$ defines a scalar product on V, which it does in the symmetric case, then we can conclude that the error is orthogonal (w.r.t. the scalar product $a(\cdot,\cdot)$) to all functions in V_h.

An important consequence of Galerkin orthogonality is

Lemma 1.29 (Céa's lemma)**.** The bilinear form $a : V \times V \to \mathbb{R}$, $V = H_0^1(\Omega)$, fulfills the properties

- continuity: $|a(v, w)| \leq C\|v\|_{1,\Omega}\,\|w\|_{1,\Omega}$ for some constant $C > 0$ and all $v, w \in V$ and
- coercivity: $a(v, v) \geq \alpha\|v\|_{1,\Omega}^2$ for some constant $\alpha > 0$ and all $v \in V$.

Then the error satisfies

$$\|u - u_h\|_{1,\Omega} \leq \frac{C}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega} \quad \forall v_h \in V_h.$$

The infimum term characterizes the best approximation of u in the subspace V_h with respect to the H^1-norm. □

Céa's lemma together with the approximation property gives the a-priori estimate.

Theorem 1.30 (A priori error estimate)**.** For the error $u - u_h$ between the exact solution $u \in V$ and the FE solution u_h with the ansatz space $V_h \subset H_0^1(\Omega)$ of order $k \geq 1$, the polynomial degree of the ansatz functions, it holds the a priori error estimation

$$\|u - u_h\|_{1,\Omega} \leq ch^{k-1}|u|_{k,\Omega},$$

whereby the dimension $n \leq 3$ and the solution is required to be in $H^k(\Omega)$. □

In the L^2-norm one can show

$$\|u - u_h\|_{0,\Omega} \leq ch^2|u|_{2,\Omega}$$

for polynomial degree 1.

Practical Implementation of the matrix A^h

In this section we want to present a systematic way to compute the entries of the stiffness matrix $A^h \in \mathbb{R}^{N_h \times N_h}$ for the elliptic problem

$$(K\nabla u, \nabla v) = (f, v) \quad \forall v \in V.$$

This process is called "matrix assembly". To assemble the linear system of equations,

$$A^h z^h = b^h,$$

we use a cell-wise computation of the necessary integrals. The definition of the matrix entry is

$$(A^h)_{ij} = a(\varphi_j^h, \varphi_i^h) = \int\limits_\Omega (K\nabla\varphi_j^h) \cdot \nabla\varphi_i^h \, dx.$$

Now we split the domain into elements to arrive at

$$(A^h)_{ij} = \sum_{t \in \mathcal{T}_h} \int\limits_t (K\nabla\varphi_j^h) \cdot \nabla\varphi_i^h \, dx.$$

We calculate the contribution of one element with the help of the reference transformation μ_t from 1.18. On element t we have the relation

$$\hat{v}(\hat{x}) = v(\mu_t(\hat{x})) \tag{1.8}$$

between the finite element function v on the element $t \in \mathcal{T}_h$ and the corresponding function on the reference element. Recall that for affine transformations we have $\mu_t(\hat{x}) = B_t\hat{x}+z_t$ and $B_t = \hat{\nabla}\mu_t(\hat{x})$ (the hat on the gradient operator means differentiation with respect to $\hat{x}$). The transformation formula for integrals

$$\int\limits_t v(x)\,dx = \int\limits_{\hat{t}} \hat{v}(\hat{x})|\det B_t|\,d\hat{x}$$

then establishes that we can calculate the required integral on the reference element.

In addition, from the chain rule applied to (1.8) it follows

$$\nabla v(\mu_t(\hat{x})) = B_t^{-T}\hat{\nabla}\hat{v}(\hat{x}).$$

Using all these relations the matrix entry can be computed as

$$(A^h)_{ij} = \sum_{t\in\mathcal{T}_h}\int\limits_{\hat{t}} [K(\mu_t(\hat{x}))\big(\hat{\nabla}\mu_t(\hat{x})\big)^{-T}\hat{\nabla}\hat{\varphi}_j(\hat{x})]\cdot\big(\hat{\nabla}\mu_t(\hat{x})\big)^{-T}\hat{\nabla}\hat{\varphi}_i(\hat{x})|\det\hat{\nabla}\mu_t(\hat{x})|\,d\hat{x}.$$

In practice the computations are organized such that all integrals on the element t contributing to different i, j are computed consecutively so that the (expensive) evaluations of μ_t (Jacobian and determinant) can be reused. Moreover, the evaluations of (gradients of) the basis functions $\hat{\varphi}_i^h$ on the reference element can be computed once and stored.

A posteriori error estimation

An important role in partial differential equations is error control. Of interest is to estimate the error between an approximate solution u_h and the exact solution u. For this purpose we have the “a posteriori error estimator”, which only depends on calculated quantities and the data f. The a priori error in the previous section is not useful to control the error, because the necessary information about higher-order derivatives of the exact solution u are not available.

Theorem 1.31. For the error $u - u_h$ there holds the psteriori error estimate

$$\|u-u_h\|_{1,\Omega} \le c\left\{\sum_{t\in\mathcal{T}_h} h_t^2\|R\|_{0,t}^2 + \sum_{\gamma\in\mathcal{F}_h^i\cup\mathcal{F}_h^N} h_\gamma\|r\|_{0,\gamma}^2\right\}^{\frac{1}{2}}$$

with the strong formulation of the elliptic operator

$$R = f + \underbrace{\nabla\cdot(K\nabla u_h)}_{=0\text{ for }\mathcal{P}_1\text{-elements}} - c\,u_h,$$

the jump terms over the edges $\gamma \in \mathcal{F}_h^i$ and the error in the Neumann boundary condition $\gamma \in \mathcal{F}_h^N$

$$r(x) = \begin{cases} [-(K\nabla u_h) \cdot \nu] & x \in \gamma \in \mathcal{F}_h^i \\ -(K\nabla u_h) \cdot \nu - j & x \in \gamma \in \mathcal{F}_h^N \end{cases}.$$

The constant c depends on the mesh and the polynomial degree and is hardly computable in practice. □

Interpolation of non-smooth functions

The Lagrange interpolation requires enough regularity of the Sobolev function. In certain situations, such as for the a-posteriori error estimate given above, on requires a finite element interpolation that can work directly on H^1 functions.

One possibility is the local "Clement" interpolation 1.32:

Definition 1.32 (Clement interpolation)**.** For every function $v \in H^1(\Omega)$ exists the "Clement" interpolation $C_h v \in V_h$:

$$C_h : H^1(\Omega) \to V_h \supset \mathcal{P}_1^n(\mathcal{T}_h),$$

which is a combination of the Lagrange interpolation and the following L^2-projection. □

Remark 1.33. The Clement interpolation is not a projection, i.e. $C_h C_h v \neq C_h v$. □

Another option is the "L^2-projection" 1.34, which is orthogonal but not local.

Definition 1.34 (L^2-projection)**.** The L^2-projection $Q_h : L^2(\Omega) \to V_h$ is defined by

$$(Q_h v, w_h)_{0,\Omega} = (v, w_h)_{0,\Omega} \quad \forall w_h \in V_h$$

with the estimate

$$\|v - Q_h v\|_{0,\Omega} \leq ch|v|_{1,\Omega}.$$ □

Chapter 2

Classical Iterative Methods

2.1 Linear Iterative Methods

The regular linear system

$$Ax = b \tag{2.1}$$

is solved by constructing a sequence $x^0, x^1, \ldots$ with *arbitrary* initial guess x^0 that converges towards the solution x. One way to construct linear iterative methods is via defect correction. For arbitrary x^k define the error as

$$e^k := x - x^k. \tag{2.2}$$

Due to linearity we have

$$Ae^k = Ax - Ax^k = b - Ax^k := d^k \tag{2.3}$$

which is called *defect*. Note that $d^k = b - Ax^k$ can be readily computed while the underlying error e^k is usually not available.

In order to arrive at an iterative method A on the left hand side of (2.3) is replaced by some approximation W, i.e. we solve $Wv^k = d^k$ and $v^k = W^{-1}d^k$ approximates e^k. This gives us the generic form of a linear iterative method:

$$x^{k+1} = x^k + W^{-1}(b - Ax^k). \tag{2.4}$$

Particular choices for W are

$$\begin{aligned}
&W_{Ric} = \omega^{-1} I && \omega \in \mathbb{R}, \text{Richardson}\\
&W_{Jac} = \operatorname{diag}(A) && \text{Jacobi}\\
&W_{GS} = L + D && A = L + D + U, \text{Gauß-Seidel}
\end{aligned}$$

Analysis of linear iterative methods is based on the error propagation equation

$$\begin{aligned}
e^{k+1} &= x - x^{k+1}\\
&= x - x^k - W^{-1}(Ax - Ax^k)\\
&= (x - x^k) - W^{-1}A(x - x^k)\\
&= e^k - W^{-1}Ae^k\\
&= (I - W^{-1}A)e^k =: Se^k
\end{aligned}$$

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

Definition 2.1.

$$\sigma(A) := \{\lambda \in \mathbb{C} : \lambda \text{ is eigenvalue of } A\}$$

is called the *spectrum* of A and

$$\rho(A) := \max\{|\lambda| : \lambda \in \sigma(A)\}$$

is called the *spectral radius* of A.

Theorem 2.2. *A, W regular matrices. Then the iterative scheme (2.4) converges if and only if $\rho(S) < 1$.*

Proof. See Hackbusch [1991]. Idea: $e^k = S^k e^0$ and show $S^k \to 0$. For diagonalizable matrices this is easy to see as $S^k = TD^kT^{-1}$ and $D^k = \operatorname{diag}(\lambda_1^k, \ldots, \lambda_N^k)$ a diagonal matrix. The argument can be extended to non-diagonalizable matrices. □

In general it is difficult to determine $\rho(S)$. One option is to use a norm estimate

$$\begin{aligned} e^{k+1} &\leq (I - W^{-1}A)e^k \\ \Rightarrow \|e^{k+1}\| &\leq \|I - W^{-1}A\| \, \|e^k\| \end{aligned}$$

for any submultiplicative matrix norm. Since $\rho(S) \leq \|S\|$ for any norm and $\|S\| < 1$ is required for convergence, the norm needs to be chosen carefully.

A special case are symmetric positive definite matrices where the spectral radius can be computed exactly and related to the condition number.

Theorem 2.3. *A, B symmetric and positive-definite matrices. Then the iteration*

$$x^{k+1} = x^k + \frac{1}{\lambda_{max}(BA)} B(b - Ax^k)$$

converges with the rate

$$\rho = 1 - \frac{1}{\kappa(BA)}$$

where

$$\kappa(BA) = \frac{\lambda_{max}(BA)}{\lambda_{min}(BA)}$$

is the spectral condition number and $\lambda_{max}(BA), \lambda_{min}(BA)$ are the extreme eigenvalues of BA.

Proof. A is symmetric positive definite, so there is an unitary matrix Q such that $A = QDQ^T$ with $D = \operatorname{diag}(\lambda_1, \dots, \lambda_N)$ with $\lambda_i \in \sigma(A) \subset \mathbb{R}_+$. Set $D^{\frac{1}{2}} := \operatorname{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_N})$ and $A^{\frac{1}{2}} := QD^{\frac{1}{2}}Q^T$. Then we have $\sigma(BA) = \sigma(A^{\frac{1}{2}}BAA^{-\frac{1}{2}}) = \sigma(A^{\frac{1}{2}}BA^{\frac{1}{2}})$. Since $T := A^{\frac{1}{2}}BA^{\frac{1}{2}}$ is symmetric and positive definite all eigenvalues of BA are real and positive. Now T is also diagonalizable and has a complete set of eigenvectors. Since $\sigma(S) = \sigma(I - \frac{1}{\omega}T) = \{\mu_i : \mu_i = 1 - \frac{\lambda_i}{\omega} \text{ for } \lambda_i \in \sigma(T) = \sigma(BA)\}$, setting $\omega = \lambda_{max}(T)$ we get $\mu_i \in [0, 1 - \frac{\lambda_{min}(T)}{\lambda_{max}(T)}]$. So $\rho(S) = 1 - \frac{1}{\kappa(T)}$. □

For $B = I$ we obtain the Richardson iteration $W = \frac{1}{\lambda_{max}(A)}I$. For $B = A^{-1}$ we have $\kappa(BA) = 1$ and $\rho(S) = 0$.

The matrix B is supposed to reduce the condition number of A and is therefore often called a *preconditioner*.

Now what is the condition number of A?

Observation 2.4 (Raleigh Quotient)**.** Let $A \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. Then the extreme eigenvalues can be characterized by

$$\lambda_{min}(A) = \inf_{x \neq 0} \frac{\langle Ax, x\rangle}{\langle x, x\rangle}, \qquad \lambda_{max}(A) = \sup_{x \neq 0} \frac{\langle Ax, x\rangle}{\langle x, x\rangle},$$

where $\langle ., . \rangle$ is *any* scalar product in $\mathbb{R}^n$.

Proof.

1. Let $\langle ., . \rangle$ be the Euclidean scalar product. There exists Q with $A = Q^T DQ$ and $QQ^T = I$. Then

$$\frac{\langle Ax, x\rangle}{\langle x, x\rangle} = \frac{\langle DQx, Qx\rangle}{\langle Qx, Qx\rangle} = \frac{\sum_{i=1}^N \lambda_i (Qx)_i^2}{\langle Qx, Qx\rangle}.$$

   From $\lambda_{min}\langle Qx, Qx\rangle \leq \sum_{i=1}^N \lambda_i (Qx)_i^2 \leq \lambda_{max}\langle Qx, Qx\rangle$ we conclude the result.

2. Extend to $\langle u, v\rangle_M = \langle Mu, v\rangle = \langle M^{\frac{1}{2}}u, M^{\frac{1}{2}}v\rangle$. □

Lemma 2.5. Let A_h be obtained from a Finite Element discretization of the Poisson equation, i.e. $(A_h)_{ij} = a(\phi_j, \phi_i)$, using Lagrange basis functions of P_1 on a mesh of size h. Then there exists a constant c such that

$$\kappa(A_h) \leq ch^{-2}$$

and the estimate is sharp.

Proof. Let $\langle ., . \rangle$ be the Euclidean scalar product. We write $\Omega_{ij} := \operatorname{supp}(\phi_i) \cap \operatorname{supp}(\phi_j)$.

$$
\begin{aligned}
\langle A_h x, x\rangle &= \sum_{i=1}^{N_h}\sum_{j=1}^{N_h} x_i x_j a(\phi_j, \phi_i) \\
&= \sum_{i,j=1}^{N_h} x_i x_j \int\limits_{\Omega_{ij}} \nabla\phi_j \cdot \nabla\phi_i \, dx \\
&= \sum_{i,j=1}^{N_h} x_i x_j \sum_{t\in\Omega_{ij}} \int\limits_{\hat{t}} (B_t^{-T}\nabla\hat{\phi}_j)\cdot(B_t^{-T}\nabla\hat{\phi}_i)|\det B_t|\, d\hat{x} \qquad (*) \\
&\leq \sum_{i=1}^{N_h} x_i \left(\sum_{j=1}^{N_h} x_j \sum_{t\in\Omega_{ij}} ch^{d-2} \right) \\
&= ch^{d-2}\langle Ex, x\rangle \\
&\leq ch^{d-2}\|Ex\|\,\|x\| \\
&\leq ch^{d-2}\|E\|_2\,\|x\|^2 \\
&= ch^{d-2}\|E\|_2\,\langle x, x\rangle
\end{aligned}
$$

where

$$
E_{ij} := \begin{cases} 1 & \Omega_{ij} = \operatorname{supp}(\phi_i)\cap\operatorname{supp}(\phi_j) \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}
$$

Note that $\|E\|_2 \leq K$ when E symmetric and $\|E\|_\infty = K$. In (*) we used the estimates

$$
\|B_t^{-T}\| \leq c\frac{1}{h}, \quad |\det B_t| \leq ch^d \quad \text{and } \|\nabla\hat{\phi}_i\| \leq 1.
$$

Dividing by $\langle x, x\rangle$ and taking the supremum then shows

$$
\lambda_{max}(A_h) = \sup_{x\neq 0}\frac{\langle A_h x, x\rangle}{\langle x, x\rangle} \leq ch^{d-2}.
$$

Now we give an estimate for λ_{min} and recognize that based on the Lagrange basis functions we have for any function $u_h = \sum_{i=1}^{N_h} x_i \phi_i$:

$$\begin{aligned}
\langle A_h x, x\rangle &= a(u_h, u_h) \\
&\geq \alpha \|u_h\|_{1,\Omega}^2 \\
&= \alpha(\|u_h\|_{0,\Omega}^2 + |u_h|_{1,\Omega}^2) \\
&\geq \alpha(\|u_h\|_{0,\Omega}^2 + \frac{1}{s^2}\|u_h\|_{0,\Omega}^2) \quad \text{(Friedrich inequality: assume } \Gamma_D \neq 0) \\
&\geq \alpha \frac{1+s^2}{s^2}\|u_h\|_{0,\Omega}^2 \\
&\geq \alpha \frac{1+s^2}{s^2} h^d \langle x, x\rangle \qquad \text{(not shown here)}
\end{aligned}$$

and thus

$$\lambda_{min}(A_h) = \inf_{x\neq 0} \frac{\langle A_h x, x\rangle}{\langle x, x\rangle} \geq \alpha \frac{1+s^2}{s^2} h^d$$

Together we obtain

$$\kappa(A_h) = \frac{\lambda_{max}(A_h)}{\lambda_{min}(A_h)} \leq ch^{-2}. \qquad \square$$

2.2 Block Iterative Methods

These are precursor to overlapping Schwarz methods.

The following notation is handy when displaying block methods and describing the parallel implementation of iterative methods.

Index sets

An index set I is a finite subset of $\mathbb{N}_0$. In particular index sets need not be consecutive or starting with 0 or 1. $x \in \mathbb{R}^I$ is the vector having components $(x)_i$ for all $i \in I$. Alternatively identify a vector $x \in \mathbb{R}^I$ with the map $x : I \to \mathbb{R}$.

Analogously, for any two index sets $I, J \subset \mathbb{N}_0$: $A \in \mathbb{R}^{I\times J}$ is the matrix with entries $(A)_{i,j}$ for all $(i, j) \in I \times J$. Alternatively: $A : I \times J \to \mathbb{R}$.

Subvectors and submatrices

Let $\tilde{I} \subset I$ and $\tilde{J} \subset J$. Then, for $x \in \mathbb{R}^I$, $x_{\tilde{I}}$ is given by $(x_{\tilde{I}})_i = (x)_i$ for all $i \in \tilde{I}$ and for $A \in \mathbb{R}^{I\times J}$, $A_{\tilde{I},\tilde{J}}$ is given by $(A_{\tilde{I},\tilde{J}})_{i,j} = (A)_{i,j}$ for all $(i, j) \in \tilde{I} \times \tilde{J}$.

Displaying a representation of a vector or matrix requires an *ordering* of the index sets, e.g. the lexicographic ordering. Also, certain iterative methods, e.g. Gauß-Seidel, require an ordering of the index set.

Partitioning

Block methods are based on a partitioning of the index set $I \subset \mathbb{N}_0$. Let $P = \{1, \dots, p\}$ be the index set of the blocks and choose $I_i \subseteq I$ for $i \in P$ such that

$$\bigcup_{i \in P} I_i = I \quad \text{and} \quad I_i \cap I_j = \emptyset \text{ for all } i \neq j.$$

Block-Jacobi and Block-Gauß-Seidel

Then the Block-Jacobi and Block-Gauß-Seidel methods are defined by

$$(W_{BJac})_{i,j} = \begin{cases} (A)_{i,j} & \text{if } i, j \in I_k \text{ for a } k \in P \\ 0 & \text{else,} \end{cases}$$
$$(W_{BGS})_{i,j} = \begin{cases} (A)_{i,j} & \text{if } i \in I_k, j \in I_l \text{ for } l \leq k \text{ and } k, l \in P \\ 0 & \text{else .} \end{cases}$$

Assume that the index set I is ordered such that $i < j$ whenever $\text{block}(i) < \text{block}(j)$ where $\text{block}(i) = k :\Leftrightarrow i \in I_k$. Then

$$W_{BJac} = \begin{pmatrix} A_{I_1,I_1} & 0 & \dots & 0 \\ 0 & A_{I_2,I_2} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots\dots\dots & & A_{I_p,I_p} \end{pmatrix}, \quad W_{BGS} = \begin{pmatrix} A_{I_1,I_1} & 0 & \dots & 0 \\ A_{I_2,I_1} & A_{I_2,I_2} & & \vdots \\ \vdots & & \ddots & \vdots \\ A_{I_p,I_1} & \dots\dots\dots & & A_{I_p,I_p} \end{pmatrix}.$$

Both methods require the solution of the p smaller systems A_{I_i,I_i}.

Algorithmic Formulation

Define the rectangular restriction matrix

$$R_{I_i} : \mathbb{R}^I \to \mathbb{R}^{I_i}, \qquad (R_{I_i} x)_\alpha := (x)_\alpha \, \forall \alpha \in I_i.$$

R_{I_i} is a $|I_i| \times |I|$ matrix with exactly one 1 per row. All 1s are in different columns, so $\operatorname{rank} R_{I_i} = |I_i|$.

With this we can write

$$A_{I_i,I_i} = R_{I_i} A R_{I_i}^T$$

and get for the Block-Jacobi method

$$x^{k+1} = x^k + \sum_{i \in P} R_{I_i}^T A_{I_i,I_i}^{-1} R_{I_i} (b - Ax^k)$$

where the computations can be done in parallel. In case of the Block-Gauß-Seidel method we get

$$\text{For } i = 1, \dots, p: \qquad x^{k+\frac{i}{p}} = x^{k+\frac{i-1}{p}} + R_{I_i}^T A_{I_i,I_i}^{-1} R_{I_i}(b - Ax^{k+\frac{i-1}{p}}).$$

Without further assumptions on I_i and A these corrections have to be computed sequentially!

For the convergence of the block variants one can prove:

Theorem 2.6. *A let be symmetric positive definite.*

1. *$2\,W_{BJac} - A$ be symmetric and positive definite then $\|S_{BJac}\|_A < 1$.*
2. *$\|S_{BGS}\|_A < 1$ where $\|S\|_A$ is the matrix norm associated to the energy norm $\|x\|_A := \sqrt{\langle Ax, x\rangle}$.*

Proof. See [Hackbusch, 1991, Satz 4.5.4 and 4.5.6]. □

2.3 Descent Methods

These are nonlinear iterative methods based on minimizing the functional

$$F(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle.$$

Theorem 2.7. *A symmetric and positive definite. Then the unique minimum x^* of F coincides with the solution of the linear system $Ax = b$.*

Proof. For any $x = x^* + v$ show $F(x) = F(x^*) + \frac{1}{2}\langle Av, v\rangle > F(x^*)$ if $v \neq 0$. Uniqueness is proven by contradiction. □

1D-minimization

Given an iterate x^k and a "search direction" p^k one can easily solve the problem

$$\text{Find } \alpha \in \mathbb{R} \text{ such that } F(x^k + \alpha p^k) \to \min$$

by

$$\alpha = \frac{(p^k)^T(b - Ax^k)}{(p^k)^T A p^k}. \tag{2.5}$$

Gradient descent method: Choose $p^k = -\nabla F(x^k) = b - Ax^k$.

Theorem 2.8. *A symmetric and positive definite. Then, with x being the solution of $Ax = b$, the gradient descent method satisfies*

$$\|x - x^k\|_A \leq \frac{\kappa(A) - 1}{\kappa(A) + 1}\|x - x^{k-1}\|_A.$$

Algorithm 2.1 Gradient Descent Method

```
Given: Initial guess $x$, right-hand side $b$ and tolerance $\epsilon < 1$
  $d := b - Ax$
  $\delta := \delta_0 := \|d\|$
  while $\delta > \epsilon\delta_0$ do
      $q := Ad$                                  ▷ matrix vector product
      $\alpha := \langle d, d\rangle / \langle d, q\rangle$        ▷ scalar products
      $x := x + \alpha d$                        ▷ $x$ update
      $d := d - \alpha q$         ▷ $d = b - A(x+v) = b - Ax - Av = d - Av$
      $\delta := \|d\|$                          ▷ recompute norm
  end while
```

Proof. See [Hackbusch, 1991, Theorem 9.2.3]. □

The convergence factor can be written has

$$\frac{\kappa(A)-1}{\kappa(A)+1} \leq 1 - \frac{1}{\kappa(A)+1}.$$

So for large $\kappa(A)$ the convergence factor nearly the same as that of the damped Richardson method.

Preconditioning

Idea: Choose M regular and multiply $Ax = b$ to the left with M^{-1} to obtain the equivalent system $M^{-1}Ax = M^{-1}b$ (left preconditioning). If $\kappa(M^{-1}A) \ll \kappa(A)$ then the convergence of the gradient method applied to this system is better.

However, in general, $M^{-1}A$ is not symmetric even when M and A are symmetric. Assume M and A are symmetric and positive definite. Then $M^{\frac{1}{2}}$ is well defined and

$$\sigma(M^{-1}A) = \sigma(M^{\frac{1}{2}}M^{-1}AM^{-\frac{1}{2}}) = \sigma(M^{-\frac{1}{2}}AM^{-\frac{1}{2}}).$$

Now transform $Ax = b$ from left and right by

$$\begin{aligned} & Ax = b \\ \Leftrightarrow\quad & M^{-\frac{1}{2}}AM^{-\frac{1}{2}}M^{\frac{1}{2}}x = M^{-\frac{1}{2}}b \\ \Leftrightarrow\quad & \hat{A}\hat{x} = \hat{b} \end{aligned}$$

with $\hat{A} := M^{-\frac{1}{2}}AM^{-\frac{1}{2}}$, $\hat{x} := M^{\frac{1}{2}}x$ and $\hat{b} := M^{-\frac{1}{2}}b$.

Obviously $\sigma(\hat{A}) = \sigma(M^{-1}A)$, $\hat{A}$ is symmetric and positive definite and the gradient method can formally be applied to this transformed system.

Unfortunately, the matrix $\hat{A}$ is in general not sparse and performing operations with $\hat{A}$ is too costly. Instead of transforming the linear system once at the beginning, we transform instead *every single step* of the method. This means

$$\begin{aligned}
\hat{d} &= \hat{b} - \hat{A}\hat{x}^k = M^{-\frac{1}{2}}b - M^{-\frac{1}{2}}AM^{-\frac{1}{2}}M^{\frac{1}{2}}x^k = M^{-\frac{1}{2}}(b - Ax^k) = M^{-\frac{1}{2}}d,\\
\hat{q} &= \hat{A}\hat{d} = M^{-\frac{1}{2}}AM^{-\frac{1}{2}}M^{-\frac{1}{2}}d = M^{-\frac{1}{2}}AM^{-1}d =: M^{-\frac{1}{2}}Av,\\
\hat{\alpha} &= \frac{\langle \hat{d}, \hat{d}\rangle}{\langle \hat{d}, \hat{q}\rangle} = \frac{\langle M^{-\frac{1}{2}}d, M^{-\frac{1}{2}}d\rangle}{\langle M^{-\frac{1}{2}}d, M^{-\frac{1}{2}}AM^{-1}d\rangle} = \frac{\langle d, M^{-1}d\rangle}{\langle M^{-1}d, AM^{-1}d\rangle} = \frac{\langle d, v\rangle}{\langle v, Av\rangle} = \alpha,\\
\hat{x} &= \hat{x} + \hat{\alpha}\hat{d} = M^{\frac{1}{2}}x + \hat{\alpha}M^{-\frac{1}{2}}d = M^{\frac{1}{2}}(x + \alpha M^{-1}d) = M^{\frac{1}{2}}(x + \alpha v),\\
\hat{d} &= \hat{d} - \hat{\alpha}\hat{q} = M^{-\frac{1}{2}}d - \alpha M^{-\frac{1}{2}}Av = M^{-\frac{1}{2}}(d - \alpha Av).
\end{aligned}$$

where $v := M^{-1}d$ is the result of the preconditioner.

Ultimately we are interested in $x = M^{-\frac{1}{2}}\hat{x}$, so the transformed quantities need never be computed.

Algorithm 2.2 Preconditioned Gradient Descent Method

```
Given: Initial guess $x$, right-hand side $b$ and tolerance $\epsilon < 1$
  $d := b - Ax$
  $\delta := \delta_0 := \|d\|$
  while $\delta > \epsilon\delta_0$ do
      procedure Solve
          $Mv = d$
      end procedure
      $q := Av$
      $\alpha := \langle d, v\rangle / \langle v, Av\rangle$
      $x := x + \alpha v$
      $d := d - \alpha q$
      $\delta := \|d\|$
  end while
```

Conjugate Gradient Method

The conjugate gradient method is very similar to the gradient descent method. It ensures in addition that the search directions are A-orthogonal:

$$\langle Ap^k, p^i\rangle = 0 \quad \forall i < k. \tag{2.6}$$

The convergence rate can be estimated by

$$\|x - x^k\|_A \leq \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1}\|x - x^{k-1}\|_A. \tag{2.7}$$

Proof. See [Hackbusch, 1991, Theorem 9.4.12]. □

Algorithm 2.3 Conjugate Gradient Method

```
Given: Initial guess $x$, right-hand side $b$ and tolerance $\epsilon < 1$
  $d := b - Ax$                          ▷ initial defect
  $p := d$                               ▷ initial search direction
  $\delta := \delta_0 := \|d\|$          ▷ initial norm
  while $\delta > \epsilon\delta_0$ do
      $q := Ap$                          ▷ matrix vector product
      $\alpha := \langle p, d\rangle / \langle p, q\rangle$    ▷ optimal step length, see (2.5)
      $x := x + \alpha p$                ▷ solution update
      $d := d - \alpha q$                ▷ defect update
      $\beta = \langle d, q\rangle / \langle p, q\rangle$
      $p = d - \beta p$                  ▷ new orthogonal search direction
      $\delta := \|d\|$                  ▷ recompute norm
  end while
```

The conjugate gradient method is given in Algorithm 2.3. Preconditioning can be applied in the same way as for the gradient descent method.

There exist also extensions of descent methods applicable to nonsymmetric matrices, e.g. the BiCGStab method and the GMRES method. All these methods also go under the name of *Krylov methods*

2.4 Parallel Implementation

Block-Jacobi without Convergence Test

We consider first the implementation of the Block-Jacobi method with the following assumptions:

- as many processors p as there are blocks and
- no termination criterion.

Assume partitioning

$$I_i \subset I, \quad \bigcup_{i=1}^{p} I_i = I, \quad I_i \cap I_j = \emptyset \ \forall i \neq j.$$

Partitioning is (usually) done on mesh level, i.e. a partitioning of $\mathcal{T}_h$ *induces* that of I. Data decomposition: Every process $i \in P = \{1, \ldots, p\}$ stores all rows of A corresponding to I_i.

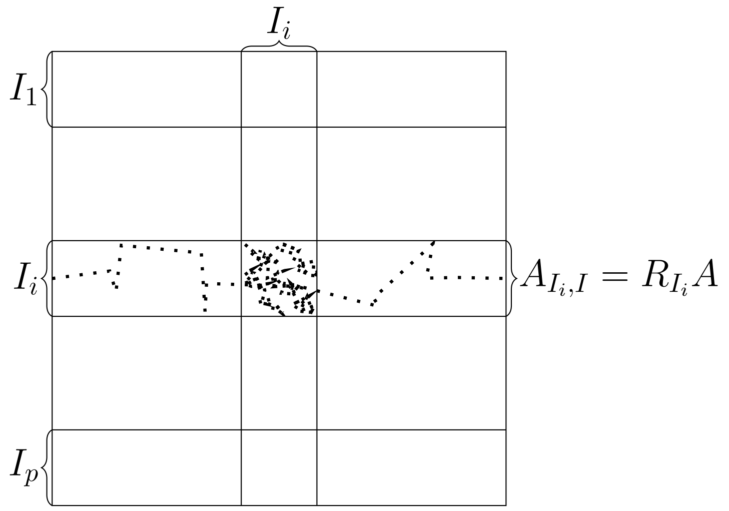


Figure 2.1: Data Decomposition for Block-Jacobi Method

Note that A is *sparse* with only few non-zero elements off the block diagonal, see fig. 2.1. In order to compute $A_{I_i,I}x$ process i does not need the whole x but only $x_{\hat{I}_i}$ with

$$\hat{I}_i := \{j \in I : (A)_{k,j} \neq 0 \text{ for } k \in I_i\}.$$

Consequently, all non-zeroes are contained in $A_{I_i,\hat{I}_i}$. Note also, that

$$I_i \subseteq \hat{I}_i$$

since $(A)_{j,j} > 0$ for A symmetric and positive definite.

We also define the restriction $R_{\hat{I}_i,I_i} : \mathbb{R}^{\hat{I}_i} \to \mathbb{R}^{I_i}$ inthe usual way by

$$\text{For } x \in \mathbb{R}^{\hat{I}_i} : \qquad (R_{\hat{I}_i,I_i}x)_j = (x)_j \qquad \forall j \in I_i.$$

The parallel Block-Jacobi Method is then given by algorithm 2.4. The communication step involves the exchange of messages with (or access to memory of) only a few other processes for each process p.

Algorithm 2.4 Parallel Block-Jacobi Method without Convergence Test

```
for all $i \in \{1, \dots, p\}$ do in parallel
    $x_{\hat{I}_i} := R_{\hat{I}_i} x^0$                                    ▷ set initial guess
    $b_{I_i} := R_{I_i} b$                                        ▷ right-hand side
    for $k = 1, \dots$ do
        $d_{I_i} := b_{I_i} - A_{I_i,\hat{I}_i} x_{\hat{I}_i}$                          ▷ local defect
        $v_{\hat{I}_i} := R^T_{\hat{I}_i,I_i} A^{-1}_{I_i,I_i} d_{I_i}$                          ▷ local solve
        $v_{\hat{I}_i} := v_{\hat{I}_i} + \sum_{j \neq i, \hat{I}_i \cap I_j \neq \emptyset} R_{\hat{I}_i} R^T_{\hat{I}_j} v_{\hat{I}_j}$     ▷ communication!
        $x_{\hat{I}_i} := x_{\hat{I}_i} + v_{\hat{I}_i}$                              ▷ local update
    end for
end for
```

Parallel Preconditioned Gradient Descent Method

The preconditioned gradient descent method (and also the (preconditioned) conjugate gradient method) can be parallelized along the same ideas.

Algorithm 2.5 Parallel Preconditioned Gradient Descent Method

```
for all $i \in \{1, \dots, p\}$ do in parallel
    $x_{\hat{I}_i} := R_{\hat{I}_i} x^0$
    $b_{I_i} := R_{I_i} b$
    $d_{I_i} := b_{I_i} - A_{I_i,\hat{I}_i} x_{\hat{I}_i}$
    $\delta := \delta^0 := \|d\| = \sqrt{\sum_{i=1}^p \|d_{I_i}\|^2}$                ▷ global communication
    while $\delta > \epsilon\delta^0$ do
        $v_{\hat{I}_i} := \mathrm{prec}(d_{I_i})$                ▷ involves communication of $v_{\hat{I}_i}$
        $q_{I_i} := A_{I_i,\hat{I}_i} v_{\hat{I}_i}$
        $\alpha := \langle d, v\rangle / \langle v, Av\rangle = \frac{\sum_{i=1}^p \langle d_i, R_{\hat{I}_i,I_i} v_{\hat{I}_i}\rangle}{\sum_{i=1}^p \langle R_{\hat{I}_i,I_i} v_{\hat{I}_i}, q_{I_i}\rangle}$                ▷ two global communications
        $x_{\hat{I}_i} := x_{\hat{I}_i} + \alpha v_{\hat{I}_i}$
        $d_{I_i} := d_{I_i} - \alpha q_{I_i}$
        $\delta := \sqrt{\sum_{i=1}^p \|d_{I_i}\|^2}$                ▷ global communication
    end while
end for
```

Chapter 3

Overlapping Domain Decomposition Methods

3.1 Overlapping versus Non-overlapping Methods

Let us start with some basic ideas and two subdomains. Consider the Poisson equation with homogeneous Dirichlet boundary conditions for simplicity:

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega \subset \mathbb{R}^d, \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{3.1}$$

Non-overlapping Methods

Assume Ω is partitioned into two non-overlapping subdomains

$$\bar{\Omega} = \bar{\Omega}_1 \cup \bar{\Omega}_2, \quad \Omega_1 \cap \Omega_2 = \emptyset, \quad \Gamma = \partial\Omega_1 \cap \partial\Omega_2$$

and measure$(\partial\Omega_1 \cap \partial\Omega) > 0$, measure$(\partial\Omega_2 \cap \partial\Omega) > 0$ and Ω_1, Ω_2 have Lipschitz-continuous boundaries, for example as in figure 3.1.

Under suitable assumptions on f (more than $f \in H^{-1}(\Omega)$!) and the Ω_i (e.g. Lipschitz boundary is sufficient) problem (3.1) is equivalent to

$$-\Delta u_1 = f \text{ in } \Omega_1, \qquad u_1 = 0 \text{ on } \partial\Omega_1 \setminus \Gamma \tag{3.2a}$$

$$-\Delta u_2 = f \text{ in } \Omega_2, \qquad u_2 = 0 \text{ on } \partial\Omega_2 \setminus \Gamma \tag{3.2b}$$

$$u_1 = u_2 \text{ on } \Gamma, \qquad -\nabla u_1 \cdot \nu = -\nabla u_2 \cdot \nu \text{ on } \Gamma \tag{3.2c}$$

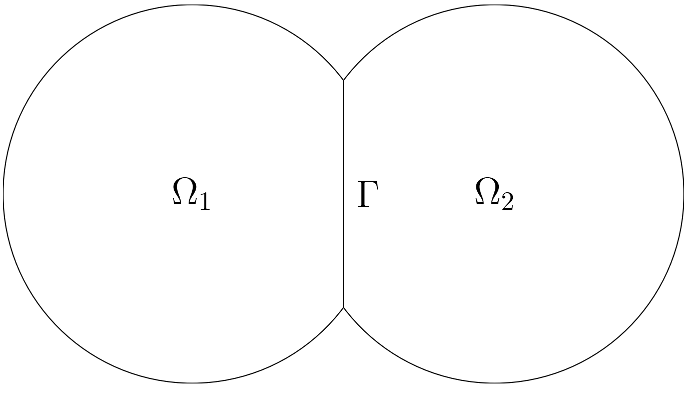


Figure 3.1: Decomposition of Ω in two non-overlapping subdomains

where ν is the normal on Γ (selected in either way) and (3.2c) is meant in the L_2-sense. (3.2c) are called interface or transmission condition. The proof is done in the variational framework, see Quarteroni and Valli [1999]. The continuity of u on Γ is a consequence of the trace theorem.

One possible algorithm to solve (3.2) is the Dirichlet-Neumann procedure: Given u_Γ^k on the boundary Γ one iteration reads

$$\begin{aligned}
-\Delta u_1^{k+\frac{1}{2}} &= f && \text{in } \Omega_1 \\
u_1^{k+\frac{1}{2}} &= 0 && \text{on } \partial\Omega_1 \setminus \Gamma && (3.3a) \\
u_1^{k+\frac{1}{2}} &= u_\Gamma^k && \text{on } \Gamma \\
-\Delta u_2^{k+\frac{1}{2}} &= f && \text{in } \Omega_2 \\
u_2^{k+\frac{1}{2}} &= 0 && \text{on } \partial\Omega_2 \setminus \Gamma && (3.3b) \\
-\nabla u_2^{k+\frac{1}{2}} \cdot \nu &= -\nabla u_1^{k+\frac{1}{2}} \cdot \nu && \text{on } \Gamma \\
u_\Gamma^{k+1} &= \theta u_2^{k+\frac{1}{2}} + (1-\theta) u_\Gamma^k && \text{on } \Gamma && (3.3c)
\end{aligned}$$

with $\theta \in (0, \theta_{max})$. This is an iteration for the values on Γ. After convergence the solution inside the subdomains can be recovered by solving two Dirichlet problems.

We will treat non-overlapping methods in chapter 5 of the lecture.

Overlapping Methods

One problem in the analysis of the non-overlapping case is that the restriction of $v \in H_0^1(\Omega)$ to Ω_i, i.e.

$$v_i(x) = \begin{cases} v(x) & \text{for } x \in \Omega_i \\ 0 & \text{for } x \notin \Omega_i, \end{cases}$$

is *not* in $H_0^1(\Omega)$. This difficulty is overcome if overlap is added. Assume $\hat\Omega_1, \hat\Omega_2$ are domains (i.e. *open*, connected subsets of $\mathbb{R}^d$) such that $\hat\Omega_1 \cup \hat\Omega_2 = \Omega$.

Set $\Gamma_1 := \partial\hat\Omega_1 \cap \hat\Omega_2$, $\Gamma_2 := \partial\hat\Omega_2 \cap \hat\Omega_1$ as in figure 3.2. Then the Schwarz alternating method reads as follows. Given u^k defined on the whole domain Ω

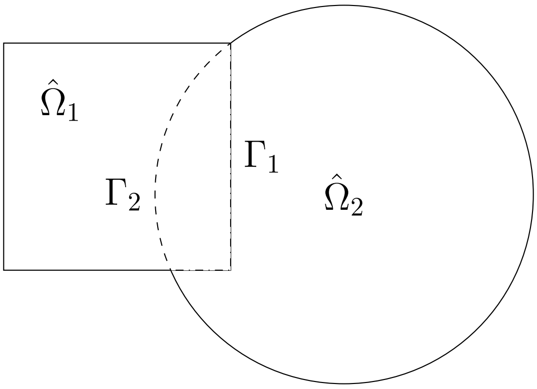


Figure 3.2: Decomposition of Ω in two overlapping subdomains

with $u^k = 0$ on $\partial\Omega$ compute

$$\begin{aligned} -\Delta u_1^{k+\frac12} &= f && \text{on } \hat\Omega_1 \\ u_1^{k+\frac12} &= 0 && \text{on } \partial\hat\Omega_1 \setminus \Gamma_1 \qquad (3.4a) \\ u_1^{k+\frac12} &= u^k && \text{on } \Gamma_1 \end{aligned}$$

$$\begin{aligned} -\Delta u_2^{k+\frac12} &= f && \text{on } \hat\Omega_2 \\ u_2^{k+\frac12} &= 0 && \text{on } \partial\hat\Omega_2 \setminus \Gamma_2 \qquad (3.4b) \\ u_2^{k+\frac12} &= u_1^{k+\frac12} && \text{on } \Gamma_2 \end{aligned}$$

$$u^{k+1} = \begin{cases} u_2^{k+\frac12} & \text{on } \hat\Omega_2 \\ u_1^{k+\frac12} & \text{on } \hat\Omega_1 \setminus (\hat\Omega_1 \cap \hat\Omega_2) \end{cases} \qquad (3.4c)$$

This procedure was used by H. A. Schwarz in 1870 to prove the existence of solutions of the Laplace equation in regions with non-smooth boundaries.

We will later show that for a variational formulation of (3.4) there exists $\rho < 1$ such that

$$\|u - u^{k+1}\|_{1,\Omega} \leq \rho \|u - u^k\|_{1,\Omega}.$$

The convergence factor ρ depends on the form of the subdomains, in particular the overlap of the subdomains. This can be easily seen in one space dimension. Consider

$$\begin{aligned} -\frac{d^2u}{dx^2} &= 0 \text{ in } \Omega = (0,1) \\ u(0) = u(1) &= 1 \end{aligned}$$

and set $\hat\Omega_1 = (0, \frac12 + a)$, $\hat\Omega_2 = (\frac12 - a, 1)$ for $0 < a < \frac12$. Obviously $u = 1$ is the exact solution.

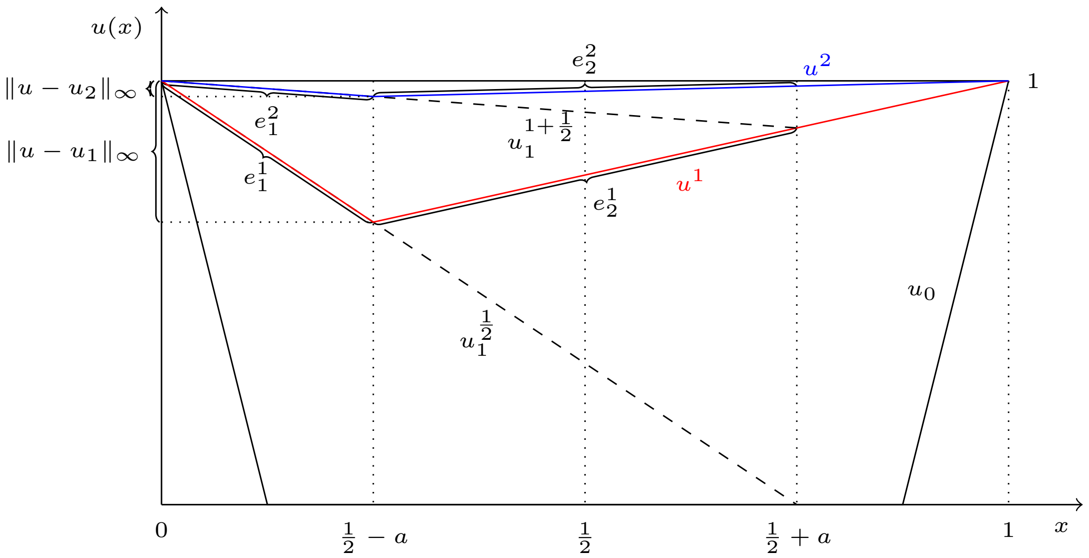


Figure 3.3: Graphical error determination for example problem

For the Schwarz alternating procedure we choose the initial guess

$$u^0(x) := \begin{cases} 1 - \frac{x}{\epsilon} & 0 \le x \le \epsilon \\ 0 & \epsilon < x < 1 - \epsilon \\ 1 - \frac{1-x}{\epsilon} & 1 - \epsilon < x \le 1 \end{cases}$$

with $\epsilon < \frac{1}{2} - a$. The error can be determined graphically as in figure 3.3. We analyze the error in the infinity norm and observe

$$\|u - u^k\|_\infty = \|e^k\|_\infty = e^k(\frac{1}{2} - a).$$

Then

$$\begin{aligned} \|e^{k+1}\|_\infty &= e^{k+1}(\frac{1}{2} - a) \\ &= e_1^{k+1}(\frac{1}{2} - a) \\ &= \frac{\frac{1}{2} - a}{\frac{1}{2} + a} e_2^k(\frac{1}{2} + a) && \text{evaluate in } (0, \frac{1}{2} + a)\text{: } \frac{x}{\frac{1}{2} + a} e_2^k(\frac{1}{2} + a) \\ &= \frac{\frac{1}{2} - a}{\frac{1}{2} + a} \frac{\frac{1}{2} - a}{\frac{1}{2} + a} e_1^k(\frac{1}{2} - a) && \text{evaluate in } (\frac{1}{2} - a, 1)\text{: } \frac{1 - x}{\frac{1}{2} + a} e_1^k(\frac{1}{2} - a) \\ &= \left(\frac{1 - 2a}{1 + 2a}\right)^2 \|e^k\|_\infty \end{aligned}$$

So here we have in the infinity norm $\rho = \left(\frac{1-2a}{1+2a}\right)^2$.

3.2 Overlapping Schwarz Methods with Many Subdomains

Now we turn to a more general construction of Schwarz methods that allows us to extend it to

1. more than two subdomains and to
2. solve the subdomain problems by the finite element method.

Step 1: Decompose the domain Ω into p *non-overlapping* subdomains

$$\bar{\Omega} = \bigcup_{i=1}^{p} \bar{\Omega}_i, \qquad \Omega_i \cap \Omega_j = \emptyset, i \neq j.$$

In practice this could be done by constructing a mesh $\mathcal{T}_H$ with at least p elements. and choosing Ω_i as a union of mesh elements.

Step 2: Add overlap around every Ω_i

$$\hat{\Omega}_i = \{x \in \Omega : \operatorname{dist}(x, \Omega_i) < \beta H\}$$

with $H = \max_i \operatorname{diam}(\Omega_i)$.
In practice this could be done by *refining* the mesh $\mathcal{T}_H$ into a mesh $\mathcal{T}_h$ and choosing $\hat{\Omega}_i$ to be a union of elements from $\mathcal{T}_h$.

Concerning the finite element discretization there are two options:

1. "Partition, then discretize": Construct subdomains $\hat{\Omega}_i$, then discretize subdomains individually as in figure 3.4.
   - Advantage: mesh can be individually adapted to subdomains.
   - Disadvantage: complicated interpolation between subdomains.
2. "Discretize, then partition": First discretize Ω into $\mathcal{T}_H$. The make $\mathcal{T}_h$ a refinement of $\mathcal{T}_H$ and choose $\hat{\Omega}_i$ to be a union of elements from $\mathcal{T}_h$ as in figure 3.5.
   - Easy interpolation.
   - $\mathcal{T}_H$ will play an important role later.

We follow the *second* approach now. Assume Ω is polygonal.

1. Construct a coarse conforming and affine triangulation $\mathcal{T}_H = \{\Omega_1, \dots, \Omega_p\}$.

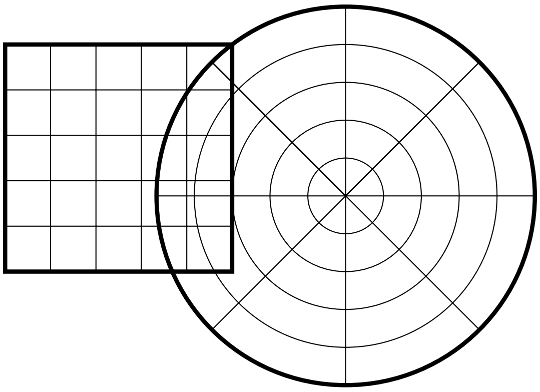


Figure 3.4: Partition, then discretize

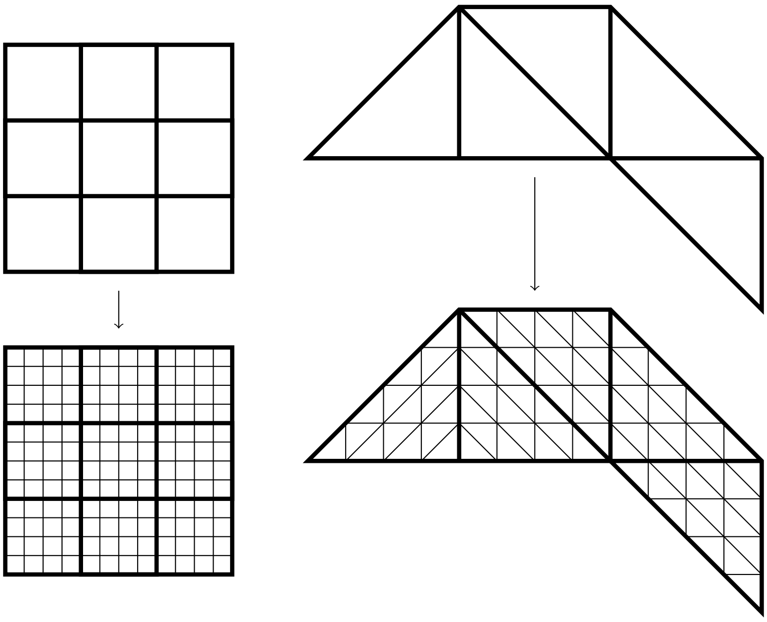


Figure 3.5: Subdomain decomposition using a grid hierarchy

2. Refine $\mathcal{T}_H$ uniformly m times to obtain fine triangulation $\mathcal{T}_h$.
3. Add overlap, i.e.
$$\hat{\Omega}_i := \{e \in T_h : e \in \Omega_i\} \cup \{e \in \mathcal{T}_h : \text{dist}(e, \Omega_i) < \beta H\}.$$
Note that
$$\mathcal{T}_{h,i} := \{e \in \mathcal{T}_h : e \subset \hat{\Omega}_i\}$$
provides a conforming and affine triangulation of $\hat{\Omega}_i$.

Then the Schwarz method can be formulated in variational form as follows. For $V = H_0^1(\Omega)$ let

$$u \in V : \qquad a(u, v) = l(v) \qquad \forall v \in V$$

be the variational formulation of (3.1). With the extension operator $\mathcal{E}_i : H_0^1(\hat{\Omega}_i) \to V$ given by

$$(\mathcal{E}_i u_i)(x) := \begin{cases} u_i(x) & x \in \hat{\Omega}_i \\ 0 & \text{else} \end{cases}$$

we define the subspaces

$$V_i := \{u \in V : u = \mathcal{E}_i u_i, u_i \in H_0^1(\hat{\Omega}_i)\} \subset V.$$

Then the alternating Schwarz method for many subdomains is given by algorithm 3.1. In the algorithm we had to solve

Algorithm 3.1 Alternating Schwarz Method for many subdomains

```
for $k = 0, 1, \ldots$ do
    for $i = 1, \ldots, p$ do
        $w_i \in V_i : a(u^{k+\frac{i-1}{p}} + w_i, v) = l(v) \forall v \in V_i$
        $u^{k+\frac{i}{p}} := u^{k+\frac{i-1}{p}} + w_i$
    end for
end for
```

$$w_i \in V_i : \qquad a(u^{k+\frac{i-1}{p}} + w_i, v) = l(v) \qquad \forall v \in V_i. \tag{3.5}$$

Note that the solution of (3.5) involves solving only local problems

$$\begin{aligned} a(w_i, v) &= \int\limits_{\Omega} \nabla w_i \cdot \nabla v \, dx = \int\limits_{\hat{\Omega}_i} \nabla w_i \cdot \nabla v \, dx = a_i(w_i, v) \\ &= \int\limits_{\hat{\Omega}_i} f v \, dx - \int\limits_{\hat{\Omega}_i} \nabla u^{k+\frac{i-1}{p}} \cdot \nabla v \, dx = l_i(v) - a_i(u^{k+\frac{i-1}{p}}, v) \;\; \forall v \in H_0^1(\hat{\Omega}_i), \end{aligned}$$

since $v = 0$ and $\nabla v = 0$ outside $\hat{\Omega}_i$.

3.3 Discrete Variational Formulation of Schwarz Methods

In order to apply (3.5) in practice we solve the variational problems approximately by the conforming finite element method. So assume $V_h \subset V$ is a finite-dimensional subspace equipped with a local basis, i.e. $P_k^d(\mathcal{T}_h)$ or $Q_k^d(\mathcal{T}_h)$. Then we set

$$V_{h,i} := \{v \in V_h : v(x) = 0 \, \forall x \in \bar{\Omega} \setminus \hat{\Omega}_i\} \subset V_h \subset V = H_0^1(\Omega).$$

Here we exploit that $\hat{\Omega}_i$ is given by the construction in 3.2, i.e. Ω_i is polygonal and the mesh resolves $\partial\hat{\Omega}_i$.

Algorithm 3.2 Discrete Schwarz Method

```
for $k = 0, \dots$ do
    for $i = 1, \dots, p$ do
        $w_i \in V_{h,i} : a_i(w_i, v) = l_i(v) - a_i(u_h^{k+\frac{i-1}{p}}, v) \, \forall v \in V_{h,i}$
        $u_h^{k+\frac{i}{p}} := u_h^{k+\frac{i-1}{p}} + w_i$
    end for
end for
```

In order to derive the algebraic formulation one needs to insert a basis representation of the discrete function spaces:

$$V_h := \operatorname{span} \Phi_h, \;\; \Phi_h := \{\phi_k : k \in I\}, \;\; V_{h,i} := \operatorname{span} \Phi_{h,i}, \;\; \Phi_{h,i} := \{\phi_k : k \in \hat{I}_i\},$$

where

$$\hat{I}_i := \{k \in I : \operatorname{supp} \phi_k \subset \hat{\Omega}_i\} \subset \tilde{I}_i, \qquad \tilde{I}_i := \{k \in I : \operatorname{supp} \phi_k \cap \hat{\Omega}_i \neq \emptyset\}.$$

Hereby we assume that a Lagrange basis has been chosen and that the Lagrange basis functions have local support. Now the local problems read

$$\begin{aligned}
& w_i \in V_{h,i} : a_i(w_i, v) = l_i(v) - a(u_h^{k+\frac{i-1}{p}}, v) \, \forall v \in V_{h,i} \\
\iff & a_i(\sum_{j \in \hat{I}_i} (z_i)_j \phi_j, \phi_m) = l_i(\phi_m) - a_i(\sum_{j \in \tilde{I}_i} (x^{k+\frac{i-1}{p}})_j \phi_j, \phi_m) \, \forall m \in \hat{I}_i \\
\iff & \sum_{j \in \hat{I}_i} (z_i)_j a_i(\phi_j, \phi_m) = l_i(\phi_m) - \sum_{j \in \tilde{I}_i} (x^{k+\frac{i-1}{p}})_j a_i(\phi_j, \phi_m) \, \forall m \in \hat{I}_i \\
\iff & A_{\hat{I}_i, \hat{I}_i} z_{\hat{I}_i} = b_{\hat{I}_i} - A_{\hat{I}_i, \tilde{I}_i} R_{\tilde{I}_i} x \qquad (3.6)
\end{aligned}$$

with $A_{\hat{I}_i, \hat{I}_i} = R_{I_i} A R_{I_i}^T$ and $b_{\hat{I}_i} = R_{\hat{I}_i} b$.

The right hand side can be written equivalently as

$$\begin{aligned} b_{\hat{I}_i} - A_{\hat{I}_i,\tilde{I}_i} R_{\tilde{I}_i} x^{k+\frac{i-1}{p}} &= R_{\hat{I}_i} b - R_{\hat{I}_i} A R^T_{\tilde{I}_i} R_{\tilde{I}_i} x^{k+\frac{i-1}{p}} \\ &= R_{\hat{I}_i} b - R_{\hat{I}_i} A x^{k+\frac{i-1}{p}} && \text{(adding zeros)} \\ &= R_{\hat{I}_i} (b - A x^{k+\frac{i-1}{p}}). && (3.7) \end{aligned}$$

The algebraic version of the update step is

$$\begin{aligned} & u_h^{k+\frac{i}{p}} = u_h^{k+\frac{i-1}{p}} + w_i \\ \iff & \sum_{j\in I} (x^{k+\frac{i}{p}})_j \phi_j = \sum_{j\in I} (x^{k+\frac{i-1}{p}})_j \phi_j + \sum_{j\in \hat{I}_i} (z_i)_j \phi_j \\ \iff & x^{k+\frac{i}{p}} = x^{k+\frac{i-1}{p}} + R^T_{\hat{I}_i} z_i && (3.8) \end{aligned}$$

So we arrive at the algebraic formulation of the alternating Schwarz method given in algorithm 3.3. We observe that this is identical to the block Gauß-Seidel method with the only difference that the index sets $\hat{I}_i$ are now overlapping!

Algorithm 3.3 Algebraic Formulation of the Alternating Schwarz Method

```
for $k = 0, \dots$ do
    for $i = 1, \dots, p$ do
        $x^{k+\frac{i}{p}} \overset{(3.8)}{:=} x^{k+\frac{i-1}{p}} + R^T_{\hat{I}_i} A^{-1}_{\hat{I}_i,\hat{I}_i} R_{\hat{I}_i} (b - A x^{k+\frac{i-1}{p}})$    (3.9)
    end for
end for
```

In complete analogy we can formulate a method that corresponds to the block Jacobi method with additional damping

$$x^{k+1} = x^k + \omega \sum_{i=1}^{p} R^T_{\hat{I}_i} A^{-1}_{\hat{I}_i,\hat{I}_i} R_{\hat{I}_i} (b - A x^k) \tag{3.10}$$

which is the algebraic version of

Algorithm 3.4 Variational Formulation corresponding to the block Jacobi method with additional damping

```
for $k = 0, 1, \dots$ do
    for $i = 1, \dots, p$ do
        $w_i \in V_i : a_i(w_i, v) = l_i(v) - a_i(u^k, v) \, \forall v \in V_i$    (3.11)
        $u^{k+1} := \omega \sum_{i=1}^{p} w_i$
    end for
end for
```

Clearly, in this version all corrections w_i can be computed in parallel. The damping factor needs to be sufficiently small to make the method convergent.

Error propagation operators

For the ease of writing let us introduce the abbreviations

$$R_i = R_{\hat{I}_i} \quad \text{and} \quad A_i = A_{\hat{I}_i,\hat{I}_i}.$$

Setting $e^{k+\frac{i}{p}} = x - x^{k+\frac{i}{p}}$ as usual, we obtain for one substep of the alternating Schwarz method with these abbreviations

$$e^{k+\frac{i}{p}} = (I - R_i^T A_i^{-1} R_i A) e^{k+\frac{i-1}{p}} = (I - P_i) e^{k+\frac{i-1}{p}}$$

where we defined the projection operator $P_i = R_i^T A_i^{-1} R_i A$. Consequently, for one complete step we obtain

$$e^{k+1} = (I - P_p) \cdots (I - P_1) e^k = \left(\prod_{i=1}^{p} (I - P_i) \right) e^k.$$

The alternating Schwarz method is therefore also called *multiplicative Schwarz method*.

Remark 3.1. Note that in each individual substep the matrix $R_i^T A_i^{-1} R_i$ does *not* have full rank and therefore can *not* be written as the inverse of some W_i.

Remark 3.2. Note also that $z_i = P_i e^{k+\frac{i-1}{p}}$ is the correction computed in substep i. Below the projection operator P_i will play an important role in the analysis.

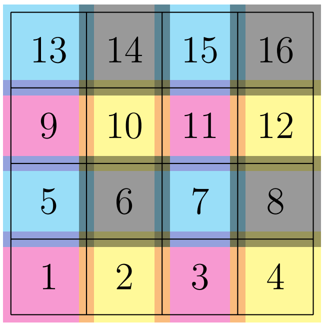
Remark 3.3. The multiplicative method needs to be symmetrized to be used as preconditioner in CG.

For the block Jacobi inspired variant we obtain the error propagation

$$e^{k+1} = \left(I - \omega \sum_{i=1}^{p} R_i^T A_i^{-1} R_i A \right) e^k = \left(I - \omega \sum_{i=1}^{p} P_i \right) e^k.$$

Due to this form the method is called *additive Schwarz method*.

It turns out $B = \sum_{i=1}^{p} R_i^T A_i^{-1} R_i$ is a symmetric positive definite preconditioner and therefore, according to Theorem 2.3, $\omega < \frac{1}{\lambda_{max}(BA)}$ is sufficient for convergence. In practice, however, one would rather employ the additive Schwarz method as a preconditioner in the CG method. Then the damping step can be omitted.



$$J_1 = \{1, 3, 9, 11\}$$
$$J_2 = \{2, 4, 10, 12\}$$
$$J_3 = \{5, 7, 13, 15\}$$
$$J_4 = \{6, 8, 14, 16\}$$

Figure 3.6: Coloring of a structured mesh in 2D

Independent corrections

The multiplicative Schwarz method can be parallelized with the following trick.

Observation 3.4. Provided $R_i A R_j^T = 0$, the order of computation of the two corrections in subdomains i and j is irrelevant.

Proof.

$$\begin{aligned}(I - P_i)(I - P_j) &= (I - R_i^T A_i^{-1} R_i A)(I - R_j^T A_j^{-1} R_j A) \\ &= I - R_i^T A_i^{-1} R_i A - R_j^T A_j^{-1} R_j A + R_i^T A_i^{-1} \underbrace{R_i A R_j^T}_{0} A_j^{-1} R_j A \\ &= I - P_i - P_j \end{aligned}$$ □

Now suppose that $J = \{1, \dots, p\}$ can be partitioned into

$$J = \bigcup_{n=1}^{c} J_n, \qquad J_i \cap J_j = \emptyset,\ i \neq j$$

such that $R_i A R_j^T = 0$ for all $i, j \in J_n$. Then all corrections in J_n can be computed in parallel.

For an appropriate decomposition of Ω into subdomains $\hat{\Omega}_i$ and sufficiently small overlap the constant c is independent of the number of subdomains p. An example for the unit square is given in figure 3.6. For a structured mesh in 2D (3D) four (eight) *colors* are sufficient.

An algorithm for computing the partitioning into the J_n is called a *coloring algorithm* and c is the *number of colors*.

The error propagation operator of the algorithm is then

$$e^{k+1} = \left(\prod_{n=1}^{c} \left(I - \sum_{i \in J_n} P_i \right) \right) e^k.$$

3.4 Coarse Grid Correction

We will prove below that the condition number of the system preconditioned by the additive or multiplicative Schwarz method defined so far is

$$\kappa(BA) \leq c(1 + \frac{H}{\delta})H^{-2}.$$

Since $H \approx \mathrm{diam}(\Omega)/p^{\frac{1}{d}}$ this is not acceptable for large p. The reason for this is that smooth errors are not reduced well: Consider an *interior* subdomain, i.e. $\partial\hat{\Omega}_i \cap \partial\Omega = \emptyset$ and an error $e^k = 1$ on $\hat{\Omega}_i$. Then

$$R_i A e^k = 0$$

since $\sum_j (A)_{lj} = 0$ for $l \in \hat{I}_i$ and consequently the correction computed in the subdomain $\hat{\Omega}_i$ is zero. The remedy is to add a so-called coarse grid correction which is constructed as follows.

Let V_H be a conforming finite element space equipped with a Lagrange basis on the coarse mesh $\mathcal{T}_H$ used to construct the subdomain decomposition. Due to the hierarchic construction we have $V_H \subset V_h$. Then for a given u^k compute the correction

$$w \in V_H : \qquad a(u^k + w, v) = l(v) \qquad \forall v \in V_H.$$

Setting $V_0 := V_H$ the variational formulation of the (damped) additive Schwarz method is given by algorithm 3.5.

Algorithm 3.5 Additive Schwarz Method with Coarse Grid Correction

```
for $k = 0, \dots$ do
    for $i = 0, \dots, p$ do
        $w_i \in V_i : a(w_i, v) = l(v) - a(u^k, v) \, \forall v \in V_i$
    end for
    $u^{k+1} := u^k + \omega \sum_{i=0}^{p} w_i$
end for
```

The algebraic formulation of the additive Schwarz method with coarse grid correction is given in algorithm 3.6.

The entries of R_H there are obtained from the representation of the coarse grid basis functions in terms of the fine grid basis functions. If

$$V_H = \mathrm{span}\{\phi_k^H : k \in I_H\} \subset V_h = \mathrm{span}\{\phi_k^h : k \in I_h\}$$

then

$$\phi_m^H = \sum_{n \in I_h} r_{mn} \phi_n^h = \sum_{n \in I_h} \phi_m^H(s_n) \phi_n^h$$

Algorithm 3.6 Additive Schwarz Method with Coarse Grid Correction (Algebraic)

```
for $k = 0, \dots$ do
    $x^{k+1} := x^k + \omega \sum_{i=0}^{p} R_i^T A_i^{-1} R_i (b - Ax^k)$
end for
```

where we extend the definition of the restriction matrices

$$R_i = \begin{cases} R_{\hat{I}_i} & \text{if } i \in \{1, \dots, p\} \\ R_H & \text{otherwise} \end{cases}$$
$$A_i = R_i A R_i^T.$$

and

$$(R_H)_{mn} = r_{mn}.$$

In the same way the multiplicative version can be extended by a coarse grid correction. Algorithm 3.7 gives a symmetrized version.

Algorithm 3.7 Multiplicative Schwarz Method with Coarse Grid Correction

```
for $k = 0, \dots$ do
    for $i = 1, \dots, p$ do
        $x^{k+\frac{i}{2p+1}} := x^{k+\frac{i-1}{2p+1}} + R_i^T A_i^{-1} R_i (b - x^{k+\frac{i-1}{2p+1}})$
    end for
    $x^{k+\frac{p+1}{2p+1}} := x^{k+\frac{p}{2p+1}} + R_0^T A_0^{-1} R_0 (b - x^{k+\frac{p}{2p+1}})$
    for i = p, . . . , 1 do
        $x^{k+\frac{2p+2-i}{2p+1}} := x^{k+\frac{2p+1-i}{2p+1}} + R_i^T A_i^{-1} R_i (b - x^{k+\frac{2p+1-i}{2p+1}})$
    end for
end for
```

We will prove below that these algorithms correspond to $\kappa(BA) \leq c(1 + \frac{H}{\delta})$ when used as preconditioners.

3.5 Complexity Considerations and Speedup

Before we analyze the Schwarz methods we present some general considerations about the parallel scalability of the method.

Reduction of sequential complexity

We first consider the Schwarz method as a *sequential* method, i.e. all corrections are computed sequentially on one processor. We make the following assumptions:

- Subdomain problems are solved with a direct solver having a complexity $O(N^\alpha)$ with $\alpha \geq 1$.
- We assume the computation time is dominated by the factorization phase. This is justified mainly if α is large.
- We assume a structured mesh in d dimensions discretizing the domain $\Omega = (0,1)^d$ with $N = n^d$, $n = \frac{1}{h}$ and $n_H = \frac{1}{H}$.
- Overlap is $\delta = \beta H$.

Then the time for the factorization phase is

$$\begin{aligned} T_S(n, n_H) &= \underbrace{n_H^{d\alpha}}_{\text{coarse grid}} + \underbrace{n_H^{d}}_{\text{\# subdomains}} \left(\frac{H + \beta H}{h}\right)^{d\alpha} \\ &= n_H^{d\alpha} + n_H^{d}\left(\frac{n}{n_H}(1+\beta)\right)^{d\alpha} \\ &= n_H^{d\alpha} + n_H^{d}\frac{n^{d\alpha}}{n_H^{d\alpha}}(1+\beta)^{d\alpha} \\ &= n_H^{d\alpha} + n_H^{d(1-\alpha)} n^{d\alpha}(1+\beta)^{d\alpha}. \end{aligned} \tag{3.12}$$

How should n_H be chosen? Minimize with respect to n_H:

$$\begin{aligned} \frac{\partial}{\partial n_H} T_S(n, n_H) &= d\alpha n_H^{d\alpha - 1} + d(1-\alpha) n_H^{d(1-\alpha)-1} n^{d\alpha}(1+\beta)^{d\alpha} \overset{!}{=} 0 \\ \iff n_H^{d\alpha - 1 - d(1-\alpha)+1} &= -\frac{d(1-\alpha)}{d\alpha} n^{d\alpha}(1+\beta)^{d\alpha} \\ \iff n_H = \left(\frac{\alpha-1}{\alpha}\right)^{\frac{1}{d(2\alpha-1)}} &(n(1+\beta))^{\frac{\alpha}{2\alpha-1}} \end{aligned}$$

Inserting the optimal n_H into (3.12):

$$\begin{aligned} T_S(n) &= c(n(1+\beta))^{\frac{\alpha d\alpha}{2\alpha-1}} + c^{d(1-\alpha)}(n(1+\beta))^{\frac{\alpha d(1-\alpha)}{2\alpha-1}}(n(1+\beta))^{d\alpha} \\ &= c(n(1+\beta))^{\frac{d\alpha^2}{2\alpha-1}} + c^{d(1-\alpha)}(n(1+\beta))^{\frac{d\alpha^2}{2\alpha-1}} \\ &= c(1+\beta)^{\frac{d\alpha^2}{2\alpha-1}} n^{\frac{d\alpha^2}{2\alpha-1}} \\ &= O(n^{\frac{d\alpha^2}{2\alpha-1}}). \end{aligned}$$

d	α	$d\alpha$	$\frac{d\alpha^2}{2\alpha-1}$	Remark
2	2	4	8/3	Banded Gauß
2	3/2	3	9/4	Nested Dissection
3	2	6	4	Nested Dissection

So

$$\frac{d\alpha^2}{2\alpha - 1} < d\alpha \iff \alpha > 1.$$

In the case $d = 3$, $\alpha = 2$ (nested dissection) this means

$$T_S(n) = n^4 = (n^3)^{\frac{4}{3}} = N^{\frac{4}{3}}$$

which is close to optimal with regard to N and much better than the direct method which is $O(N^2)$.

Optimal Coarse Grid in Parallel Case

We make the same assumptions as in the last section.

Now reserve one processor per subdomain and one *additional* processor for the coarse grid. All corrections are computed in parallel (this is optimistic as the coarse grid correction requires global communication). Then

$$T_P(n, n_H) = \max(n_H^{d\alpha}, (\frac{n}{n_H}(1+\beta))^{d\alpha}).$$

The optimal coarse grid size is obtained when the computation is balanced, i.e.

$$\begin{aligned} n_H^{d\alpha} &= (\frac{n}{n_H}(1+\beta))^{d\alpha} \\ \iff n_H &= \sqrt{n(1+\beta)}. \end{aligned}$$

Note that this fixes the number of processors to $p = n_H^d$!

The optimal run-time is then

$$n_H^{d\alpha} = (n(1+\beta))^{\frac{d\alpha}{2}}.$$

Speedup of additive Schwarz without Coarse Grid

Assumptions:

- $p = n_H^d$ processors.
- Complexity of subdomain solver is $n^{d\alpha}$.
- We consider the speedup *of one iteration* with respect to the Schwarz method used as a sequential solver.
- We analyze the influence of the communication cost.
- We consider only the communication with the nearest neighbors in coordinate directions (i.e. the overlap mus be sufficiently small and communication to diagonal neighbors is ignored).

$$
\begin{aligned}
S(n,p) &= \frac{(\frac{n}{n_H}(1+\beta))^{\alpha d}\, p\, t_f}{(\frac{n}{n_H}(1+\beta))^{\alpha d}\, t_f + (\underbrace{2d}_{\#\text{ comm}} \underbrace{\underbrace{\beta\frac{n}{n_H}}_{\text{diam. of overlap}} \quad (\frac{n}{n_H}(1+\beta))^{d-1}}_{\#\text{ dof in overlap region}})\, t_w} \\
&= \frac{p}{1 + 2d\beta\frac{n}{n_H}(\frac{n}{n_H}(1+\beta))^{d-1-\alpha d}\,\frac{t_w}{t_f}} \\
&= \frac{p}{1 + 2d\beta(1+\beta)^{d(1-\alpha)-1}(\frac{n}{n_H})^{d(1-\alpha)}\,\frac{t_w}{t_f}} \\
&= \frac{p}{1 + 2d\beta(1+\beta)^{d(1-\alpha)-1}(\frac{n_H}{n})^{d(\alpha-1)}\,\frac{t_w}{t_f}}
\end{aligned}
$$

where t_f is the time for one floating point operation and t_w is the time needed to communicate one floating point number.

Two cases need to be considered:

- $\alpha > 1$, i.e. subdomain solver has *more* than linear complexity. Then
$$\lim_{n\to\infty} S(n,p) = p$$
since $\frac{n_H}{n} \to 0$ and $d(\alpha - 1) > 0$.

- $\alpha = 1$. Then $d(\alpha - 1) = 0$ and the speedup is fixed to
$$S(n,p) = \frac{p}{1 + 2d\frac{\beta}{1+\beta}\,\frac{t_w}{t_f}}.$$

Scalability

We now investigate the case $p \to \infty$.

- No coarse grid:
As shown above any speedup for one iteration can be achieved for n sufficiently large provided a non-optimal subdomain solver is used. However the number of iterations increases as $p^{\frac{1}{d}}$!

- With coarse grid:
When the subdomain size $N_{local} := (\frac{n}{n_H}(1+\beta))^d$ is fixed and $p = n_H^d$ is increased the number of iterations stays constant, the problem size $N = pN_{local}$ increases, but also the coarse problem size $N_H = n_H^d = p$ increases and eventually needs to be parallelized. Possibility: recursive application of the Schwarz method on fewer processors.

3.6 Numerical Examples

Now let us illustrate the behavior of the overlapping Schwarz method for a concrete example. Throughout this section we solve the Poisson equation on $\Omega = (0, 1)^2$ with Dirichlet boundary conditions on a structured, axiparallel and equidistant mesh with Q_1 finite elements. The number of iterations is given for 10^{-6} reduction of the relative Euclidean norm of the residual and a random initial guess. The preconditioners are always used within a conjugate gradient method unless noted otherwise.

Strong Scaling Single Grid Additive Schwarz

This means we fix $h = 1/512$ and overlap $\delta = 4h$ and vary the subdomain size H and consequently the number of subdomains H^{-2}. Since the method is used as a preconditioner we expect an asymptotic behavior like $\#IT \sim H^{-1}$ which is not quite confirmed yet.

H	1	1/2	1/3	1/4	1/5	1/6
P	1	4	9	16	25	36
#IT	1	26	27	29	36	41

In the next experiment we fix $h = 1/512$, $H = 1/4$, i.e. $P = 16$ and just vary the size of the overlap. The results show a good improvement initially and then a saturation.

δ/h	1	2	3	4	8	16
#IT	58	46	33	29	25	19

Weak Scaling Single Grid Schwarz

Now we scale the problem size linearly with the number of subdomains, i.e. we fix $H/h = 256$ and the overlap $\delta = 4h$. The expected behavior is the same as for the first table above, i.e. $\#IT \sim H^{-1}$. The number of iterations should not depend on h.

We compare four different methods: additive Schwarz (AS), multiplicative Schwarz with lexicographic ordering f the subdomains (MS) used as preconditioner in restarted GMRES, symmetrized multiplicative Schwarz with lexicographic ordering (SMS) used with CG and symmetrized multiplicative Schwarz with coloring (SMSC) used with CG.

	H	1/2	1/3	1/4	1/5	1/6
Method	P	4	9	16	25	36
AS	#IT	26	32	38	44	50
MS	#IT	15	23	28	33	38
SMS	#IT	11	14	17	19	22
SMSC	#IT	11	14	17	19	22

We observe that the multiplicative version needs fewer iterations but shows the same asymptotic behavior with the number of subdomains. The symmetrized versions still need less iterations than the nonsymmetric version but each iteration is twice as expensive. Interestingly, the non-parallel version with lexicographic ordering shows the same convergence rate as the version with coloring which can be executed in parallel.

Weak Scaling Two-level Additive Schwarz

Now we add the coarse grid correction. Again weak scaling with $H/h = 256$ is investigated. The coarse mesh size was $H/h_0 = 2$ as our implementation does not allow for one cell on the coarse grid for one subdomain. The iteration numbers are expected to be robust in h and H and should only depend on the overlap. This is nicely confirmed by the results.

H	1/2	1/3	1/4	1/5	1/6
P	4	9	16	25	36
δ/h	#IT				
1	28	31	32	32	23
2	22	23	24	24	24
4	16	17	17	17	17
8	13	13	13	13	13

Additive Schwarz versus Direct Solver in 3D

In this section we look at run-times for a 3D problem and investigate whether additive Schwarz indeed leads to a reduction in computational complexity and run-time. All tests here have been carried out on an 2,6 GHz Intel Core i7 processor with four cores on a mesh with 40^3 elements using an overlap $\delta = 4h$. As a direct solver SuperLU was employed.

P	1	2	4	8
sequential time	149.8	100.4	145.6	199.6
wall clock time	149.8	48.1	36.0	28.0

The table shows a reduction in computing time for 2 subdomains when both subdomains are processed sequentially. A substantial reduction in wall-clock time is achieved when all four cores of the CPU are used.

Chapter 4

Abstract Schwarz Theory

This chapter is based on [Toselli and Widlund, 2005, chapter 2].

4.1 Subspace Correction Methods

The methods considered so far (and many more) can be written in an abstract, unified way. Let the following ingredients be given.

1. A conforming finite element space $V_h \subset H_0^1(\Omega)$ (non-homogeneous Dirichlet boundary conditions and Neumann boundary conditions can be treated as well).

2. A variational formulation

$$u \in V_h: \ a(u,v) = l(v) \quad \forall v \in V_h$$

   with a symmetric and coercive bilinear form $a(\cdot,\cdot)$.

3. A subspace decomposition $V_h = \sum_{i=0}^{p} V_{h,i}$ with $V_{h,i} \subset V_h$.

Introducing a basis for the (sub-)spaces

$$V_h = \operatorname{span}\{\phi_k^h : k \in I_h\}, \qquad\qquad V_{h,i} = \operatorname{span}\{\phi_k^{h,i} : k \in I_{h,i}\}$$

we arrive at symmetric positive-definite linear systems

$$\begin{aligned} u_h \in V_h: \quad & a(u_h,v) = l(v) \quad \forall v \in V_h & \iff \ & Ax = b, \\ u_i \in V_{h,i}: \quad & a(u_i,v) = l(v) \quad \forall v \in V_{h,i} & \iff \ & A_i x_i = b_i. \end{aligned}$$

We do not analyze inexact subdomain solvers here. In that case $a(\cdot,\cdot)$ is replaced by $\tilde{a}_i(\cdot,\cdot)$ in the local problems.

The prolongation operator $R_i^T : \mathbb{R}^{I_{h,i}} \to \mathbb{R}^{I_h}$ describes the change of basis from $V_{h,i}$ to V_h, i.e.

$$V_{h,i} \ni u_i = \sum_{k \in I_{h,i}} (x_i)_k \phi_k^{h,i} = \sum_{k \in I_h} (R_i^T x_i)_k \phi_k^h.$$

The matrices of the local problems are then given by $A_i = R_i A R_i^T$. Then the additive and multiplicative subspace correction methods are given by

$$\begin{aligned} x^{k+1} &= x^k + \omega \sum_{i=0}^{p} R_i^T A_i^{-1} R_i (b - A x^k), \\ x^{k+\frac{i+1}{p+1}} &= x^{k+\frac{i}{p+1}} + R_i^T A_i^{-1} R_i (b - A x^{k+\frac{i}{p+1}}) \qquad\qquad i = 0, \dots, p. \end{aligned}$$

As shown above, the corresponding error propagation operators are given by

$$E_{ad} = I + \omega \sum_{i=0}^{p} P_i, \qquad E_{mu} = (I - P_0) \dots (I - P_p) = \prod_{i=0}^{p} (I - P_i)$$

where

$$P_i = R_i^T A_i^{-1} R_i A.$$

The analysis of the additive case is based on analyzing the condition number $\kappa(\sum_{i=0}^{p} P_i A)$, i.e. the method is used as a preconditioner. Since the multiplicative method is not symmetric (of course it could be symmetrized) we will directly analyze the norm of the error propagation operator

$$\|E_{mu}\|_A^2 = \sup_{x \neq 0} \frac{\langle A E_{mu} x, x \rangle}{\langle Ax, x \rangle}.$$

The analysis will be based solely on the following two assumptions.

Assumption 4.1 (Stable splitting)**.** There exists a constant $c_0 > 0$ such that for all $x \in \mathbb{R}^{I_h}$ there exists a splitting $x = \sum_{i=0}^{p} R_i^T x_i$ with $x_i \in \mathbb{R}^{I_{h,i}}$ and

$$\sum_{i=0}^{p} \langle R_i^T x_i, R_i^T x_i \rangle_A \leq c_0 \langle x, x \rangle_A.$$

Here $\langle x, y \rangle_A = \langle Ax, y \rangle$ is the A-scalar product. □

Assumption 4.2 (Strengthened Cauchy-Schwarz Inequality)**.** There exist constants $0 \leq \epsilon_{ij} \leq 1$ for $1 \leq i, j \leq p$ such that

$$|\langle R_i^T x_i, R_j^T x_j \rangle_A| \leq \epsilon_{ij} \langle R_i^T x_i, R_i^T x_i \rangle_A^{\frac{1}{2}} \langle R_j^T x_j, R_j^T x_j \rangle_A^{\frac{1}{2}}$$

for all $x_i \in \mathbb{R}^{I_{h,i}}$ and $x_j \in \mathbb{R}^{I_{h,j}}$. Let $\mathscr{E} \in \mathbb{R}^{p \times p}$ be the matrix with coefficients $(\mathscr{E})_{ij} = \epsilon_{ij}$ and spectral radius $\rho(\mathscr{E})$. Note that $\mathscr{E}$ does *not* include the index $i = 0$ which is assumed to be a coarse grid space. □

Obviously the second assumption holds trivially with $\epsilon_{ij} = 1$. Moreover, the constants c_0 and ϵ_{ij} should be as independent as possible of the mesh size h, the number of local problems p and possibly other problem parameters (such as the diffusion coefficient).

Assumptions 4.1 and 4.2 need to be verified individually for the different schemes and then can be plugged into the general theorems proven in this chapter.

4.2 Additive Case

As mentioned, the analysis of the additive case is based on estimating the spectral condition number $\kappa(C) = \frac{\lambda_{max}(C)}{\lambda_{min}(C)}$.

Observation 4.3. Any scalar product can be used in the definition of the Raleigh quotient in Observation 2.4.

Proof. Let C, M be symmetric and positive definite matrices. Then

$$\begin{aligned}
\min_{x\neq 0} \frac{\langle Cx, x\rangle}{\langle x, x\rangle} &= \min_{0\neq x=M^{\frac{1}{2}}y} \frac{\langle CM^{\frac{1}{2}}y, M^{\frac{1}{2}}y\rangle}{\langle M^{\frac{1}{2}}y, M^{\frac{1}{2}}y\rangle} \\
&= \min_{y\neq 0} \frac{\langle M^{\frac{1}{2}}CM^{-\frac{1}{2}}M^{\frac{1}{2}}y, M^{\frac{1}{2}}y\rangle}{\langle My, y\rangle} \qquad (\sigma(C) = \sigma(M^{\frac{1}{2}}CM^{-\frac{1}{2}})) \\
&= \min_{y\neq 0} \frac{\langle MCy, y\rangle}{\langle My, y\rangle} \\
&= \min_{x\neq 0} \frac{\langle Cx, x\rangle_M}{\langle x, x\rangle_M}.
\end{aligned}$$

The same argument can be applied for the maximal eigenvalue. □

In particular we will in the following use the scalar product induced by the stiffness matrix A itself.

Setting $\omega = 1$ in the additive Schwarz iteration we identify the preconditioned system as

$$BAx = \left(\sum_{i=0}^{p} R_i^T A_i^{-1} R_i\right) Ax = \sum_{i=0}^{p} P_i x = Bb.$$

According to the considerations in Section 2.3 (set $M^{-1} = B$) we are led to analyze the condition number of the preconditioned system

$$\kappa(BA) = \kappa(\sum_{i=0}^{p} P_i).$$

Using observation 4.3 if suffices to obtain constants γ, Γ such that

$$\gamma\langle x, x\rangle_A \leq \langle\sum_{i=0}^{p} P_i x, x\rangle_A \leq \Gamma\langle x, x\rangle_A$$

which implies

$$\kappa(\sum_{i=0}^{p} P_i) \leq \frac{\Gamma}{\gamma}.$$

Lemma 4.4 (Properties of P_i)**.** $P_i = R_i^T A_i^{-1} R_i A$ is an orthogonal projection in the A-scalar product and we have

1. $P_i^2 = P_i$
2. $AP_i = P_i^T A$. This implies $\langle x, P_i y\rangle_A = \langle P_i x, y\rangle_A$, i.e. P_i is self-adjoint with respect to the A-scalar product.
3. $\langle P_i x, P_i y\rangle_A = \langle x, P_i y\rangle_A$ for all $x, y \in \mathbb{R}^{I_h}$.
4. $\langle P_i x, (I - P_i)y\rangle_A = 0$ for all $x, y \in \mathbb{R}^{I_h}$.
5. $\|x\|_A^2 = \|P_i x\|_A^2 + \|(I - P_i)x\|_A^2$ for all $x \in \mathbb{R}^{I_h}$.
6. $\|P_i x\|_A \leq \|x\|_A$

Proof.

1. $P_i^2 = R_i^T A_i^{-1} \underbrace{R_i A R_i^T}_{A_i} A_i^{-1} R_i A = R_i^T A_i^{-1} R_i A = P_i$
2. $AP_i = AR_i^T A_i^{-1} R_i A = (R_i^T A_i^{-1} R_i A)^T A = P_i^T A$
3. $\langle P_i x, P_i y\rangle_A = x^T P_i^T A P_i y \overset{2.}{=} x^T A P_i P_i y \overset{1.}{=} x^T A P_i y = \langle x, P_i y\rangle_A$
4. $\langle P_i x, (I - P_i)y\rangle_A = \langle P_i x, y\rangle_A - \langle P_i x, P_i y\rangle_A \overset{3.}{=} 0$
5. $$\begin{aligned}\|x\|_A^2 &= \|P_i x + (I - P_i)x\|_A^2 \\ &= \langle P_i x + (I - P_i)x, P_i x + (I - P_i)x\rangle_A \\ &= \langle P_i x, P_i x\rangle_A + \langle (I - P_i)x, (I - P_i)x\rangle_A \\ &= \|P_i x\|_A^2 + \|(I - P_i)x\|_A^2\end{aligned}$$
6. $\|P_i x\|_A^2 \overset{5.}{=} \|x\|_A^2 - \|(I - P_i)x\|_A^2 \leq \|x\|_A^2$ □

Lemma 4.5 (Estimate of largest eigenvalue)**.** From assumption 4.2 follows

$$\langle\sum_{i=0}^{p} P_i x, x\rangle_A \leq (1 + \rho(\mathscr{E}))\langle x, x\rangle_A.$$

Proof. 1. Since assumption 4.2 does not involve P_0, we split it off:

$$\langle\sum_{i=0}^{p} P_i x, x\rangle_A = \langle P_0 x, x\rangle_A + \langle\sum_{i=1}^{p} P_i x, x\rangle_A \leq \langle x, x\rangle_A + \langle\sum_{i=1}^{p} P_i x, x\rangle_A$$

Here we used

$$\langle P_0 x, x\rangle_A \overset{4.4\ (3.)}{=} \langle P_0 x, P_0 x\rangle_A \overset{4.4\ (6.)}{\leq} \langle x, x\rangle_A.$$

2. We have

$$\begin{aligned}
\langle\sum_{i=1}^{p} P_i x, \sum_{i=1}^{p} P_i x\rangle_A &= \sum_{i=1}^{p}\sum_{j=1}^{p}\langle P_i x, P_j x\rangle_A \\
&\overset{4.2}{\leq} \sum_{i=1}^{p}\sum_{j=1}^{p} \epsilon_{ij} \underbrace{\langle P_i x, P_i x\rangle_A^{\frac{1}{2}}}_{=:(z)_i}\underbrace{\langle P_j x, P_j x\rangle_A^{\frac{1}{2}}}_{=:(z)_j} \quad (z \in \mathbb{R}^p) \\
&= z^T \mathscr{E} z \\
&\leq \|\mathscr{E}\|_2 \langle z, z\rangle \\
&= \rho(\mathscr{E}) \sum_{i=1}^{p}\langle P_i x, P_i x\rangle_A \qquad \text{(Def. of } z) \\
&\overset{4.4\ (3.)}{=} \rho(\mathscr{E})\langle\sum_{i=1}^{p} P_i x, x\rangle_A \\
&\overset{\text{C.S.}}{\leq} \rho(\mathscr{E})\langle\sum_{i=1}^{p} P_i x, \sum_{i=1}^{p} P_i x\rangle^{\frac{1}{2}}\langle x, x\rangle_A^{\frac{1}{2}}
\end{aligned}$$

Dividing by $\langle\sum_{i=1}^{p} P_i x, \sum_{i=1}^{p} P_i x\rangle^{\frac{1}{2}}$ gives

$$\langle\sum_{i=1}^{p} P_i x, \sum_{i=1}^{p} P_i x\rangle^{\frac{1}{2}} \leq \rho(\mathscr{E})\langle x, x\rangle_A^{\frac{1}{2}}.$$

Note that we used

$$\|\mathscr{E}\|_2 = \sup_{x\neq 0}\frac{\|\mathscr{E}x\|_2}{\|x\|_2} = \sup_{x\neq 0}\sup_{y\neq 0}\frac{|\langle \mathscr{E}x, y\rangle|}{\|x\|_2\|y\|_2}$$

which implies

$$|\langle \mathscr{E}x, y\rangle| \leq \|\mathscr{E}\|_2\|x\|_2\|y\|_2.$$

3. Finally

$$\begin{aligned}\langle\sum_{i=1}^{p} P_i x, x\rangle_A &\overset{\text{C.S.}}{\leq} \langle\sum_{i=1}^{p} P_i x, \sum_{i=1}^{p} P_i x\rangle^{\frac{1}{2}}\langle x, x\rangle_A^{\frac{1}{2}} \\ &\overset{2.}{\leq} \rho(\mathscr{E})\langle x, x\rangle_A.\end{aligned}$$

Now combine 1. and 3. to conclude. □

Remark 4.6. In overlapping domain decomposition each subdomain overlaps only with a maximum number N of other subdomains. Setting

$$\epsilon_{ij} = \begin{cases} 0 & R_i A R_j^T = 0 \text{ (when } \partial\Omega_i \cap \partial\Omega_j = \emptyset) \\ 1 & \text{otherwise} \end{cases}$$

this means

$$\|\mathscr{E}\|_\infty = \max_i |\{(i,j) : \epsilon_{ij} \neq 0\}| = N.$$

Since $\rho(\mathscr{E}) \leq \|\mathscr{E}\|_\infty$ we can conclude that the largest eigenvalue is bounded by $N + 1$ independent of p.

Lemma 4.7 (Partitioning Lemma, Lions' Lemma)**.** The stable splitting from assumption 4.1 implies

$$c_0^{-1}\langle x, x\rangle_A \leq \langle\sum_{i=0}^{p} P_i x, x\rangle_A,$$

i.e. c_0^{-1} is an estimate of the smallest eigenvalue.

Proof.

$$\begin{aligned}\langle x, x\rangle_A &= \langle x, \sum_{i=0}^{p} R_i^T x_i\rangle_A = \sum_{i=0}^{p}\langle x, R_i^T x_i\rangle_A = \sum_{i=0}^{p}\langle x, P_i R_i^T x_i\rangle_A \qquad (*) \\ &= \sum_{i=0}^{p} x^T A P_i R_i^T x_i = \sum_{i=0}^{p}(P_i x)^T A R_i^T x_i = \sum_{i=0}^{p}\langle P_i x, R_i^T x_i\rangle_A \\ &\leq \sum_{i=0}^{p}\|P_i x\|_A \|R_i^T x_i\|_A \qquad \text{(Cauchy-Schwarz)} \\ &\leq \left(\sum_{i=0}^{p}\|P_i x\|_A^2\right)^{\frac{1}{2}}\left(\sum_{i=0}^{p}\|R_i^T x_i\|_A^2\right)^{\frac{1}{2}} \qquad \text{(Cauchy-Schwarz)}\end{aligned}$$

For (*) we used that $P_i^2 = P_i$ and $\text{range}(P_i) = \text{range}(R_i^T)$.

Squaring both sides and inserting assumption 4.1:

$$\begin{aligned}\langle x, x\rangle_A^2 &\leq \left(\sum_{i=0}^{p} \|P_i x\|_A^2\right)\left(\sum_{i=0}^{p} \|R_i^T x_i\|_A^2\right)\\ &\leq \left(\sum_{i=0}^{p} \langle P_i x, P_i x\rangle_A\right) c_0 \langle x, x\rangle_A.\end{aligned}$$

Dividing through and using 3. from Lemma 4.4:

$$c_0^{-1}\langle x, x\rangle_A \leq \sum_{i=0}^{p} \langle P_i x, x\rangle_A. \qquad \square$$

Theorem 4.8 (Condition number of additive Schwarz)**.** Assumptions 4.1 and 4.2 imply

$$\kappa(\sum_{i=0}^{p} P_i) \leq c_0(\rho(\mathscr{E}) + 1).$$

Proof. Use Lemma 4.5 and 4.7. □

Note that from Remark 4.6 follows that the upper bound $\rho(\mathscr{E})+1$ for overlapping Schwarz methods is independent of p and h under very mild assumptions. So the main difficulty is to ensure assumption 4.1.

4.3 Multiplicative Case

We now aim to estimate $\|E_{mu}\|_A$ directly as E_{mu} is not symmetric. Let us start with a technical lemma.

Lemma 4.9. Assume a strengthened Cauchy-Schwarz inequality holds. Then we have for $0 \leq i, j \leq p$ and all $x, y \in \mathbb{R}^{I_h}$

$$\begin{aligned}\langle P_i x, y\rangle_A &\leq \langle P_i x, x\rangle_A^{\frac{1}{2}} \langle P_i y, y\rangle_A^{\frac{1}{2}} \qquad \text{(Note the } P_i \text{ in the second factor.)}\\ \langle P_i x, P_j y\rangle_A &\leq \epsilon_{ij} \langle P_i x, x\rangle_A^{\frac{1}{2}} \langle P_j y, y\rangle_A^{\frac{1}{2}}.\end{aligned}$$

Proof. Using Lemma 4.4 we have

$$\langle P_i x, y\rangle_A = \langle P_i x, P_i y\rangle_A \leq \langle P_i x, P_i x\rangle_A^{\frac{1}{2}} \langle P_i y, P_i y\rangle_A^{\frac{1}{2}} = \langle P_i x, x\rangle_A^{\frac{1}{2}} \langle P_i y, y\rangle_A^{\frac{1}{2}}.$$

Since $\text{range}(P_i) = \text{range}(R_i^T)$ we can use the strengthened Cauchy-Schwarz inequality:

$$\langle P_i x, P_j y\rangle_A \leq \epsilon_{ij} \langle P_i x, P_i x\rangle_A^{\frac{1}{2}} \langle P_j y, P_j y\rangle_A^{\frac{1}{2}} = \epsilon_{ij} \langle P_i x, x\rangle_A^{\frac{1}{2}} \langle P_j y, y\rangle_A^{\frac{1}{2}}. \qquad \square$$

The main theorem for the multiplicative subspace correction method then reads:

Theorem 4.10. Let assumptions 4.1 and 4.2 hold. Then

$$\|E_{mu}\|_A^2 \leq 1 - \frac{1}{c_0(1+2\rho^2(\mathscr{E}))} < 1.$$

Observe that $\rho(\mathscr{E}) \geq 1$ since $\epsilon_{ii} = 1$ for $1 \leq i \leq p$ and $c_0 \geq 1$ since $x = R_i^T x_i$, $x_j = 0$, $i \neq j$, gives a contradiction to $c_0 < 0$.

Proof. The proof is carried out in several steps.

1. Some definitions:

$$E_{-1} := I$$

$$E_j := (I - P_j) \cdot \ldots \cdot (I - P_0) = \prod_{k=0}^{j} (I - P_k) \qquad \text{for } 0 \leq j \leq p.$$

   Obviously $E_{mu} = E_p$.

2. We need to cope with the fact that E_j is not symmetric. This is achieved by introducing the adjoin E^* of E with respect to the A-scalar product:

$$\langle Ex, y\rangle_A = \langle x, E^* y\rangle_A \quad \forall x, y \in \mathbb{R}^{I_h}.$$

   (Observe that $E^* = A^{-1}EA$). Then we have the following recursion:

$$E_{j-1}^* E_{j-1} - E_j^* E_j = E_{j-1}^* P_j E_{j-1} \quad 0 \leq j \leq p. \tag{4.1}$$

   In order to prove this, first observe that $P_j^* = P_j$ since $\langle P_j x, y, \rangle_A = \langle x, P_j y, \rangle_A$ according to Lemma 4.4(2). Therefore

$$\begin{aligned}
E_j^* E_j &= E_{j-1}^* (I - P_j)^* (I - P_j) E_{j-1} && \text{(Definition of } E_j\text{)} \\
&= E_{j-1}^* (I - P_j^* - P_j + P_j^* P_j) E_{j-1} \\
&= E_{j-1}^* (I - P_j) E_{j-1} && (P_j = P_j^*,\ P_j^2 = P_j) \\
&= E_{j-1}^* E_{j-1} - E_{j-1}^* P_j E_{j-1}.
\end{aligned}$$

   This holds also for $j = 0$ since $E_{-1} = I$.

3. Now use eq. (4.1) in a telescoping sum:

$$\begin{aligned}
\sum_{j=1}^{p} E_{j-1}^* P_j E_{j-1} &= \sum_{j=1}^{p} (E_{j-1}^* E_{j-1} - E_j^* E_j) = E_0^* E_0 - E_p^* E_p \\
&= I - P_0 - E_p^* E_p
\end{aligned}$$

which results in the additive representation

$$I - E_p^* E_p = \sum_{j=0}^{p} E_{j-1}^* P_j E_{j-1}$$

where we used the fact $E_{-1} = I$.

Since the P_j are positive semidefinite we have

$$\langle (I - E_p^* E_p)x, x\rangle_A = \langle (\sum_{i=0}^{p} E_{j-1}^* P_j E_j)x, x\rangle_A = \sum_{j=0}^{p} \langle P_j E_j x, E_j x\rangle_A \geq 0.$$

If we *could* show an estimate of the form

$$\langle (I - E_p^* E_p)x, x\rangle_A \geq \alpha \langle x, x\rangle_A \tag{4.2}$$

with $\alpha > 0$ this implies

$$(1-\alpha)\langle x, x\rangle_A \geq \langle E_p^* E_p x, x\rangle_A$$

and thus

$$\|E_{mu}\|_A^2 = \|E_p\|_A^2 = \sup_{x \neq 0} \frac{\langle E_p x, E_p x\rangle_A}{\langle x, x\rangle_A} = \sup_{x \neq 0} \frac{\langle E_p^* E_p x, x\rangle_A}{\langle x, x\rangle_A} \leq 1 - \alpha.$$

So our goal now is to show (4.2).

4. From the definition of the error we get the following recursive relation:

$$E_j = I - \sum_{k=0}^{j} P_k E_{k-1}, \quad 0 \leq j \leq p$$

which we show by induction:

$$\begin{aligned} E_0 &= I - P_0 = I - \sum_{k=0}^{0} P_0 E_{-1} \\ E_j &= (I - P_j)E_{j-1} = E_{j-1} - P_j E_{j-1} \\ &= (I - \sum_{k=0}^{j-1} P_k E_{k-1}) - P_j E_{j-1} = I - \sum_{k=0}^{j} P_k E_{k-1}. \end{aligned}$$

5. With 4. in the form $I = E_{j-1} + \sum_{k=0}^{j-1} P_k E_{k-1}$ we get using Lemma 4.4 and 4.9:

$$\begin{aligned}
\langle P_j x, x\rangle_A &= \langle P_j x, (E_{j-1} + \sum_{k=0}^{j-1} P_k E_{k-1})x\rangle_A \\
&= \langle P_j x, E_{j-1}x\rangle_A + \langle P_j x, P_0 x\rangle_A + \sum_{k=1}^{j-1}\langle P_j x, P_k E_{k-1}x\rangle_A \\
&\le \langle P_j x, x\rangle_A^{\frac{1}{2}} \Bigg(\langle P_j E_{j-1}x, E_{j-1}x\rangle_A^{\frac{1}{2}} + \langle P_j P_0 x, P_0 x\rangle_A^{\frac{1}{2}} \\
&\qquad + \sum_{k=1}^{j-1} \epsilon_{jk} \underbrace{\langle P_k E_{k-1}x, E_{k-1}x\rangle_A^{\frac{1}{2}}}_{=:c_k} \Bigg) \\
&= \langle P_j x, x\rangle_A^{\frac{1}{2}} \left(\langle P_j P_0 x, P_0 x\rangle_A^{\frac{1}{2}} + \sum_{k=1}^{j} \epsilon_{jk} c_k \right)
\end{aligned}$$

Squaring both sides, dividing by $\langle P_j x, x\rangle_A$ and estimating with $(a+b)^2 \le 2a^2 + 2b^2$ yields:

$$\begin{aligned}
\langle P_j x, x\rangle_A &\le 2\langle P_j P_0 x, P_0 x\rangle_A + 2\left(\sum_{k=1}^{j} \epsilon_{jk} c_k\right)^2 \\
&\le 2\langle P_j P_0 x, P_0 x\rangle_A + 2(\mathscr{E}c)_j^2
\end{aligned}$$

(Note: $\epsilon_{jk}, c_k \ge 0$)

Now sum $j = 1, \dots, p$ and add $\langle P_0 x, x\rangle_A$ to both sides:

$$\begin{aligned}
\langle (\sum_{j=0}^{p} P_j)x, x\rangle_A &\le \langle P_0 x, x\rangle_A + \sum_{j=1}^{p} \left(2\langle P_j P_0 x, P_0 x\rangle_A + 2(\mathscr{E}c)_j^2\right) \\
&\le (1 + 2\rho(\mathscr{E}))\langle P_0 x, x\rangle_A + 2\rho^2(\mathscr{E})\|c\|^2 \qquad (*) \\
&= (1 + 2\rho(\mathscr{E}))\langle P_0 x, x\rangle_A + 2\rho^2(\mathscr{E}) \sum_{k=1}^{p}\langle P_k E_{k-1}x, E_{k-1}x\rangle_A \\
&\le \max(1 + 2\rho(\mathscr{E}), 2\rho^2(\mathscr{E})) \sum_{k=0}^{p}\langle P_k E_{k-1}x, E_{k-1}x\rangle_A \\
&\le (1 + 2\rho^2(\mathscr{E})) \sum_{k=0}^{p}\langle P_k E_{k-1}x, E_{k-1}x\rangle_A
\end{aligned}$$

For (*) we used that

$$\langle (\sum_{j=1}^{p} P_j) P_0 x, P_0 x\rangle_A \le \rho(\mathscr{E})\langle P_0 x, x\rangle_A$$

according to Lemma 4.5 and that

$$\sum_{j=1}^{p} (\mathscr{E}c)_j^2 = \sum_{j=1}^{p} (\mathscr{E}c)_j(\mathscr{E}c)_j = \langle \mathscr{E}c, \mathscr{E}c\rangle = c^T\mathscr{E}^T\mathscr{E}c \le \rho^2(\mathscr{E})\|c\|^2.$$

The last step is due to $\rho(\mathscr{E}) \ge 1$ since $\mathscr{E} = I + \tilde{\mathscr{E}}$ where $\tilde{\mathscr{E}}$ is symmetric positive semi-definite.

Using the lower bound for the ASM from Lemma 4.7 we obtain

$$\begin{aligned} c_0^{-1}\langle x, x\rangle_A &\le \langle (\sum_{j=0}^{p} P_j)x, x\rangle_A \\ &\le (1 + 2\rho^2(\mathscr{E}))\sum_{k=0}^{p}\langle E_{k-1}^* P_k E_{k-1} x, x\rangle_A \\ &= (1 + 2\rho^2(\mathscr{E}))\langle (I - E_p^* E_p)x, x\rangle_A \qquad \text{(from 3.)} \end{aligned}$$

From this we get

$$\langle (I - E_p^* E_p)x, x\rangle_A \ge \frac{1}{c_0(1 + 2\rho^2(\mathscr{E}))}\langle x, x\rangle_A.$$

From this we conclude as in 3. with $1 - \alpha = 1 - \frac{1}{c_0(1+2\rho^2(\mathscr{E}))}$. □

Chapter 5

Convergence Theory for Overlapping Schwarz

This chapter follows the presentation in Toselli and Widlund [2005]. The goal is to verify the stable splitting assumption 4.1 for the two-level Schwarz method presented in the previous chapter. We will restrict ourselves to exact subdomain solvers. On the other hand, the theory will allow for the more general case where the fine mesh $\mathcal{T}_h$ is *not* a uniform refinement of the coarse mesh $\mathcal{T}_H$.

5.1 Technical Preliminaries

In order to verify the existence of a stable splitting (assumption 4.1) a specific splitting is constructed and then analyzed. The splitting is

$$u_h = \mathscr{I}^h \tilde{\mathscr{I}}^H u_h + \sum_{i=1}^{p} \mathscr{I}^h \theta_i (u_h - \mathscr{I}^h \tilde{\mathscr{I}}^H u_h)$$

where $\tilde{\mathscr{I}}^H : V_h \to V_H = V_{h,0}$ maps into the coarse grid space, $\mathscr{I}^h$ is the standard Lagrange interpolation operator[1] and θ_i is a partition of unity.

We recall the Friedrich and Poincaré inequalities.

Lemma 5.1 (Friedrich Inequality)**.** Suppose $\Omega \subset \mathbb{R}^d$ is a bounded domain with Lipschitz-continuous boundary[2] $\partial\Omega$ and $\Gamma \subset \partial\Omega$ has non-vanishing $(d-1)$-dimensional measure. Then for all $u \in H^1(\Omega)$

$$\|u\|_{0,\Omega}^2 \leq c_1 |u|_{1,\Omega}^2 + c_2 \|u\|_{0,\Gamma}^2$$

with constants c_1, c_2 only depending on Ω and Γ.

Proof. [Toselli and Widlund, 2005, A.14] □

Lemma 5.2 (Poincaré Inequality)**.** Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then we have for all $u \in H^1(\Omega)$

$$\|u\|_{0,\Omega}^2 \leq c_1 |u|_{1,\Omega}^2 + c_2 \left(\int_{\Omega} u \, \mathrm{d}x \right)^2$$

[1] Note that it is only applied to finite element functions — no regularity issue here.

[2] That is in the vicinity of a point on the boundary it can be expressed as the graph of a Lipschitz-continuous function.

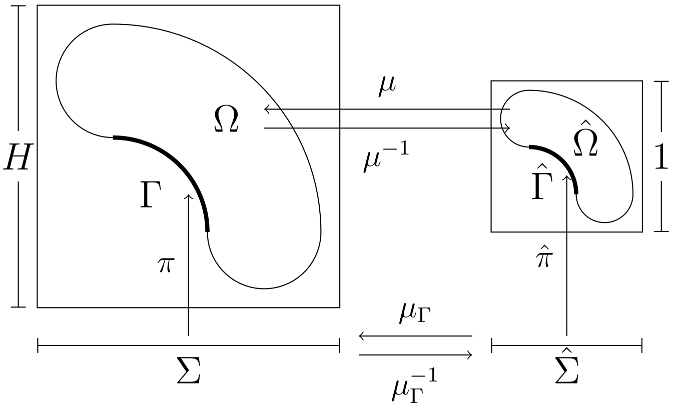


Figure 5.1: Scaling argument used in proof of Cor. 5.3

where c_1, c_2 only depend on Ω.

Note that for $\bar{u} := \int_\Omega u \, \mathrm{d}x$ we have

$$\|u - \bar{u}\|_{0,\Omega} \leq \sqrt{c_1}|u|_{1,\Omega}.$$

Moreover, the dependence of c_1, c_2 on Ω can be made more explicit. We use a scaling argument here to show this (but also a direct proof is possible).

Corollary 5.3. Let Ω be a bounded Lipschitz domain with diameter H. Then there exist constants $\hat{c}_1$ and $\hat{c}_2$ only depending on the shape of Ω (but not its diameter) such that

$$\|u\|_{0,\Omega}^2 \leq \hat{c}_1 H^2 |u|_{1,\Omega}^2 + \hat{c}_2 H \|u\|_{0,\Gamma}^2.$$

Proof. Consider the map $\mu(\hat{x}) = H\hat{x} + x_0$ such that Ω is contained in a box with side length H and origin x_0. The map $\mu^{-1}(x) = \frac{x - x_0}{H}$ maps Ω into the unit cell, see figure 5.1. We set $\hat{u}(\hat{x}) := u(\mu(\hat{x}))$, then

$$\begin{aligned} \frac{\partial \hat{u}}{\partial \hat{x}_i}(\hat{x}) &= \frac{\partial u(\mu(\hat{x}))}{\partial \hat{x}_i} = \frac{\partial u}{\partial x_i}(\mu(\hat{x}))H \\ \Longrightarrow \nabla_{\hat{x}} \hat{u} &= H \nabla_x u \\ \Longleftrightarrow \nabla_x u &= H^{-1} \nabla_{\hat{x}} \hat{u} \end{aligned}$$

and

$$\nabla \mu(\hat{x}) = HI \quad \Longleftrightarrow \quad \nabla \mu^{-1}(x) = H^{-1} I.$$

Now

$$\begin{aligned}
\|u\|^2_{0,\Omega} &= \int\limits_{\Omega} u^2(x)\,\mathrm{d}x = \int\limits_{\hat\Omega} u^2(\mu(\hat x))H^d\,\mathrm{d}\hat x = H^d\|\hat u\|^2_{0,\hat\Omega} \\
&\le H^d(\hat c_1|\hat u|^2_{1,\hat\Omega} + \hat c_2\|\hat u\|^2_{0,\hat\Gamma}) \\
&= H^d\left(\hat c_1\int\limits_{\hat\Omega}\nabla_{\hat x}\hat u\cdot\nabla_{\hat x}\hat u\,\mathrm{d}\hat x + \hat c_2\int\limits_{\hat\Gamma}\hat u^2(\hat s)\,\mathrm{d}\hat s\right) \\
&= H^d\left(\hat c_1\int\limits_{\hat\Omega}(H\nabla_x u(\mu(\hat x)))\cdot(H\nabla_x u(\mu(\hat x)))\,\mathrm{d}\hat x + \hat c_2\int\limits_{\hat\Gamma}u^2(\mu_\Gamma(\hat s))\,\mathrm{d}\hat s\right) \\
&= H^d\left(\hat c_1 H^2\int\limits_{\Omega}\nabla_x u(x)\cdot\nabla_x u(x)H^{-d}\,\mathrm{d}x + \hat c_2\int\limits_{\Gamma}u^2(s)H^{-(d-1)}\,\mathrm{d}s\right) \\
&= \hat c_1 H^2|u|^2_{1,\Omega} + c_2 H\|u\|^2_{0,\Gamma}.
\end{aligned}$$

For the second term write the surface integral as a volume integral:

$$\begin{aligned}
\int\limits_{\Gamma} f(s)\,\mathrm{d}s &= \int\limits_{\Sigma} f(\pi(\xi))\phi(\xi)\,\mathrm{d}\xi, &&\pi:\Sigma\to\Gamma,\quad \Sigma\subset\mathbb{R}^{d-1}, \\
\int\limits_{\hat\Gamma} \hat f(\hat s)\,\mathrm{d}\hat s &= \int\limits_{\hat\Sigma} \hat f(\hat\pi(\hat\xi))\hat\phi(\hat\xi)\,\mathrm{d}\hat\xi, &&\hat\pi:\hat\Sigma\to\hat\Gamma,\quad \hat\Sigma\subset\mathbb{R}^{d-1}.
\end{aligned}$$

Now $\operatorname{diam}(\Sigma) = \operatorname{diam}(\Gamma)$, $\operatorname{diam}(\hat\Sigma) = \operatorname{diam}(\hat\Gamma)$ and all the scaling is contained in the map $\mu_\Gamma : \hat\Sigma \to \Sigma$ with $\xi = \mu_\Gamma(\hat\xi) = H\hat\xi + \sigma_0$ and $\hat\xi, \xi \in \mathbb{R}^{d-1}$. As a result $\hat\phi(\hat\xi) = \phi(\mu_\Gamma(\hat\xi))$. □

5.2 Coarse Grid Contribution

The interpolation to the coarse grid space $V_H = V_{h,0}$ is defined as follows.

Definition 5.4 (Quasi-Interpolation Operator)**.** Let V_H be the P_1 or Q_1 finite element space on $\mathcal{T}_H$. Then define $\tilde{\mathscr{I}}_H : H^1_0(\Omega) \to V_H$ as

$$(\tilde{\mathscr{I}}_H u)(s_i) := \begin{cases} 0 & s_i \in \partial\Omega \\ |\omega_{s_i}|^{-1}\int_{\omega_{s_i}} u(x)\,\mathrm{d}x & \text{otherwise} \end{cases}$$

where s_i is a vertex in the mesh $\mathcal{T}_H$, $\bar\omega_{s_i}$ is the union of all elements $T \in \mathcal{T}_H$ having s_i as a vertex and $|w_{s_i}| = \int_{\omega_{s_i}} 1\,\mathrm{d}x$ is the volume of ω_{s_i}.

We now prove some important properties of the quasi-interpolation operator $\tilde{\mathscr{I}}_H$.

Lemma 5.5. Let $\mathcal{T}_H$ be a shape regular mesh. Then there exists $c > 0$ such that for all $T \in \mathcal{T}_H$

$$\|u - \tilde{\mathscr{I}}_H u\|_{0,T} \leq c H_T |u|_{1,\omega_T} \tag{5.1}$$

$$|\tilde{\mathscr{I}}_H u|_{1,T} \leq c |u|_{1,\omega_T} \tag{5.2}$$

where $\bar{\omega}_T$ is a union of elements including T itself such that

- $\bigcup_{T' \in \mathcal{T}_H, \bar{T}' \cap \bar{T} \neq \emptyset} \subseteq \bar{\omega}_T$,
- either $\partial\omega_T \cap \partial\Omega = \emptyset$ or $\partial\omega_T \cap \partial\Omega$ has non-vanishing $(d-1)$-dimensional measure and
- the number of elements making up ω_T is finite and independent of H.

Estimate (5.1) is an error estimate for the coarse space interpolation operator, while the estimate (5.2) provides the stability of the interpolation operator in the H^1-norm.

Proof. We restrict ourselves to P_1 finite elements in three dimensions (of course the argument can be transfered to other cases).

1. Let $\{\phi_i^H : i \in I_H\}$ be the Lagrange basis of V_H and let $i \in I_H$ be the index of any vertex of $T \in \mathcal{T}_H$. Then since $\phi_i(x) \leq 1$ we have

$$\|\phi_i^H\|_{0,T} = \int\limits_T |\phi_i^H(x)|^2 \,\mathrm{d}x \leq \int\limits_T 1 \,\mathrm{d}x \leq c\, H_T^3.$$

2. For $T \in \mathcal{T}_H$ let s_i be a vertex of T such that $s_i \notin \partial\Omega$. Then

$$\begin{aligned}
|(\tilde{\mathscr{I}}_H u)(s_i)| &= |\omega_{s_i}|^{-1} \,|\int\limits_{\omega_{s_i}} u(x)\,\mathrm{d}x| && \text{(Def. 5.4)}\\
&\leq |\omega_{s_i}|^{-1} \left(\int\limits_{\omega_{s_i}} |u(x)|^2 \,\mathrm{d}x\right)^{\frac{1}{2}} \left(\int\limits_{\omega_{s_i}} 1 \,\mathrm{d}x\right)^{\frac{1}{2}} && \text{(C.S.)}\\
&= \|u\|_{0,\omega_{s_i}} |\omega_{s_i}|^{-\frac{1}{2}}\\
&\leq \|u\|_{0,\omega_{s_i}} c H_T^{-\frac{3}{2}}.
\end{aligned}$$

For the last inequality we used the shape regularity of the mesh which implies $|\omega_{s_i}| \leq c H_T^3$ (we need to know the diameter of the neighbors).

3. Let A, B, C, D be the indices of the vertices of T. Then

$$\begin{aligned}
\|\tilde{\mathscr{I}}_H u\|_{0,T}^2 &= \| \sum_{i\in\{A,B,C,D\}} (\tilde{\mathscr{I}}_H u)(s_i)\phi_i^H \|_{0,T}^2 \\
&\leq 4 \sum_{i\in\{A,B,C,D\}} |(\tilde{\mathscr{I}}_H u)(s_i)|^2 \|\phi_i^H\|_{0,T}^2 \qquad \text{(triangle ineq.)} \\
&\leq 4 \sum_{i\in\{A,B,C,D\}} \|u\|_{0,\omega_{s_i}}^2 c\, H_T^{-3} H_T^3 \\
&\leq c\|u\|_{0,\omega_T}^2.
\end{aligned}$$

So we have shown the stability in the L^2-norm $\|\tilde{\mathscr{I}}_H u\|_{0,T} \leq c\|u\|_{0,\omega_T}$. Note that all constants are generic, i.e. c may have different values at different occurences.

In the first estimate we used in addition to the triangle inequality the estimate

$$\left(\sum_{i=1}^N a_i\right)^2 \leq 2^{\lceil \operatorname{ld} N\rceil} \sum_{i=1}^N a_i^2$$

for any numbers $a_i \in \mathbb{R}$. $2^{\lceil \operatorname{ld} N\rceil}$ is the smallest power of two greater or equal to N. Let us proof this. From the binomial formula $(a-b)^2 = a^2 - 2ab + b^2$ follows $2ab \leq a^2 + b^2$ and together with the other binomial formula we get the well-known estimate $(a+b)^2 = a^2 + 2ab + b^2 \leq 2a^2 + 2b^2$, i.e. we have proven $N = 2$. Now assume the estimate has been proven up to size $n \in \mathbb{N}$ and let $N = 2n - m$ where $m \in \{0, 1\}$. Then

$$\begin{aligned}
\left(\sum_{i=1}^N a_i\right)^2 &= \left(\sum_{i=1}^n a_i + \sum_{i=n+1}^N a_i\right)^2 \\
&\leq 2\left(\sum_{i=1}^n a_i\right)^2 + 2\left(\sum_{i=n+1}^N a_i\right)^2 \\
&\leq 2\cdot 2^{\lceil \operatorname{ld} n\rceil} \sum_{i=1}^n a_i^2 + 2\cdot 2^{\lceil \operatorname{ld}(n-m)\rceil} \sum_{i=n+1}^N a_i^2 \\
&\leq 2\cdot 2^{\lceil \operatorname{ld} n\rceil} \sum_{i=1}^N a_i^2 \leq 2^{\lceil \operatorname{ld} N\rceil} \sum_{i=1}^N a_i^2.
\end{aligned}$$

For the last estimate $2\cdot 2^{\lceil \operatorname{ld} n\rceil} \leq 2^{\lceil \operatorname{ld} N\rceil}$ consider first the case $N = 2n$. Then $N = 2n \iff \operatorname{ld} N = \operatorname{ld}(2n) = 1 + \operatorname{ld} n \iff \operatorname{ld} n = \operatorname{ld} N - 1$. Now

$2 \cdot 2^{\lceil \operatorname{ld} n \rceil} = 2^{1+\lceil \operatorname{ld} n \rceil} = 2^{1+\lceil \operatorname{ld} N \rceil - 1} = 2^{\lceil \operatorname{ld} N \rceil}$ where we used that $\lceil x + k \rceil = \lceil x \rceil + k$ for $k \in \mathbb{Z}$. Now consider the second case $N = 2n - 1 \iff 2n = N + 1$. Then $\operatorname{ld}(2n) = \operatorname{ld}(N+1) \iff 1 + \operatorname{ld} n = \operatorname{ld}(N+1) \iff \operatorname{ld} n = \operatorname{ld}(N+1) - 1$. Taking the ceiling of the last equality gives $\lceil \operatorname{ld} n \rceil = \lceil \operatorname{ld}(N+1) - 1 \rceil = \lceil \operatorname{ld}(N+1) \rceil - 1 = \lceil \operatorname{ld} N \rceil - 1$. The last step follows from the fact that $\operatorname{ld} N$ is not integer since N is odd and therefore not a power of two. Now we can conclude $2 \cdot 2^{\lceil \operatorname{ld} n \rceil} = 2^{\lceil \operatorname{ld} N \rceil}$ as in the case N even.

Finally observe that all natural numbers $N \geq 2$ can be generated by a unique sequence of doubling or doubling and subtracting one.

4. First consider the case $\partial\omega_T \cap \partial\Omega = \emptyset$, i.e. an *interior subdomain.* Set $\hat{u}(x) = u(x) - |\omega_T|^{-1} \int_{\omega_T} u \,\mathrm{d}x$ on ω_T. Now we have

$$\begin{aligned}
\|u - \tilde{\mathscr{I}}_H u\|_{0,T}^2 &= \|u - |\omega_T|^{-1} \int\limits_{\omega_T} u \,\mathrm{d}x + |\omega_T|^{-1} \int\limits_{\omega_T} u \,\mathrm{d}x - \tilde{\mathscr{I}}_H u\|_{0,T}^2 \\
&= \|\hat{u} - \tilde{\mathscr{I}}_H \hat{u}\|_{0,T}^2 \qquad (*) \\
&\leq (\|\hat{u}\|_{0,T} + \|\tilde{\mathscr{I}}_H \hat{u}\|_{0,T})^2 \\
&\overset{(3.)}{\leq} c\|\hat{u}\|_{0,\omega_T}^2 \\
&\leq cH_T^2 |\hat{u}|_{1,\omega_T}^2 \\
&= cH_T^2 |u|_{1,\omega_T}^2.
\end{aligned}$$

For (*) we used that $\tilde{\mathscr{I}}_H$ reproduces constants as long as we are away from $\partial\Omega$. For the last equality note that the 1-semi-norm of a constant is 0.

Now to the case $\partial\omega_T \cap \partial\Omega \neq \emptyset$. Through enlargement of ω_T, $\Gamma = \partial\omega_T \cap \partial\Omega$ has non-zero $(d-1)$-dimensional measure. By construction $\tilde{\mathscr{I}}_H u$ is zero on Γ for $u \in H_0^1(\Omega)$. Therefore

$$\|u - \tilde{\mathscr{I}}_H u\|_{0,T}^2 \leq c\|u\|_{0,\omega_T}^2 \leq cH_T^2 |u|_{1,\omega_T}^2.$$

Here we used the Friedrich inequality.

5. Now we are in the position to prove the estimate eq. (5.2). Here we need to use a so-called *local inverse inequality.* It states that for a finite element function $v \in V_H$ there exists a constant c depending only on the mesh such that for all $T \in \mathcal{T}_H$

$$|v|_{1,T} \leq cH_T^{-1} \|v\|_{0,T}.$$

Now consider first again the case $\partial\omega_T \cap \partial\Omega = \emptyset$:

$$\begin{aligned}
|\tilde{\mathscr{I}}_H u|_{1,T}^2 &= |\tilde{\mathscr{I}}_H u - |\omega_T|^{-1} \int\limits_{\omega_T} u \,\mathrm{d}x|_{1,T}^2 && \text{(constant has zero seminorm)}\\
&= |\tilde{\mathscr{I}}_H \hat{u}|_{1,T}^2 \\
&\le cH_T^{-2} \|\tilde{\mathscr{I}}_H \hat{u}\|_{0,T}^2 && \text{(inverse ineq.)}\\
&= cH_T^{-2} \|\tilde{\mathscr{I}}_H \hat{u} - \hat{u} + \hat{u}\|_{0,T}^2 \\
&\le cH_T^{-2} (\|\hat{u} - \tilde{\mathscr{I}}_H \hat{u}\|_{0,T} + \|\hat{u}\|_{0,T})^2 && \text{(triangle ineq.)}\\
&\le cH_T^{-2} (c' H_T^2 |\hat{u}|_{1,\omega_T}^2 + c'' H_T^2 |\hat{u}|_{1,\omega_T}^2) && \text{((5.1) and Poincaré)}\\
&\le c|u|_{1,\omega_T}^2 && (|\hat{u}|_{1,\omega_T} = |u|_{1,\omega_T})
\end{aligned}$$

And for the boundary case $\partial\omega_T \cap \partial\Omega \neq \emptyset$ we get

$$\begin{aligned}
|\tilde{\mathscr{I}}_H u|_{1,T}^2 &\le cH_T^{-2} \|\tilde{\mathscr{I}}_H u\|_{0,T}^2 && \text{(inverse ineq.)}\\
&\le cH_T^{-2} \|u\|_{0,\omega_T}^2 \\
&\le cH_T^{-2} H_T^2 |u|_{1,\omega_T}^2 && \text{(Friedrich, shape reg.)}\\
&\le c|u|_{1,\omega_T}^2 .
\end{aligned}$$

□

Now let $\{\phi_i^h : i \in I_h\}$ be the Lagrange basis for V_h. For $u \in C^0(\bar{\Omega})$ we define the Lagrange interpolation operator as usual:

$$\mathscr{I}_h u := \sum_{i \in I_h} u(s_i)\phi_i^h$$

where s_i is the Lagrange point associated with the basis function ϕ_i.

The next lemma investigates the fine grid interpolation operator applied to coarse grid functions.

Lemma 5.6. There exists $c > 0$ independent of h and H such that

$$|u_H - \mathscr{I}_h u_H|_{s,t}^2 \le c h_t^{2(1-s)} |u_H|_{1,\omega_t}^2 \tag{5.3}$$

for all $t \in \mathcal{T}_h, u_H \in V_H$ and $s \in \{0, 1\}$. Note that $|v|_{0,t} = \|v\|_{0,t}$.

Proof. 1. Consider $t \in \mathcal{T}_h$ such that t is *completely inside* some $T \in \mathcal{T}_H$.

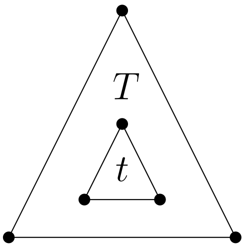


Figure 5.2: $t \in \mathcal{T}_h$ such that t is completely inside some $T \in \mathcal{T}_H$

Then the Lagrange interpolation on t is exact, i.e.

$$u_H - \mathscr{I}_h u_H = 0.$$

If $V_H \subseteq V_h$ (which is true when $\mathcal{T}_h$ is a refinement of $\mathcal{T}_H$) then the proof of (5.3) is complete. The remaining part is only necessary when $V_H \nsubseteq V_h$.

We consider again the case of P_1 in 3D, i.e. tetrahedral elements.

2. For $\phi_i^h \in V_h$ we have

$$|\phi_i^h|_{1,t}^2 = \int_t \sum_{j=1}^d (\partial_j \phi_i^h)^2 \,\mathrm{d}x \leq c h_t^{-2} h_t^3 = c h_t$$

(needs shapre regularity $\rho \geq c h_t$).

3. With that we get

$$\begin{aligned} |\mathscr{I}_h u_H|_{1,t}^2 &= |\sum_{i \in \{A,B,C,D\}} u_H(s_i) \phi_i^h|_{1,t}^2 \\ &\leq c \sum_{i \in \{A,B,C,D\}} |u_H(s_i)|^2 |\phi_i^h|_{1,t}^2 \quad \text{(as before } (\sum_{i=1}^N a_i)^2 \leq N \sum_{i=1}^N a_i^2\text{)} \\ &\leq c h_t \sum_{i \in \{A,B,C,D\}} |u_H(s_i)|^2. \end{aligned}$$

Now we need to estimate $|u_H(s_i)|$.

4. Consider $t \in \mathcal{T}_h$ with corner indices $i \in \{A, B, C, D\}$. Then for any pair (t, i) we can find a tetrahedron t_i', not necessarily an element of $\mathcal{T}_h$ (!), such that:

   - t_i' has s_i (i.e. one of the vertices of tetrahedron t) as one of its corners.
   - t_i' has diameter $h_{t_i'} \leq c h_t$ and $\rho_{t_i'} \geq c' h_t$ (shape regularity).
   - $t_i' \subseteq \omega_t$
   - $t_i' \subseteq T_i \in \mathcal{T}_H$ (i.e. it is completely inside one $T \in \mathcal{T}_H$).

   This is due to the shape regularity of $\mathcal{T}_H$.

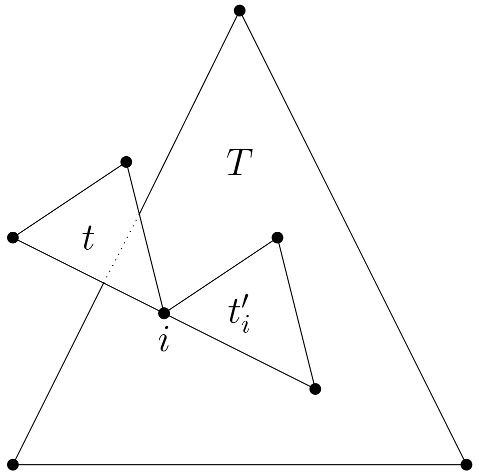


Figure 5.3: tetrahedron t_i'

Without loss of generality consider now one corner $i = A$ of $t \in \mathcal{T}_h$. Then choose t_A' with properties as described above and vertex positions $\{x_A, x_{B'}, x_{C'}, x_{D'}\}$. Let μ be the map that transforms the reference tetrahedron $\hat{t}$ to t_A' and set $\hat{\mu}(\hat{x}) := u_H(\mu(\hat{x}))$ as usual. Then, on the reference element, we have with $x_A = \mu(\hat{x}_A)$

$$\begin{aligned}
|u_H(x_A)| = |\hat{u}(\hat{x}_A)| &\leq \sum_{i\in\{A,B,C,D\}} |\hat{u}(\hat{x}_i)| \\
&= c\frac{4}{|\hat{t}|}\left(|\hat{t}| \sum_{i\in\{A,B,C,D\}} \frac{1}{4}|\hat{u}(\hat{x}_i)|\right) && |\hat{t}| = \int_{\hat{t}} 1\,\mathrm{d}x \\
&= c\int_{\hat{t}} |\hat{u}(\hat{x})|\,\mathrm{d}\hat{x} && \text{(quadrature rule)} \\
&\leq c\left(\int_{\hat{t}} |\hat{u}(\hat{x})|^2\,\mathrm{d}\hat{x}\right)^{\frac{1}{2}}\left(\int_{\hat{t}} 1\,\mathrm{d}\hat{x}\right)^{\frac{1}{2}} && \text{(Cauchy-Schwarz)} \\
&\leq c\|\hat{u}\|_{0,\hat{t}}.
\end{aligned}$$

Now use a scaling argument to transform to the real element:

$$\begin{aligned}
\|\hat{u}\|_{0,\hat{t}}^2 = \int_{\hat{t}} |\hat{u}(\hat{x})|^2\,\mathrm{d}\hat{x} &= \int_{t_i'} |\underbrace{\hat{u}(\mu^{-1}(x))}_{=u_H(x)}|^2 \det B^{-1}\; dx \\
&\leq ch_{t_i'}^{-3}\int_{t_i'} |u_H(x)|^2\,\mathrm{d}x \leq ch_t^{-3}\|u_H\|_{0,t_i'}^2 && \text{(assumption on } t_i'\text{)} \\
&\leq ch_t^{-3}\|u_H\|_{0,\omega_t}^2. && \text{(enlarge domain of int.)}
\end{aligned}$$

5. Going now back to 3. we have

$$\begin{aligned}
|\mathscr{I}_h u_H|_{1,t}^2 &\le ch_t \sum_{i\in\{A,B,C,D\}} |u_H(s_i)|^2 && \text{(result from 3.)}\\
&\le ch_t \sum_{i\in\{A,B,C,D\}} h_t^{-3}\|u_H\|_{0,\omega_t}^2\\
&= ch_t^{-2}\|u_H\|_{0,\omega_t}^2
\end{aligned}$$

6. Now again we use either the Friedrich or Poincaré inequality.

   In the case that $\partial\omega_t \cap \partial\Omega = \Gamma$ has non-zero measure, we have

$$\begin{aligned}
|u_H - \mathscr{I}_h u_H|_{1,t}^2 &\le 2|u_H|_{1,t}^2 + 2|\mathscr{I}_h u_H|_{1,t}^2\\
&\le 2|u_H|_{1,t}^2 + ch_t^{-2}\|u_H\|_{0,\omega_t}^2 && \text{(using 5.)}\\
&\le 2|u_H|_{1,t}^2 + ch_t^{-2}h_t^2|u_H|_{1,\omega_t}^2 && \text{(Friedrich)}\\
&\le c|u_H|_{1,\omega_t}^2
\end{aligned}$$

   For the case $\partial\omega_t \cap \partial\Omega = \emptyset$ the operator $\mathscr{I}_h$ reproduces constants and we set $\bar{u}_H$ to the average of u_H on ω_t. Then

$$\begin{aligned}
|u_H - \mathscr{I}_h u_H|_{1,t}^2 &= |u_H + \mathscr{I}_h(\bar{u}_H - u_H)|_{1,t}^2 && \text{(since } \nabla\bar{u}_H = 0)\\
&\le 2|u_H|_{1,t}^2 + 2|\mathscr{I}_h(\bar{u}_H - u_H)|_{1,t}^2\\
&\le 2|u_H|_{1,t}^2 + ch_t^{-2}\|\bar{u}_H - u_H\|_{0,\omega_t}^2 && \text{(using 5.)}\\
&\le 2|u_H|_{1,t}^2 + ch_t^{-2}h_t^2|\bar{u}_H - u_H|_{1,\omega_t}^2 && \text{(Poincaré ineq.)}\\
&\le c|u_H|_{1,\omega_t}^2 && (\nabla\bar{u}_H = 0)
\end{aligned}$$

   This proves (5.3) for the case $s = 1$.

7. Now the case $s = 0$ in (5.3). We use the L_2-norm in 3.:

$$\begin{aligned}
\|\mathscr{I}_h u_H\|_{0,t}^2 &= \|\sum_{i\in\{A,B,C,D\}} u_H(s_i)\phi_i^h\|_{0,t}^2\\
&\le c \sum_{i\in\{A,B,C,D\}} |u_H(s_i)|^2\|\phi_i^h\|_{0,t}^2\\
&\le ch_t^3 \sum_{i\in\{A,B,C,D\}} |u_H(s_i)|^2\\
&\le c\|u_H\|_{0,\omega_t}^2 && \text{(using 4.)}
\end{aligned}$$

   Now we can proceed as in 6.

First we consider $\partial\omega_t \cap \partial\Omega = \Gamma$ with non-zero measure:

$$\begin{aligned}
\|u_H - \mathscr{I}_h u_H\|_{0,t}^2 &\le 2\|u_H\|_{0,t}^2 + 2\|\mathscr{I}_h u_H\|_{0,t}^2 \\
&\le c\|u_H\|_{0,\omega_t}^2 && \text{(7. and enlargement)} \\
&\le ch_t^2 |u_H|_{1,t}^2 && \text{(Friedrich ineq.)}
\end{aligned}$$

In the case $\partial\omega_t \cap \partial\Omega = \emptyset$ we have

$$\begin{aligned}
\|u_H - \mathscr{I}_h u_H\|_{0,t}^2 \le\ & 2\|u_H - \bar{u}_H\|_{0,t}^2 \\
&+ 2\|\mathscr{I}_h(\bar{u}_H - u_H)\|_{0,t}^2 \\
\le\ & c\|u_H - \bar{u}_H\|_{0,\omega_t}^2 && \text{(use 7., enlarge first, combine)} \\
\le\ & ch_t^2 |u_H - \bar{u}_H|_{1,\omega_t}^2 && \text{(Poincaré ineq.)} \\
=\ & ch_t^2 |u_H|_{1,\omega_t}^2 && (\nabla \bar{u}_H = 0)
\end{aligned}$$

This ends the proof of Lemma 5.6. □

5.3 Localization to the Subdomains

For $u_h, v_h \in P_1(\mathcal{T}_h)$ the product function $u_h v_h$ is a piecewise quadratic finite element function.

Lemma 5.7. Assume u_h is a P_2 finite element function and $\mathscr{I}_h$ is the Lagrange-interpolation operator into P_1 on the same grid. Then there exists $c > 0$ independent of h such that

$$|\mathscr{I}_h u_h|_{1,t} \le c|u_h|_{1,t} \quad \forall t \in \mathcal{T}_h.$$

Proof.

$$\begin{aligned}
|\mathscr{I}_h u_h|_{1,t}^2 &= |\mathscr{I}_h u_h - u_h + u_h|_{1,t}^2 \\
&\le 2|u_h - \mathscr{I}_h u_h|_{1,t}^2 + 2|u_h|_{1,t}^2 \\
&\le ch_t^2 |u_h|_{2,t}^2 + 2|u_h|_{1,t}^2 && \text{(approximation error)} \\
&\le ch_t^2 h_t^{-2} |u_h|_{1,t}^2 + 2|u_h|_{1,t}^2 && \text{(inverse ineq., } u_h \in P_2\text{)} \\
&\le c|u_h|_{1,t}^2
\end{aligned}$$

□

Next we look in detail at the L_2-norm of a function in the vicinity of a subdomain boundary.

Let $\hat{\Omega}_i$ denote one of the (overlapping) subdomains, δ_i the overlap of this subdomain and $H_i = \operatorname{diam}(\hat{\Omega}_i)$. Then set

$$\hat{\Omega}_{i,\delta_i} := \{x \in \hat{\Omega}_i : \operatorname{dist}(x, \partial\hat{\Omega}_i \backslash \partial\Omega) < \delta_i\}.$$

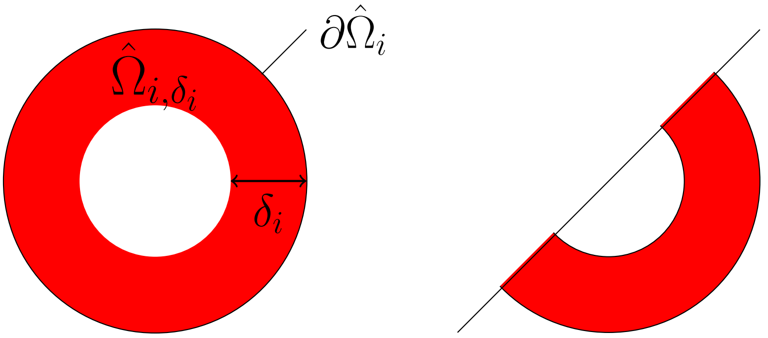


Figure 5.4: Interior (left) and boundary subdomain (right)

Lemma 5.8. There exists $c > 0$ such that for all $u \in H^1(\hat{\Omega}_i)$:

$$\|u\|^2_{0,\hat{\Omega}_{i,\delta_i}} \leq c\delta_i^2 \left((1 + \frac{H_i}{\delta_i})|u|^2_{1,\hat{\Omega}_i} + \frac{1}{H_i\delta_i}\|u\|^2_{0,\hat{\Omega}_i} \right).$$

Proof. 1. According to the trace theorem 1.15 there exists

$$\gamma : H^1(\omega) \to L^2(\partial\omega)$$

such that $\|\gamma u\|_{0,\partial\omega} \leq c\|u\|_{1,\omega}$. Through a scaling argument we make explicit the dependence on the size of the domain (refer to corollary 5.3):

$$\begin{aligned}
\|u\|^2_{0,\partial\omega} &= \int\limits_{\partial\omega} |u(s)|^2 \, \mathrm{d}s && (\hat{\omega} \text{ is of size } 1) \\
&= \int\limits_{\partial\hat{\omega}} |\hat{u}(\hat{s})|^2 H^{d-1} \, \mathrm{d}\hat{s} && (H = \mathrm{diam}(\omega)) \\
&\leq \hat{c}H^{d-1}\|\hat{u}\|^2_{1,\hat{\omega}} && (\text{trace thm. on } \hat{\omega}) \\
&= \hat{c}H^{d-1}\left(\int\limits_{\hat{\omega}} \hat{u}^2 \, \mathrm{d}\hat{x} + \int\limits_{\hat{\omega}} \nabla_{\hat{x}}\hat{u} \cdot \nabla_{\hat{x}}\hat{u} \, \mathrm{d}\hat{x} \right) \\
&= \hat{c}H^{d-1}\left(\int\limits_{\omega} u^2 H^{-d} \, \mathrm{d}x + \int\limits_{\omega} H^2 \nabla_x u \cdot \nabla_x u H^{-d} \, \mathrm{d}x \right) \\
&= \hat{c}\left(\frac{1}{H}\|u\|^2_{0,\omega} + H|u|^2_{1,\omega} \right)
\end{aligned}$$

i.e. $\|u\|^2_{0,\partial\omega} \leq \frac{\hat{c}}{H}\|u\|^2_{0,\omega} + H\hat{c}|u|^2_{1,\omega}$.

2. Now assume there exists a triangulation $\mathcal{T}_{i,\delta_i}$ of $\hat{\Omega}_{i,\delta_i}$ into shape regular patches of size δ_i:

$$\bar{\hat{\Omega}}_{i,\delta_i} = \bigcup_{k \in \mathcal{T}_{i,\delta_i}} \bar{k}.$$

Note that the patches don't need to be simplices or cubes since we will not construct finite element functions.

Moreover set $\Gamma_i = \partial\hat{\Omega}_i \backslash \partial\Omega$. Then

$$\begin{aligned}
\|u\|^2_{0,\hat{\Omega}_{i,\delta_i}} &= \sum_{k\in\mathcal{T}_{i,\delta_i}} \|u\|^2_{0,k} \\
&\le \sum_{k\in\mathcal{T}_{i,\delta_i}} c\left(\delta_i^2 |u|^2_{1,k} + \delta_i \|u\|^2_{0,\partial k\cap\Gamma_i}\right) && \text{(Friedrich ineq.)} \\
&= c\left(\delta_i^2 |u|^2_{1,\hat{\Omega}_{i,\delta_i}} + \delta_i \|u\|^2_{0,\partial\hat{\Omega}_i}\right) && \text{(enlarge } \Gamma_i \to \partial\hat{\Omega}_i) \\
&\le c\left(\delta_i^2 |u|^2_{1,\hat{\Omega}_{i,\delta_i}} + \delta_i \hat{c}\left(\frac{1}{H_i}\|u\|^2_{0,\hat{\Omega}_i} + H_i |u|^2_{1,\hat{\Omega}_i}\right)\right) && \text{(using 1.)} \\
&\le c\delta_i^2 \left(\left(1 + \frac{H_i}{\delta_i}\right) |u|^2_{1,\hat{\Omega}_{i,\delta_i}} + \frac{1}{H_i\delta_i}\|u\|^2_{0,\hat{\Omega}_i}\right)
\end{aligned}$$

□

Next we need two assumptions that let us prove properties of the partition of unity.

Assumption 5.9 (Minimal Distance)**.** Let $\{\hat{\Omega}_i\}_{i=1}^p$ be a decomposition of Ω into overlapping subdomains. Then we require that for $i \in \{1, ..., p\}$ exists $\delta_i > 0$ such that

$$\forall x \in \hat{\Omega}_i \quad \exists j(x) \in \{1, \ldots, p\} : \quad x \in \hat{\Omega}_{j(x)} \wedge \operatorname{dist}(x, \partial\hat{\Omega}_{j(x)} \backslash \partial\Omega) \ge \delta_i.$$

This means that $x \in \hat{\Omega}_i$ is away from the (interior) boundary in at least one subdomain. (Note: $j(x) = i$ is possible).

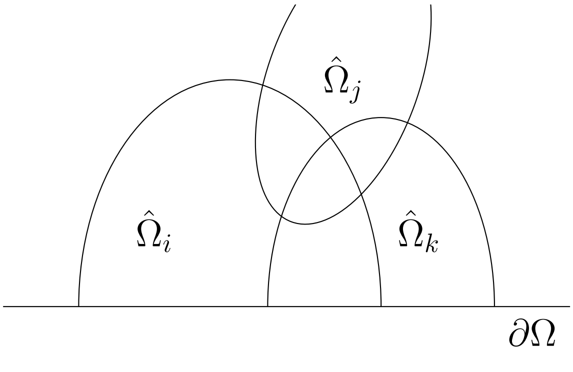


Figure 5.5: Example

Assumption 5.10 (Finite Covering)**.** For $\{\hat{\Omega}_i\}_{i=1}^p$ exists a coloring with at most N^c colors in the following sense:

$$c(i) = c(j) \quad \Longrightarrow \quad \hat{\Omega}_i \cap \hat{\Omega}_j = \emptyset$$

where $c : \{1, \dots, p\} \to \{1, \dots, N^c\}$ is the coloring map.

From Assumption 5.10 we can deduce that any $x \in \Omega$ is contained in at most N^c subdomains:
Set $J_x := \{j \in 1, \dots, p : x \in \hat{\Omega}_j\}$. Then for all $i, j \in J_x$ we have $x \in \hat{\Omega}_i \cap \hat{\Omega}_j$, i.e. $\hat{\Omega}_i \cap \hat{\Omega}_j \neq \emptyset \quad \Longrightarrow \quad c(i) \neq c(j)$.
Since the number of colors is N^c we have $|J_x| \leq N^c$.

Note also that assumption 5.10 does *not* bound the number of neighbors of a single subdomain.

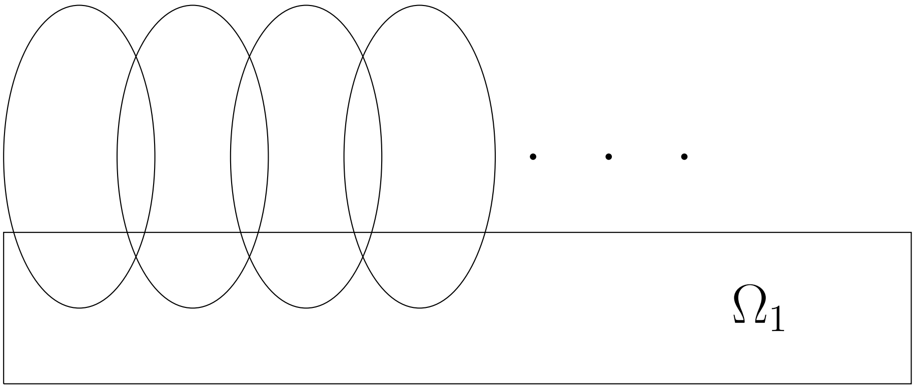


Figure 5.6: Only 3 colors are needed but Ω_1 overlaps with all subdomains

Finally we need some properties of the partition of unity.

Lemma 5.11 (Partition of Unity). Let $\{\hat{\Omega}_i\}_{i=1}^p$ a decomposition of Ω into overlapping subdomains such that assumptions 5.9 and 5.10 are satisfied. Then there exist functions $\{\tilde{\theta}_i\}_{i=1}^p$ from $W^{1,\infty}(\Omega)$ such that

(a) $0 \leq \tilde{\theta}_i(x) \leq 1, \quad x \in \bar{\Omega}$,

(b) $\operatorname{supp}(\tilde{\theta}_i) = \{x \in \Omega : \tilde{\theta}_i(x) \neq 0\} \subset \overline{\hat{\Omega}}_i$,

(c) $\sum_{i=1}^p \tilde{\theta}_i(x) = 1 \quad \forall x \in \bar{\Omega}$

(d) $\|\nabla \tilde{\theta}_i\|_\infty \leq \frac{c}{\delta_i}$ or rather $\|\nabla \tilde{\theta}_i(x)\|_\infty \leq \frac{c}{\delta_i} \quad \forall i = 1, \dots, p$.

$W^{1,\infty}(\Omega)$ is the space of functions with bounded derivatives almost everywhere (i.e. up to a set of zero measure).

Proof. [Toselli and Widlund [2005], Lemma 3.4] The proof is constructive.
For all $i \in \{1, \dots, p\}$ set

$$d_i(x) := \begin{cases} \operatorname{dist}(x, \partial\hat{\Omega}_i \backslash \partial\Omega) & x \in \hat{\Omega}_i \cup \{\partial\hat{\Omega}_i \cap \partial\Omega\} \\ 0 & \text{else} \end{cases}$$

and

$$\tilde{\theta}_i(x) := \frac{d_i(x)}{\sum_{k=1}^{p} d_k(x)}.$$

$d_i(x)$ and $\tilde{\theta}_i$ are well defined on $\bar{\Omega}$ and $\tilde{\theta}_i \in C^0(\bar{\Omega})$.

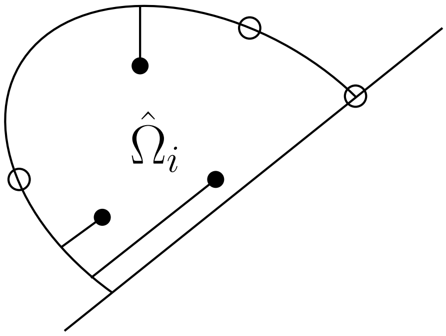


Figure 5.7: Example of $d_i(x)$

Clearly,

- $\tilde{\theta}_i \geq 0$ since $d_i \geq 0$
- $\sum_{i=1}^{p} \tilde{\theta}_i(x) = \sum_{i=1}^{p} \frac{d_i(x)}{\sum_{k=1}^{p} d_k(x)} = 1$
- $\tilde{\theta}_i(x) = 1 - \sum_{k \neq i} \tilde{\theta}_k(x) \leq 1$ since $\tilde{\theta}_k(x) \geq 0$.

So the only difficult property to prove is (d) which we do in several steps.

1. We aim to show that $\tilde{\theta}_i$ is Lipschitz continous, i.e.

$$|\tilde{\theta}_i(x) - \tilde{\theta}_i(y)| \leq \frac{c}{\delta_i}|x-y|$$

provided $|x-y|$ is sufficiently small. This would bound the gradient since $|\partial_j \tilde{\theta}_i(x)| \leq \frac{c}{\delta_i}$.

2. Abbreviate for ease of writing $\delta_k(x,y) := d_k(x) - d_k(y)$ for $1 \leq k \leq p$. Now we show

$$|\delta_k(x,y)| \leq |x-y|.$$

Case I $x, y \in \hat{\Omega}_k$. Then choose $z \in \partial\hat{\Omega}_k \backslash \partial\Omega$ such that

$$d_k(y) = \operatorname{dist}(y, \partial\hat{\Omega}_k \backslash \partial\Omega) = |y-z|$$

i.e. z is a closest point to the boundary for y.

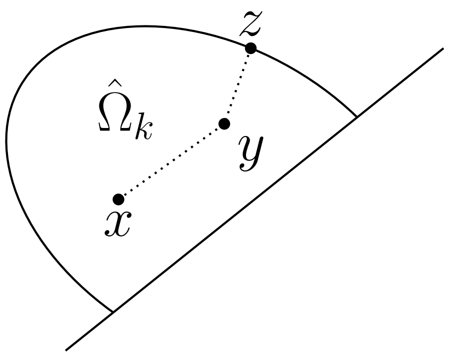


Figure 5.8: Case I in proof of Lemma 5.11

Then

$$\begin{aligned} d_k(x) &= \operatorname{dist}(x, \partial\hat{\Omega}_k \backslash \partial\Omega) \\ &\le |x - z| && (\text{since } z \in \partial\hat{\Omega}_k \backslash \partial\Omega) \\ &= |x - y + y - z| \\ &\le |x - y| + |y - z| \\ &= |x - y| + d_k(y) \end{aligned}$$

$$\iff \delta_k(x, y) = d_k(x) - d_k(y) \le |x - y|$$

In the same way choose $z' \in \partial\hat{\Omega}_k \backslash \partial\Omega$ such that

$$d_k(x) = \operatorname{dist}(x, \partial\hat{\Omega}_k \backslash \partial\Omega) = |x - z'|.$$

The same argument shows

$$\begin{aligned} d_k(y) &= \operatorname{dist}(y, \partial\hat{\Omega}_k \backslash \partial\Omega) \\ &\le |y - z'| \\ &= |y - x + x - z'| \\ &\le |y - x| + d_k(x) \end{aligned}$$

$$\iff -\delta_k(x, y) = d_k(y) - d_k(x) \le |y - x|$$

and therefore $|\delta_k(x, y)| \le |x - y|$.

Case II $x, y \notin \overline{\hat{\Omega}}_k$. i.e. x, y are outside $\hat{\Omega}_k$. Then we have

$$|d_k(x) - d_k(y)| = |0 - 0| = 0 \le |x - y|.$$

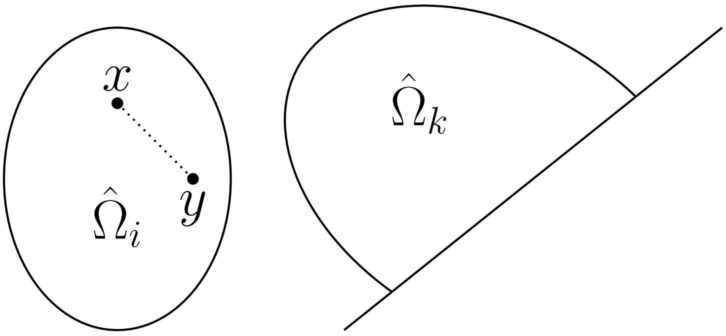


Figure 5.9: Case II in proof of Lemma 5.11

Case III $x \in \partial\hat{\Omega}_k$. We need to consider two subcases.

a) $x \in \partial\hat{\Omega}_k \backslash \partial\Omega$. Let $y \in \hat{\Omega}_k$.

1.) $z \in \partial\hat{\Omega}_k \backslash \partial\Omega$ such that $\operatorname{dist}(y, \partial\hat{\Omega}_k \backslash \partial\Omega) = |y - z|$

$$d_k(x) = 0 \le |x - z| = |x - y + y - z| \le |x - y| + d_k(y)$$

$$\iff -d_k(y) \le |x - y|$$

2.) $d_k(y) = |y - z| \le |y - x|$ since z was the closest point on the boundary from y.

Since $d_k(x) = 0$ we get from 2(III)a1 and 2(III)a2:

$$|\delta_k(x, y)| = |d_k(x) - d_k(y)| = |d_k(y)| \le |x - y|.$$

b) $x \in \partial\hat{\Omega}_k \cap \partial\Omega$: This is very similar to 2I above. Let $y \in \hat{\Omega}_k$.

1.) And $z \in \partial\hat{\Omega}_k \backslash \partial\Omega$ such that $d_k(y) = |y - z|$:

$$d_k(x) \le |x - z| \le |x - y| + |y - z| = |x - y| + d_k(y)$$

$$\iff \delta_k(x, y) \le |x - y|.$$

2.) And $z' \in \partial\hat{\Omega}_k \backslash \partial\Omega$ such that $d_k(x) = |x - z'|$:

$$d_k(y) \le |y - z'| \le |y - x| + |x - z'| = |y - x| + d_k(x)$$

$$\iff -\delta_k(x, y) \le |x - y|.$$

3. From the minimal distance assumption 5.9 we get

$$\sum_{k=1}^{p} d_k(x) \ge d_{j(x)}(x) \ge \delta_i \qquad \text{(one distance is at least } \delta_i\text{)}.$$

4. Now we are able to estimate

$$\begin{aligned}
|\tilde{\theta}_i(x) - \tilde{\theta}_i(y)| &= \left| \frac{d_i(x)}{\sum_{k=1}^p d_k(x)} - \frac{d_i(y)}{\sum_{k=1}^p d_k(y)} \right| && \text{(insert definition)} \\
&= \left| \frac{d_i(x) \sum_{k=1}^p d_k(y) - d_i(y) \sum_{k=1}^p d_k(x)}{\left(\sum_{k=1}^p d_k(x)\right) \left(\sum_{k=1}^p d_k(y)\right)} \right| && \text{(common denominator)} \\
&= \left| \frac{d_i(x) \sum_{k\neq i}^p d_k(y) - d_i(y) \sum_{k\neq i}^p d_k(x)}{\left(\sum_{k=1}^p d_k(x)\right) \left(\sum_{k=1}^p d_k(y)\right)} \right| && \text{($d_i(x)d_i(y)$ drops out)}
\end{aligned}$$

$$\begin{aligned}
&= \left| \frac{d_i(x) \sum_{k\neq i}^p d_k(y) \overbrace{- d_i(x) \sum_{k\neq i}^p d_k(x) + d_i(x) \sum_{k\neq i}^p d_k(x)}^{0} - d_i(y) \sum_{k\neq i}^p d_k(x)}{\left(\sum_{k=1}^p d_k(x)\right) \left(\sum_{k=1}^p d_k(y)\right)} \right| \\
&= \Bigg(\underbrace{\sum_{k=1}^p d_k(y)}_{>0} \Bigg)^{-1} \left| \underbrace{\frac{d_i(x)}{\sum_{k=1}^p d_k(x)}}_{\tilde{\theta}_i(x)} \sum_{k\neq i}^p \underbrace{d_k(y) - d_k(x)}_{\delta_k(x,y)} + \underbrace{\frac{\sum_{k\neq i}^p d_k(x)}{\sum_{k=1}^p d_k(x)}}_{1-\tilde{\theta}_i(x)} \underbrace{(d_i(x) - d_i(y))}_{\delta_i(x,y)} \right| \\
&\leq \underbrace{\frac{1}{\delta_i}}_{\text{(use 3.)}} \left[\underbrace{\tilde{\theta}_i(x)}_{\leq 1} \underbrace{\sum_{k\neq i}^p |\delta_k(x,y)|}_{\leq N^c |x-y|} + \underbrace{\left(1 - \tilde{\theta}_i(x)\right)}_{\leq 1} \underbrace{|\delta_i(x,y)|}_{\leq |x-y|} \right] \\
&\leq \frac{N^c + 1}{\delta_i} |x - y|
\end{aligned}$$

□

5.4 Proof of Assumptions of Abstract Schwarz Theory

First consider the strengthened Cauchy-Schwarz inequaltiy 4.2. This enters into the upper bound for $\kappa(P_{ad})$ as well as in the bound for $\|E_{mu}\|_A$. In the latter the spectral radius $\rho(\mathscr{E})$ is required. Remark 4.6 shows that

$$\rho(\mathscr{E}) \leq \|\mathscr{E}\|_\infty = \hat{N}^c$$

where $\hat{N}^c$ is the maximum number of subdomains with wich a single subdomain overlaps (including itself).

This is generally not the same as N^c from assumption 5.10 which allows a direct estimate on the upper bound of the additive Operator.

Lemma 5.12. Let assumption 5.10 hold. Then

$$\langle \sum_{i=0}^{p} P_i x, x\rangle_A \leq (N^c + 1)\,\langle x, x\rangle_A.$$

Proof. Set $J_k := \{i \in \{1, \ldots, p\} : c(i) = k\}$ for $1 \leq k \leq N^c$.

$$\begin{aligned}
\langle \sum_{i=0}^{p} P_i x, x\rangle_A &= \langle P_0 x, x\rangle_A + \langle \sum_{k=1}^{N^c} \sum_{i \in J_k} P_i x, x\rangle_A \\
&= \langle P_0 x, x\rangle_A + \sum_{k=1}^{N^c} \sum_{i \in J_k} \langle P_i x, x\rangle_A \\
&= \langle P_0 x, x\rangle_A \\
&\quad + \sum_{k=1}^{N^c} \sum_{i \in J_k} \langle P_i x, \underbrace{P_i x}_{4.4-3.} + \underbrace{\sum_{j \in J_k, j \neq i}}_{\hat{\Omega}_i \cap \hat{\Omega}_j = \emptyset} P_j x\rangle_A \\
&= \langle P_0 x, x\rangle_A + \sum_{k=1}^{N^c} \langle \sum_{i \in J_k} P_i x, \sum_{i \in J_k} P_i x\rangle_A \\
&\leq \langle x, x\rangle_A + \sum_{k=1}^{N^c} \langle x, x\rangle_A \qquad \text{(sum of orth. proj. = orth. proj.)} \\
&= (N^c + 1)\langle x, x\rangle_A
\end{aligned}$$

□

In general assumption 5.10 does not deliver a satisfactory upper bound for $\rho(\mathscr{E})$.

Lemma 5.13. Let $\mathcal{T}_H, \mathcal{T}_h$ be shape regular, quasi-uniform triangulations and let assumptions 5.9 and 5.10 hold. Then for every $x \in \mathbb{R}^{I_h}$ there exists a decomposition $x = \sum_{i=0}^{p} R_i^T x_i$ such that

$$\sum_{i=0}^{p} \langle R_i^T x_i, R_i^T x_i\rangle_A \leq c\left(1 + \frac{H}{\delta}\right)\langle x, x\rangle_A$$

with $c > 0$ independent of H, h and $\delta := \min \delta_i$.

Proof. 1. With each $R_i^T x_i$ we identify a finite element function

$$V_h \supset V_{h,i} \ni u_i = \sum_{j \in I_h} (R_i^T x_i)_j \phi_j^h$$

and $u = \sum_{j \in I_h} (x)_j \phi_j^h$ is the function to be decomposed. We construct $u_i \in V_{h,i}$ by setting

$$u_0 := \mathscr{I}_h(\tilde{\mathscr{I}}_h u),$$

$$u_i := \mathscr{I}_h\left(\theta_i(u - u_0)\right), \quad i = 1, \ldots, p, \theta_i = \mathscr{I}_h \tilde{\theta}_i.$$

By construction the u_i are in the correct subspaces such that they can be represented by the appropriate coefficient vectors (i.e. the equation $u_i = \sum_{j \in I_h} (R_i^T x_i)_j \phi_j^h$ can be solved for x_i). Obviously

$$\sum_{i=0}^{p} \langle R_i^T x_i, R_i^T x_i \rangle_A = \sum_{i=0}^{p} a(u_i, u_i)$$

and the rest of the proof we consider only finite element functions.

2. . We consider u_0 first. Set $u_H := \tilde{\mathscr{I}}_H u$, i.e. $u_0 = \mathscr{I}_h u_H$.

$$\begin{aligned}
a(u_0, u_0) &= a(\mathscr{I}_h u_H, \mathscr{I}_h u_H) \\
&\leq c\,|\mathscr{I}_h u_H|_{1,\Omega}^2 && \text{(continuity of BLF)} \\
&= c\,|\mathscr{I}_h u_H - u_H + u_H|_{1,\Omega}^2 \\
&\leq c\left(|u_H - \mathscr{I}_h u_H|_{1,\Omega}^2 + |u_H|_{1,\Omega}^2\right) \\
&= c\left(\sum_{t \in \mathcal{T}_h} |u_H - \mathscr{I}_h u_H|_{1,t}^2 + |u_H|_{1,\Omega}^2\right) \\
&\leq c\left(c \sum_{t \in \mathcal{T}_h} |u_H|_{1,\omega_t}^2 + |u_H|_{1,\Omega}^2\right) && \text{(Corollary 5.3, } s = 1\text{)} \\
&\leq c\,|u_H|_{1,\Omega}^2 && (t \text{ appears in finitely many } \omega_t) \\
&= c\left|\tilde{\mathscr{I}}_H u\right|_{1,\Omega}^2 && \text{(``finite covering argum.'')} \\
&\leq c\,|u|_{1,\Omega}^2 && \text{(5.3, (5.2), finite covering arg.)} \\
&\leq c\,a(u, u) && \text{(coercivity of BLF)}
\end{aligned}$$

3. Now we look at the subdomains $i \in \{1, \dots, p\}$.

$$\begin{aligned} a(u_i, u_i) &\leq c|u_i|^2_{1,\Omega} && \text{(continuity)} \\ &= c \left| \mathscr{I}_h \left(\theta_i \underbrace{(u - u_0)}_{=:\omega} \right) \right|^2_{1,\Omega} && \text{(Definition)} \\ &\leq c|\theta_i \omega|^2_{1,\hat{\Omega}_i} && (5.7,\ \theta_i\omega \in P_2(\mathcal{T}_h), \operatorname{supp}(\theta_i\omega) \subset \hat{\Omega}_i) \\ &= c \int\limits_{\hat{\Omega}_i} \nabla(\theta_i\omega) \cdot \nabla(\theta_i\omega) \,\mathrm{d}x \\ &\leq c \int\limits_{\hat{\Omega}_i} \underbrace{\|\omega \nabla\theta_i\|_2^2}_{=:(1)} + \underbrace{\|\theta_i \nabla\omega\|_2^2}_{=:(2)} \,\mathrm{d}x \end{aligned}$$

Product rule:

$$\begin{aligned} \nabla(\theta_i\omega) \cdot \nabla(\theta_i\omega) &= \sum_{j=1}^d \left(\partial_j(\theta_i\omega)\right)^2 \\ &= \sum_{j=1}^d \left((\partial_j\theta_i)\omega + \theta_i\partial_j\omega)\right)^2 \\ &\leq \sum_{j=1}^d 2\left(\partial_j\theta_i\omega\right)^2 + 2\left(\theta_i\partial_j\omega\right)^2 \end{aligned}$$

We will estimate the two terms (1) and (2) later.

4. For $s = 0, 1$ we have

$$\begin{aligned} |\omega|^2_{s,\hat{\Omega}_i} &= |u - \mathscr{I}_h u_H|^2_{s,\hat{\Omega}_i} && \text{(Def. } \omega) \\ &= |u - u_H + u_H - \mathscr{I}_h u_H|^2_{s,\hat{\Omega}_i} \\ &\leq 2|u - \tilde{\mathscr{I}}_H u|^2_{s,\hat{\Omega}_i} + 2|u_H - \mathscr{I}_h u_H|^2_{s,\hat{\Omega}_i}. \end{aligned}$$

$$\begin{aligned} |u - \tilde{\mathscr{I}}_H u|^2_{s,\hat{\Omega}_i} &= \sum_{T \in \mathcal{T}_H, T \cap \hat{\Omega}_i \neq \emptyset} |u - \tilde{\mathscr{I}}_H u|^2_{s,T} \\ &\leq \sum_{T \in \mathcal{T}_H, T \cap \hat{\Omega}_i \neq \emptyset} c\, H_T^{2(1-s)} |u|^2_{1,\omega_T} && \text{(Lemma 5.5)} \end{aligned}$$

$$\begin{aligned}
|u_H - \mathscr{I}_h u_H|^2_{s,\hat\Omega_i} &= \sum_{t\in\mathcal{T}_h, t\cap\hat\Omega_i\neq\emptyset} |u_H - \mathscr{I}_h u_H|^2_{s,t} \\
&\le \sum_{t\in\mathcal{T}_h, t\cap\hat\Omega_i\neq\emptyset} c\, h_t^{2(1-s)} |u_H|^2_{1,\omega_t} \\
&\le c \sum_{t\in\mathcal{T}_h, t\cap\hat\Omega_i\neq\emptyset} h_t^{2(1-s)} |u_H|^2_{1,t} && \text{(reorganize, enlarge possibly)} \\
&\le c \underbrace{\sum_{T\in\mathcal{T}_H, T\cap\overline{\hat\Omega}_i\neq\emptyset}}_{\text{larger}} H_T^{2(1-s)} |\tilde{\mathscr{I}}_H u|^2_{1,T} && (h_t \le H_T \quad \forall t\cap T \neq \emptyset) \\
&\le c \sum_{T\in\mathcal{T}_H, T\cap\overline{\hat\Omega}_i\neq\emptyset} H_T^{2(1-s)} |u|^2_{1,\omega_T} && \text{(Lemma 5.5)}
\end{aligned}$$

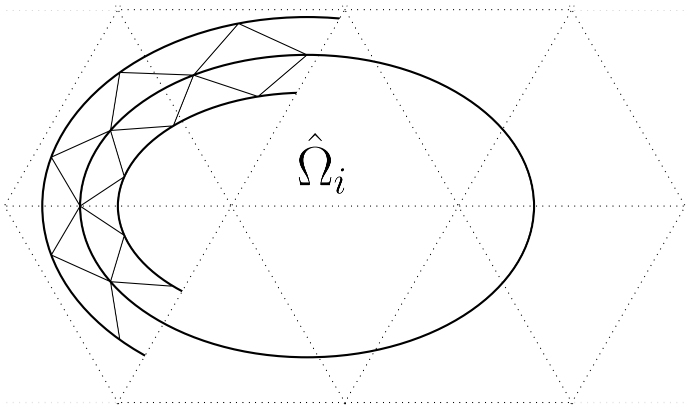


Figure 5.10: $\mathcal{T}_H$ and $\mathcal{T}_h$ shape regular, quasi-uniform triangulations

Therefore we have for $s = 0, 1$:

$$|\omega|^2_{s,\hat\Omega_i} \le c \sum_{T\in\mathcal{T}_H, T\cap\overline{\hat\Omega}_i\neq\emptyset} H_T^{2(1-s)} |u|^2_{1,\omega_T}.$$

5. Now we continue with 3.

Term (2):

$$
\begin{aligned}
\int\limits_{\hat{\Omega}_i} \|\theta_i \nabla\omega\|_2^2 \,\mathrm{d}x &\leq \int\limits_{\hat{\Omega}_i} \|\nabla\omega\|_2^2 \,\mathrm{d}x && (\theta_i \leq 1)\\
&= |u - u_0|^2_{1,\hat{\Omega}_i}\\
&\leq c \sum_{T\in\mathcal{T}_H, T\cap\overline{\hat{\Omega}}_i\neq\emptyset} |u|^2_{1,\omega_T} && (4.\text{ above } s = 1)
\end{aligned}
$$

Term (1):

$$
\begin{aligned}
\int\limits_{\hat{\Omega}_i} \|\omega\nabla\theta_i\|_2^2 \,\mathrm{d}x &= \int\limits_{\hat{\Omega}_{i,\delta_i}} |\omega|^2 \|\nabla\theta_i\|_2^2 \,\mathrm{d}x && (\nabla\theta_i = 0 \text{ interior})\\
&\leq \left(\frac{c}{\delta_i}\right)^2 \int\limits_{\hat{\Omega}_{i,\delta_i}} |\omega|^2 \,\mathrm{d}x\\
&\leq \left(\frac{c}{\delta_i}\right)^2 \delta_i^2 \left(\left(1 + \frac{H_i}{\delta_i}\right) |\omega|^2_{1,\hat{\Omega}_i} + \frac{1}{H_i\delta_i} \|\omega\|^2_{0,\hat{\Omega}_i}\right) && (\text{Lemma } 5.8)\\
&\leq c\left(1 + \frac{H_i}{\delta_i}\right) c \sum_{T\in\mathcal{T}_H, T\cap\overline{\hat{\Omega}}_i\neq\emptyset} |u|^2_{1,\omega_T}\\
&\qquad + c\frac{1}{H_i\delta_i} c' \sum_{T\in\mathcal{T}_H, T\cap\overline{\hat{\Omega}}_i\neq\emptyset} H_T^2 |u|^2_{1,\omega_T} && (\text{using } 4.)
\end{aligned}
$$

Use that $H_i \subseteq H_T \subseteq H_i$, i.e. $\mathcal{T}_H$ is *quasi-uniform with* $H_T \sim H_i$ (the subdomain diameter). So we get in continuation of 3.:

$$
a(u_i, u_i) \leq c\left(1 + \frac{H_i}{\delta_i}\right) \sum_{T\in\mathcal{T}_H, T\cap\overline{\hat{\Omega}}_i\neq\emptyset} |u|^2_{1,\omega_T}.
$$

6. Finally, sum over subdomains and coarse grid.

$$
\begin{aligned}
\sum_{i=1}^{p} a(u_i,u_i) &\leq c \sum_{i=1}^{p} \left[\left(1+\frac{H_i}{\delta_i}\right) \sum_{T\in\mathcal{T}_H, T\cap\overline{\hat{\Omega}}_i\neq\emptyset} |u|_{1,\omega_T}^2 \right] && \Big(\frac{H_i}{\delta_i} \sim \frac{H}{\delta}\Big) \\
&\leq c\left(1+\frac{H}{\delta}\right) \sum_{T\in\mathcal{T}_H} |u|_{1,\omega_T}^2 && (T\cap\overline{\hat{\Omega}}_i \neq \emptyset \text{ finitely many } \hat{\Omega}_i) \\
&\leq c\left(1+\frac{H}{\delta}\right) \sum_{T\in\mathcal{T}_H} |u|_{1,T}^2 && (T\cap\omega_T \text{ for finitely many}) \\
&\leq c\left(1+\frac{H}{\delta}\right) |u|_{1,\Omega}^2 \\
&\leq c\left(1+\frac{H}{\delta}\right) a(u,u) && (\text{coercivity})
\end{aligned}
$$

Together with 2. and the fact that $\frac{H}{\delta} > 0$ we get

$$
\begin{aligned}
\sum_{i=1}^{p} a(u_i,u_i) &\leq c\left(1+\frac{H}{\delta}\right) a(u,u) \\
&= c\left(1+\frac{H}{\delta}\right) \langle x,x\rangle_A .
\end{aligned}
$$

□

This ends the proof for the additive and multiplicative two-level Schwarz method.

Chapter 6

Multigrid Methods

6.1 Multilevel Finite Element Spaces

In this chapter we consider a hierarchy of finite element spaces that are obtained on a sequence of nested meshes obtained by uniform refinement of an initial mesh $\mathcal{T}_H = \mathcal{T}_0$ as shown in figure 6.1.

The mesh levels are denoted by

$$\mathcal{T}_H = \mathcal{T}_0, \mathcal{T}_1, \dots, \mathcal{T}_L = \mathcal{T}_h,$$

and the corresponding lowest order (bi-, tri-, ...) linear finite element spaces are

$$V_H = V_0 \subset V_1 \subset \cdots \subset V_L = V_h. \tag{6.1}$$

The spaces are spanned by the corresponding Lagrange basis functions:

$$V_l = \operatorname{span}\{\phi_i^l : i \in I_l\}, \qquad 0 \le l \le L.$$

The index sets $I_l = \{1, \dots, N_l\}$ are chosen in such a way that

$$i \in I_l \wedge \phi_i^l(s_i) = 1 \implies \forall k > l \wedge k \le L : \phi_i^k(s_i) = 1,$$

i.e. the basis functions ϕ_i^k correspond to the same vertex position s_i at all levels $k \ge l$ when $i \in I_l$.

Since $V_l \subset V_k$ for all $k > l$ any basis function on level l can be represented on all finer levels, i.e. there exist coefficients $\theta_{i,j}^{l,k}$ such that

$$\phi_i^l = \sum_{j \in I_k} \theta_{i,j}^{l,k} \phi_j^k, \qquad i \in I_l, l < k \le L.$$

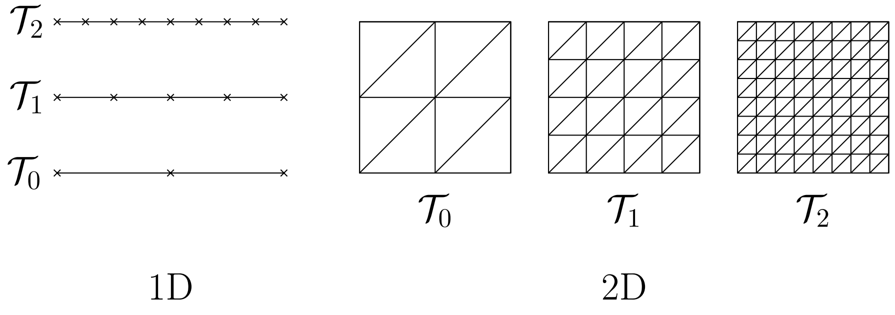


Figure 6.1: Uniform refinement

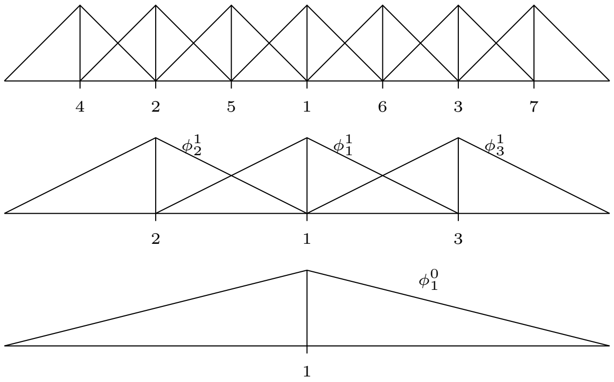


Figure 6.2: Basis functions on multiple levels

As an example consider the $1D$ case in figure 6.2.

Remark 6.1. The coefficients $\theta_{i,j}^{l,k}$ are given by

$$\theta_{i,j}^{l,k} = \phi_i^l(s_j)$$

and therefore $\theta_{i,j}^{l,k} \neq 0$ if and only if $s_j \in \operatorname{supp} \phi_i^l$. On a *cube* mesh in d space dimensions we have at most $(2^{k-l+1}-1)^d$ nonzero coefficients to represent ϕ_i^l on level $k > l$.

As a consequence the interpolation (restriction) of $v_l \in V_l$ to level $l+1$ $(l-1)$ can be done in $O(N_l)$ operations. We will see below that in fact the interpolation to any level $k > l$ can be done in $O(N_k)$ operations.

6.2 Multilevel Subspace Correction Methods

Jacobi and Gauß-Seidel as subspace corrections

Consider $V_h = \operatorname{span}\{\phi_i^h : i \in I_h\}$ and set $V_i := \operatorname{span}\{\phi_i^h\}$ for $i \in I_h$. Then $V_i \subset V_h$ and $V_h = \bigoplus_{i=1}^{N_h} V_i$ (direct sum of vector spaces), meaning that there is a *unique* representation of any $u_h \in V_h$ as a sum of elements from the V_i's.

In that case the additive subspace correction method reads in function space formulation

$$u_h^{k+1} := u_h^k + \sum_{i \in I_h} \frac{(f, \phi_i^h)_{0,\Omega} - a(u_h^k, \phi_i^h)}{a(\phi_i^h, \phi_i^h)} \phi_i^h$$

since the subspace problem is

$$\begin{aligned} & a(u_h^k + z_i \phi_i^h, \phi_i^h) = (f, \phi_i^h)_{0,\Omega} \\ \iff & z_i a(\phi_i^h, \phi_i^h) = (f, \phi_i^h)_{0,\Omega} - a(u_h^k, \phi_i^h). \end{aligned}$$

This corresponds to the Jacobi Method.

The multiplicative method (aka Gauß-Seidel) reads

$$u_h^{k+\frac{i}{N_h}} = u_h^{k+\frac{i-1}{N_h}} + \frac{(f,\phi_i^h)_{0,\Omega} - a(u_h^{k+\frac{i-1}{N_h}},\phi_i^h)}{a(\phi_i^h,\phi_i^h)}\phi_i^h.$$

Hierarchical Basis (Multigrid)

In this method $V_h = V_L$ is decomposed as a direct sum by

$$V_L = V_0 \oplus \bigoplus_{l=1}^{L} \bigoplus_{i\in I_l\setminus I_{l-1}} \operatorname{span}\{\phi_i^l\}.$$

Again $V_i^l := \operatorname{span}\{\phi_i^l\}$ is one-dimensional and the additive version reads

$$u_h^{k+1} := u_h^k + v_0 + \sum_{l=1}^{L} \sum_{i\in I_l\setminus I_{l-1}} \frac{(f,\phi_i^l)_{0,\Omega} - a(u_h^k,\phi_i^l)}{a(\phi_i^l,\phi_i^l)}\phi_i^l \tag{6.2}$$

where v_0 solves $a(u_h^k+v_0,w) = (f,w)_{0,\Omega} \quad \forall w \in V_0$, i.e. the standard coarse grid solution. The method given by (6.2) is called *Hierarchical Basis Method* and was introduced in Yserentant [1986]. The corresponding multiplicative version is known as the *Hierarchical Basis Multigrid Method* introduced in Bank et al. [1988].

The multiplicative method requires the fixation of an ordering of the indices and is typically symmetrized in order to use it as a preconditioner. This involves visiting some subspaces more than once, for example in the order

$$I_L \setminus I_{L-1}, I_{L-1} \setminus I_{L-2}, \ldots, I_1 \setminus I_0, I_0, \widetilde{I_1 \setminus I_0}, \ldots, \widetilde{I_L \setminus I_{L-1}} \tag{6.3}$$

where tilde means that the ordering is reversed within the set.

Finally, the *efficient* implementation of the sum (6.2) requires some thought because in $a(u_h^k,\phi_i^l)$, $u_h^k \in V_L$ is a fine-grid function and ϕ_i^l is a coarse-grid function.

The convergence properties of these methods are

$$\kappa(B_{HB}A) = \begin{cases} O(\log\frac{H}{h}) & d=2 \\ O(\frac{H}{h}) & d=3, \end{cases}$$

$$\|E_{HBMG}\|_A \le \begin{cases} 1-\frac{1}{O(\log\frac{H}{h})} & d=2 \\ 1-\frac{1}{O(\frac{H}{h})} & d=3. \end{cases}$$

True Multigrid Methods

True multigrid methods work with a non-unique decomposition of the subspaces:

$$V_L = V_0 + \sum_{l=1}^{L} \sum_{i \in I_l} V_i^l.$$

Obviously the representation

$$u_h = u_0 + \sum_{l=1}^{L} \sum_{i \in I_l} u_i^l, \qquad u_0 \in V_0, u_i^l \in V_i^l$$

is *not* unique.

The additive version is then very similar to (6.2) and reads

$$u_h^{k+1} := u_h^k + v_0 + \sum_{l=1}^{L} \sum_{i \in I_l} \frac{(f, \phi_i^l)_{0,\Omega} - a(u_h^k, \phi_i^l)}{a(\phi_i^l, \phi_i^l)} \phi_i^l \tag{6.4}$$

Due to the non-uniqueness of the decomposition the method needs to be used as a preconditioner and is known as *Multilevel Diagonal Scaling.* A simpler version where $a(\phi_i^l, \phi_i^l) \approx h_l^{d-2}$ (valid on quasi-uniform grids for the Laplace operator) is known as the *BPX-Method* introduced by Bramble et al. [1990].

The appropriately symmetrized version, e.g. using the ordering from (6.3), is a special case of the standard multigrid method using a V-cycle ($\gamma = 1$), one forward Gauß-Seidel step as pre-smoother and one backward Gauß-Seidel step as post-smoother. The multigrid method was introduced in Brandt [1977] and Hackbusch [1978].

The convergence properties of these methods are

$$\kappa(B_{MDS}A) \leq c$$

and

$$\|E_{MDS}\|_A \leq c' < 1$$

with c, c' independent of H, h. The first proofs of this have been given by Oswald [1991] and Dahmen and Kunoth [1992], see also Yserentant [1993].

6.3 Sequential Implementation

We now consider the implementation of (6.4) on a sequential machine.

First, observe that the sum

$$W_L := \sum_{l=0}^{L} N_l = N_L(1 + \frac{1}{\eta} + \frac{1}{\eta^2} + \cdots + \frac{1}{\eta^L}) \leq N_L \sum_{l=0}^{\infty} \eta^{-l} = O(N_L) \tag{6.5}$$

when $\eta^{-1} < 1$. Since $\eta = 2^d$ for uniform refinement this is the case. This motivates that we would like to implement (6.4) with work proportional to N_L.

However, a naive implementation of either (f_h, ϕ_i^l) or $a(u_h^k, \phi_i^l)$ requires $O(\eta^{L-l})$ operations which yields

$$\sum_{l=0}^{L} \frac{N_L}{\eta^{L-l}} \eta^{L-l} = \sum_{l=0}^{L} N_L = O(N_L L) = O(N_L \log N_L)$$

since $N_l = N_0 \eta^l$.

An optimal implementation requires to compute the numbers

$$\nu_i^l = (f_h, \phi_i^l) - a(u_h^k, \phi_i^l)$$

recursively proceeding from fine to coarse:

$$\begin{aligned} \nu_i^l &= (f_h, \phi_i^l) - a(u_h^k, \phi_i^l) \\ &= (f_h, \sum_{j \in I_{l+1}} \theta_{i,j}^{l,l+1} \phi_j^{l+1}) - a(u_h^k, \sum_{j \in I_{l+1}} \theta_{i,j}^{l,l+1} \phi_j^{l+1}) \\ &= \sum_{j \in I_{l+1}} \theta_{i,j}^{l,l+1} \left((f_h, \phi_j^{l+1}) - a(u_h^k, \phi_j^{l+1})\right) \\ &= \sum_{j \in I_{l+1}} \theta_{i,j}^{l,l+1} \nu_j^{l+1} \end{aligned} \tag{6.6}$$

The computation of (6.6) for all ν_i^l, $i \in I_l$ has complexity $O(N_l)$ and consequently the computation of ν_i^l for all $i \in I_l, 0 \leq l < L$ takes $O(N_L)$ according to (6.5).

As a consequence only a level-wise computation of (6.4) as in the standard multigrid method is computationally efficient.

This implies the following sequential algorithms (6.1, 6.2) in terms of coefficients. By A_l and b_l denote the stiffness matrix and right-hand side on level l and $R_l : \mathbb{R}^{I_{l+1}} \to \mathbb{R}^{I_l}$ is the restriction matrix given by

$$(R_l)_{i,j} = \theta_{i,j}^{l,l+1} \tag{6.7}$$

as in section 3.4.

The multiplicative algorithm can be extended by visiting subspaces multiple times in different order. By $S_{pre}^{\nu_1}$ and $S_{post}^{\nu_2}$ we denote two generic iterative

Algorithm 6.1 Multilevel Diagonal Scaling

```
Given: $x_L^k$
  $d_L := b_L - A_L x_L^k$
  for $l = L-1, \dots, 0$ do
      $d_l := R_l d_{l+1}$
  end for
  $v_0 := A_0^{-1} d_0$
  for $l = 1, \dots, L$ do
      $v_l := D_l^{-1} d_l$                                  ▷ $D_l := \mathrm{diag}(A_l)$
      $v_l := v_l + R_{l-1}^T v_{l-1}$
  end for
  $x_L^{k+1} := x_L^k + v_L$
```

schemes applying ν_1 (ν_2) iterations. Typically S_{pre}, S_{post} are either symmetric or $S_{post} \circ S_{pre}$ is symmetric (but this is not required). The method introduced in section 6.2 assumes S_{pre} and S_{post} to be the forward and backward Gauß-Seidel method. Other choices are SSOR or ILU.

Algorithm 6.2 Recursive version of the multiplicative algorithm with parameters ν_1, ν_2 and γ

```
  function MGC($l, x_l, b_l$)                  ▷ $b_l$ input, $x_l$ input/output parameter
      if l==0 then
          $x_0 := A_0^{-1} b_0$
      end if
      $x_l := S_{pre}^{\nu_1}(x_l, b_l)$     ▷ apply $\nu_1$ iterations on $A_l x_l = b_l$ with initial guess $x_l$
      $d_l := b_l - A_l x_l$
      $d_{l-1} := R_{l-1} d_l$
      $v_{l-1} := 0$
      for $j = 0, \dots, \gamma - 1$ do
          function MGC($l-1, v_{l-1}, b_{l-1}$)
          end function
      end for
      $x_l := x_l + R_{l-1}^T v_{l-1}$
      $x_l := S_{post}^{\nu_2}(x_l, b_l)$
  end function
```

The value $\gamma = 1$ is called V-cycle, $\gamma = 2$ is called W-cycle.

The cost of one iteration is optimal, i.e. $O(N_L)$ if $\gamma < \eta$, where η is the growth factor from level to level.

6.4 Parallel Implementation of Multigrid Methods

The parallelization is based on the following ideas:

- The coarse grid problem is either solved sequentially or in a distributed fashion as in the two level Schwarz method.
- The solution of the one-dimensional subspaces in the additive scheme is trivially parallel.
- In the multiplicative scheme the smoothers S_{pre}, S_{post} need to be parallelized: Either use the damped Jacobi method or a hybrid method that is additive between processors and multiplicative within processors. (Aka Block-Jacobi with Gauß-Seidel as inexact block solver.)
- Prolongation and restriction between grid levels is trivially parallel.
- Note that in multigrid N_0 is in principle *independent* of the number of processors p. If $p \gg N_0$ then the coarse levels need to be treated by less than p processors.

The parallelization of the grid transfer is essential. We first treat the case where:

- $\hat{\Omega}_i^l$ is a union of elements of $\mathcal{T}_l$ associated with processor i on level l.
- Every grid level is decomposed in an *overlapping* fashion where $\hat{\Omega}_i^l \supseteq \hat{\Omega}_i^{l+1}$ for $0 \leq l < L$ and the overlap is at least one element on each level.
- The domain decomposition results in an overlapping decomposition of the index sets
$$I_l = \bigcup_{i=1}^{p} I_{l.i}, \qquad 0 \leq l \leq L$$
and we require that the decomposition is such that
$$\forall \alpha \in I_l \; \forall \beta \in I_{l+1,i} : \quad \theta_{\alpha,\beta}^{l,l+1} \neq 0 \implies \alpha \in I_{l,i}, \tag{6.8}$$
i.e. the interpolation can be done locally.
- The two conditions above imply that $|\mathcal{T}_0| \geq p$.

Remember, by $R_l : \mathbb{R}^{I_{l+1}} \to R^{I_l}$ we denote the sequential multigrid restriction operator and by $r_{l,i} : \mathbb{R}^{I_l} \to \mathbb{R}^{I_{l,i}}$ the subdomain restriction operator, i.e. $(r_{l,i}x)_j = (x)_j \, \forall j \in I_{l,i}$ known from the Schwarz method. Finally,

$R_{l,i} : \mathbb{R}^{I_{l+1,i}} \to \mathbb{R}^{I_{l,i}}$ is the local restriction operator that can be carried out in each processor and which is given by

$$(R_{l,i}\, x_{l+1,i})_\alpha = \sum_{\beta \in I_{l+1,i}} \theta^{l,l+1}_{\alpha,\beta} (x_{l+1,i})_\beta. \tag{6.9}$$

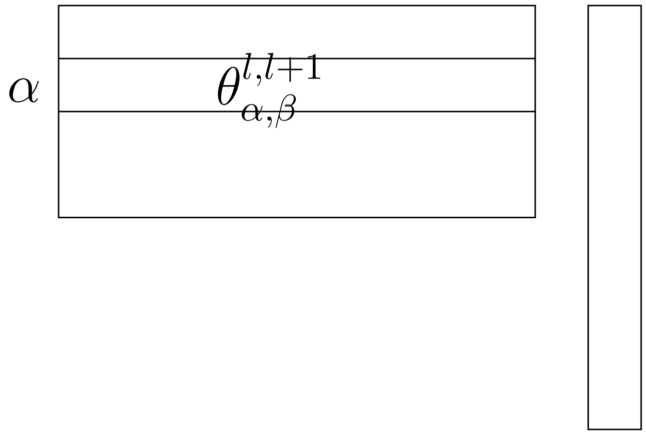


Figure 6.3: Local restriction operator

Observation 6.2 (Local Restriction)**.** For all $0 \le l < L$, $1 \le i \le p$ and $x_{l+1,i} \in \mathbb{R}^{I_{l+1,i}}$:

$$R_l\, r^T_{l+1,i}\, x_{l+1,i} = r^T_{l,i}\, R_{l,i}\, x_{l+1,i} \tag{6.10}$$

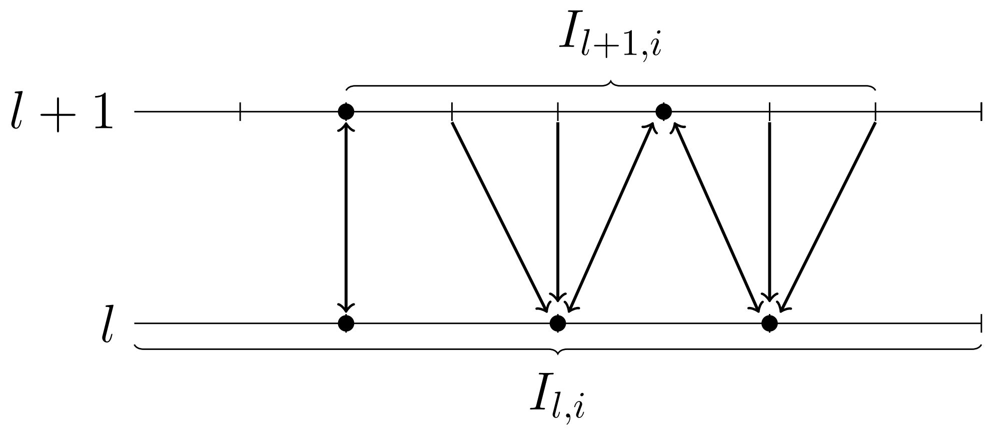


Figure 6.4: Local restriction

Proof.

- For $\alpha \in I_{l,i}$:

$$\begin{aligned}(r_{l,i}^T R_{l,i}\, x_{l+1,i})_\alpha &= (R_{l,i}\, x_{l+1,i})_\alpha && \text{(Def. of } r_{l,i}, \alpha \in I_{l,i}\text{)}\\ &= \sum_{\beta\in I_{l+1,i}} \theta_{\alpha,\beta}^{l,l+1} (x_{l+1,i})_\beta && \text{(Def. of } R_{l,i}\text{, (6.9))}\\ &= \sum_{\beta\in I_{l+1}} \theta_{\alpha,\beta}^{l,l+1} (r_{l+1,i}^T\, x_{l+1,i})_\beta && \text{(Def. of } r_{l+1,i}\text{)}\\ &= (R_l\, r_{l+1,i}^T\, x_{l+1,i})_\alpha\end{aligned}$$

  Here we used that

$$(r_{l+1,i}^T x_{l+1,i})_\beta = \begin{cases}(x_{l+1,i})_\beta & \beta \in I_{l+1,i}\\ 0 & \beta \notin I_{l+1,i}.\end{cases}$$

- For $\alpha \notin I_{l,i}$: In this case (6.8) implies

$$\alpha \notin I_{l,i} \implies \forall \beta \in I_{l+1,i} : \ \theta_{\alpha,\beta}^{l,l+1} = 0.$$

  So

$$(R_l\, r_{l+1,i}^T\, x_{l+1,i})_\alpha = \sum_{\beta\in I_{l+1,i}} \theta_{\alpha,\beta}^{l,l+1} (x_{l+1,i})_\beta = 0$$

  and

$$(r_{l,i}^T\, R_{l,i}\, x_{l+1,i})_\alpha = 0$$

  since $\alpha \notin I_{l,i}$. □

Observation 6.3 (Local prolongation)**.** For all $0 \le l < L$, $1 \le i \le p$ and $x_l \in \mathbb{R}^{I_l}$:

$$\underbrace{r_{l+1,i} \underbrace{R_l^T x_l}_{\text{global, seq. prolong.}}}_{\text{on subdom. } i} = \underbrace{R_{l,i}^T \underbrace{r_{l,i}\, x_l}_{\text{on subdom. } i}}_{\text{local prolongation}} \tag{6.11}$$

Proof. Consider $\beta \in I_{l+1,i}$.

$$\begin{aligned}(R_{l,i}^T r_{l,i} x_l)_\beta &= \sum_{\alpha\in I_{l,i}} \theta_{\alpha,\beta}^{l,l+1} (r_{l,i} x_l)_\alpha && \text{(Def. of } R_{l,i}^T\text{)}\\ &= \sum_{\alpha\in I_{l,i}} \theta_{\alpha,\beta}^{l,l+1} (x_l)_\alpha && \text{(Def. of } r_{l,i}\text{, } \alpha \in I_{l,i}\text{)}\\ &= \sum_{\alpha\in I_l} \theta_{\alpha,\beta}^{l,l+1} (x_l)_\alpha && \text{(due to assumption (6.8))}\end{aligned}$$

□

Now these two observations can be used as follows. Given any $x_{l+1} \in \mathbb{R}^{I_{l+1}}$, decompose it into $x_{l+1,i}$ such that

$$x_{l+1} = \sum_{i=1}^{p} r_{l+1,i}^T \, x_{l+1,i}.$$

Then

$$\begin{aligned} R_l \, x_{l+1} &= R_l \sum_{i=1}^{p} r_{l+1,i}^T \, x_{l+1,i} \\ &= \sum_{i=1}^{p} R_l \, r_{l+1,i}^T \, x_{l+1,i} \\ &= \sum_{i=1}^{p} r_{l,i}^T \, R_{l,i} \, x_{l+1,i} \end{aligned}$$

where the right-hand side is again an additive decomposition on the coarser grid. This can be iterated over all levels to get

$$R_0 R_1 \cdots R_l \, x_{l+1} = \sum_{i=1}^{p} r_{0,i}^T R_{0,i} R_{1,i} \cdots R_{l,i} \, x_{l+1,i}. \tag{6.12}$$

For the prolongation we get for any $l < L$ by recursive application of observation 6.3:

$$\begin{aligned} r_{L,i} R_{L-1}^T \cdots R_{l+1}^T R_l^T x_l &= R_{L-1,i}^T \, r_{L-1,i} R_{L-2}^T \cdots R_l^T x_l \\ &= R_{L-1,i}^T \cdots R_{l,i}^T \, r_{l,i} \, x_l. \end{aligned} \tag{6.13}$$

Coarse Grid correction in two level Schwarz

Eq. (6.12) and (6.13) immediately show how to implement the coarse grid correction in two level Schwarz:

1. Compute defect.
2. Compute additive splitting.
3. Restrict locally over all levels until $l = 0$ is reached.
4. Sum defect over all partitions, requires nearest neighbor. Communication.
5. Solve coarse grid problem (either parallel or sequentially).
6. Communicate required parts of corrections to each processor.
7. Prolongate corrections over all levels until $l = L$ is reached.

Note that *no* communication is needed at intermediate levels.

Application to Multigrid

On each level do:

1. Smooth (includes communication).
2. Compute defect, do additive decomposition.
3. Restrict locally.
4. Sum defect over partitions (communication required).
5. Solve coarse grid problem recursively (includes communication).
6. Provided correction is available on partition prolongate locally.
7. Smooth (includes communication).

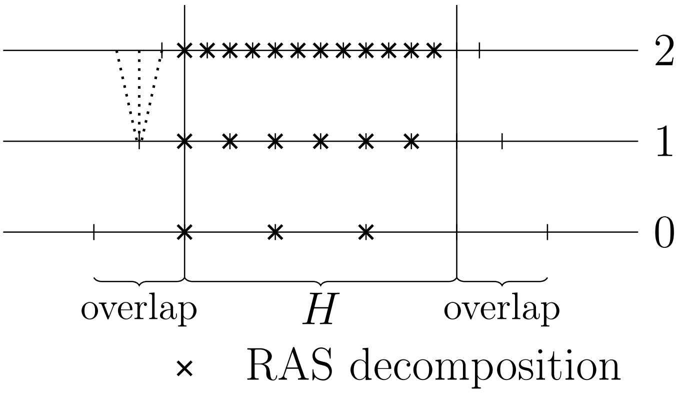


Figure 6.5: Data decomposition with minimal overlap and RAS smoother.

Observe that condition (6.8) is satisfied.

Non-overlapping Multigrid Implementation

The data decomposition from figure 6.6 also satisfies (6.8). Note that there is *no* overlap with the other subdomains, except in codimension 1.

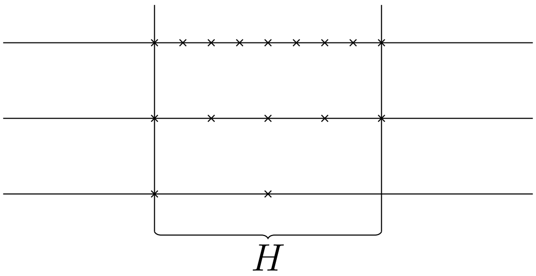


Figure 6.6: Data decomposition without overlap

This data decomposition corresponds to a non-overlapping decomposition

$$\bar{\Omega} = \bigcup_{i=1}^{p} \bar{\Omega}_i, \qquad \Omega_i \cap \Omega_j = \emptyset \, \forall i \neq j.$$

We may define the linear and bilinear forms

$$l_i(v) := \int\limits_{\Omega_i} fv \, \mathrm{d}x, \qquad a_i(u,v) := \int\limits_{\Omega_i} (K\nabla u) \cdot \nabla v \, \mathrm{d}x, \qquad i = 1, \dots, p. \quad (6.14)$$

With the index sets

$$\tilde{I}_{l,i} := \{\alpha \in I_l : \operatorname{supp}(\phi_\alpha^l) \cap \Omega_i \neq \emptyset\}, \qquad i = 1, \dots, p \quad (6.15)$$

we define the right-hand side vectors and matrices

$$\begin{aligned} b_{l,i} &\in \mathbb{R}^{\tilde{I}_{l,i}} & (b_{l,i})_\alpha &:= l_i(\phi_\alpha^l), \\ A_{l,i} &\in \mathbb{R}^{\tilde{I}_{l,i} \times \tilde{I}_{l,i}} & (A_{l,i})_{\alpha,\beta} &:= a_i(\phi_\beta^l, \phi_\alpha^l). \end{aligned}$$

Together with the index sets we have the restriction operators $\tilde{R}_{l,i} : \mathbb{R}^{I_l} \to \mathbb{R}^{\tilde{I}_{l,i}}$ given as usual by

$$(\tilde{R}_{l,i} \, x_l)_\alpha := (x)_\alpha, \quad \forall \alpha \in \tilde{I}_{l,i}.$$

Since the Ω_i form a partitioning of Ω we have

$$l(v) = \sum_{i=1}^{p} l_i(v), \qquad a(u,v) = \sum_{i=1}^{p} a_i(u,v)$$

and as a consequence

$$b_l = \sum_{i=1}^{p} \tilde{R}_{l,i}^T b_{l,i}, \qquad A_l = \sum_{i=1}^{p} \tilde{R}_{l,i}^T A_{l,i} \tilde{R}_{l,i}.$$

This representation is similar to the local stiffness matrix in finite element assembly.

Now if we consider the computation of the defect on level l for a given iterate x_l^k, we obtain

$$\begin{aligned} d_l^k = b_l - A_l x_l^k &= \sum_{i=1}^{p} \tilde{R}_{l,i}^T b_{l,i} - \sum_{i=1}^{p} \tilde{R}_{l,i}^T A_{l,i} \tilde{R}_{l,i} x_l^k \\ &= \sum_{i=1}^{k} \tilde{R}_{l,i}^T (b_{l,i} - A_{l,i}(\tilde{R}_{l,i} x_l^k)) = \sum_{i=1}^{p} \tilde{R}_{l,i}^T d_{l,i}^k. \end{aligned}$$

This means:

- The quantity $d_{l,i}^k = b_{l,i} - A_{l,i}\, x_{l,i}^k$ can be computed locally, without communication provided $x_{l,i}^k = \tilde{R}_{l,i} x_l^k$ is available.
- $d_l^k = \sum_{i=1}^p d_{l,i}^k$ is an additive representation of the defect which is exactly what is needed to apply observation 6.2 for parallel restriction.

Thus, in the non-overlapping version, no communication is needed in restriction. However, there is an additional problem in the smoother: For $\alpha \in I_{l,i} \cap I_{l,j} \neq \emptyset$ and $i \neq j$ neither processor can compute the correction implied by the Jacobi or Gauß-Seidel smoothers since $a(\phi_\alpha^l, \phi_\alpha^l)$ is not available. There are two options:

- Use additional storage and a single communication during the setup phase to store the missing entries $a(\phi_\beta^l, \phi_\alpha^l)$. (For a hybrid Gauß-Seidel smoother an additional non-overlapping decomposition of the index set $I_l = \bigcup_{i=1}^p \hat{I}_{l,i}$, $\hat{I}_{l,i} \cap \hat{I}_{l,j} = \emptyset\, \forall i \neq j$, $\hat{I}_{l,i} \subseteq \tilde{I}_{l,i}$ is needed.)
- For the BPX method $a(\phi_\alpha^l, \phi_\alpha^l) \approx h_l^{d-2}$ is assumed and no additional storage and communication is necessary.

6.5 A Convergence Proof for Multilevel Methods

This proof applies to MDS and to multiplicative V-cycle multigrid with *one* Gauß-Seidel pre-smoothing step. It uses the Schwarz theory and follows the presentation in Smith et al. [1996].

Define the *a-projection* $\mathscr{P}_l : H^1(\Omega) \to V_l$ to level l as

$$a(\mathscr{P}_l u, v) = a(u, v) \quad \forall v \in V_l.$$

Practically a computation of $\mathscr{P}_l u$ would involve the solution of a linear system with the matrix A_l.

With the help of the a-projection we can define the splitting

$$\begin{aligned} u_0 &:= \mathscr{P}_0 u, \\ u_l &:= (\mathscr{P}_l - \mathscr{P}_{l-1})u \text{ for } l > 0. \end{aligned}$$

Lemma 6.4 (Properties of $\mathscr{P}_l$)**.** Assume a is symmetric. Given $u \in V_h = V_L$ the projection $\mathscr{P}_l$ satisfies

1. $u = \sum_{l=0}^L u_l$ and
2. $a(u_l, u_k) = \begin{cases} 0 & l > k, \\ a(u, u_l) & l = k. \end{cases}$

Proof. 1. $$\sum_{l=0}^{L} u_l = \mathscr{P}_0 u + (\mathscr{P}_1 u - \mathscr{P}_0 u) + (\mathscr{P}_2 u - \mathscr{P}_1 u) \text{ The last equality}$$ $$+ \cdots + (\mathscr{P}_L u - \mathscr{P}_{L-1} u)$$ $$= \mathscr{P}_L u = u$$
holds as $u \in V_L$.

2. Assume $l \geq k > 0$.

$$\begin{aligned} a(u_l, u_k) &= a(\mathscr{P}_l u - \mathscr{P}_{l-1} u, \mathscr{P}_k u - \mathscr{P}_{k-1} u) \\ &= a(\mathscr{P}_l u, \mathscr{P}_k u) - a(\mathscr{P}_l u, \mathscr{P}_{k-1} u) \\ &\quad - a(\mathscr{P}_{l-1} u, \mathscr{P}_k u) + a(\mathscr{P}_{l-1} u, \mathscr{P}_{k-1} u) \\ &= a(\mathscr{P}_l u, \mathscr{P}_k u) - a(\mathscr{P}_{l-1} u, \mathscr{P}_k u) \qquad (*) \\ &= \begin{cases} a(u, \mathscr{P}_k u) - a(u, \mathscr{P}_k u) = 0 & \text{if } l-1 \geq k \iff l > k, \\ a(u, \mathscr{P}_l u) - a(\mathscr{P}_{l-1} u, u) = a(u, u_l) & \text{if } l = k. \end{cases} \end{aligned}$$

For (*) we used that $a(\mathscr{P}_l u, \mathscr{P}_{k-1} u) = a(u, \mathscr{P}_{k-1} u)$ since $\mathscr{P}_{k-1} u \in V_l$ and $a(\mathscr{P}_{l-1} u, \mathscr{P}_{k-1} u) = a(u, \mathscr{P}_{k-1} u)$ as $\mathscr{P}_{k-1} u \in V_{l-1}$.

Assume $l > k = 0$.

$$\begin{aligned} a(u_l, u_0) &= a(\mathscr{P}_l u - \mathscr{P}_{l-1} u, \mathscr{P}_0 u) \\ &= a(\mathscr{P}_l u, \mathscr{P}_0 u) - a(\mathscr{P}_{l-1} u, \mathscr{P}_0 u) \\ &= a(u, \mathscr{P}_0 u) - a(u, \mathscr{P}_0 u) \qquad (*) \\ &= 0. \end{aligned}$$

For (*) we used that $\mathscr{P}_0 \in V_0 \subseteq V_{l-1} \subset V_l$.

Assume $l = k = 0$.

$$a(u_0, u_0) = a(\mathscr{P}_0 u, \mathscr{P}_0 u) = a(u, \mathscr{P}_0 u)$$

since $\mathscr{P}_0 u \in V_0$. □

For any function $w \in V_l$, $l \geq 0$, we can write

$$w = \sum_{i \in I_l} w(s_i^l) \phi_i^l = \sum_{i \in I_l} \mathscr{I}_l(\phi_i^l w)$$

with the Lagrange points s_i^l and the Lagrange interpolation operator $\mathscr{I}_l$. Here the ϕ_i^l act as a "partition of unity" (up to the Dirichlet boundary and $\mathscr{I}_l(\phi_i^l w) = w(s_i)$, $\phi_i^l \in V_{l,i} = \operatorname{span}\{\phi_i^l\}$). Then for $u \in V_h = V_L$

$$u = u_0 + \sum_{l=1}^{L} \sum_{i \in I_l} \underbrace{\mathscr{I}_l(\phi_i^l u_l)}_{=:u_{l,i}} = u_0 + \sum_{l=1}^{L} \sum_{i \in I_l} u_{l,i}$$

forms a splitting of u since

$$u_0 + \sum_{l=1}^{L} \sum_{i \in I_l} \mathscr{I}_l(\phi_i^l u_l) = u_0 + \sum_{l=1}^{L} \underbrace{\sum_{i \in I_l} u_l(s_i^l)\phi_i^l}_{=u_l} = \sum_{l=0}^{L} u_l = u.$$

Now we show that this splitting is stable.

Lemma 6.5. Assume the variational problem is H^2-regular and the meshes $\mathcal{T}_l$ are quasi-uniform. Then there exists a constant $c > 0$ independent of h such that for all $u \in V_h$:

$$a(u_0, u_0) + \sum_{l=1}^{L} \sum_{i \in I_l} a(u_{l,i}, u_{l,i}) \le c\, a(u, u).$$

Proof. 1. Let $l > 0$.

$$\begin{aligned}
\sum_{i \in I_l} a(u_{l,i}, u_{l,i}) &\le \sum_{i \in I_l} c|u_{l,i}|_{1,\Omega}^2 && \text{(cont., equiv. of } |.|_1,\ \|.\|_1) \\
&\le c \sum_{i \in I_l} |u_{l,i}|_{1,\Omega_{l,i}}^2 && (\Omega_{l,i} := \operatorname{supp} \phi_i^l) \\
&= c \sum_{i \in I_l} |\mathscr{I}_l(\phi_i^l u_l)|_{1,\Omega_{l,i}}^2 && \text{(Def. of } u_{l,i}) \\
&\le c \sum_{i \in I_l} |\phi_i^l u_l|_{1,\Omega_{l,i}}^2 && \text{(Lemma 5.7)} \\
&= c \sum_{i \in I_l} \int_{\Omega_{l,i}} \|\nabla(\phi_i^l u_l)\|_2^2 \,\mathrm{d}x \\
&\le c \sum_{i \in I_l} \int_{\Omega_{l,i}} \|\phi_i^l \nabla u_l\|_2^2 + \|u_l \nabla \phi_i^l\|_2^2 \,\mathrm{d}x \\
&\le c \sum_{i \in I_l} \left(|u_l|_{1,\Omega_{l,i}}^2 + \frac{1}{h_{l,i}^2} \|u_l\|_{0,\Omega_{l,i}}^2 \right) \\
&\le c \left(|u_l|_{1,\Omega}^2 + \frac{1}{h_l^2} \|u_l\|_{0,\Omega}^2 \right) && (\mathcal{T}_l \text{ quasi-uniform})
\end{aligned}$$

2. We need to estimate $\|u_l\|_{0,\Omega}$. The continuous problem

$$U \in H_0^1(\Omega) :\ a(U, v) = l(v) \quad \forall v \in H_0^1(\Omega)$$

is H^2-regular, i.e. $U \in H^2(\Omega)$. Due to Aubin-Nitsche, see for example [Braess, 2003, Lemma 7.6], the FE solution

$$U_h \in V_h : \ a(U_h, v) = l(v) \quad \forall v \in V_h$$

satisfies the estimate

$$\|U - U_h\|_{0,\Omega} \leq ch|U - U_h|_{1,\Omega}.$$

We apply this to the a-projection in the following way: For given $u_l \in V_l$, $l > 0$, we define a linear form $l(v) := a(u_l, v)$. Then

- $U \in H_0^1(\Omega) : \ a(U, v) = a(u_l, v) \, \forall v \in H_0^1(\Omega)$ is solved by $U = u_l$.
- $\mathscr{P}_{l-1} u \in V_{l-1} : \ a(\mathscr{P}_{l-1} u_l, v) = a(u_l, v) \, \forall v \in V_{l-1}$ is the a-projection of u_l.

And we have the estimate

$$\|u_l - \mathscr{P}_{l-1} u_l\|_{0,\Omega} \leq ch_{l-1} |u_l - \mathscr{P}_{l-1} u_l|_{1,\Omega}.$$

3. The $\mathscr{P}_l$ are projections, i.e. $\mathscr{P}_l^2 = \mathscr{P}_l$. Moreover, for $\mathscr{P}_{l-1}\mathscr{P}_l u$ we get

$$a(\mathscr{P}_{l-1}(\mathscr{P}_l u), v) = a(\mathscr{P}_l u, v) = a(u, v) \quad \forall v \in V_{l-1} \subset V_l$$

and on the other hand

$$a(\mathscr{P}_{l-1} u, v) = a(u, v) \quad \forall v \in V_{l-1}.$$

So we have

$$\mathscr{P}_{l-1}\mathscr{P}_l u = \mathscr{P}_{l-1} u.$$

As a consequence, for $l > 0$:

$$\begin{aligned} \mathscr{P}_{l-1} u_l &= \mathscr{P}_{l-1}(\mathscr{P}_l u - \mathscr{P}_{l-1} u) \\ &= \mathscr{P}_{l-1}\mathscr{P}_l u - \mathscr{P}_{l-1}^2 u \\ &= \mathscr{P}_{l-1} u - \mathscr{P}_{l-1} u \\ &= 0. \end{aligned}$$

Combining that with 2. we obtain

$$\|u_l\|_{0,\Omega} = \|u_l - \mathscr{P}_{l-1} u_l\|_{0,\Omega} \leq ch_{l-1}|u_l - \mathscr{P}_{l-1} u_l|_{1,\Omega} = ch_{l-1}|u_l|_{1,\Omega}.$$

4. Now back to 1.:

$$\begin{aligned}
\sum_{i\in I_l} a(u_{l,i},u_{l,i}) &\le c\left(|u_l|^2_{1,\Omega}+\frac{1}{h_l^2}\|u_l\|^2_{0,\Omega}\right)\\
&\le c\left(|u_l|^2_{1,\Omega}+(\underbrace{\frac{h_{l-1}}{h_l}}_{=2})^2|u_l|^2_{1,\Omega}\right)\\
&\le c|u_l|^2_{1,\Omega}\\
&\le c\,a(u_l,u_l) && \text{(Ellipticity).}
\end{aligned}$$

5. Now sum over all levels.

$$\begin{aligned}
a(u_0,u_0)&+\sum_{l=1}^{L}\sum_{i\in I_l} a(u_{l,i},u_{l,i})\\
&\le a(u_0,u_0)+c\sum_{l=1}^{L}a(u_l,u_l) && \text{(4.)}\\
&\le c\sum_{l=0}^{L}a(u_l,u_l)\\
&= c\sum_{l=0}^{L}a(u,u_l) && \text{(Lemma 6.4)}\\
&= c\,a(u,\sum_{l=0}^{L}u_l)\\
&= c\,a(u,u) && \text{(Lemma 6.4)}
\end{aligned}$$ □

Now we search the upper bound. Define the restriction operators

$$\begin{aligned}
R_0 &: \mathbb{R}^{I_L}\to\mathbb{R}^{I_0}, && (R_0)_{\alpha,\beta} := \theta^{0,L}_{\alpha,\beta} = \phi^0_\alpha(s_\beta),\\
R_{l,i} &: \mathbb{R}^{I_l}\to\mathbb{R}^{\{i\}}, && (R_{l,i})_{i,\beta} := \theta^{l,L}_{i,\beta} = \phi^l_i(s_\beta)
\end{aligned}$$

and the Schwarz projection operators

$$\begin{aligned}
A_0 &:= R_0A_LR_0^T, && P_0 := R_0^TA_0^{-1}R_0A,\\
A_{l,i} &:= R_{l,i}A_LR_{l,i}^T, && P_{l,i} := R_{l,i}^TA_{l,i}^{-1}R_{l,i}A.
\end{aligned}$$

Lemma 6.6. With the definitions above there exists $c>0$ independent of the mesh size h such that

$$\langle(P_0+\sum_{l=1}^{L}\sum_{i\in I_l}P_{l,i})x,x\rangle_{A_L}\le(1+Lc)\langle x,x\rangle_{A_L}.$$

Proof. On every level l there exists a coloring of the index set into at most N_c colors

$$I_l = \bigcup_{c=1}^{N_c} I_{l,c}, \qquad I_{l,c} \cap I_{l,c'} = \emptyset \; \forall c \neq c',$$

such that

$$\langle P_{l,i}x, P_{l,j}x \rangle_{A_L} = 0 \text{ when } c(i) = c(j).$$

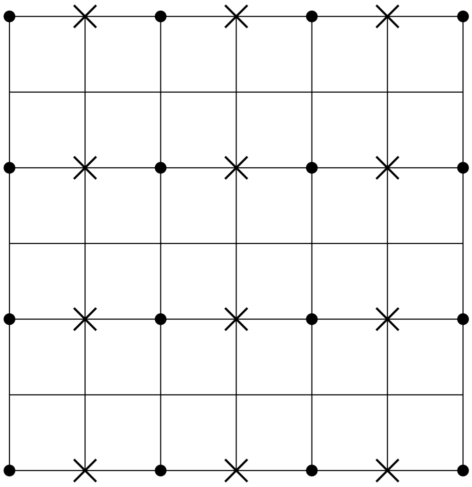


Figure 6.7: Coloring of index set

Set $\tilde{P}_{l,c} := \sum_{i \in I_{l,c}} P_{l,i}$. Then $\tilde{P}_{l,c}$ is also an A_L-orthogonal projection.

$$\begin{aligned}
&\langle (P_0 + \sum_{l=1}^{L} \sum_{i \in I_l} P_{l,i})x, x \rangle_{A_L} \\
&= \langle (P_0 + \sum_{l=1}^{L} \sum_{c=1}^{N_c} \tilde{P}_{l,c})x, x \rangle_{A_L} \\
&= \langle P_0 x, x \rangle_{A_L} + \sum_{l=1}^{L} \sum_{c=1}^{N_c} \langle \tilde{P}_{l,c} x, x \rangle_{A_L} \\
&= \langle P_0 x, P_0 x \rangle_{A_L} + \sum_{l=1}^{L} \sum_{c=1}^{N_c} \langle \tilde{P}_{l,c} x, \tilde{P}_{l,c} x \rangle_{A_L} \qquad \text{(Lemma 4.4)} \\
&\leq (1 + LN_c) \langle x, x \rangle_{A_L}. \qquad \square
\end{aligned}$$

In order to handle the multiplicative method a bound for $\rho(\mathscr{E})$ in the strengthened Cauchy-Schwarz inequality (assumption 4.2) is necessary. We will prove this in the form of an optimal bound.

Lemma 6.7 (Zhang [1992])**.** Let $A \in \mathbb{R}^{m \times n}$ be a matrix with at most c nonzero entries per *column*. Then

$$\|A\|_2 \leq c^{\frac{1}{2}} \max_i (\sum_j (A)^2_{i,j})^{\frac{1}{2}}.$$

Proof. Recall the Raleigh quotient from observation 2.4:

$$\|A\|_2 = \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

Set

$$\theta_{ij} := \begin{cases} 1 & (A)_{i,j} \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} \|Ax\|_2^2 &= \sum_i \left(\sum_j (A)_{i,j} (x)_j \right)^2 \\ &= \sum_i \left(\sum_j (A)_{i,j} \theta_{ij} (x)_j \right)^2 \\ &\leq \sum_i \left((\sum_j (A)^2_{i,j})(\sum_j \theta_{ij} (x)^2_j) \right) && \text{(Cauchy-Schwarz)} \\ &\leq \max_i (\sum_j (A)^2_{i,j})(\sum_i \sum_j \theta_{ij} (x)^2_j) \\ &= \max_i (\sum_j (A)^2_{i,j}) \sum_j (x)^2_j \underbrace{\sum_i \theta_{ij}}_{\leq c} \\ &\leq c \max_i (\sum_j (A)^2_{i,j}) \|x\|_2^2. \end{aligned}$$

Finally, combine this with the definition of $\|A\|_2$. □

Lemma 6.8. Let $\mathcal{T}_0$ be shape regular and quasi-uniform. $\mathcal{T}_l$ is obtained from $\mathcal{T}_{l-1}$ through regular subdivision of each element into 2^d elements. K is piecewise constant on $\mathcal{T}_0$. Then there exist constants c_1 and c_2 independent of h and L such that

1. $a(u, v) \leq c_1 \, 2^{-d|k-l|/2} \, a(u, u)^{\frac{1}{2}} a(v, v)^{\frac{1}{2}} \quad \forall u \in V_{l,i}, v \in V_{k,j},$
2. $\rho(\mathscr{E}) \leq c_2.$

Proof. 1. Let $u \in V_{l,i}$, $v \in V_{k,j}$ with $k \geq l$.

$$
\begin{aligned}
a(u,v) &= \int_\Omega (K\nabla u)\cdot\nabla v \,\mathrm{d}x \\
&= \int_{\Omega_{l,i}\cap\Omega_{k,j}} (K\nabla u)\cdot\nabla v \,\mathrm{d}x \\
&\leq \left(\int_{\Omega_{l,i}\cap\Omega_{k,j}} (K\nabla u)\cdot\nabla u \,\mathrm{d}x\right)^{\frac{1}{2}} \left(\int_{\Omega_{l,i}\cap\Omega_{k,j}} (K\nabla v)\cdot\nabla v \,\mathrm{d}x\right)^{\frac{1}{2}} \\
&\leq \left(\int_{\Omega_{l,i}\cap\Omega_{k,j}} (K\nabla u)\cdot\nabla u \,\mathrm{d}x\right)^{\frac{1}{2}} a(v,v)^{\frac{1}{2}}
\end{aligned}
$$

In the last step we enlarged the integration domain for the second integral. Further

$$
\begin{aligned}
&\int_{\Omega_{l,i}\cap\Omega_{k,j}} (K\nabla u)\cdot\nabla u \,\mathrm{d}x \\
&\leq c \int_{\Omega_{l,i}\cap\Omega_{k,j}} \|\nabla u\|_2^2 \,\mathrm{d}x \\
&= c \sum_{T\in\mathcal{T}_l : T\cap\Omega_{l,i}\neq\emptyset} \int_{T\cap\Omega_{k,j}} \|\nabla u\|_2^2 \,\mathrm{d}x \\
&= c \sum_{T\in\mathcal{T}_l : T\cap\Omega_{l,i}\neq\emptyset} \|\nabla u\|_2^2 \int_{T\cap\Omega_{k,j}} 1 \,\mathrm{d}x & (*) \\
&\leq c \sum_{T\in\mathcal{T}_l : T\cap\Omega_{l,i}\neq\emptyset} \left(\underbrace{\|\nabla u\|_2^2 \,|T|}_{=|u|_{1,T}^2} \frac{c'}{2^{d|k-l|}} \right) & (**) \\
&\leq \frac{c}{2^{d|k-l|}} \sum_{T\in\mathcal{T}_l : T\cap\Omega_{l,i}\neq\emptyset} |u|_{1,T}^2 \\
&= \frac{c}{2^{d|k-l|}} |u|_{1,\Omega}^2 & (***) \\
&\leq \frac{c}{\alpha} 2^{-d|k-l|} a(u,u)
\end{aligned}
$$

For (*) we used that ∇u is constant on T. For (**) we used that $\Omega_{k,j}$

consists of up to c' elements $t \in \mathcal{T}_k$ and that the volume of $t \in \mathcal{T}_k$, being a refinement of T, is $\frac{|T|}{2^{d|k-l|}}$. For (***) note that $u = 0$ outside of $\Omega_{l,i}$.

Finally take the square roots and use the symmetry.

2. Let $u \in V_{l,i}$, $v \in V_{k,j}$ with $k \geq j$, then

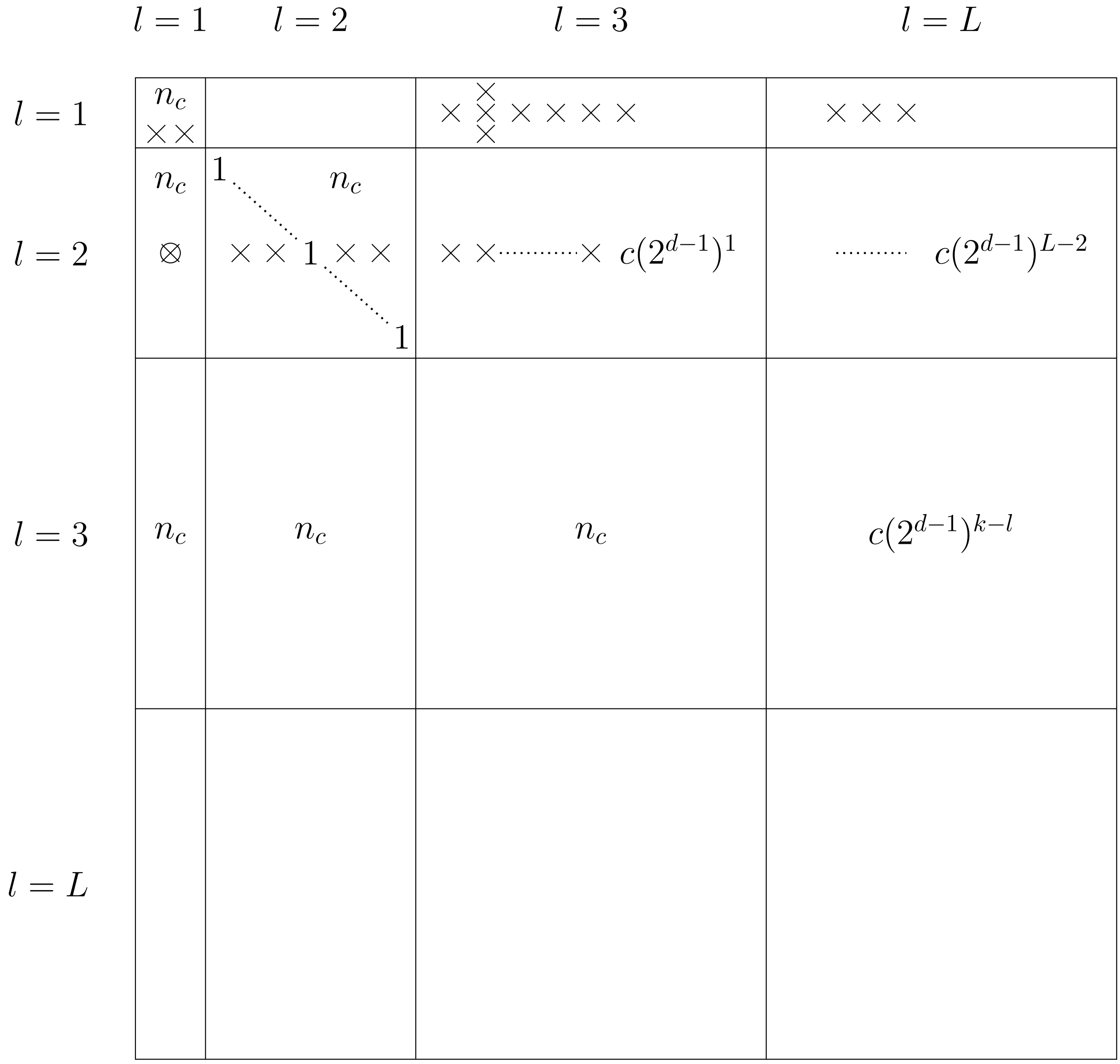
   - $\Omega_{l,i} \cap \Omega_{k,j} = \emptyset \implies a(u,v) = 0$

   - $\Omega_{k,j} \subseteq T \in \mathcal{T}_l \implies a(u,v) = 0$ because

   $$\begin{aligned}\int\limits_{\Omega_{l,i}\cap\Omega_{k,j}} (K\nabla u)\cdot\nabla v \,\mathrm{d}x &= \int\limits_T (K\nabla u)\cdot\nabla v\,\mathrm{d}x \\ &= -\int\limits_T \nabla\cdot(K\nabla u)v\,\mathrm{d}x + \int\limits_{\partial T}(K\nabla u)\cdot nv\,\mathrm{d}s \\ &= 0\end{aligned}$$

   as $\nabla \cdot (K\nabla u) = 0$ since $u \in P_1$ or Q_1 and $v = 0$ on ∂T since $\operatorname{supp}\phi_k^l = \Omega_{k,j} \subseteq T$.

3. The structure of $\mathscr{E}$:



$\otimes \quad a(\phi_i^2, \phi_j^1) = a(\phi_j^1, \phi_i^2)$

Figure 6.8: Structure of $\mathscr{E}$

As a consequence of 2. $a(\phi_i^l, \phi_j^k) \neq 0$ for at *most*

$$\left(2^{d-1}\right)^{|k-l|} c = c\, 2^{(d-1)|k-l|}$$

basis functions, where 2^{d-1} stands for the surface, $|k-l|$ is the number of refinements and c is the number of finitely many edges/faces of $T \in \mathcal{T}_l$ in $\operatorname{supp} \Omega_{l,i}$.

Introduce the level-wise block structure of $\mathscr{E}$:

$$\mathscr{E}_{l.k} \in \mathbb{R}^{I_l \times I_k},\; 1 \leq l,\, k \leq L$$

with

$$(\mathscr{E}_{l,k})_{i,j} = (\mathscr{E})_{i,j}.$$

$\mathscr{E}_{l,k}$ has the following properties:

- All entries have the size at most $c_1\, 2^{-d|k-l|/2}$.
- $\mathscr{E}_{l,k} = \mathscr{E}_{l,k}^T$
- $k \ge l$: at most $c_2\, 2^{(d-1)|k-l|}$ entries *per row* are non-zero.
- $k \ge l$: at most n_c entries *per column* are non-zero.

So, together with Lemma 6.7 we obtain for $k \ge l$

$$\begin{aligned} \|\mathscr{E}_{l,k}\|_2 &\le n_c^{\frac{1}{2}} \max_i \underbrace{\left(c_2\, 2^{(d-1)|k-l|}\left(c_1\, 2^{-d|k-l|/2}\right)^2\right)^{\frac{1}{2}}}_{\text{independent of } i} \\ &= n_c^{1/2} c_2^{1/2} c_1 2^{-|k-l|/2} \\ &= c\, 2^{-|k-l|/2}. \end{aligned}$$

4. Finally define $E \in \mathbb{R}^{L \times L}$ as
$$(E)_{l,k} := \|\mathscr{E}_{l,k}\|_2.$$
E is symmetric since $\mathscr{E}_{l,k} = \mathscr{E}_{l,k}^T$ and $\|A\|_2 = \|A^T\|_2$ for any matrix A ([Hackbusch, 1991, Folgerung 2.9.4]).

Now we show
$$\rho(\mathscr{E}) = \|\mathscr{E}\|_2 \le \|E\|_2.$$
This follows from
$$\begin{aligned} \|\mathscr{E}\|_2^2 &= \sup_{\|x\|_2^2=1} \|\mathscr{E}x\|_2^2 \\ &= \sup_{\|x\|_2^2=1} \sum_{l=1}^{L} \left(\| \sum_{k=1}^{L} \mathscr{E}_{l,k} x_k \|_2 \right)^2 && \text{(exploit block structure } x = (x_1, \dots, x \\ &\le \sup_{\|x\|_2^2=1} \sum_{l=1}^{L} \left(\sum_{k=1}^{L} \|\mathscr{E}_{l,k} x_k\|_2 \right)^2 && \text{(triangle ineq.)} \\ &\le \sup_{\|x\|_2^2=1} \sum_{l=1}^{L} \left(\sum_{k=1}^{L} \|\mathscr{E}_{l,k}\|_2 \underbrace{\|x_k\|_2}_{=:(z)_k,\, z \in \mathbb{R}^L} \right)^2 && \text{(associated matrix norm)} \\ &= \sup_{\|z\|_2^2=1} \sum_{l=1}^{L} \left(\sum_{k=1}^{L} (E)_{l,k} (z)_k \right)^2 && (\|x\|_2^2 = \sum_k \|x_k\|_2^2 = \sum_k (z)_k^2 = \|z\|_2^2) \\ &= \sup_{\|z\|_2^2=1} \|Ez\|_2^2 \\ &= \|E\|_2^2. \end{aligned}$$

5. Now combine 3. and 4.:

$$\|E\|_2 = \rho(E) \leq \|E\|_\infty = \max_l c \sum_{k=1}^{L} (\frac{1}{2})^{\frac{|k-l|}{2}} \leq \frac{c}{1-(\frac{1}{2})^{\frac{1}{2}}}$$

which is independent of L. For the last estimate we used the geometric series. □

6.6 Algebraic Multigrid (AMG)

The Need for AMG

Convergence of standard geometric multigrid (GMG) and many other methods such as additive Schwarz is often not robust with respect to problem parameters of the PDE. Consider the general elliptic problem

$$\nabla \cdot (bu - K\nabla u) + cu = f \quad \text{in } \Omega.$$

A method where the convergence rate does *not* depend on b, K and c is called robust.

Typical problems are

- $c < 0$ and $|c|$ large: indefinite Helmholtz problem
- $\|b\| \gg \|K\|$: convection-dominated problem
- $K = k(x)I$ where $k(x)$ is discontinuous with large “jumps”
- $K = Q^T DQ$ where $\max d_{ii} \gg \min d_{ii}$: anisotropic problem
- irregular/anisotropic meshes can have similar effects.

Typical remedies in GMG invented to solve these problems:

- robust smoothers: line and plane relaxation, ILU (anisotropic problems), streamline ordering (convective problem)
- semi-coarsening (anisotropic problem)
- matrix-dependent prolongations and restrictions, Galerkin coarse grid product (variable coefficient problem)

Complex geometries: generating coarse grids *of good quality* resolving complex geometries is an enormous challenge.

Software issue:

- many commercial simulators have matrix-vector interface to linear solver and only a single grid.
- geometric multigrid requires grid hierarchy and solver is intertwined with discretization.

AMG Introduction

AMG mimics the GMG method by

- Construct a hierarchy of linear systems A_l, $0 \leq l < L$.
- A_l is constructed using solely information from A_{l+1} (there are exceptions).
- Usually A_l is obtained by $A_l = R_l A_{l+1} R_l^T$ where R_l^T is a prolongation operator (rectangular matrix, sparse, full rank).
- Then a usual multiplicative (V-, W-, ...) cycle is performed.
- Intentionally only simple smoothers such as Jacobi, Gauß-Seidel, SSOR are used.
- All operations are local and parallelizable similar as in GMG.

Algebraic smoothness

Smoothing and coarse grid correction are complementary: coarse grid correction should reduce errors that are *not* reduced by the smoother. For the Poisson problem and standard FE this corresponds to low- and high-frequency sine waves. In general elliptic problems this is not so easy. Alternatively define algebraic smooth errors as those where $Se \approx e$ with S the iteration matrix of the smoother. Theoretical investigations suggest a more rigorous definition. We introduce the following scalar product and corresponding vector norms:

$$\begin{aligned}
\langle x, y\rangle_0 &:= \langle Dx, y\rangle, \\
\langle x, y\rangle_1 &:= \langle Ax, y\rangle, \\
\langle x, y\rangle_2 &:= \langle D^{-1}Ax, Ay\rangle, \\
\|x\|_i &:= \sqrt{\langle x, x\rangle_i}
\end{aligned}$$

where $\langle\cdot,\cdot\rangle$ is the Euclidean scalar product, A a symmetric positive-definite matrix and $D := \operatorname{diag}(A)$.

Observation 6.9. The scale of scalar products and norms satisfies the following relations:

$$\begin{aligned}\langle x, y\rangle_1 &\le \|x\|_0\|y\|_2,\\ \|x\|_2^2 &\le \rho(D^{-1}A)\|x\|_1^2,\\ \|x\|_1^2 &\le \rho(D^{-1}A)\|x\|_0^2.\end{aligned}$$

Proof. The first inequality follows from the Definition and Cauchy Schwarz:

$$\begin{aligned}\langle x, y\rangle_1 &= \langle x, Ay\rangle\\ &= \langle D^{\frac{1}{2}}x, D^{-\frac{1}{2}}Ay\rangle\\ &\le \sqrt{\langle D^{\frac{1}{2}}x, D^{\frac{1}{2}}x\rangle}\sqrt{\langle D^{-\frac{1}{2}}Ay, D^{-\frac{1}{2}}Ay\rangle}\\ &= \|x\|_0\|y\|_2.\end{aligned}$$

For the second inequality holds

$$\begin{aligned}\sup_{x\neq 0}\frac{\|x\|_2^2}{\|x\|_1^2} = \sup_{x\neq 0}\frac{\langle D^{-1}Ax, Ax\rangle}{\langle Ax, x\rangle}\\ &= \sup_{x=A^{-\frac{1}{2}}y\neq 0}\frac{\langle A^{\frac{1}{2}}D^{-1}A^{\frac{1}{2}}y, y\rangle}{\langle y, y\rangle}\\ &= \rho(A^{\frac{1}{2}}D^{-1}A^{\frac{1}{2}})\\ &= \rho(D^{-1}A).\end{aligned}$$

The third can be proven in the same way. □

Definition 6.10. The smoother with iteration matrix S is said to have the algebraic smoothing property if there exists $\sigma > 0$ such that

$$\|Se\|_1^2 \le \|e\|_1^2 - \sigma\|e\|_2^2.$$

From $\frac{\|Se\|_1^2}{\|e\|_1^2} \le 1 - \sigma\frac{\|e\|_2^2}{\|e\|_1^2}$ we deduce that the reduction of the error is small if $\|e\|_2 \ll \|e\|_1$ (provided σ is reasonably large). We characterize algebraically smooth errors as those where

$$\|e\|_2 \ll \|e\|_1.$$

Lemma 6.11. Provided $0 < \omega < \frac{2}{\rho(D^{-1}A)}$, the damped Jacobi iteration $S = I - \omega D^{-1}A$ has the algebraic smoothing property with $\sigma = \omega(2 - \omega\rho(D^{-1}A))$.

Proof.

$$\begin{aligned}\|Se\|_1^2 &= \langle Se, Se\rangle_1\\ &= \langle A(I-\omega D^{-1}A)e, (I-\omega D^{-1}A)e\rangle\\ &= \langle Ae, e\rangle - 2\omega\langle AD^{-1}Ae, e\rangle + \omega^2\langle AD^{-1}Ae, D^{-1}Ae\rangle\\ &\le \|e\|_1^2 - 2\omega\|e\|_2^2 + \omega^2\langle D^{-\frac{1}{2}}AD^{-\frac{1}{2}}D^{-\frac{1}{2}}Ae, D^{-\frac{1}{2}}Ae\rangle\\ &\le \|e\|_1^2 - 2\omega\|e\|_2^2 + \omega^2\rho(D^{-1}A)\|e\|_2^2\\ &= \|e\|_1^2 - \omega(2-\omega\rho(D^{-1}A))\|e\|_2^2\end{aligned}$$

From the assumption $0 < \omega < \frac{2}{\rho(D^{-1}A)}$ follows that $\omega(2-\omega\rho(D^{-1}A)) > 0$. □

Interpretation of algebraic smoothness

e is algebraically smooth if $\|e\|_2 \ll \|e\|_1$. Since $\|e\|_1^2 \le \|e\|_0\|e\|_2 \ll \|e\|_0\|e\|_1$ this implies also $\|e\|_1 \ll \|e\|_0$.

Observation 6.12. Suppose A is a symmetric positive-definite M-matrix. Then

$$\|e\|_1^2 = \frac{1}{2}\sum_i\sum_j -(A)_{i,j}\left((e)_i-(e)_j\right)^2 + \sum_i (e)_i^2\left(\sum_j (A)_{i,j}\right).$$

Proof.

$$\begin{aligned}\|e\|_1^2 &= \langle Ae, e\rangle\\ &= \sum_i\left(\sum_j (A)_{i,j}(e)_j\right)(e)_i\\ &= \sum_i\sum_j\left((A)_{i,j}\frac{1}{2}\left[(e)_i^2+(e)_j^2-\left((e)_i-(e)_j\right)^2\right]\right) \qquad ((a-b)^2 = a^2-2ab+b^2)\\ &= \frac{1}{2}\sum_i\sum_j\left(-(A)_{i,j}\left((e)_i-(e)_j\right)^2\right) + \sum_i (A)_{i,i}(e)_i^2\\ &\quad + \sum_i\sum_{j>i}\left((A)_{i,j}(e)_i + (A)_{j,i}(e)_j\right)\\ &= \frac{1}{2}\sum_i\sum_j\left(-(A)_{i,j}\left((e)_i-(e)_j\right)^2\right)\\ &\quad + \sum_i\left(\sum_j (A)_{i,j}\right)(e)_i^2\end{aligned}$$

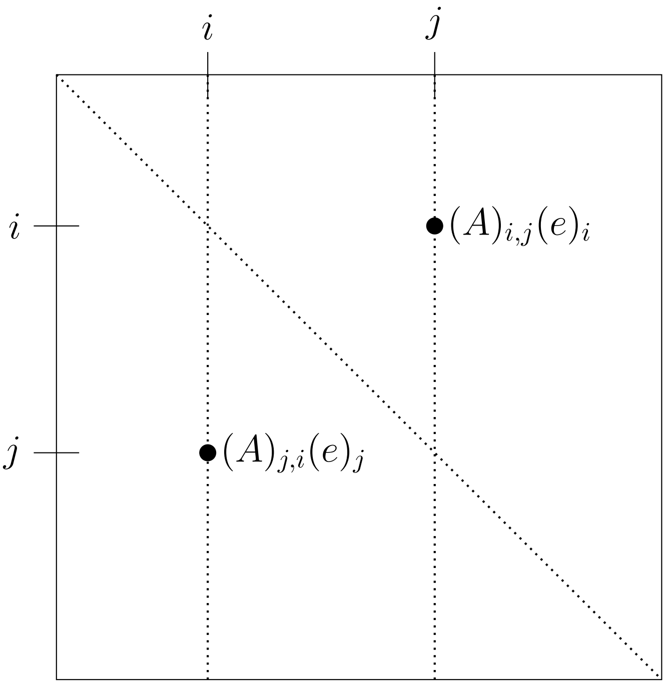


Figure 6.9: The structure of matrix A

□

From the observation and $\|e\|_1 \ll \|e\|_0$ we conclude

$$\sum_i \left(\frac{1}{2} \sum_j \left(-(A)_{i,j}((e)_i - (e)_j)^2 \right) + \left(\sum_j (A)_{i,j} \right) (e)_i^2 \right) \ll \sum_i (A)_{i,i}(e)_i^2.$$

Assuming row sum zero in all rows (true if there is no zero order term) then *on average* for every i (sufficient condition):

$$\sum_j \left(-(A)_{i,j}((e)_i - (e)_j)^2 \right) \ll (A)_{i,i}(e)_i^2$$

$$\iff \sum_{j \neq i} \frac{|(A)_{i,j}|}{(A)_{i,i}} \frac{((e)_i - (e)_j)^2}{(e)_i^2} \ll 1.$$

In the sum the term for $i = j$ is zero and since A is a M-matrix $(A)_{i,j} < 0$ for $i \neq j$. Therefore all terms are nonnegative and

$$\frac{|(A)_{i,j}|}{(A)_{i,i}} = O(1) \implies \frac{|(e)_i - (e)_j|}{|(e)_i|} \text{ small, and}$$

$$\frac{|(e)_i - (e)_j|}{|(e)_i|} \text{ large} \implies \frac{|(A)_{i,j}|}{(A)_{i,i}} \text{ small.}$$

From this follows: if $\frac{|(A)_{i,j}|}{(A)_{i,i}} = O(1)$ the errors in unknown i and j are about equal and $(e)_i$ can be interpolated from $(e)_j$.

Algebraic Multigrid (AMG) Algorithm

1. Given $A_h \in \mathbb{R}^{I_h \times I_h}$ a symmetric positive-definite M-matrix. Set

$$S_i := \{j \in I_h : j \neq i \text{ and } \frac{|(A)_{i,j}|}{(A)_{i,i}} \geq \alpha\}$$

   with, for example, $\alpha \in [\frac{1}{4}, \frac{1}{2}]$. S_i is the set of strongly connected neighbors of $i \in I_h$.

2. Partition $I_h = I_H \cup (I_h \setminus I_H) = C \cup F$ such that
   - $\forall i \in F\ \forall j \in S_i : j \in C \vee (\exists k \in S_j : k \in C)$.
     Direct coupling: $S_i \cap C \neq \emptyset$.
   - $|I_H|$ is not too large, preferably $\frac{|I_H|}{|I_h|} \approx 2^{-d}$.

3. Define interpolation as

$$(R_H^T e)_i = \begin{cases} (e)_i & i \in C \\ \sum_j \omega_{ij} (e)_j & i \in F \end{cases}$$

   where $\omega_{ij} := \frac{(A)_{i,j}}{\sum_{k \in S_i \cap C} (A)_{i,k}}$.

4. Set $A_H := R_H A_h R_H^T$ and proceed recursively.

A further objective is that A_H is as sparse as A_h. This is ensured heuristically in the partitioning strategy in step 2.

Agglomeration-based AMG

1. As above.

2. Build partitioning

$$I_h = \bigcup_{i=1}^{N} C_i, \qquad C_i \cap C_j = \emptyset \quad \forall i \neq j$$

   such that for every $1 \leq i \leq N$, $C_i = \{\alpha_1, \ldots, \alpha_{|C_i|}\}$ there exists a chain

$$\alpha_{j_1}, \alpha_{j_2}, \ldots, \alpha_{j_{|C_i|}}$$

   such that $\alpha_{j_{k+1}} \in S_{\alpha_{j_k}}$, $k < |C_i|$.

   Set $I_H := \{1, \ldots, N\}$ and $m : I_h \to I_H$ as $m(i) = j \iff i \in C_j$.

3. Define piecewise constant interpolation
$$(R_H^T e)_i = (e)_{m(i)}.$$
4. Set $A_H := R_H A_h R_H^T$.

Again it needs to be ensured that A_H is as sparse as A_h and it needs to be used within AMLI cycle or with coarse grid extrapolation
$$x_h^{new} = x_h^{old} + \omega R_H^T A_H^{-1} R_H (b_h - A_h x_h^{old}).$$

Table 6.1: Iteration numbers for various domain decomposition and multigrid methods. The Laplace equation with Dirichlet boundary conditions and C^∞-solution is solved in a square domain with Q_1 finite elements. Weak scaling was employed with $768^2 = 589824$ elements per processor on the finest grid. The coarse grid was $3^2 = 9$ elements per processor. We compare the single grid additive Schwarz method (SASM), the single grid multiplicative Schwarz method (SMSM), the two-grid additive Schwarz method, the multilevel diagonal scaling method (additive multigrid with one step Jacobi smoothing) and the multigrid method with one step of hybrid, symmetric Gauß-Seidel as pre- and post-smoother. All domain decomposition methods used exact subdomain solves with SuperLU. The numbers behind the domain decomposition acronyms denote the overlap in mesh cells on the finest grid. Residual reduction was 10^{-6} with random initial guess.

$\sqrt{P}$	1	2	3	4	5	8	10	15	16	20	32
SASM 1	1	67	83	116	131	186	221	310	328	391	-
SASM 2	-	48	58	81	94	133	156	221	235	278	-
SASM 4	-	34	43	58	68	95	112	154	168	200	273
SASM 8	-	26	31	44	48	69	81	114	120	142	196
SMSM 1	1	37		68		114			227		
SMSM 2	-	30		49		81			197		
SMSM 4	-	23		40		60			109		
SMSM 8	-	17		29		45			78		
TASM 1	1	36	38	39	39	38	37	37	37	36	36
TASM 2	-	27	28	28	28	27	27	27	27	27	27
TASM 4	-	20	21	21	20	20	20	20	20	20	20
TASM 8	-	15	16	16	16	15	15	15	15	15	15
MDS	15	15	15	15	15	15	15	15	15	15	15
MGC	3	3	3	3	3	3	3	3	3	3	3

Table 6.2: Total computation time in seconds for various domain decomposition and multigrid methods. This table corresponds to the iteration numbers given in Table 6.1. Computations were done on a cluster consisting of 32 nodes with four 8 Core AMD Opteron 6212 processors at 2.6 GHz connected by an infiniband network (40G QDR). The time given does *not* include the time for the factorization of the subdomain problem. *Note that for the multigrid methods we provide two digits after the comma!*

$\sqrt{P}$	1	2	3	4	5	8	10	15	16	20	32
SASM 1		212.8	330.9	407.1	708.7	1105.2	1460.7	2073.5	2176.9	2616.0	-
SASM 2		148.5	237.3	291.6	519.4	806.8	1024.5	1497.6	1589.1	1875.8	-
SASM 4		110.0	177.4	207.5	378.9	585.6	768.2	1062.8	1155.8	1377.7	1886.2
SASM 8		81.7	129.6	160.6	273.3	426.6	554.0	786.9	821.1	971.5	1350.5
TASM 1		116.3	127.4	218.5	220.1	234.3	227.9	261.9	267.0	251.0	266.6
TASM 2		85.6	99.5	162.4	167.5	172.8	174.3	204.9	203.8	202.9	220.4
TASM 4		66.1	71.7	121.2	116.4	127.7	127.7	150.3	153.4	142.3	156.1
TASM 8		48.2	57.6	98.3	99.0	97.2	97.2	116.5	119.9	114.3	123.8
MDS		2.54	2.88	3.22	3.29	3.34	6.78	7.04	7.19	7.53	13.64
MGC		1.77	1.91	2.08	2.12	2.12	3.57	3.62	3.64	3.64	3.80

Chapter 7

Nonoverlapping Domain Decomposition Methods

7.1 Introduction to Iterative Substructuring

We consider the case where Ω is subdivided into two non-overlapping simply connected subdomains, i.e.

$$\bar{\Omega}_1 \cup \bar{\Omega}_2 = \bar{\Omega}, \quad \Omega_1 \cap \Omega_2 = \emptyset.$$

We set

$$\Gamma := \partial\Omega_1 \cap \partial\Omega_2.$$

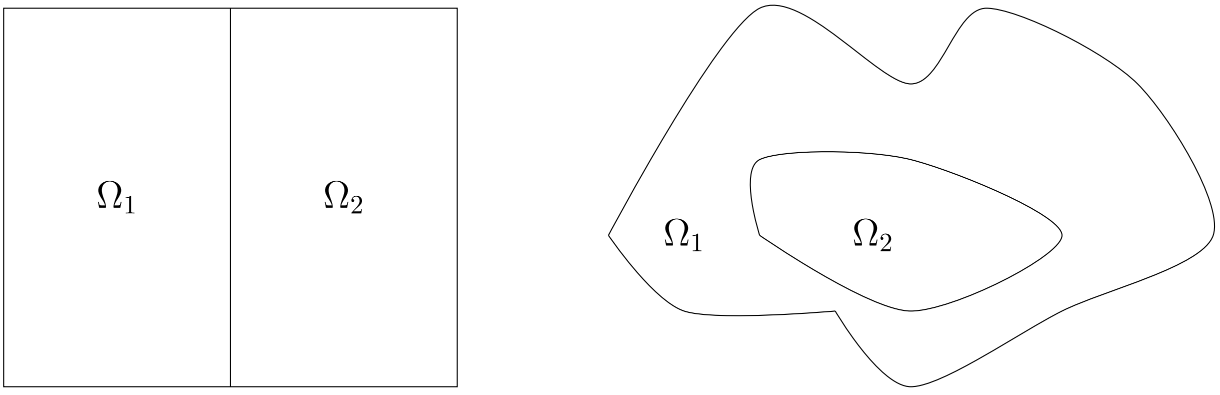


Figure 7.1: model case (left) and general case (right)

Lemma 7.1. The weak formulation of the elliptic problem in Ω is equivalent to the following two subdomain formulations

$$a_1(u_1, v_1) = \int\limits_{\Omega_1} (K\nabla u_1) \cdot \nabla v_1 \, \mathrm{d}x = (f, v_1)_{0,\Omega_1} \quad \forall v_1 \in H_0^1(\Omega_1) \tag{7.1a}$$

$$a_2(u_2, v_2) = \int\limits_{\Omega_2} (K\nabla u_2) \cdot \nabla v_2 \, \mathrm{d}x = (f, v_2)_{0,\Omega_2} \quad \forall v_2 \in H_0^1(\Omega_2) \tag{7.1b}$$

$$u_1 = u_2 \qquad \text{on } \Gamma \tag{7.1c}$$

$$a_1(u_1, R_1\mu) + a_2(u_2, R_2\mu) = (f, R_1\mu)_{0,\Omega_1} + (f, R_2\mu)_{0,\Omega_2} \, \forall \mu \in \Lambda \tag{7.1d}$$

where

$$\Lambda = \{\eta \in H^{\frac{1}{2}}(\Gamma) : \eta = v|_\Gamma \text{ for suitable } v \in H_0^1(\Omega)\}$$

and

$$R_i : \Lambda \to V_i = \{v \in H^1(\Omega_i) : v|_{\partial\Omega_i \cap \partial\Omega} = 0\}$$

are extension operators.

Proof. [Quarteroni and Valli, 1999, Lemma 1.2.1] □

If $\Gamma \cap \partial\Omega = \emptyset$ we have $\Lambda = H^{\frac{1}{2}}(\Gamma)$ and if $\Gamma \cap \partial\Omega \neq \emptyset$, Λ is denoted by $H_{00}^{\frac{1}{2}}(\Gamma)$. Note that (7.1d) is a weak formulation of the *strong* interface condition

$$(K\nabla u_1) \cdot n = (K\nabla u_2) \cdot n \quad \text{ on } \Gamma$$

with n the normal to Γ.

The extension operator R_i is not unique. One particular choice is the so called *harmonic* extension $H_i : \Lambda \to V_i$ given as follows:

$$H_i\lambda = R_i\lambda + w$$

with

$$w \in H_0^1(\Omega_i) : a_i(w, v) = -a(R_i\lambda, v) \quad \forall v \in H_0^1(\Omega_i).$$

This means that $u_i = H_i\lambda$ solves the weak formulation of

$$\begin{aligned} -\nabla \cdot (K\nabla u_i) &= 0 && \text{in } \Omega_i, \\ u_i &= \lambda && \text{on } \Gamma, \\ u_i &= 0 && \text{on } \partial\Omega_i \cap \partial\Omega. \end{aligned}$$

With this extension one can formulate an operator $\mathscr{S} : \Lambda \to \Lambda'$, the so-called Poincaré-Steklov operator, given by

$$\langle \mathscr{S}\eta, \mu \rangle = \sum_{i=1}^{2} a_i(H_i\eta, H_i\mu) \quad \forall \eta, \mu \in \Lambda. \tag{7.2}$$

It can be shown that $\mathscr{S}$ is symmetric, continuous and coercive, i.e. systems of the form

$$\mathscr{S}\eta = \xi \tag{7.3}$$

have a unique solution.

On the discrete level this procedure can be formulated as follows. Discretize the domain with conforming finite elements such that the interface is resolved by the mesh, as shown in figure 7.2.

Partition the index set into

$$I_h = I_{h,1} \cup I_{h,2} \cup I_{h,\Gamma}$$

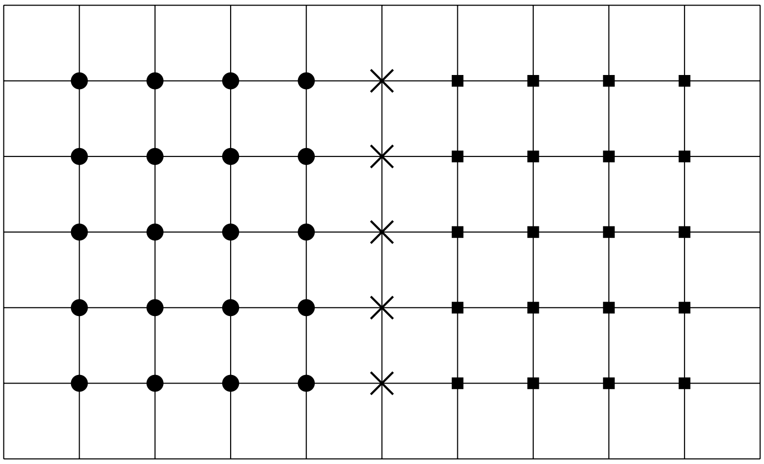


Figure 7.2: Discretization in two subdomains with the mesh resolving the interface

then the linear system exhibits the 3 × 3 block structure

$$\begin{pmatrix} A_{11} & 0 & A_{1\Gamma} \\ 0 & A_{22} & A_{2\Gamma} \\ A_{\Gamma 1} & A_{\Gamma 2} & A_{\Gamma\Gamma} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_\Gamma \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_\Gamma \end{pmatrix}. \tag{7.4}$$

Block-Gauß elimination of the blocks $A_{\Gamma i}$ results in the block triangular system

$$\begin{pmatrix} A_{11} & 0 & A_{1\Gamma} \\ 0 & A_{22} & A_{2\Gamma} \\ 0 & 0 & S \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_\Gamma \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ g \end{pmatrix} \tag{7.5}$$

with

$$\begin{aligned} S &= A_{\Gamma\Gamma} - A_{\Gamma 1} A_{11}^{-1} A_{1\Gamma} - A_{\Gamma 2} A_{22}^{-1} A_{2\Gamma}, \\ g &= b_\Gamma - A_{\Gamma 1} A_{11}^{-1} b_1 - A_{\Gamma 2} A_{22}^{-1} b_2. \end{aligned} \tag{7.6}$$

A general procedure to solve (7.4) would consist of the following steps:

1. Compute S, g.
2. Solve $Sx_\Gamma = g$.
3. Backsolve (7.5) for x_1 and x_2.

The advantages are

- $|I_\Gamma| \ll |I_h|$
- The backsolves can be done in parallel.

A disadvantage is that S is in general not sparse.

It can be shown that the matrix S is actually a discretization of the Poincaré-Steklov operator $\mathscr{S}$. Since $\mathscr{S} : H^{\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$ it is better conditioned than the original system $\mathscr{A} : H_0^1(\Omega) \to H^{-1}(\Omega)$ and $\kappa(S) = O(h^{-1})$ can be shown.

This suggests to solve $Sx_\Gamma = g$ *iteratively* with the conjugate gradient method which requires only matrix-vector products Sx. These products can be done "on the fly" without explicitly assembling S and exploiting (7.6) instead. The corresponding CG-method requires $O(h^{\frac{1}{2}})$ steps and efficient subdomain solvers.

Note that the subdomain solves $A_{ii}x_i = b_i - A_{i\Gamma}x_\Gamma$ involve the harmonic extensions into the subdomains.

A disadvantage of this iterative approach is that the subdomain problems must be solved exactly. Consider for simplicity the Richardson-iteration for the interface problem

$$x_\Gamma^{k+1} = x_\Gamma^k + \omega(g - Sx_\Gamma^k) \tag{7.7}$$

then $g - Sy = 0 \iff y = x_\Gamma$ is a necessary condition for the iteration to converge to x_Γ.

An alternative formulation that overcomes this problem is based on the LDL^T decomposition of A given by

$$A = \underbrace{\begin{pmatrix} I_{1,1} & 0 & 0 \\ 0 & I_{2,2} & 0 \\ A_{\Gamma,1}A_{1,1}^{-1} & A_{\Gamma,2}A_{2,2}^{-2} & I_{\Gamma,\Gamma} \end{pmatrix}}_{=:L} \underbrace{\begin{pmatrix} A_{1,1} & 0 & 0 \\ 0 & A_{2,2} & 0 \\ 0 & 0 & S \end{pmatrix}}_{=:D} \underbrace{\begin{pmatrix} I_{1,1} & 0 & A_{1,1}^{-1}A_{1,\Gamma} \\ 0 & I_{2,2} & A_{2,2}^{-1}A_{1,\Gamma} \\ 0 & 0 & I_{\Gamma,\Gamma} \end{pmatrix}}_{=L^T}. \tag{7.8}$$

Now iteratively solve the original $Ax = b$ by using an approximation of A from (7.7) as a preconditioner where $A_{1,1}^{-1}$, $A_{2,2}^{-1}$ and S are replaced by approximations.

Remark 7.2. Setting

$$\tilde{D} := \begin{pmatrix} A_{1,1} & 0 & 0 \\ 0 & A_{2,2} & 0 \\ 0 & 0 & I_{\Gamma,\Gamma} \end{pmatrix} \quad \text{and} \quad \tilde{A} := L\tilde{D}L^T$$

yields the same iterates as (7.7) for the interface unknowns x_Γ.

Proof. Straightforward calculation for the iteration matrix $I - \omega\tilde{A}^{-1}A$. □

7.2 Two Subdomain Preconditioners

The iterative solution of the Schur complement system is not efficient enough for larger problems due to the growth in condition number. Therefore, preconditioners for the Schur complement are required. In this section some preconditioners for the case of two subdomains are introduced.

J-Operator

For the model problem $-\Delta u = f$ in $\Omega = (0,2) \times (0,1)$, discretized on a structured quadrilateral mesh with $h = (n+1)^{-1}$ the Schur complement can be explicitly diagonalized

$$S = F\Lambda F$$

where $F \in \mathbb{R}^{n \times n}$ is

$$(F)_{i,j} = \sqrt{\frac{2}{n+1}} \sin\left(\frac{ij\pi}{n+1}\right), \qquad i,j \in \{1,\dots,n\}$$

and $\Lambda \in \mathbb{R}^{n \times n}$ can be *approximated* by

$$\Lambda \approx \Sigma, \quad (\Sigma)_{i,i} = 2\sin\left(\frac{i\pi}{2(n+1)}\right).$$

It can be shown that the preconditioner

$$J := F\Sigma^{-1}F$$

is spectrally equivalent to S, for details see Chan and Mathew [1994].

In case of non-uniform grids mappings to/from the model region and mesh are used. The quality of the preconditioner then strongly depends on the mesh and the form of the domain.

The multiplication with F can be done with $O(n \log n)$ operations using FFT.

Neumann-Dirichlet Preconditioner

The matrix block $A_{\Gamma,\Gamma}$ from (7.4) can be split into

$$(A_{\Gamma,\Gamma})_{i,j} = a(\phi_j,\phi_i) = a_1(\phi_j,\phi_i) + a_2(\phi_j,\phi_i) = (A^{(1)}_{\Gamma,\Gamma})_{i,j} + (A^{(2)}_{\Gamma,\Gamma})_{i,j}, \quad i,j \in I_\Gamma,$$

which gives rise to a splitting of the Schur complement S into

$$S = A^{(1)}_{\Gamma,\Gamma} - A_{\Gamma,1}A^{-1}_{1,1}A_{1,\Gamma} + A^{(2)}_{\Gamma,\Gamma} - A_{\Gamma,2}A^{-1}_{2,2}A_{2,\Gamma} =: S^{(1)} + S^{(2)}.$$

In addition, $S^{(i)}$ occurs as the Schur complement of the 2×2 block matrix

$$A^{(i)} = \begin{pmatrix} A_{i,i} & A_{i,\Gamma} \\ A_{\Gamma,i} & A^{(i)}_{\Gamma,\Gamma} \end{pmatrix} \tag{7.9}$$

when eliminating the block $A_{\Gamma,i}$. Note that $A^{(i)}$ is a discretization of the variational problem with Neumann conditions on Γ.

The Neumann-Dirichlet preconditioner is based on the idea that in the case of symmetry of the domain, K, and mesh with respect to Γ, we have

$$S^{(1)} = S^{(2)} \implies S = S^{(1)} + S^{(2)} = 2S^{(1)}.$$

In that case $B_{ND} = {S^{(1)}}^{-1}$ is a spectrally equivalent preconditioner. Since S is not directly available we apply the CG method directly to the (right) preconditioned system $S{S^{(1)}}^{-1}$. This amounts to be able to compute matrix-vector products

$$S{S^{(1)}}^{-1}x = (S^{(1)} + S^{(2)}){S^{(1)}}^{-1}x = x + S^{(2)}{S^{(1)}}^{-1}x.$$

$y = S^{(2)}{S^{(1)}}^{-1}x$ can be computed in two steps:

1. Compute $v = {S^{(1)}}^{-1}x$.
2. Multiply $y = S^{(2)}v$.

Realization of Step 1: For $A^{(1)}$ from (7.9) we have the LDL^T decomposition:

$$A^{(1)} = \begin{pmatrix} I & 0 \\ A_{\Gamma,1}A_{1,1}^{-1} & I \end{pmatrix} \begin{pmatrix} A_{1,1} & 0 \\ 0 & S^{(1)} \end{pmatrix} \begin{pmatrix} I & A_{1,1}^{-1}A_{1,\Gamma} \\ 0 & I \end{pmatrix}$$

from which the following representation of the inverse ${A^{(1)}}^{-1}$ can be obtained:

$${A^{(1)}}^{-1} = \begin{pmatrix} A_{1,1}^{-1} + A_{1,1}^{-1}A_{1,\Gamma}{S^{(1)}}^{-1}A_{\Gamma,1}A_{1,1}^{-1} & -A_{1,1}^{-1}A_{1,\Gamma}{S^{(1)}}^{-1} \\ -{S^{(1)}}^{-1}A_{\Gamma,1}A_{1,1}^{-1} & {S^{(1)}}^{-1} \end{pmatrix}$$

with $R^{(1)} : \mathbb{R}^{I_1 \cup I_\Gamma} \to \mathbb{R}^{I_\Gamma}$ the usual restriction $(R^{(1)}x)_i = (x)_i \; \forall i \in I_\Gamma$, we obtain

$$\begin{aligned} R^{(1)}{A^{(1)}}^{-1}{R^{(1)}}^T &= \begin{pmatrix} 0 & I \end{pmatrix} \begin{pmatrix} * & * \\ * & {S^{(1)}}^{-1} \end{pmatrix} \begin{pmatrix} 0 \\ I \end{pmatrix} \\ &= \begin{pmatrix} 0 & I \end{pmatrix} \begin{pmatrix} * \\ {S^{(1)}}^{-1} \end{pmatrix} \\ &= {S^{(1)}}^{-1}. \end{aligned}$$

So we can complete step 1 by

$$v = {S^{(1)}}^{-1}x = R^{(1)}{A^{(1)}}^{-1}{R^{(1)}}^T x$$

which involves the solution of a linear system of the form

$$A^{(1)}w = {R^{(1)}}^T x$$

involving the *Neumann data* x.

Realization of step 2:

$$y = S^{(2)}v = (A^{(2)}_{\Gamma,\Gamma} - A_{\Gamma,2}A^{-1}_{2,2}A_{2,\Gamma})v$$

involves the solution of the linear system

$$A_{2,2}z = A_{2,\Gamma}v$$

involving Dirichlet data on Γ.

Comments:

- Left preconditioning can be done in the same way and leads to the Dirichlet-Neumann preconditioner.
- Convergence is independent of h, but depends on how well the symmetry condition is satisfied (domains, coefficients).
- One iteration of the CG method to the preconditioned system obviously requires *two* subdomain solves that need to be executed *sequentially*.

Neumann-Neumann Preconditioner

Assuming again that $S^{(1)} = S^{(2)}$, $B_{NN} := {S^{(1)}}^{-1} + {S^{(2)}}^{-1}$ is a good preconditioner because

$$B_{NN}S = ({S^{(1)}}^{-1} + {S^{(2)}}^{-1})(S^{(1)} + S^{(2)}) = I + {S^{(1)}}^{-1}S^{(2)} + {S^{(2)}}^{-1}S^{(1)} + I \approx 4I.$$

The application of this preconditioner in the way described above requires

- the solution of two Neumann and two Dirichlet problems,
- but two problems can be solved in parallel at a time.

7.3 Many Subdomains

Let Ω be subdivided into p non-overlapping subdomains of each diameter $O(H)$. The interface in two space dimensions

$$\Gamma = \bigcup_{i=1}^{p} \partial\Omega_i \cap \Omega = \bigcup_{i,j} E_{i,j} \cup V$$

consists of edges $E_{i,j} = \partial\Omega_i \cap \partial\Omega_j$ and vertices $v_k \in V$ where more than two subdomains meet. The $v_k \in V$ are denoted as "cross points".

The variational problem is discretized using conforming finite elements of lowest order, where the mesh resolves the interface Γ. Then the index set can be partitioned as

$$I_h = I_I \cup I_\Gamma$$

into interior (I_I) and interface (I_Γ) degrees of freedoms. This induces a 2×2 block structure on the FE system in the form

$$\begin{pmatrix} A_{I,I} & A_{I,\Gamma} \\ A_{\Gamma,I} & A_{\Gamma,\Gamma} \end{pmatrix} \begin{pmatrix} x_I \\ x_\Gamma \end{pmatrix} = \begin{pmatrix} b_I \\ b_\Gamma \end{pmatrix}$$

where

$$A_{I,I} = \begin{pmatrix} A_{1,1} & & 0 \\ & \ddots & \\ 0 & & A_{p,p} \end{pmatrix}, \quad A_{\Gamma,I} = \begin{pmatrix} A_{\Gamma,1} & \dots & A_{\Gamma,p} \end{pmatrix}, \quad A_{I,\Gamma} = A_{\Gamma,I}^T.$$

Eliminating the $A_{\Gamma,I}$ block results in the Schur complement system

$$Sx_\Gamma = g$$

with

$$\begin{aligned} S &= A_{\Gamma,\Gamma} - \sum_{i=1}^{p} A_{\Gamma,i} A_{i,i}^{-1} A_{i,\Gamma}, \\ g &= b_\Gamma - \sum_{i=1}^{p} A_{\Gamma,i} A_{i,i}^{-1} b_i. \end{aligned}$$

7.4 Hierarchical Basis for the Schur Complement System

As pointed out in chapter 6, the hierarchical basis method is an additive subspace correction method based on the direct sum

$$\Psi^h = \Phi^0 \cup \bigcup_{l=1}^{L} \bigcup_{i \in I_l \setminus I_{l-1}} \{\phi_i^l\}.$$

Setting $\psi_i := \phi_i^l$ when $i \in I_0$ for $l = 0$ or $i \in I_l \setminus I_{l-1}$ for $l > 0$ we observe $V^h = V^L = \operatorname{span} \Phi^L = \operatorname{span} \Psi^h$.

The linear system

$$\hat{A}\hat{x} = \hat{b}$$

with

$$(\hat{A})_{i,j} = a(\psi_j, \psi_i), \qquad (\hat{b})_i = l(\psi_i)$$

has the condition number

$$\kappa(\hat{A}) = O\left(H^{-2}\left(1+\log\frac{H}{h}\right)^2\right),$$

see Smith et al. [1996].

Using the simple matrix

$$(\hat{D})_{i,j} = \begin{cases} (\hat{A})_{i,j} & i = j \vee i,j \in I_0 \\ 0 & \text{otherwise} \end{cases}$$

results in a preconditioned system with condition number

$$\kappa(\hat{D}^{-1}A) = O\left(\left(1+\log\frac{H}{h}\right)^2\right)$$

for $d = 2$.

Now we review how this method can be implemented with optimal complexity. Since $\operatorname{span}\Psi^h = \operatorname{span}\Phi^L$ there exist coefficients ω_{ij} such that

$$\psi_i = \sum_{j\in I_L} \omega_{ij}\phi_j^L.$$

Then

$$u_h = \sum_{i\in I_L}(\hat{x})_i\psi_i = \sum_{i\in I_L}(\hat{x})_i\left(\sum_{j\in I_L}\omega_{ij}\phi_j^L\right) = \sum_{j\in I_L}\underbrace{\left(\sum_{i\in I_L}\omega_{ij}(\hat{x})_i\right)}_{=:(x)_j}\phi_j^L.$$

By introducing $H \in \mathbb{R}^{I_L\times I_L}$ with $(H)_{i,j} := \omega_{ji}$,

$$x = H\hat{x}$$

transforms the coefficients w.r.t. the hierarchical basis into the coefficients w.r.t. the standard Lagrange basis. The evaluation of a linear functional is transformed by

$$(\hat{b})_i = l(\psi_i) = l(\sum_{j\in I_L}\omega_{ij}\phi_j^L) = \sum_{j\in I_L}\underbrace{\omega_{ij}}_{(H^T)_{i,j}}\underbrace{l(\phi_j^L)}_{=(b)_j} = (H^Tb)_i.$$

Finally, the stiffness matrix in the hierarchical basis is related to the original matrix by

$$\begin{aligned}(\hat{A})_{i,j} &= a(\psi_j, \psi_i) \\ &= a(\sum_{r\in I_L} \omega_{jr}\phi_r^L, \sum_{s\in I_L} \omega_{is}\phi_s^L) \\ &= \sum_{s\in I_L} \omega_{is} \sum_{r\in I_L} \underbrace{a(\phi_r^L, \phi_s^L)}_{=(A)_{s,r}} \underbrace{\omega_{jr}}_{(H)_{r,j}} \\ &= \sum_{s\in I_L} \underbrace{\omega_{is}}_{(H^T)_{i,s}} (AH)_{s,j} \\ &= (H^T AH)_{i,j}.\end{aligned}$$

Therefore

$$\hat{A}\hat{x} = \hat{b} \iff H^T AH\hat{x} = H^T b.$$

One iteration of a linear iterative method using $\hat{D}$ as approximate inverse is then

$$\hat{x}^{k+1} = \hat{x}^k + \hat{D}^{-1}(\hat{b} - \hat{A}\hat{x}^k).$$

Transforming from hierarchical to nodal basis reads

$$\begin{aligned}x^{k+1} &= H\hat{x}^{k+1} \\ &= H\hat{x}^k + H\hat{D}^{-1}(\hat{b} - \hat{A}\hat{x}^k) \\ &= H\hat{x}^k + H\hat{D}^{-1}(H^T b - H^T AH\hat{x}^k) \\ &= x^k + H\hat{D}^{-1}H^T(b - Ax^k).\end{aligned}$$

The multiplication with H, H^T can be realized with $O(n_L)$ steps and is very similar to a multigrid prolongation (H) and restriction (H^T).

Application to the Schur complement system

Partition the index set I_L into interior and interface unknowns:

$$I_L = I_I \cup I_\Gamma.$$

This implies a corresponding 2×2 block structure on A, $\hat{A}$ and H such that

$$\begin{pmatrix} \hat{A}_{I,I} & \hat{A}_{I,\Gamma} \\ \hat{A}_{\Gamma,I} & \hat{A}_{\Gamma,\Gamma} \end{pmatrix} = \begin{pmatrix} H_{I,I}^T & 0 \\ H_{I,\Gamma}^T & H_{\Gamma,\Gamma}^T \end{pmatrix} \begin{pmatrix} A_{I,I} & A_{I,\Gamma} \\ A_{\Gamma,I} & A_{\Gamma,\Gamma} \end{pmatrix} \begin{pmatrix} H_{I,I} & H_{I,\Gamma} \\ 0 & H_{\Gamma,\Gamma} \end{pmatrix}.$$

This is because in the transformation from hierarchical basis to nodal basis (multiplication by H) the coefficients on the interface do not depend on interior coefficients (as in multigrid prolongation). The 2×2 structure of $\hat{D}$ is then

$$\hat{D} = \begin{pmatrix} \hat{D}_{I,I} & 0 \\ 0 & \hat{D}_{\Gamma,\Gamma} \end{pmatrix}$$

with $\hat{D}_{I,I}$ a diagonal matrix and $\hat{D}_{\Gamma,\Gamma}$ containing the coarse grid system.

Observation 7.3. Let A be a 2×2 block matrix w.r.t. the partitioning $I = I_I \cup I_\Gamma$ of the index set. By $\mathrm{Schur}(A) := A_{\Gamma,\Gamma} - A_{\Gamma,I}A_{I,I}^{-1}A_{I,\Gamma}$ we denote the Schur complement w.r.t. $A_{\Gamma,\Gamma}$. Furthermore let

$$T = \begin{pmatrix} T_{I,I} & T_{I,\Gamma} \\ 0 & T_{\Gamma,\Gamma} \end{pmatrix}$$

be an upper triangular block matrix. Then

$$\mathrm{Schur}(T^TAT) = T_{\Gamma,\Gamma}^T \,\mathrm{Schur}(A) T_{\Gamma,\Gamma}.$$

Proof. Straightforward calculation. □

With respect to the hierarchical basis we solve the preconditioned system

$$\hat{D}^{-1}H^TAH\hat{x} = \hat{D}^{-1}H^Tb,$$

which we can symmetrize to

$$\hat{D}^{-\frac{1}{2}}H^TAH\hat{D}^{-\frac{1}{2}}\hat{y} = \hat{D}^{-\frac{1}{2}}H^Tb \tag{7.10}$$

by setting $\hat{x} = \hat{D}^{-\frac{1}{2}}\hat{y}$ and multiplying by $\hat{D}^{\frac{1}{2}}$ from the left. Using the block decomposition and lemma 7.3 we obtain

$$\mathrm{Schur}(\hat{D}^{-\frac{1}{2}}H^TAH\hat{D}^{-\frac{1}{2}}) = \hat{D}_{\Gamma,\Gamma}^{-\frac{1}{2}}H_{\Gamma,\Gamma}^T \,\mathrm{Schur}(A) H_{\Gamma,\Gamma}\hat{D}_{\Gamma,\Gamma}^{-\frac{1}{2}}.$$

Therefore (7.10) can be interpreted as a preconditioned Schur complement system

$$\hat{D}_{\Gamma,\Gamma}^{-\frac{1}{2}}H_{\Gamma,\Gamma}^T S H_{\Gamma,\Gamma}\hat{D}_{\Gamma,\Gamma}^{-\frac{1}{2}}\hat{y}_\Gamma = \hat{D}_{\Gamma,\Gamma}^{-\frac{1}{2}}H_{\Gamma,\Gamma}^T g.$$

Undoing the symmetrizing transformation by $\hat{y}_\Gamma = \hat{D}_{\Gamma,\Gamma}^{\frac{1}{2}}\hat{x}_\Gamma$ we obtain

$$\hat{D}_{\Gamma,\Gamma}^{-1}H_{\Gamma,\Gamma}^T S H_{\Gamma,\Gamma}\hat{x}_\Gamma = \hat{D}_{\Gamma,\Gamma}^{-1}H_{\Gamma,\Gamma}^T g.$$

With the transformation $x_\Gamma = H_{\Gamma,\Gamma}\hat{x}_\Gamma$ we obtain an iteration for the Schur complement system in the nodal basis:

$$x_\Gamma^{k+1} = x_\Gamma^k + H_{\Gamma,\Gamma}D_{\Gamma,\Gamma}^{-1}H_{\Gamma,\Gamma}^T(g - Sx_\Gamma^k),$$

i.e. the preconditioner is $B_{HB} = H_{\Gamma,\Gamma}D_{\Gamma,\Gamma}^{-1}H_{\Gamma,\Gamma}^T$.

In order to estimate the condition number we can utilize the result for the original hierarchical basis method.

Lemma 7.4. Let A be a symmetric and positive definite matrix with 2×2 block structure as introduced above. Then

$$\kappa(\mathrm{Schur}(A)) \leq \kappa(A).$$

Proof. Set $S := \mathrm{Schur}(A)$. Any $x \in \mathbb{R}^{I_L}$ can be decomposed as

$$x = \begin{pmatrix} x_I \\ x_\Gamma \end{pmatrix} = \begin{pmatrix} Ex_\Gamma \\ x_\Gamma \end{pmatrix} + \begin{pmatrix} x_I - Ex_\Gamma \\ 0 \end{pmatrix} =: x' + x'',$$

where $E := -A_{I,I}^{-1} A_{I,\Gamma}$ is the discrete harmonic extension.

Then

$$\begin{aligned}
\langle x, Ax \rangle =& \langle x' + x'', A(x' + x'') \rangle \\
=& \langle x', Ax' \rangle + 2\langle x'', Ax' \rangle + \langle x'', Ax'' \rangle \\
=& \langle \begin{pmatrix} Ex_\Gamma \\ x_\Gamma \end{pmatrix}, \begin{pmatrix} 0 \\ Sx_\Gamma \end{pmatrix} \rangle \\
& + 2\langle \begin{pmatrix} x_I - Ex_\Gamma \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ Sx_\Gamma \end{pmatrix} \rangle \\
& + \langle \begin{pmatrix} x_I - Ex_\Gamma \\ 0 \end{pmatrix}, \begin{pmatrix} A_{I,I}(x_I - Ex_\Gamma) \\ A_{\Gamma,I}(x_I - Ex_\Gamma) \end{pmatrix} \rangle \\
=& \langle x_\Gamma, Sx_\Gamma \rangle + \langle x_I - Ex_\Gamma, A_{I,I}(x_I - Ex_\Gamma) \rangle.
\end{aligned}$$

So

$$\begin{aligned}
\lambda_{max}(S) &= \sup_{x_\Gamma \neq 0} \frac{\langle x_\Gamma, Sx_\Gamma \rangle}{\langle x_\Gamma, x_\Gamma \rangle} \\
&\leq \sup_{x_\Gamma \neq 0,\, x_I = 0} \frac{\langle x_\Gamma, Sx_\Gamma \rangle + \overbrace{\langle x_I - Ex_\Gamma, A_{I,I}(x_I - Ex_\Gamma) \rangle}^{\geq 0 \text{ since } A_{I,I} \text{ s.p.d.}}}{\langle x_\Gamma, x_\Gamma \rangle + \underbrace{\langle x_I, x_I \rangle}_{=0}} \\
&= \sup_{x = (x_I, x_\Gamma)^T,\, x_\Gamma \neq 0,\, x_I = 0} \frac{\langle x, Ax \rangle}{\langle x, x \rangle} \\
&\leq \sup_{x \neq 0} \frac{\langle x, Ax \rangle}{\langle x, x \rangle} \\
&= \lambda_{max}(A)
\end{aligned}$$

and

$$\begin{aligned}
\lambda_{min}(S) &= \inf_{x_\Gamma \neq 0} \frac{\langle x_\Gamma, S x_\Gamma \rangle}{\langle x_\Gamma, x_\Gamma \rangle} \\
&\geq \inf_{x_\Gamma \neq 0,\, x_I = E x_\Gamma} \frac{\langle x_\Gamma, S x_\Gamma \rangle + \overbrace{\langle x_I - E x_\Gamma, A_{I,I}(x_I - E x_\Gamma)\rangle}^{=0}}{\langle x_\Gamma, x_\Gamma \rangle + \underbrace{\langle x_I, x_I \rangle}_{\geq 0}} \\
&\geq \inf_{x \neq 0} \frac{\langle x, Ax \rangle}{\langle x, x \rangle} \\
&= \lambda_{min}(A).
\end{aligned}$$

□

Theorem 7.5. The Schur complement system preconditioned by the hierarchical basis method has condition number

$$\kappa(\hat{D}_{\Gamma,\Gamma}^{-1} \hat{S}) = \kappa(\hat{D}_{\Gamma,\Gamma}^{-1} H_{\Gamma,\Gamma}^T S H_{\Gamma,\Gamma}) = O((1 + \log\left(\frac{H}{h}\right))^2).$$

Proof. Symmetrize and apply lemma 7.4:

$$\begin{aligned}
\kappa(\hat{D}_{\Gamma,\Gamma}^{-1} \hat{S}) &= \kappa(\hat{D}_{\Gamma,\Gamma}^{-\frac{1}{2}} H_{\Gamma,\Gamma}^T S H_{\Gamma,\Gamma} \hat{D}_{\Gamma,\Gamma}^{-\frac{1}{2}}) \\
&= \kappa(\text{Schur}(\hat{D}^{-\frac{1}{2}} H^T A H \hat{D}^{-\frac{1}{2}})) && \text{(Obs. 7.3)} \\
&\leq \kappa(\hat{D}^{-\frac{1}{2}} H^T A H \hat{D}^{-\frac{1}{2}}) && \text{(Lemma 7.4)} \\
&= \kappa(\hat{D}^{-1} \hat{A}) \\
&= O((1 + \log\left(\frac{H}{h}\right))^2) && \text{(Result of Yserentant)}
\end{aligned}$$

□

7.5 Bramble-Pasciak-Schatz Method (BPS)

The Bramble-Pasciak-Schatz is also referred to as BPS method or “iterative substructuring”.

In $d = 2$ we may partition the interface index set as

$$I_\Gamma = I_E \cup I_V = I_{E^1} \cup I_{E^2} \cup \cdots \cup I_{E^{n_E}} \cup I_V,$$

where I_V are the cross points and I_{E^i} are the interiors of the edges. Then the Schur complement S can be block structured accordingly:

$$S = \begin{pmatrix} S_{EE} & S_{EV} \\ S_{VE} & S_{VV} \end{pmatrix} = \begin{pmatrix} S_{E^1E^1} & \dots & S_{E^1E^{n_E}} & S_{E^1V} \\ \vdots & & \vdots & \vdots \\ S_{E^{n_E}E^1} & \dots & S_{E^{n_E}E^{n_E}} & S_{E^{n_E}V} \\ S_{VE^1} & \dots & S_{VE^{n_E}} & S_{VV} \end{pmatrix}.$$

We switch to a partial hierarchical basis. In Observation 7.3 we saw that the basis transformation of the Schur complement is independent from the transformation in the interior which is not considered at all here. As transformation use

$$\bar{H}_{\Gamma\Gamma} = \begin{pmatrix} I_{EE} & \bar{H}_{EV} \\ & I_{VV} \end{pmatrix}$$

where $\bar{H}_{EV}$ is the linear interpolation on an edge and I_{EE}, I_{VV} are identities. Partial hierarchical basis means that

$$\bar{\Psi} = \Phi_0 \cup \bigcup_{k=1}^{n_E} \bigcup_{i \in I_{E^i}} \{\phi_i^L\}.$$

Then the transformed Schur complement is

$$\bar{S} = \bar{H}_{\Gamma\Gamma}^T S \bar{H}_{\Gamma\Gamma} = \begin{pmatrix} S_{EE} & \bar{S}_{EV} \\ \bar{S}_{VE} & \bar{S}_{VV} \end{pmatrix}.$$

As a preconditioner for $\bar{S}$ we may use

$$\bar{D}_{\Gamma\Gamma}^{-1} = \begin{pmatrix} S_{E^1E^1}^{-1} & & & 0 \\ & \ddots & & \\ & & S_{E^{n_E}E^{n_E}}^{-1} & \\ 0 & & & \bar{S}_{VV}^{-1} \end{pmatrix}.$$

Now what is $\bar{S}_{VV}$?

Assume that edges of the subdomains are straight lines, the subdomains have either quadrilateral or triangular shape and that permeability is scalar and constant in each subdomain. Then, the harmonic extension of piecewise linear boundary data is a function in $V_H = \operatorname{span} \Phi_0$, or, in terms of coefficients

$$R_H^T x_V = \begin{pmatrix} -A_{II}^{-1} A_{I\Gamma} \\ I_{\Gamma\Gamma} \end{pmatrix} \bar{H}_{\Gamma\Gamma} R_V^T x_V$$

where

$R_H : \mathbb{R}^{I_h} \to \mathbb{R}^{I_H}$ standard two-grid Schwarz restriction operator
$R_V : \mathbb{R}^{I_\Gamma} \to \mathbb{R}^{I_V}$ is $(R_V x_\Gamma)_i = (x_\Gamma)i \quad \forall i \in I_V \subset I_\Gamma$.

With this, we obtain

$$\begin{aligned} A_H &= R_H A R_H^T \\ &= R_V \bar{H}_{\Gamma\Gamma}^T \begin{pmatrix} -A_{II}^{-1}A_{I\Gamma} & I_{\Gamma\Gamma} \end{pmatrix} \underbrace{A \begin{pmatrix} -A_{II}^{-1}A_{I\Gamma} \\ I_{\Gamma\Gamma} \end{pmatrix}}_{\begin{pmatrix} 0 \\ S \end{pmatrix}} \bar{H}_{\Gamma\Gamma} R_V^T \\ &= R_V \bar{H}_{\Gamma\Gamma}^T S \bar{H}_{\Gamma\Gamma} R_V^T \\ &= R_V \bar{S} R_V^T \\ &= \bar{S}_{VV}. \end{aligned}$$

The BPS preconditioner can be written in the following form, due to the block diagonal form of $D_{\Gamma\Gamma}$:

$$B_{BPS} = \bar{H}_{\Gamma\Gamma} R_V^T A_H^{-1} R_V \bar{H}_{\Gamma\Gamma}^T + \sum_{i=1}^{n_E} R_{E^i}^T S_{E^i E^i}^{-1} R_{E^i}. \tag{7.11}$$

where

$$R_{E_i} : \mathbb{R}^{I_\Gamma} \to \mathbb{R}^{I_{E^i}}, \quad (R_{E^i} x_\Gamma)_j = (x_\Gamma)_j \quad \forall j \in I_{E^i}.$$

Note that $\bar{H}_{\Gamma\Gamma}$ transforms correction to standard basis, $\bar{H}_{\Gamma\Gamma}^T$ transforms the right-hand side into partial hierarchical basis and it holds $R_{E^i} \bar{H}_{\Gamma\Gamma}^T = R_{E^i}$.

Interpretation as a subspace correction method

We assume that $K(x) = k_i I$ for $x \in \Omega_i$ and denote by $\mathscr{P} : \mathbb{R}^{I_h} \to V_h$ the finite element isomorphism

$$\mathscr{P} x = \sum_{i \in I_h} (x)_i \phi_i^h$$

on the fine mesh. Now define the following subspaces of V_h:

$$\begin{aligned} \tilde{V}_h &:= \{u \in V_h : \mathscr{P}^{-1} u = \begin{pmatrix} -A_{II}^{-1} A_{I\Gamma} x_\Gamma \\ x_\Gamma \end{pmatrix}\} \\ \hat{V}_h &:= \{u \in V_h : \mathscr{P}^{-1} u = \begin{pmatrix} x_I \\ 0 \end{pmatrix}\}. \end{aligned}$$

$\tilde{V}_h$ is the subspace of discrete harmonic functions and the two subspaces provide a direct decomposition

$$V_h = \tilde{V}_h \oplus \hat{V}_h. \tag{7.12}$$

Lemma 7.6. We have

1. $a(u,v) = x_\Gamma^T S y_\Gamma$ for $u, v \in \tilde{V}_h$ and

$$u = \mathscr{P}\left(\begin{pmatrix} -A_{II}^{-1}A_{I\Gamma}x_\Gamma \\ x_\Gamma \end{pmatrix}\right), \quad v = \mathscr{P}\left(\begin{pmatrix} -A_{II}^{-1}A_{I\Gamma}y_\Gamma \\ y_\Gamma \end{pmatrix}\right).$$

2. $a(u,v) = 0$ for $u \in \hat{V}_h$, $v \in \tilde{V}_h$.

3. $a(u,v) = x_I^T A_{II} y_I$ for $u, v \in \hat{V}_h$ and

$$u = \mathscr{P}\left(\begin{pmatrix} x_I \\ 0 \end{pmatrix}\right), \quad v = \mathscr{P}\left(\begin{pmatrix} y_I \\ 0 \end{pmatrix}\right).$$

Proof. Just insert:

1. $a(u,v) = \langle A\left(\begin{smallmatrix} -A_{II}^{-1}A_{I\Gamma}x_\Gamma \\ x_\Gamma \end{smallmatrix}\right), \left(\begin{smallmatrix} -A_{II}^{-1}A_{I\Gamma}y_\Gamma \\ y_\Gamma \end{smallmatrix}\right)\rangle = \langle\left(\begin{smallmatrix} 0 \\ Sx_\Gamma \end{smallmatrix}\right), \left(\begin{smallmatrix} * \\ y_\Gamma \end{smallmatrix}\right)\rangle = \langle Sx_\Gamma, y_\Gamma\rangle$.

2. $a(u,v) = \langle\left(\begin{smallmatrix} x_I \\ 0 \end{smallmatrix}\right), A\left(\begin{smallmatrix} -A_{II}^{-1}A_{I\Gamma}y_\Gamma \\ y_\Gamma \end{smallmatrix}\right)\rangle = \langle\left(\begin{smallmatrix} x_I \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 0 \\ Sy_\Gamma \end{smallmatrix}\right)\rangle = 0$.

3. $a(u,v) = \langle A\left(\begin{smallmatrix} x_I \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} y_I \\ 0 \end{smallmatrix}\right)\rangle = \langle A_{II}x_I, y_I\rangle$. □

Due to (7.12) we can split $u, v \in V_h$ into

$$\begin{aligned} u &= \tilde{u} + \hat{u} \qquad & \tilde{u} \in \tilde{V}_h, \hat{u} \in \hat{V}_h \\ v &= \tilde{v} + \hat{v} \qquad & \tilde{v} \in \tilde{V}_h, \hat{v} \in \hat{V}_h \end{aligned}$$

and

$$\begin{aligned} & a(u,v) = l(v) \quad \forall v \in V_h \\ \Longleftrightarrow\ & a(\tilde{u}+\hat{u}, \tilde{v}+\hat{v}) = l(\tilde{v}+\hat{v}) \quad \forall \tilde{v} \in \tilde{V}_h, \hat{v} \in \hat{V}_h \\ \Longleftrightarrow\ & a(\tilde{u},\tilde{v}) + a(\hat{u},\hat{v}) = l(\tilde{v}) + l(\hat{v}) \quad \forall \tilde{v} \in \tilde{V}_h, \hat{v} \in \hat{V}_h \\ \Longleftrightarrow\ & \begin{cases} a(\tilde{u},\tilde{v}) = l(\tilde{v}) \quad \forall \tilde{v} \in \tilde{V}_h \text{ (set } \hat{v} = 0) \\ a(\hat{u},\hat{v}) = l(\hat{v}) \quad \forall \hat{v} \in \hat{V}_h \text{ (set } \tilde{v} = 0). \end{cases} \end{aligned}$$

So, the FE problem in this basis is split into two completely independent subproblems.

Due to Lemma 7.6 (1.) the problem in $\tilde{V}_h$ corresponds algebraically to the Schur complement, i.e. solving the Schur complement problem is equivalent to solve the original FE problem in the space of discrete harmonic functions.

Due to (3.) the problem in $\hat{V}_h$ corresponds to the p independent subdomain solves.

With

$$\begin{aligned} u &= \tilde{u} + \hat{u} \\ &= \mathscr{P}\left(\begin{pmatrix} -A_{II}^{-1}A_{I\Gamma}x_\Gamma \\ x_\Gamma \end{pmatrix} + \begin{pmatrix} A_{II}^{-1}b_I \\ 0 \end{pmatrix}\right) \\ &= \mathscr{P}\left(\begin{pmatrix} A_{II}^{-1}(b_I - A_{I\Gamma}x_\Gamma) \\ x_\gamma \end{pmatrix}\right) \end{aligned}$$

we see that the summation corresponds to the back substitution.

Now the BPS preconditioner. As noted above we have $V_H \subset \tilde{V}_h$, i.e. coarse grid functions are discrete harmonic. From (7.11) we observe that the transformation to partial hierarchic basis is irrelevant for the edges. Therefore, BPS corresponds to an additive Schwarz method corresponding to the direct sum

$$\tilde{V}_h = V_H \oplus \bigoplus_{i=1}^{n_E} \tilde{V}_h^{E_i} \tag{7.13}$$

where

$$\tilde{V}_h^{E_i} = \{u \in V_h : \mathscr{P}^{-1}u = \begin{pmatrix} -A_{II}^{-1}A_{I\Gamma}x_\Gamma \\ x_\Gamma \end{pmatrix} \wedge (x_\Gamma)_j = 0 \quad \forall j \notin I_{E_i}\}.$$

Let us now turn to the convergence proof of the BPS method.

The upper bound is based on a coloring argument as usual:

Lemma 7.7. Assume there exists $c > 0$, independent of n_E, such that

$$I_E = \bigcup_{k=1}^{c} C_k, \quad C_i \cap C_j = \emptyset \, \forall i \neq j$$

and

$$\forall k = 1, \ldots, c \, \forall i, j \in C_k, i \neq j : R_{E^iE^i} S R_{E^jE^j}^T = 0.$$

Then

$$\langle B_{BPS} S x, x\rangle_S \leq (c+1)\langle x, x\rangle_S.$$

Proof. From

$$B_{BPS}S = \bar{H}_{\Gamma\Gamma} R_V^T A_H^{-1} R_V H_{\Gamma\Gamma}^T S + \sum_{i=1}^{n_E} R_{E^i}^T S_{E^iE^i}^{-1} R_{E^i} S$$

we see that the $n_E + 1$ contributions are S-orthogonal projections. Also

$$P_k = \sum_{i \in C_k} R_{E^i}^T S_{E^iE^i}^{-1} R_{E^i} S$$

is S-orthogonal. □

Note that in this method the edges are colored and the number of colors for the model problem is $c = 4$ as in the Schwarz method.

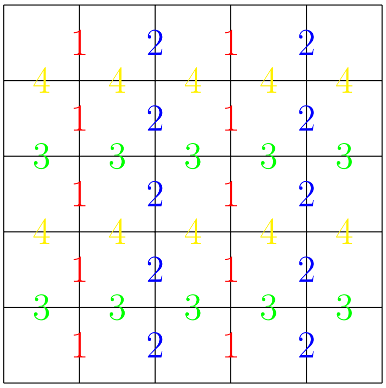


Figure 7.3: Coloring of the edges

The lower bound is based on proving the existence of a stable splitting. The following Lemma plays a crucial role in the proof.

Lemma 7.8. Let u be a piecewise linear FE function on a mesh of size h and a domain Ω of diameter H. Let U be the value of u at any point in Ω. Then there exists $c > 0$ such that

$$\|u - U\|^2_{L^\infty(\Omega)} \leq c \left(1 + \log \frac{H}{h}\right) |u|^2_{1,\Omega}.$$

Proof. See Bramble [1966]. □

Lemma 7.9. Assume the subdomains Ω_i are triangles and that the bilinear form reads

$$a(u, v) = \sum_{i=1}^{p} k_i (\nabla u, \nabla v)_{0,\Omega_i},$$

i.e. the diffusion coefficient $k_i > 0$ is constant on each of the non-overlapping subdomains Ω_i. Then there exists $c > 0$ independent of h, H, p and the k_i and a decomposition $\mathbb{R}^{I_\Gamma} \ni x_\Gamma = R_V^T x_v + \sum_{i=1}^{n_E} R_{E^i}^T x_{E^i}$ such that

$$\langle R_V^T x_V, R_V^T x_V \rangle_S + \sum_{i=1}^{n_E} \langle R_{E^i}^T x_{E^i}, R_{E^i}^T x_{E^i} \rangle_S \leq c' \langle x_\Gamma, x_\Gamma \rangle_S$$

with $c' = c \left(1 + \log \frac{H}{h}\right)^2$.

Proof. This proof follows [Smith et al., 1996, Section 5.3.2].

1. First, we observe that the decomposition is unique due to (7.13). Moreover, due to Lemma 7.6 we have equivalently

$$a(u_V, u_V) + \sum_{i=1}^{n_E} a(u_{E^i}, u_{E^i}) \leq c\, a(u_\Gamma, u_\Gamma) \tag{7.14}$$

where u_V, u_{E^i} are the functions in the space of discrete harmonic extensions $\tilde{V}_h$ that correspond to x_V and x_{E^i}.

2. Let
$$\tilde{V}_h^i := \{v \in H^1(\Omega_i) : \exists w \in \tilde{V}_h : \forall x \in \Omega_i : v(x) = w(x)\}$$
be the restriction of $\tilde{V}_h$ to subdomain Ω_i and $u^i \in \tilde{V}_h^i$ the corresponding restriction of $u \in \tilde{V}_h$.

   Assume we *could* prove that for *all* $\tilde{V}_h^i \ni u^i = u_V^i + \sum_{j=1}^{n_E} u_{E^j}^i$ we have

$$(\nabla u_V^i, \nabla u_V^i)_{0,\Omega_i} + \sum_{j=1}^{n_E} (\nabla u_{E^j}^i, \nabla u_{E^j}^i)_{0,\Omega_i} \leq c^i (\nabla u^i, \nabla u^i)_{0,\Omega} \tag{7.15}$$

   then we have (7.14) with $c = \max_{i=1,\dots,p} c^i$.

   Note that the functions in (7.15) are not necessarily zero at the boundary. In particular, for a so-called “floating subdomain” $\partial\Omega_i \cap \partial\Omega = \emptyset$ we have for u^i a constant function that $\nabla u^i = 0$ and the right-hand side is zero. But in that case the decomposition yields $u_V^i = u^i$, $u_{E^j} = 0$ and the left-hand side is zero as well and (7.15) is satisfied. This is called the “null space property”.

3. Consider now $u^i \in \tilde{V}_h^i$. Since Ω_i is triangular, we have $u_V^i = \mathscr{I}^H u^i|_{\Omega_i}$ the Lagrange interpolation of u^i on the coarse grid. With $u_{max}^i := \sup_{x\in\Omega_i} u^i$, $u_{min}^i := \min_{x\in\Omega_i} u^i$ we estimate

$$\begin{aligned} |u_V^i|_{1,\Omega_i}^2 &\leq c \left(\underbrace{\frac{u_{max}^i - u_{min}^i}{H}}_{\text{estimates } \nabla u_V^i} \right)^2 \cdot \underbrace{H^2}_{\text{integration}} \\ &\leq c(u_{max}^i - u_{min}^i)^2 \\ &\leq c\left(1 + \log\frac{H}{h}\right)^2 |u^i|_{1,\Omega_i}^2 \end{aligned}$$

   In the last step we used Lemma 7.8 with $U = u_{min}^i$.

4. We now need to consider $u^i_{E^j}$, i.e. the discrete harmonic extension of data on edge E^j in subdomain Ω_i. Let us consider a triangular subdomain with edges numbered E^1, E^2 and E^3.

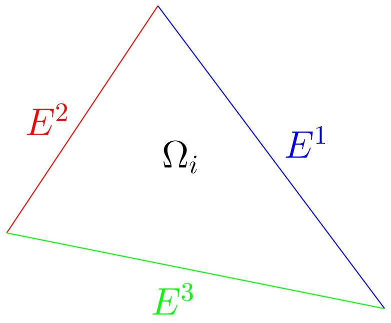


Figure 7.4: Triangular subdomain with numbered edges

u_{E^j} is difficult to handle and we want to estimate a simpler function.
We observe: For all FE functions with the same values on $\partial\Omega_i$ the harmonic extension has minimal energy.
Assume $u^i \in \tilde{V}^i_h$ and $w^i \in V^i_h$ ($= V_h$ restricted to Ω_i) with $w^i = u^i$ on $\partial\Omega_i$. Then $w^i = \tilde{w}^i + \hat{w}^i$ with $\tilde{w}^i \in \tilde{V}^i_h$, $\hat{w}_i \in \hat{V}^i_h$ ($= \hat{V}_h$ restricted to Ω_i). Since $w^i = u^i$ on $\partial\Omega_i$ we have $\tilde{w}^i = u^i$. Therefore, due to Lemma 7.6 with $a(u,v) = (\nabla u, \nabla v)_{0,\Omega_i}$:

$$\begin{aligned}(\nabla w^i, \nabla w^i)_{0,\Omega_i} &= (\nabla u^i + \nabla\hat{w}^i, \nabla u^i + \nabla\hat{w}^i)_{0,\Omega_i}\\ &= (\nabla u^i, \nabla u^i)_{0,\Omega_i} + (\nabla\hat{w}^i, \nabla\hat{w}^i)_{0,\Omega_i}\\ &\geq (\nabla u^i, \nabla u^i)_{0,\Omega_i}.\end{aligned}$$

To estimate the effect of the harmonic extension we define for each edge E^j, $j = 1,2,3$ a *finite element function* on Ω_i as follows:

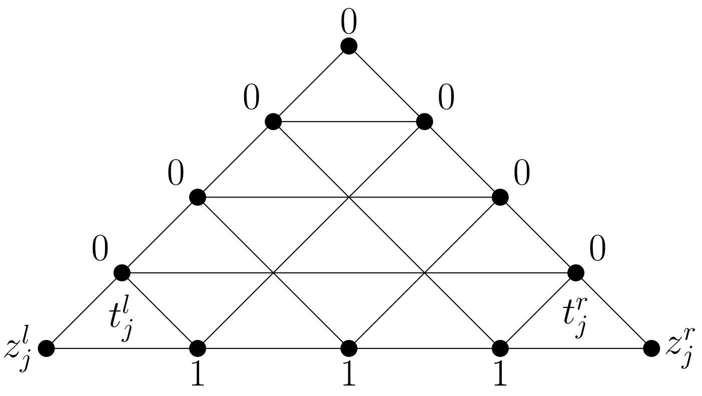


Figure 7.5: Finite element function $\theta_j(s_k)$

$$\theta_j(s_k) = \begin{cases} 1 & s_k \in E^j \text{ (excluding end points)} \\ 0 & s_k \in \partial\Omega_i \backslash E^j \text{ (includes } z_j^1, z_j^2) \\ \xi_l(s_k) & s_k \in \Omega_i^{j,l} \text{(see figure 7.5)} \\ \xi_r(s_k) & s_k \in \Omega_i^{j,r} \text{(see figure 7.5)} \end{cases}$$

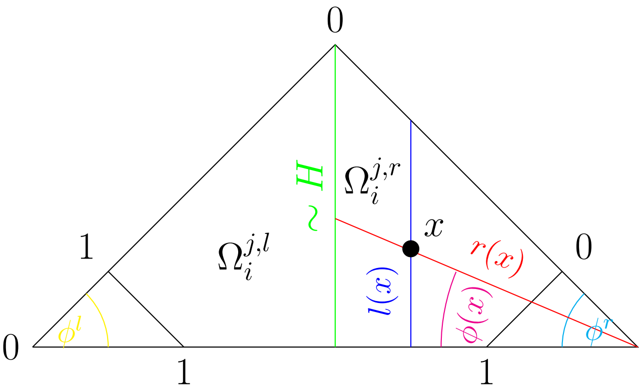


Figure 7.6: Finite element function $\xi_r(x)$

$$\xi_r(x) = 1 - \frac{\phi(x)}{\phi^r(x)}$$
$$\nabla\xi_r(x) \leq \frac{c}{l(x)}$$

Since $l(x) \sim r(x)$ we also have

$$\nabla\xi_r(x) \leq \frac{c}{r(x)}.$$

5. Let

$$I_{\Gamma_i} := \{j \in I_\Gamma : s_j \in \partial\Omega_i\}$$

and define $\mathscr{P}_i : \mathbb{R}^{I_{\Gamma_i}} \to \tilde{V}_h^i$, the FE isomorphism mapping degrees of freedom on $\partial\Omega_i$ to discrete harmonic functions on Ω_i. Then, given any $u^i \in \tilde{V}_h^i$ we wish to analyze the decomposition

$$u^i = u_V^i + \sum_{i=1}^{3} \mathscr{P}_i R_j^{\Gamma_i} \mathscr{P}_i^{-1}(u^i - u_V^i)$$

where $R_j^{\Gamma_i} : \mathbb{R}^{I_{\Gamma_i}} \to \mathbb{R}^{I_{E^j}}$, $(R_j^{\Gamma_i} x)_k = (x)_k \; \forall k \in I_{E^j}$ picks of the DOFs on edge j.

Now due to 4. we estimate $w_j = \mathscr{I}^h(\theta_j(u^i - u^i_V))$, i.e.

$$|\mathscr{P}_i R_j^{\Gamma_i} \mathscr{P}^{-1}(u^i - u^i_V)|_{1,\Omega_i} \le |\mathscr{I}^h(\theta_j(u^i - u^i_V))|_{1,\Omega_i}$$

because they have the same boundary data.

6. First split off the two small triangles next to the end points of edge E^j:

$$\begin{aligned} |w_j|^2_{1,t_j^\alpha} &\le c\left(\frac{u^i_{max} - u^i_{min}}{h}\right)^2 h^2 && \text{(similar to 3.)} \\ &\le c\left(1 + \log\frac{H}{h}\right) |u|^2_{1,\Omega_i} && \text{(Lemma 7.8)} \end{aligned}$$

7. Now for the rest $\mathcal{T}^j := \mathcal{T}(\Omega_i)\backslash\{t_j^1, t_j^2\}$.

$$\begin{aligned} \sum_{t\in\mathcal{T}^j} |w_j|^2_{1,t} &= \sum_{t\in\mathcal{T}^j} |\mathscr{I}^h\left(\theta_j(u^i - u^i_V)\right)|^2_{1,t} \\ &\le c\sum_{t\in\mathcal{T}^j} \|\nabla\left(\theta_j(u^i - u^i_V)\right)\|^2_{0,t} \quad (\theta_j(u^i - u^i_V) \text{ p. w. quadratic}) \\ &\le c\sum_{t\in\mathcal{T}^j} \{\|\theta_j\nabla(u^i - u^i_V)\|^2_{0,t} \\ &\quad + \|(u^i - u^i_V)\nabla\theta_j\|^2_{0,t}\} \end{aligned}$$

For the first term:

$$\begin{aligned} \sum_{t\in\mathcal{T}^j} \|\theta_j\nabla(u^i - u^i_V)\|^2_{0,t} &\le |u^i - u^i_V|^2_{1,\Omega_i} && (\theta_j \le 1, \text{ add } t_j^1, t_j^2) \\ &\le 2|u^i|^2_{1,\Omega_i} + 2|u^i_V|^2_{1,\Omega_i} && \text{(triangle)} \\ &\le c\left(1 + \log\frac{H}{h}\right)^2 |u^i|^2_{1,\Omega_i} && (3.) \end{aligned}$$

and the second term:

$$\begin{aligned} \sum_{t\in\mathcal{T}^j} \|(u^i - u^i_V)\nabla\theta_j\|^2_{0,t} &= \sum_{t\in\mathcal{T}^j} \int_t (u^i - u^i_V)^2 \|\nabla\theta_j(x)\|^2 \,\mathrm{d}x \\ &\le \sum_{t\in\mathcal{T}^j} (u^i_{max} - u^i_{min})^2 \int_t \frac{c}{r^2(x)} \,\mathrm{d}x \\ &= c\,(u^i_{max} - u^i_{min})^2 \int_{\Omega_i\backslash\{t_j^l, t_j^r\}} r^{-2} \,\mathrm{d}x \\ &\le c\left(1 + \log\frac{H}{h}\right)^2 |u^i|^2_{1,\Omega_i}, \end{aligned}$$

where we used again Lemma 7.8 and

$$\int\limits_{\Omega_i\setminus\{t_j^l,t_j^r\}} r^{-2}\,\mathrm{d}x = \int\limits_{\Omega_i^{j,l}\setminus\{t_j^l\}} r^{-2}\,\mathrm{d}x + \int\limits_{\Omega_i^{j,r}\setminus\{t_j^r\}} r^{-2}\,\mathrm{d}x,$$

$$\begin{aligned}\int\limits_{\Omega_i^{j,l}\setminus\{t_j^l\}} r^{-2}\,\mathrm{d}x &= \int\limits_0^{\phi^l}\int\limits_{c_1 h}^{c_2 H} r^{-2} r\,\mathrm{d}r\,\mathrm{d}\phi \\ &= \phi^l\,[\log r]_{c_1 h}^{c_2 H} \\ &\le c\log\left(\frac{H}{h}\right).\end{aligned}$$

8. Combining 6. and 7. yields

$$|w_j|_{1,\Omega_i}^2 \le c\left(1+\log\frac{H}{h}\right)^2 |u^i|_{1,\Omega_i}^2.$$

Together with 5. and 3. we obtain

$$|u_V^i|_{1,\Omega_i}^2 + \sum_{i=1}^3 |\mathscr{P}_i R_j^{\Gamma_i}\mathscr{P}_i^{-1}(u^i - u_V^i)|_{1,\Omega_i}^2 \le c\left(1+\log\frac{H}{h}\right)^2 |u^i|_{1,\Omega_i}^2$$

which is (7.15) from which the desired result follows. □

BPS is historically an interesting method and it can be analyzed relatively easily. In practice, however, it is seldomly used now. The reasons are:

- It needs to be modified to cover 3D applications.
- It requires too many subdomain solves, e.g. 4 solves per subdomain for the model problem with quadrilateral subdomains and the Dirichlet-Neumann method as edge preconditioner.

7.6 Outlook: Balancing Neumann-Neumann and FETI-DP

Notation for many subdomains is extended by:

I_Γ : DoFs on Γ
$I_{\Gamma_i} := \{j \in I_\Gamma : s_j \in \partial\Omega_i \cap \Gamma\} \subset I_\Gamma$ DoFs on Γ being part of $\partial\Omega_i$
$I_0 := \{i \in 1,\dots,p : \partial\Omega_i \cap \partial\Omega_D = \emptyset\} \subset \{1,\dots,p\}$

and restrictions

$$R_{\Gamma_i} : \mathbb{R}^{I_\Gamma} \to \mathbb{R}^{I_{\Gamma_i}}, \quad (R_{\Gamma_i} x)_j = (x)_j \quad \forall j \in I_{\Gamma_i}$$
$$R_{0,i} : \mathbb{R}^{I_\Gamma} \to \mathbb{R}^{I_0}, \quad (R_{0,i} x)_j = \begin{cases} \sum_{k \in I_{\Gamma_i}} (x)_k & j = i \\ 0 & j \neq i. \end{cases}$$

I_{Γ_i} corresponds to *tearing apart the subdomains.*

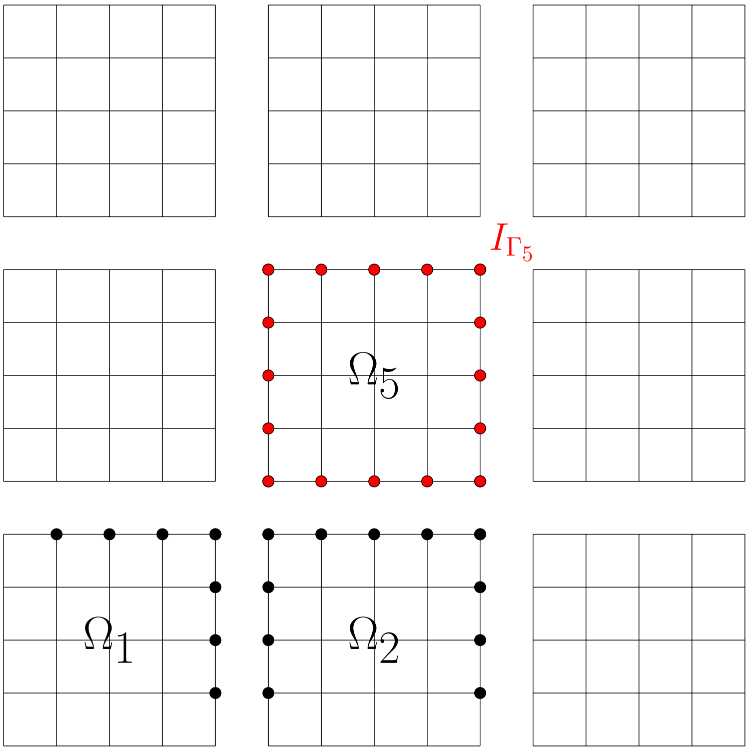


Figure 7.7: Example for I_{Γ_i}

Then we need diagonal scaling matrices per subdomain:
$\forall i = 1, \ldots, p : \quad D^{(i)} : \mathbb{R}^{\pm\Gamma_i} \to \mathbb{R}^{\Gamma_i}$ such that

$$\forall j \in I_\Gamma : \quad \sum_{i \in N(j)} (D^{(i)})_{jj} = 1$$

where $N(j) := \{i \in \{1, \ldots, p\} : j \in I_{\Gamma_i}\}$. Examples for $D^{(i)}$ are:

$$(D^{(i)})_{jj} = \frac{1}{|N(j)|}$$

or

$$(D^{(i)})_{jj} = \frac{k_i^\gamma}{\sum_{l \in N(j)} k_l^\gamma}, \quad \gamma \in [\frac{1}{2}, \infty)$$

and k_i is the diffusion coefficient.

Coarse grid correction (balancing step)

Define $R_0^T : \mathbb{R}^{I_0} \to \mathbb{R}^{I_\Gamma}$ as

$$R_0^T := \sum_{i \in I_0} R_{\Gamma_i}^T D^{(i)} R_{0,i}^T$$

and

$$R_0 := \sum_{i \in I_0} R_{0,i} D^{(i)} R_{\Gamma_i}.$$

Then define the subspace correction on the Schur complement system via

$$S_0 := R_0 S R_0^T$$

and

$$x_\Gamma^{k+1} := x_\Gamma^k + R_0^T S_0^{-1} R_0 (g - S x_\Gamma^k)$$

which has the error propagation

$$e_\Gamma^{k+1} = (I - P_0) e^k, \quad P_0 = R_0^T S_0^{-1} R_0 S.$$

P_0 and $I - P_0$ are S-orthogonal projections.

Subdomain corrections

As before $A^{(i)} = \begin{pmatrix} A_{I,I}^{(i)} & A_{I,\Gamma}^{(i)} \\ A_{\Gamma,I}^{(i)} & A_{\Gamma,\Gamma}^{(i)} \end{pmatrix}$ is the matrix arising from

$$a_i(u,v) = \int_{\Omega_i} (K \nabla u) \cdot \nabla v \, \mathrm{d}x$$

on subdomain Ω_i partitioned w.r.t. I_i (subdomain interior DoFs) and I_{Γ_i}. $A^{(i)}$ has the Schur complement

$$S^{(i)} = A_{\Gamma,\Gamma}^{(i)} - A_{\Gamma,I}^{(i)} {A_{I,I}^{(i)}}^{-1} A_{I,\Gamma}^{(i)}$$

and

$$S = \sum_{i=1}^{p} R_{\Gamma_i}^T S^{(i)} R_{\Gamma_i}.$$

Note that $S^{(i)}$ is singular for $i \in I_0$ and $\ker(S^{(i)}) = \operatorname{span}\{\mathbb{1}_i\}$.

Now a preconditioner for S is constructed by solving local problems with $S^{(i)}$ in the following way:

$$x_\Gamma^{k+1} = x_\Gamma^k + \sum_{i=1}^{p} R_{\Gamma_i}^T D^{(i)} {S^{(i)}}^{-1} D^{(i)} R_{\Gamma_i} (g - S x_\Gamma^k) \tag{7.16}$$

with error propagation

$$e^{k+1} = (I - \sum_{i=1}^{p} P_i)e^k, \quad P_i = R_{\Gamma_i}^T D^{(i)} {S^{(i)}}^{-1} D^{(i)} R_{\Gamma_i} S.$$

Here application of ${S^{(i)}}^{-1} r_i$ is understood as *one* solution of a system

$$S^{(i)} v_i = r_i. \tag{7.17}$$

In order to make v_i well-defined we must ensure that $r_i \in \text{range}(S^{(i)})$.
Then the Balancing Neumann-Neumann (BNN) method is given by

$$E_{BNN} = (I - P_0)(I - \sum_{i=1}^{p} P_i)(I - P_0)$$

which, exploiting $E_{BNN} = I - P_{BNN}$, results in

$$P_{BNN} = P_0 + (I - P_0) \sum_{i=1}^{p} P_i (I - P_0).$$

Note: In practice, since $(I - P_0)^2 = (I - P_0)$ only one coarse grid solve for iteration is needed.

How do we ensure that the local problems (7.17) are solvable?
The following observation is helpful.

Lemma 7.10. Given some matrix $A \in \mathbb{R}^{n \times n}$ with $\dim(\ker(A)) \geq 0$.
Then $Ax = b$ has a solution if and only if

$$\langle b, v \rangle = 0 \quad \forall v \in \ker(A^T).$$

Proof. Set $U = \mathbb{R}^n$ and $V = \text{range}(A) \subset U$. Then

$$\begin{aligned} V^\perp &= \{u \in U : \langle u, v \rangle = 0 \quad \forall v \in V\} \\ &= \{u \in U : \langle u, Aw \rangle = 0 \quad \forall w \in U\} \\ &= \{u \in U : \langle A^T u, w \rangle = 0 \quad \forall w \in U\} \\ &= \{u \in U : A^T u = 0\} \\ &= \ker(A^T) \end{aligned}$$

Then

$$Ax = b \iff \langle Ax, v \rangle = \langle b, v \rangle \quad \forall v \in U$$

split

$$b = b_V + b_{V^\perp}, \quad b_V \in V, \ b_{V^\perp} \in V^\perp.$$

$$\begin{aligned} &\langle Ax, v\rangle = \langle b, v\rangle \qquad \forall v \in U = V \oplus V^{\perp} \\ \iff &\begin{cases} \langle Ax, v\rangle &= \langle b_V, v\rangle \quad \forall v \in V \qquad (1) \\ \langle Ax, v\rangle &= \langle b_{V^\perp}, v\rangle \quad \forall v \in V^{\perp} \qquad (2) \end{cases} \end{aligned}$$

Now (1) has a unique solution $x \in U/_{\ker(A)}$ due to the fundamental theorem of homomorphisms.
In the second equation, since $Ax \in V$, the left-hand side is zero for all $v \in V^{\perp}$. Therefore (2) is equivalent to

$$\langle b_{V^\perp}, v\rangle = \langle b_V + b_{V^\perp}, v\rangle = \langle b, v\rangle = 0 \quad \forall v \in V^{\perp}.$$

Since $V^{\perp} = \ker(A^T)$ we have

$$\langle b, v\rangle = 0 \quad \forall v \in \ker(A^T).$$

□

Now we apply this result to the subdomain problems in BNN.
From (7.16) we see these problems are:

$$\begin{aligned} \mathbb{R}^{I_{\Gamma_i}} \ni z_i &:= D^{(i)} {S^{(i)}}^{-1} D^{(i)} R_{\Gamma_i} S e \\ \iff {D^{(i)}}^{-1} S^{(i)} {D^{(i)}}^{-1} z_i &= R_{\Gamma_i} S e \\ \iff \langle \tilde{S}^{(i)} z_i, v\rangle &= \langle R_{\Gamma_i} S e, v\rangle \quad \forall v \in U_i = \mathbb{R}^{I_{\Gamma_i}}, \end{aligned} \tag{7.18}$$

since $\ker(S^{(i)}) = \operatorname{span}\{\mathbb{1}_i\}$.
We have $S^{(i)} {D^{(i)}}^{-1} D^{(i)} \mathbb{1}_i = 0$, i.e. $\ker(\tilde{S}^{(i)}) = \operatorname{span}\{D^{(i)}\mathbb{1}_i\}$. Since a balancing step is applied before the subdomain solve we have $e = (I - P_0)\tilde{e}$. Accordingly to Lemma 7.10 one needs to check the right-hand side of (7.18) on $\ker(\tilde{S})$:

$$\langle R_{\Gamma_i} S (I - P_0)\tilde{e}, D^{(i)}\mathbb{1}_i\rangle = \langle S(I - P_0)\tilde{e}, R_{\Gamma_i}^T D^{(i)} \mathbb{1}_i\rangle = 0$$

since $(I - P_0)\tilde{e}$ is S-orthogonal to

$$\operatorname{range}(P_0) = \operatorname{range}(R_0^T) = \operatorname{range}(\sum_{i \in I_0} R_{\Gamma_i}^T D^{(i)} R_{0,i}^T)$$

and $R_{\Gamma_i}^T D^{(i)} \mathbb{1}_i = \sum_{j \in I_0} R_{\Gamma_j}^T D^{(j)} \mathbb{1}_j \delta_{ij}$.

Bibliography

R. E. Bank, T. F. Dupont, and H. Yserentant. The hierarchical basis multigrid method. *Numerische Mathematik*, 52(4):427–458, 1988. ISSN 0029-599X. doi: 10.1007/BF01462238. URL `http://dx.doi.org/10.1007/BF01462238`.

P. Bastian. Lecture notes on scientific computing with partial differential equations. `http://conan.iwr.uni-heidelberg.de/teaching/numerik2_ss2014/num2.pdf`, 2014.

D. Braess. *Finite Elemente.* Springer, 3rd edition, 2003.

J. H. Bramble, J. E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comput.*, 55:1–22, 1990.

James H. Bramble. A second order finite difference analog of the first biharmonic boundary value problem. *Numerische Mathematik*, 9(3):236–249, 1966. ISSN 0029-599X. doi: 10.1007/BF02162087. URL `http://dx.doi.org/10.1007/BF02162087`.

A. Brandt. Multi-level adaptive solutions to boundary-value problems. *Math. Comput.*, 31:333–390, 1977.

S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods.* Springer, 1994.

Tony F. Chan and Tarek P. Mathew. Domain decomposition algorithms. *Acta Numerica*, 3:61–143, 1994. ISSN 1474-0508. doi: 10.1017/S0962492900002427. URL `http://dx.doi.org/10.1017/S0962492900002427`.

P. G. Ciarlet. *The finite element method for elliptic problems.* Classics in Applied Mathematics. SIAM, 2002.

Wolfgang Dahmen and Angela Kunoth. Multilevel preconditioning. *Numerische Mathematik*, 63(1):315–344, 1992. ISSN 0029-599X. doi: 10.1007/BF01385864. URL `http://dx.doi.org/10.1007/BF01385864`.

K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational Differential Equations.* Cambridge University Press, 1996.

A. Ern and J.-L. Guermond. *Theory and practice of finite element methods.* Springer, 2004.

W. Hackbusch. A fast iterative method for solving poisson's equation in a general region. In R. Bulirsch, R. D. Griegorieff, and J. Schröder, editors, *Numerical Treatment of Differential Equation*, number 631 in Lecture Notes in Math. Springer, 1978.

Wolfgang Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme*. Number 69 in Leitfäden der angewandten Mathematik und Mechanik. Teubner, Stuttgart, 1991. ISBN 3-519-02372-5.

P. Oswald. On discrete norm estimates related to multilevel preconditioners in the finite element method. In *Proc. Int. Conf. Constr. Theory of Functions*. 1991.

A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations*. Oxford University Press, Oxford, 1999.

R. Rannacher. Einführung in die Numerische Mathematik II (Numerik partieller differentialgleichungen). `http://numerik.iwr.uni-heidelberg.de/~lehre/notes`, 2006.

B. Smith, P. Bjørstad, and W. Gropp. *Domain Decomposition – Parallel Multilevel Methods for Elliptic Partial Differential Equations*. Cambridge University Press, 1996.

A. Toselli and O. Widlund. *Domain Decomposition Methods – Algorithms and Theory*. Springer, Berlin Heidelberg, 2005.

H. Yserentant. Old and new convergence proof for multigrid methods. *Acta Numerica*, pages 285–326, 1993.

Harry Yserentant. On the multi-level splitting of finite element spaces. *Numerische Mathematik*, 49(4):379–412, 1986. ISSN 0029-599X. doi: 10.1007/BF01389538. URL `http://dx.doi.org/10.1007/BF01389538`.

X. Zhang. Multilevel schwarz methods. *Numerische Mathematik*, 63(1):521–539, 1992.