Iterative Solvers for PDEs

D. Gallistl

Lecture notes
Sommersemester 2023
Universität Jena

Last update: 14th July 2023

Contents

1. Finite differences for Poisson's equation

§1. Basic notions

In this lecture we study a class of partial differential equations (PDEs) and their numerical approximation. We confine ourselves to linear equations of second order. The most prominent example is *Poisson's equation* $-\Delta u = f$, where u is the unknown function, Δ is the Laplacian, and f is some given function, usually referred to as *right-hand side.* The equation is called *partial* differential equation because it involves partial derivatives of the solution (in contrast to *ordinary differential equations (ODEs)*, which only depend on one scalar variable. The notion of *2nd order* describes that the highest involved derivative of u has order 2. In order to get started with a fairly simple setting, we will consider Poisson's equation in the first lectures. Generally, we pose the questions of *existence* of a solution to a PDE and its *uniqueness.* Clearly, solutions to Poisson's equation are not unique without any further constraints being imposed. For instance, any solution can be shifted by an arbitrary affine function and will still remain a solution. We will thus consider the *Dirichlet problem*, which imposes a zero boundary condition on the solution. This PDE is posed on a domain $\Omega \subseteq \mathbb{R}^n$ which is open, bounded, and connected.

Definition 1.1 (Dirichlet problem for the Laplacian)**.** Let $\Omega \subseteq \mathbb{R}^n$ be open, bounded, and connected. A function $u \in C^2(\Omega) \cap C(\bar{\Omega})$ is said to be a classical solution to the Dirichlet problem (for the Laplacian) with right-hand side $f \in C(\Omega)$ and boundary values $g \in C(\partial\Omega)$ if it satisfies

$$-\Delta u = f \text{ in } \Omega \quad \text{und} \quad u = g \text{ on } \partial\Omega.$$

♦

The question under which circumstances solutions to the Dirichlet problem exist is difficult to answer in general. At this stage, we confine ourselves to a basic statement on uniqueness.

Lemma 1.2 (maximum principle)**.** *Let $\Omega \subseteq \mathbb{R}^n$ be open, bounded, and connected and let $u \in C^2(\Omega) \cap C(\bar{\Omega})$ satisfy $\Delta u \geq 0$ in Ω. Then the maximum of u is attained on the boundary, i.e.,*

$$\max_{\bar{\Omega}} u = \max_{\partial\Omega} u.$$

Proof. We note that $\bar{\Omega}$ is compact and thus the maximum of u is attained in $\Omega \cup \partial\Omega$. Let us first assume the strict inequality $\Delta u > 0$ in Ω. At any point $x_0 \in \Omega$ with $u(x_0) = \max_{\bar{\Omega}} u$, the Hessian is necessarily negative-semidefinite, written $D^2u(x_0) \leq 0$, and so has only non-positive eigenvalues. In particular its trace (the sum of all eigenvalues) is non-positive, whence $\operatorname{tr}(D^2u(x_0)) = \Delta u(x_0) \leq 0$. In view of the assumed inequality $\Delta u > 0$, such a point $x_0 \in \Omega$ cannot exist, which implies that the maximum is attained on $\partial\Omega$. In the general case of $\Delta u \geq 0$ in Ω we let $\varepsilon > 0$ and define $u_\varepsilon(x) = u(x) + \varepsilon|x|^2$ where $|\cdot|$ denotes the Euclidean norm. We then have for any $\varepsilon > 0$ that $\Delta u_\varepsilon > 0$ in Ω and the above argument shows that

$$\max_{\bar{\Omega}} u_\varepsilon = \max_{\partial\Omega} u_\varepsilon.$$

We observe for any $x \in \bar{\Omega}$ that

$$u(x) \leq u_\varepsilon(x) \leq \max_{\bar{\Omega}} u_\varepsilon = \max_{\partial\Omega} u_\varepsilon = \max_{x\in\partial\Omega} u(x) + \varepsilon|x|^2 \leq \max_{\partial\Omega} u + \varepsilon R^2$$

for $R := \max_{x\in\bar{\Omega}} |x|^2$. The assertion then follows from letting $\varepsilon \to 0$. ■

Corollary 1.3 (uniqueness). *There is at most one classical solution to the Dirichlet problem from Definition 1.1.*

Proof. Let u_1, u_2 be two classical solutions. Then, $w := u_1 - u_2$ satisfies $w \in C^2(\Omega) \cap C(\bar{\Omega})$ and solves $\Delta w = 0$ in Ω with $w = 0$ on $\partial\Omega$. The maximum principle implies that w attains its maximum on $\partial\Omega$ and thus $w \leq 0$ in Ω. On the other hand, $\Delta w = 0$ also implies $\Delta(-w) \geq 0$. The maximum principle applied to $-w$ thus proves $-w \leq 0$. In consequence $w = 0$ in Ω and thus $u_1 = u_2$. ■

Corollary 1.4 (comparison principle). *Let $u, v \in C^2(\Omega) \cap C(\bar{\Omega})$ be such that $u \leq v$ on $\partial\Omega$ and $\Delta u \geq \Delta v$ in Ω. Then $u \leq v$ in Ω.*

Proof. The difference $w := u - v$ satisfies $\Delta w \geq 0$ and by the maximum principle w attains its maximum on $\partial\Omega$. But there we have $w \leq 0$. Therefore $w \leq 0$ in Ω or equivalently $u \leq v$ in Ω. ■

§2. Finite difference discretization of the Laplacian

We want to design a numerical method to approximately solve the Dirichlet problem. For the sake of a short presentation, we confine ourselves to the case of Ω being the two-dimensional square domain $\Omega = (0,1)^2$ and to homogeneous boundary conditions, i.e., $g = 0$ in Definition 1.1. Generalizations will be discussed later (problem sessions).

The idea of the so-called Finite Difference Method (FDM) is to replace partial derivatives by difference quotients.

Definition 1.5 (first-order difference quotients). Given a step size $h > 0$ and a sequence of elements $(U_j)_{j=0,\dots,J}$ of some vector space, we define

$$\partial^+ U_j := \frac{U_{j+1} - U_j}{h}, \quad (j = 0, \dots, J-1) \quad \textit{(forward difference quotient)}$$

and

$$\partial^- U_j := \frac{U_j - U_{j-1}}{h}, \quad (j = 1, \dots, J) \quad \textit{(backward difference quotient).}$$

♦

Definition 1.6 (second-order central difference quotient). Given a step size $h > 0$ and a sequence $(U_j)_{j=0,\dots,J}$ of elements of some vector space, the quantity

$$\partial^+\partial^- U_j = \frac{U_{j+1} - 2U_j + U_{j-1}}{h^2},$$

is called the *second-order central difference quotient.* ♦

For a function u over $[0,1]$ we let

$$\partial^+ u(x) = \frac{u(x+h) - u(x)}{h}$$

with analogous notation for ∂^-. The following approximation properties can be proven via Taylor expansion.

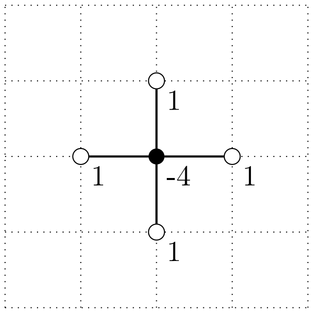


Figure 1.1.: Scematic diagram of the 5-point stencil with weights.

Lemma 1.7. *Given $u \in C^2([0,1])$, we have for ∂_x^+ and ∂_x^- that*

$$|\partial_x^+ u(x) - \partial_x u(x)| \leq \frac{h}{2}\|\partial_{xx}^2 u\|_{C([0,1])} \quad \textit{for all } x \in [0, 1-h]$$
$$|\partial_x^- u(x) - \partial_x u(x)| \leq \frac{h}{2}\|\partial_{xx}^2 u\|_{C([0,1])} \quad \textit{for all } x \in [h, 1].$$

Given $u \in C^4([0,1])$, we have for ∂_x^+ and ∂_x^- that

$$|\partial_x^+ \partial_x^- u(x) - \partial_x^2 u(x)| \leq \frac{h^2}{12}\|\partial_{xxxx}^4 u\|_{C([0,1])} \quad \textit{for all } x \in [h, 1-h].$$

Proof. Problem A.3. ■

Let $J \geq 0$ and $h = 1/J$. We set up a grid with $J+1$ points in every coordinate direction by letting

$$x_{j,k} = (jh, kh) \quad j, k = 0, \ldots, J.$$

We wish to approximate the solution u by a grid function U whose value at $x_{j,k}$ we denote by $U_{j,k}$. For interior points we define a discrete version of the Laplacian $\Delta = \partial_{x_1 x_1}^2 + \partial_{x_2 x_2}^2$ through central differences

$$\Delta_h U_{j,k} = \partial_{x_1}^+ \partial_{x_1}^- U_{j,k} + \partial_{x_2}^+ \partial_{x_2}^- U_{j,k} \quad \text{for } j, k = 1, \ldots, J-1.$$

It is straightforward to compute the representation

$$\Delta_h U_{j,k} = \frac{1}{h^2}(U_{j+1,k} + U_{j,k+1} - 4U_{j,k} + U_{j-1,k} + U_{j,k-1}). \tag{1.1}$$

We see that the value $\Delta_h U_{j,k}$ depends on the point $x_{j,k}$ and its four neighbours in the grid. The stencil is called *five-point stencil*, see Figure 1.1.

Definition 1.8. Let $\Omega = (0,1)^2$ and $f \in C(\Omega)$. The discretized Poisson problem (with zero boundary conditions) seeks $(U_{j,k} : j, k = 0, \ldots, J)$ such that

$$\begin{cases} -\Delta_h U_{j,k} = f(x_{j,k}) & \text{for } j, k = 1, \ldots, J-1 \\ U_{0,k} = U_{J,k} = U_{j,0} = U_{j,J} = 0 & \text{for } j, k = 0, \ldots, J. \end{cases}$$

♦

We briefly comment on the implementation. In order to represent U as a vector, we choose the *lexicographic enumeration* and identify $\{0,\dots,J\}^2$ with $\{1,\dots,L\}$ (where $L=(J+1)^2$) through the map

$$(j,k)\mapsto j+k(J+1)+1=:\ell.$$

Loosely speaking, we enumerate the grid by taking rows from left to right starting on the left bottom. We see from (1.1) that the discrete Laplacian takes the form

$$\Delta_h U_\ell = \frac{1}{h^2}(U_{\ell+1}+U_{\ell+(J+1)}-4U_\ell+U_{\ell-1}+U_{\ell-(J+1)})$$

for any interior point x_ℓ. We see that $U_{j,k}$ for j or k in $\{0,J\}$ are no unknowns because they are known through the boundary condition. We are therefore merely interested in computing $U_{j,k}$ for $j,k\in\{1,\dots,J-1\}$.

We consider the sub-list $(\mathring{U}_1,\dots\mathring{U}_N)$ corresponding to the interior points and define the matrix

$$X := \begin{bmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix}.$$

This results in the system

$$\begin{bmatrix} X & -I & & \\ -I & \ddots & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & X \end{bmatrix} \begin{bmatrix} \mathring{U}_1 \\ \vdots \\ \vdots \\ \mathring{U}_N \end{bmatrix} = h^2 \begin{bmatrix} f_1 \\ \vdots \\ \vdots \\ f_N \end{bmatrix}.$$

Here $f_\ell = f(x_\ell)$ for every interior node. We note that this is a system of the type $Ax=b$ for a sparse matrix A. In an implementation, a sparse matrix format should be used. In a practical implementation, the easiest choice is to first set up a system matrix for all points including the boundary points (imagining U being zero for all “ghost points” outside $\bar\Omega$) because this matrix has constant diagonals. In a second step, the system is restricted to the interior points, the so-called *degrees of freedom*, the global indices corresponding to the values $\mathring{U}_j$.

§3. Basic error analysis of the finite difference method

We want to quantify the error $u-U$ between the true solution u to the Dirichlet problem and its finite difference approximation U. The fundamental tool is a discrete version of the maximum principle for Δ_h.

Lemma 1.9 (discrete maximum principle). *Let Ω be the unit square. Let the mesh function U satisfy $\Delta_h U_{j,k}\geq 0$ for all $j,k\in\{1,\dots,J-1\}$. Then, U attains its maximum at a boundary point (i.e., at some $x_{j,k}$ with $j\in\{0,J\}$ or $k\in\{0,J\}$.*

Proof. Let $x_{j,k}$ with $j,k\in\{1,\dots,J-1\}$ be an interior point. From the definition of Δ_h we obtain

$$U_{j,k}=\frac{1}{4}(U_{j-1,k}+U_{j+1,k}+U_{j,k+1}+U_{j,k-1})-\frac{h^2}{4}\Delta_h U_j.$$

From $\Delta_h U_{j,k} \geq 0$ we thus infer

$$U_{j,k} \leq \frac{1}{4}(U_{j-1,k} + U_{j+1,k} + U_{j,k+1} + U_{j,k-1}).$$

Assume $U_{j,k}$ is the maximum of U. Then it is not smaller than any of the four neighbouring values. Hence, equality holds in the foregoing estimate. In particular

$$U_{j,k} = U_{j-1,k} = U_{j+1,k} = U_{j,k+1} = U_{j,k-1}.$$

Iterating this argument up to the boundary shows that U is constant and therefore the maximum is attained at the boundary. ■

The foregoing lemma was formulated for the unit square. It is clear how to generalize it to other geometries.

We denote the set of boundary points of the grid by Γ. For mesh functions V we use the following notation on maximum norms

$$\begin{aligned}
|V|_{\infty,\bar{\Omega}} &:= \max_{\substack{j,k=0,\dots,J \\ \text{s.t. } x_{j,k} \in \Omega \cup \Gamma}} |V_{j,k}| \\
|V|_{\infty,\Omega} &:= \max_{\substack{j,k=0,\dots,J \\ \text{s.t. } x_{j,k} \in \Omega}} |V_{j,k}| \\
|V|_{\infty,\Gamma} &:= \max_{\substack{j,k=0,\dots,J \\ \text{s.t. } x_{j,k} \in \Gamma}} |V_{j,k}|
\end{aligned}$$

For the unit square we have $\Gamma \subseteq \partial\Omega$. Note, however, that for more complicated geometries the 'boundary points' of the grid need not lie on $\partial\Omega$.

The discrete maximum principle implies the following stability estimate.

Lemma 1.10 (stability)**.** *Let Ω be the unit square. There exists a constant $C > 0$ with the following property. Given a mesh over Ω and a mesh function U, we have*

$$|U|_{\infty,\bar{\Omega}} \leq |U|_{\infty,\Gamma} + C|\Delta_h U|_{\infty,\Omega}.$$

Proof. We define the mesh function

$$W_{j,k} = \frac{1}{4}|x_{j,k}|^2 \quad \text{(squared Euclidean norm)}.$$

Then $\Delta_h W_{j,k} = 1$ for any pair (j,k). Let $r := |\Delta_h U|_{\infty,\Omega}$ and define the mesh functions $V^{\pm} := \pm U + rW$. Then

$$\Delta_h V^{\pm} = \pm\Delta_h U + r \geq 0.$$

By the discrete maximum principle, $V^{\pm}$ attains its maximum on the boundary. This means

$$\pm U + rW \leq |\pm U + rW|_{\infty,\Gamma} \quad \text{over } \bar{\Omega}.$$

The triangle inequality on the right-hand side and $W \geq 0$ on the left hand side thus prove

$$|U|_{\infty,\bar{\Omega}} \leq |U|_{\infty,\Gamma} + r|W|_{\infty,\Gamma}.$$

This proves the assertion with $C = |W|_{\infty,\Gamma}$. ■

Remark 1.11. The generalization of the stability estimate to domains different from the square is immediate. ♦

Corollary 1.12. *The finite difference method has a unique solution U.*

Proof. We have already seen that the finite difference system is a quadratic finite-dimensional system of linear equations. Thus, uniqueness implies existence. Suppose there exist two solutions U, V satisfying $-\Delta_h U = F = \Delta_h V$ (where F is the mesh function interpolating f at the grid points) and $U|_\Gamma = 0 = V_\Gamma$. Then $\Delta_h(U-V) = 0$, and the stability estimate implies $|U-V|_{\infty,\bar\Omega} = 0$. Thus $U = V$. ■

Remark 1.13. For a mesh function F we denote by $-\Delta_h^{-1}F$ the solution to the finite difference system with zero boundary conditions. The stability estimate can then be written as follows

$$|\Delta_h^{-1}F|_{\infty,\bar\Omega} \leq C|F|_{\infty,\Omega}.$$

We thus see that $-\Delta_h^{-1}$ has a uniformly bounded continuity constant (C is independent of the grid size h). ♦

When operating on grids we identify u with the mesh function having values $u(x_{j,k})$.

Lemma 1.14 (consistency). *Assume $u \in C^4(\bar\Omega)$. Then*

$$|\Delta_h u - \Delta u|_{\infty,\Omega} \leq \frac{1}{2}h^2 \sum_{j=1,2} \|\partial_{x_j}^4 u\|_{C(\bar\Omega)}.$$

Proof. This is an immediate consequence of Lemma 1.7. ■

Theorem 1.15 (FDM convergence). *Assume the solution u to the Poisson problem $-\Delta u = f$ over the unit square Ω with homogeneous boundary conditions satisfies $u \in C^4(\bar\Omega)$. Then the finite difference error satisfies*

$$|u - U|_{\infty,\bar\Omega} \leq Ch^2 \sum_{j=1,2} \|\partial_{x_j}^4 u\|_{C(\bar\Omega)}$$

with a constant C independent of the mesh size and f.

Proof. Stability implies

$$|u - U|_{\infty,\bar\Omega} \leq C|\Delta_h(u - U)|_{\infty,\Omega} = C|\Delta_h u - \Delta u|_{\infty,\Omega}$$

because $-\Delta_h U = F = -\Delta u$ at the grid points. The right-hand side is then estimated with the consistency estimate, which concludes the proof. ■

Remark 1.16. The simple proof of convergence shows the general principle of convergence proofs for finite difference methods:

$$\text{stability } + \text{consistency} \implies \text{convergence.}$$

This can be formalized in a general framework (Lax–Richtmyer theorem), but we confine ourselves to formulating this rule of thumb. The above convergence proof contains the whole essence of the reasoning behind. ♦

§4. Nine-point stencils and complements on FDM

The five-point stencil studied so far is somehow a minimal choice. One can think of improving accuracy by increasing the dependence on neighbouring grid points. In two dimensions, nine-point stencil take into account the diagonal neighbours as well. We note that the distance of a point $x_{j,k}$ to its diagonal neighbour is $\sqrt{2}h$. We then have the central differences

$$\frac{U_{j+1,k} - 2U_{j,k} + U_{j-1,k}}{2h}$$
$$\frac{U_{j,k+1} - 2U_{j,k} + U_{j,k-1}}{2h}$$
$$\frac{U_{j+1,k-1} - 2U_{j,k} + U_{j-1,k+1}}{2\sqrt{2}h}$$
$$\frac{U_{j+1,k+1} - 2U_{j,k} + U_{j-1,k-1}}{2\sqrt{2}h},$$

see also Figure 1.2.

We next discuss how to design a linear combination that consistently discretizes the Laplacian and has higher-order convergence properties.

Consider the function $u(x_{j,k} + te_m)$ where $m \in \{1, 2\}$ is the mth cartesian unit vector. Taylor expansion of fourth order results in

$$\begin{aligned} u(x_{j,k} + te_m) = u(x_{j,k}) + \partial_m u(x_{j,k})t + \frac{1}{2}\partial_m^{(2)} u(x_{j,k})t^2 + \frac{1}{6}\partial_m^{(3)} u(x_{j,k})t^3 \\ + \frac{1}{24}\partial_m^{(4)} u(x_{j,k})t^4 + \frac{1}{120}\partial_m^{(5)} u(x_{j,k})t^6 + O(t^5). \end{aligned}$$

If we evaluate this expression for $t = \pm h$ and add the results, the odd-order terms cancel and we obtain

$$u(x_{j,k} + he_m) + u(x_{j,k} - he_m) = 2u(x_{j,k}) + \partial_m^{(2)} u(x_{j,k})h^2 + \frac{1}{12}\partial_m^{(4)} u(x_{j,k})h^4 + O(h^6).$$

Adding this identity for $m = 1, 2$ results in the well known relation of the 5-point stencil

$$\begin{aligned} u(x_{j+1,k}) + u(x_{j-1,k}) + u(x_{j,k+1}) + u(x_{j,k-1}) \\ = 4u(x_{j,k}) + \Delta u(x_{j,k})h^2 + \frac{1}{12}(\partial_{xxxx} + \partial_{yyyy})u(x_{j,k})h^4 + O(h^6). \end{aligned} \tag{1.2}$$

We can apply similar arguments to the diagonal directions

$$d_1 = 2^{-1/2}(1, -1) \quad \text{and} \quad d_1 = 2^{-1/2}(1, 1)$$

and obtain with $t = \pm\sqrt{2}h$ and analogous computations

$$\begin{aligned} u(x_{j+1,k-1}) + u(x_{j-1,k+1}) + u(x_{j+1,k+1}) + u(x_{j-1,k-1}) \\ = 4u(x_{j,k}) + 2\Delta u(x_{j,k})h^2 + \frac{1}{6}(\partial_{xxxx} + \partial_{yyyy} + 6\partial_{xxyy})u(x_{j,k})h^4 + O(h^6). \end{aligned} \tag{1.3}$$

Here, various sums of mixed derivatives have cancelled out. We now add 4 times (1.2) to (1.3) and obtain

$$\begin{aligned} 4u(x_{j+1,k}) + 4u(x_{j-1,k}) + 4u(x_{j,k+1}) + 4u(x_{j,k-1}) \\ + u(x_{j+1,k-1}) + u(x_{j-1,k+1}) + u(x_{j+1,k+1}) + u(x_{j-1,k-1}) \\ = 20u(x_{j,k}) + 6\Delta u(x_{j,k})h^2 + \frac{1}{2}(\partial_{xxxx} + \partial_{yyyy} + 2\partial_{xxyy})u(x_{j,k})h^4 + O(h^6) \end{aligned}$$

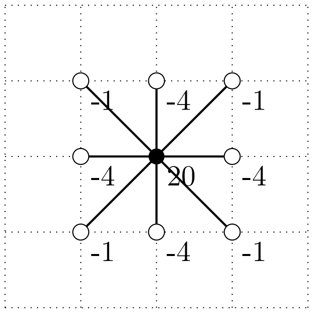


Figure 1.2.: Scematic diagram of the 9-point stencil with weights.

We use that

$$-\Delta u(x_{j,k}) = f(x_{j,k})$$

and

$$-(\partial_{xxxx} + \partial_{yyyy} + 2\partial_{xxyy})u(x_{j,k}) = -(\partial_{xx} + \partial_{yy})\Delta u(x_{j,k}) = (\partial_{xx} + \partial_{yy})f(x_{j,k})$$

and derive the relations

$$S^{9p}_{(j,k)}u = 6h^2 f(x_{j,k}) + \frac{1}{2}h^4 \Delta f(x_{j,k}) + O(h^6).$$

for the 9-point stencil $S^{9p}_{(j,k)}$ symbolized as follows

$$\begin{bmatrix} -1 & -4 & -1 \\ -4 & 20 & -4 \\ -1 & -4 & -1 \end{bmatrix}.$$

The corresponding finite difference equations are then

$$\begin{aligned} 20U_{j,k} - 4U_{j+1,k} - 4U_{j-1,k} - 4U_{j,k+1} - 4U_{j,k-1} - U_{j+1,k-1} - U_{j-1,k+1} - U_{j+1,k+1} - U_{j-1,k-1} \\ = 6h^2 f(x_{j,k}) + \frac{1}{2}h^4 \Delta f(x_{j,k}). \end{aligned}$$

The Laplacian of f on the right-hand side can be either directly computed from f or alternatively approximated by the 5-point stencil.

Remark 1.17. We expect U to converge at a better order than the ordinary 5-point stencil provided the exact solution is sufficiently regular. We will not provide a detailed proof in this lecture but remark that it can in principle be worked out with the basic tools from the previous section. ⧫

Remark 1.18. The 9-point stencil can be viewed as a weighted average of two (rotated) 5-point stencils. From the above derivation it is clear that any convex combination of the stencils yields a first-order scheme. The special choice 4 : 1 and a modification of the right-hand side, however, result in an even higher-order scheme. ⧫

We know that convergence of any finite difference scheme follows from stability and consistency. We do not work out an error analysis of the nine-point stencil here; it will be part of the problem sessions.

Curved geometries. We end this section by commenting on practical aspects of the FDM (be it the 5- or 9-point stencil). Wor convenience, we formulated many results for the square where the domain could be exactly covered by a cartesian mesh. For more complicated situations with possibly curved geometries this is no longer possible. Assume for example that domain $\Omega \subseteq [0,1]^2$ can be embedded in the unit square (or any other box after appropriate scaling). Generally we cannot expect that the boundary $\partial\Omega$ has a meaningful intersection with the gridpoints. Instead, we define

$$\Omega_h := \{x_{j,k} : x_{j,k} \in \Omega \text{ and all neighbours belong to } \bar{\Omega}\}$$

and

$$\Gamma_h := \{x_{j,k} : x_{j,k} \in \Omega \text{ and a neighbour does not belong to } \bar{\Omega}\}.$$

By neighbour we mean a gridpoint belonging to the stencil at $x_{j,k}$. The FDM equations then read $\Delta_h U_{j,k} = F_{j,k}$ for all $x_{j,k} \in \Omega_h$. The results proven in the foregoing sections transfer to this situation.

More general elliptic operators. We can reduce a PDE of the form

$$\operatorname{tr}(AD^2u) = f$$

with a (constant) positive definite and symmetric matrix A to an equation involving only the diagonal entries of D^2 by diagonalizing $A = R\Lambda R^T$ with an orthogonal matrix R and a diagonal matrix Λ. Since the trace is independent of the chosen coordinate system we see that the above PDE is equivalent to

$$\operatorname{tr}(\Lambda R^T D^2 u R) = f.$$

It is easy to check that this PDE only depends on ∂_{r_1,r_1} and ∂_{r_2,r_2} where r_1, r_2 are the chosen eigenvectors of A. Thus, after rotating the coordinate system, a (weighted) 5-point stencil can be used.

When lower-order terms are present, for instance as

$$\operatorname{tr}(AD^2u) + b \cdot \nabla u + cu = f$$

for a vector b and a constant c, these can be included as well. The zero-order term is simply discretized by cU. The term involving the gradient can be discretized through first-order difference quotients.

2. Iterative methods

§1. Elementary perturbation analysis

We denote by $\mathbb{K}$ one of the fields $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$. In what follows, V is a vector space over $\mathbb{K}$ and $n \in \mathbb{N}$. A *norm* on V is a homogeneous, positive definite, and subadditive real-valued function, typically denoted by $\|\cdot\|$.

Example 2.1. Important norms on the finite dimensional spave $\mathbb{K}^n$ are the Euclidean (ℓ_2) norm $\|x\|_2 := \sqrt{\sum_{j=1}^n |x_j|^2}$, the maximum ($\ell_\infty$) norm $\|x\|_\infty := \max_{j=1,\dots n} |x_j|$, or the the ℓ_p norm (for $1 \le p < \infty$) $\|x\|_p := \left(\sum_{j=1}^n |x_j|^p\right)^{1/p}$.

Linear maps between normed spaces can be measured with the *operator norm.*

Definition 2.2 (operator norm)**.** Given normed spaces $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$, and a linear map $A : V \to W$, the *operator norm* of A is given by

$$\|A\|_{L(V,W)} := \sup_{x \in \mathbb{V} \setminus \{0\}} \frac{\|Ax\|_W}{\|x\|_V}.$$

♦

If the norms are fixed and there is no risk of confusion, we will often write $\|A\|$ with out index for the operator norm. Note that $\|A\|_{L(V,W)}$ can be ∞ in the above definition. We will denote by $L(V, W)$ the space of linear maps between V and W with bounded norm, referred to as the *bounded linear operators.* This is not critical if V is finite-dimensional because the unit sphere in $\mathbb{K}^n$ is compact. If $V = \mathbb{K}^m$ and $W = \mathbb{K}^n$ and the linear maps are interpreted as matrices, the norms over V and W are referred to as *vector norms* and the operator norm is sometimes called the *induced matrix norm.*

The operator norm is submultiplicative $\|AB\| \le \|A\|\,\|B\|$ and in particular compatible in the sense that $\|Ax\| \le \|A\|\|x\|$.

Remark 2.3. Given norms on $\mathbb{K}^m$ and $\mathbb{K}^n$, there may exists other compatible matrix norms on $\mathbb{K}^{m\times n}$ which are not the operator norm. ♦

Example 2.4. a) The *Frobenius norm* $\|A\|_{\mathrm{Fr}} = \sqrt{\sum_{j,k=1}^n |A_{jk}|^2}$ is submultiplicative and compatible with the Euclidean norm, but not the corresponding operator norm.

b) The operator norm corresponding to $\|\cdot\|_\infty$ is the maximal row sum

$$\|A\|_\infty := \max_{1\le j\le n} \sum_{k=1}^n |A_{jk}|.$$

c) The operator norm corresponding to $\|\cdot\|_1$ is the maximal column sum.

$$\|A\|_1 := \max_{1\le k\le n} \sum_{j=1}^n |A_{jk}|.$$

The operator norm corresponding to the Euclidean norm is called the *spectral norm.* It is computed in the following theorem.

Theorem 2.5. *The operator norm corresponding to the Euclidean norm is given by*

$$\|A\|_2 = \max\{|\lambda|^{1/2} : \lambda \text{ eigenvalue of } A^*A\}.$$

Proof. The matrix $M := A^*A$ is Hermitian and possesses n real eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ (counted with multiplicities) with a corresponding orthonormal systems $w_1, \ldots, w_n$ of eigenvectors. Every $x \in \mathbb{K}^n$ has an expansion

$$x = \sum_{j=1}^{n} \alpha_j w_j \quad \text{with } \alpha_j = \langle x, w_j \rangle_2.$$

From the orthogonality one computes $\|x\|_2^2 = \sum_{j=1}^n |\alpha_j|^2$ and

$$\|Mx\|_2^2 = \langle Mx, Mx \rangle_2 = \sum_{j,k=1}^{n} \alpha_j \lambda_j \bar{\alpha}_k \bar{\lambda}_k \langle w_j, w_k \rangle_2 = \sum_{j=1}^{n} |\lambda_j|^2 |\alpha_j|^2.$$

Therefore

$$\|M\|_2^2 = \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\sum_{j=1}^n |\lambda_j|^2 |\alpha_j|^2}{\sum_{j=1}^n |\alpha_j|^2} \leq \max_{j=1\ldots n} |\lambda_j|^2$$

and so

$$\|A\|_2^2 = \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\langle Mx, x \rangle_2}{\|x\|_2^2} = \|M\|_2 \leq \max_{j=1\ldots n} |\lambda_j|.$$

The bound is sharp because for any $j \in \{1, \ldots, n\}$ we have

$$|\lambda_j| = |\lambda_j| \, \|w_j\|_2 = \|\lambda_j w_j\|_2 = \|M w_j\|_2 \leq \|M\|_2 \leq \|A\|_2^2.$$

■

We recall that a normed vector space V is said to be complete if every Cauchy sequence in V has a limit in V. It is an elementary result from linear functional analysis that $L(V, V)$ is then complete under the operator norm (Exercise A.11). In such case, an operator B is guaranteed to be invertible if it does not deviate too much from the identity I.

Lemma 2.6 (Neumann series)**.** *Let V be a complete normed linear space and let $\| \cdot \|$ denote a submultiplicative norm on $L(V, V)$ that is equivalent to the operator norm. If $B \in L(V, V)$ satifies $\|B\| < 1$, then $(I - B)$ is invertible with $(I - B)^{-1} = \sum_{j=0}^{\infty} B^j$ und $\|(1 - B)^{-1}\| \leq (1 - \|B\|)^{-1}$.*

Proof. The series converges because

$$\Big\| \sum_{j=0}^{n} B^j - \sum_{j=0}^{m} B^j \Big\| \leq \sum_{j=m+1}^{n} \|B\|^j \qquad \text{für } n \geq m \geq 0.$$

The geometric series yields convergence with the limit

$$\Big\| \sum_{j=0}^{\infty} B^j \Big\| \leq \sum_{j=0}^{\infty} \|B\|^j = (1 - \|B\|)^{-1}.$$

The inverse property can be directly checked

$$\sum_{j=0}^{\infty} B^j(1-B) = \sum_{j=0}^{\infty} B^j - \sum_{j=0}^{\infty} B^{j+1} = I.$$

■

In the case of $\mathbb{K}^n$, the Neumann series is valid for any submultiplicative matrix norm because all norms over a given finite-dimensional space are equivalent.

When solving a linear matrix-vector system $Ax = b$ it is important to understand the response of the solution $\tilde{x} = x + \Delta x$ of the system $\tilde{A}\tilde{x} = \tilde{b}$ with perturbed data $\tilde{A} = A + \Delta A$ and $\tilde{b} = b + \Delta b$.

Definition 2.7. Given a norm on $\mathbb{K}^n$, the number $\kappa(A) := \|A\|\|A^{-1}\|$ is called the *condition number* von A. If A is singular, we interpret the expression as $\kappa(A) = \infty$. ♦

Theorem 2.8 (perturbation result). *Let $\|\cdot\|$ be any norm on $\mathbb{K}^n$, let $A \in \mathbb{K}^{n\times n}$ be regular, and let the $\Delta A \in \mathbb{K}^{n\times n}$ satisfy*

$$\|\Delta A\| < \frac{1}{\|A^{-1}\|}.$$

Then $\tilde{A} = A + \Delta A$ is regular and the relative error of the perturbed system satisfies

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\|\Delta A\|\|A\|^{-1}} \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right).$$

Proof. Because $A + \Delta A = A(I + A^{-1}\Delta A)$ and $\|A^{-1}\Delta A\| \leq \|A^{-1}\|\|\Delta A\| < 1$ by assumption, we can apply the Neumann series and see that $\tilde{A}$ is regular. We directly compute that $\tilde{A}\Delta x = \Delta b - \Delta A x$. Therefore

$$\|\Delta x\| = \|\tilde{A}^{-1}(\Delta b - \Delta A x)\| \leq \|\tilde{A}^{-1}\|(\|\Delta b\| + \|\Delta A\|\|x\|).$$

With the norm bound from the Neumann series we infer

$$\|\tilde{A}^{-1}\| = \|(A(I + A^{-1}\Delta A))^{-1}\| \leq \|(I + A^{-1}\Delta A)^{-1}\|\|A^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\Delta A\|}.$$

We combine these bounds and obtain (excluding the trivial case $x = 0$) that

$$\|\Delta x\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\Delta A\|}(\|\Delta b\| + \|\Delta A\|\|x\|) \leq \frac{\|A^{-1}\|\,\|A\|\,\|x\|}{1 - \|A^{-1}\|\|\Delta A\|} \left(\frac{\|\Delta b\|}{\|A\|\|x\|} + \frac{\|\Delta A\|}{\|A\|} \right).$$

The relations $\|b\| \leq \|A\|\|x\|$ and $\|A^{-1}\| = \kappa(A)/\|A\|$ finally lead to

$$\|\Delta x\| \leq \frac{\kappa(A)\|x\|}{1 - \kappa(A)\|\Delta A\|/\|A\|} \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right).$$

■

The result states that small perturbations of the data are at most amplified by a factor proportional to $\kappa(A)$. The bound can be considered to be sharp. Note that the condition number depends on the choice of norm. For the Euclidean norm, Hermitian matrices satisfy

$$\kappa_2(A) := \|A\|_2\|A^{-1}\|_2 = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$$

the the eigenvalues $\lambda_{\max}$ resp. $\lambda_{\min}$ of A of maximal resp. minimal modulus. This follows from Theorem 2.5. Therefore, $\kappa_2(A)$ is sometimes called *spectral condition number*.

§2. Classical iterative methods

We are interested in characterizing the convergence of a linear stationary fixed-point iteration $x_{k+1} = Mx_k$. The principal tool will be the *spectral radius* of M.

Definition 2.9. The set of eigenvalues of a given matrix $M \in \mathbb{C}^{n\times n}$ is denoted by $\sigma(M)$. The number $\rho(M) := \max_{\lambda\in\sigma(M)} |\lambda|$ is called the *spectral radius* of M. ♦

It is immediately verified that $\rho(M) \leq \|M\|$ for any norm $\|\cdot\|$ on $\mathbb{C}^n$ because the largest-in-modulus eigenvalues λ and any corresponding eigenvector x satisfy

$$\rho(M) = \|\lambda x\| = \|Mx\| \leq \|M\|.$$

Conversely, the following holds.

Lemma 2.10. *Given $M \in \mathbb{C}^{n\times n}$ and $\varepsilon > 0$ there exists a norm $\|\cdot\|$ on $\mathbb{C}^n$ with the property*

$$\|M\| \leq \rho(M) + \varepsilon.$$

Proof. Without loss of generality we may assume that M has the format of a single Jordan block

$$M = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \lambda & 1 \\ & & & \lambda \end{bmatrix}$$

(the justification is left as an exercise). A transformation with the diagonal matrix D_ε with diagonal entries $1, \varepsilon, \varepsilon^2, \dots, \varepsilon^{n-1}$ yields

$$D_\varepsilon^{-1} M D_\varepsilon = \begin{bmatrix} \lambda & \varepsilon & & \\ & \ddots & \ddots & \\ & & \lambda & \varepsilon \\ & & & \lambda \end{bmatrix}$$

A direct computation shows

$$\|D_\varepsilon^{-1} M D_\varepsilon\|_\infty \leq \rho(M) + \varepsilon$$

and the left-hand side is the operator norm corresponding to $\|x\| := \|D_\varepsilon^{-1}x\|_\infty$. ■

Lemma 2.11. *A matrix $M \in \mathbb{C}^{n\times n}$ satisfies $\rho(M) < 1$ if and only if $\lim_{k\to 0} M^k x = 0$ for all $x \in \mathbb{C}^n$.*

Proof. If $\rho(M) < 1$, then by the foregoing lemma there exists a norm $\|\cdot\|$ on $\mathbb{C}^n$ such that $\|M\| < 1$, so that the sequence $\|M^k x\| \leq \|M\|^k \|x\|$ converges to zero as $k \to \infty$. Conversely, any eigenpair (λ, x) satisfies $M^k x = \lambda^k x$, so that convergence for all x necessarily requires $\rho(M) < 1$. ■

The spectral radius can be viewed as a measure for convergence speed.

Lemma 2.12. *Any norm $\|\cdot\|$ on $\mathbb{C}^n$ and any $M \in \mathbb{C}^{n\times n}$ satisfy $\lim_{k\to\infty} \|M^k\|^{1/k} = \rho(M)$.*

Proof. For any $\varepsilon > 0$ we can define the scaled matrix

$$N := \frac{1}{\rho(M) + \varepsilon} M$$

satisfying $\rho(N) < 1$. In particular

$$\lim_{k\to\infty} N^k = 0 \quad \text{und} \quad \sup_{k\to\infty} \|N^k\| \leq c < \infty.$$

We observe

$$\|M^k\| \leq c|\rho(M) + \varepsilon|^k$$

and, since ε was arbitrary,

$$\limsup_{k\to\infty} \|M^k\|^{1/k} \leq \rho(M).$$

Testing with a normalized eigenvector x corresponding to the eigenvalue realizing the spectral radius yields

$$\rho(M)^k = \|M^k x\| \leq \|M^k\|,$$

which implies the stated identity. ■

As a consequence we have the following convergence of the relative error

$$\frac{\|M^k y\|}{\|y\|} \leq \rho(M)^k.$$

for any nontrivial starting vector $y \in \mathbb{C}^n$.

We want to use the foregoing results for deriving and justifying iterative methods for solving $Ax = b$. Throughout the whole course we will tacitly assume that A is regular, so that the linear problem is well posed.

The basic idea is to (additively) split A in an easy-to-invert part and some remainder. The most basic choice is the unit matrix I, and the splitting $A = I + (A - I)$ leads to the equivalent system

$$x = b + (I - A)x.$$

This motivates the fixed-point iteration

$$x_{k+1} = b + (I - A)x_k, \qquad k = 0, 1, 2, \ldots \tag{2.1}$$

with any given $x_0 \in \mathbb{C}^n$.

Definition 2.13. The scheme (2.1) is called *Richardson iteration.* ♦

It is easy to verify that the Richardson iteration satisfies $x_{k+1} = \sum_{j=0}^{k}(I - A)^j b + (I - A)^k x_0$, which reveals that the iteration describes the inversion process of the Neumann series. If $\rho(I-A) < 1$, the iteration converges. This is a consequence of the Neumann series or alternatively follows from the fact that the error $e_k = x_k - x$, satisfies the relation

$$e_{k+1} = (I - A)e_k = \cdots = (I - A)^{k+1} e_0.$$

In this case we can apply the lemma with $M := (I - A)$.

In order to obtain better methods, we need to detract more information from A than just the unit matrix. Since we think of easy-to-solve systems in diagonal or echelon form, we consider the splitting $A = D + L + R$, where D is a diagonal matrix, L is a strict lower (left) triangular matrix and R is a strict upper (right) triangular matrix. By taking D as the invertible part, we obtain the fixed-point equation

$$x = D^{-1}(b - (L + R)x)$$

and the corresponding iteration

$$x_{k+1} = D^{-1}b + Jx_k, \qquad k = 0, 1, 2, \ldots, \quad \text{for } J := -D^{-1}(L + R). \tag{2.2}$$

Definition 2.14. The scheme (2.2) is called *Jacobi iteration.* ♦

Obviously, the Jacobi iteration requires that all diagonal entries of A be nonzero.

By considering the full upper triangular part $D + L$ ab, analogous considerations lead to the iteration

$$x_{k+1} = (D + L)^{-1}b + H_1 x_k, \qquad k = 0, 1, 2, \ldots, \quad \text{for } H_1 := -(D + L)^{-1}R. \tag{2.3}$$

Definition 2.15. The scheme (2.3) is called *Gauss-Seidel* method. ♦

To be well defined, also the *Gauss-Seidel* method requires that all diagonal entries of A be nonzero.

The substitution $Ax = b$ shows that the errors $e_k := x_k - x$ of the iterations satisfy the recurrence relations

$$e_{k+1} := Je_k \quad \text{resp.} \quad e_{k+1} := H_1 e_k.$$

Convergence properties will therefore hinge on the spectral radii $\rho(J)$ for the Jacobi method and $\rho(H_1)$ for the Gauss-Seidel method. We formulate a simple criterion for bounding these.

Definition 2.16. A matrix $A \in \mathbb{C}^{n\times n}$ is said to be *irreducible*, if for $J \in \{1, \ldots, n\}$ the property $A_{j,k} = 0$ for all $j \in J$ and $k \in J^c$ implies $J = \emptyset$ or $J^c = \emptyset$. ♦

Irreducibility is a coupling condition on A. It states that there is no permutation P of the indices that could lead to $\tilde{A} = P^*AP$ having a zero block $\tilde{A}_{J,J^c}$.

Theorem 2.17 (Gershgorin)**.** *Let $A \in \mathbb{C}^{n\times n}$. Any eigenvalue $\lambda \in \mathbb{C}$ of A belongs to the union of the Gershgorin disks*

$$\lambda \in \bigcup_{j=1}^{n} \overline{B}_{r_j}(A_{jj}) \qquad \textit{with the radii } r_j := \sum_{\substack{k=1\\k\neq j}}^{n} |A_{jk}|.$$

If A is irreducible and $\lambda \in \partial(\bigcup_{j=1}^{n} \overline{B}_{r_j}(A_{jj}))$, then $\lambda \in \bigcap_{j=1}^{n} \overline{B}_{r_j}(A_{jj})$.

Proof. Let $x \in \mathbb{C}^n$ with $|x|_\infty$ be an eigenvector corresponding to λ. By separating the components in $Ax = \lambda x$ we obtain with the triangle inequality that

$$|(\lambda - A_{jj})||x_j| \leq \sum_{k\neq j} |A_{jk}||x_k| \leq \sum_{k\neq j} |A_{jk}| = R_j.$$

If j is an index such that $|x_j| = 1$ (which exists thanks to the normalization of x), we have $|\lambda - A_{jj}| \leq R_j$ and, thus, λ belongs to one of the Gershgorin disks.

For the proof of the second assertion, let A be irreducible and let λ belong to the boundary of the union of disks. Assume for contradiction that $\lambda \notin \overline{B}_{r_\ell}(A_{\ell\ell})$ for some ℓ. We define the index set

$$J = \{j \in \{1, \dots, n\} : |x_j| = 1\}.$$

By normalization of x, the set J is nonempty, and so is its complement J^c by our assumption on the index ℓ, otherwise the above chain of inequalities would imply $R_\ell < R_\ell$. We observe that for any $j \in J$ equality holds in the above chain of inequalities. Comparing the sums therein then yields $A_{jk} = 0$ for all $J \in J$ and $k \in J^c$. This contradicts the irreducibility of A. Therefore λ must belong to all disks simultaneously. ■

Definition 2.18. A matrix $A \in \mathbb{C}^{n \times n}$ is said to be *strictly diagonally dominant* if

$$|A_{jj}| > \sum_{\substack{k=1 \\ k \neq j}}^{n} |A_{jk}| \quad \text{for all } j \in \{1, \dots, n\}.$$

It is *weakly diagonally dominant* if

$$|A_{jj}| \geq \sum_{\substack{k=1 \\ k \neq j}}^{n} |A_{jk}| \quad \text{for all } j \in \{1, \dots, n\}$$

and strict inequality holds for at least one j. ♦

Theorem 2.19. *If A is strictly diagonally dominant, both the Jacobi and Gauss-Seidel iterations are convergent and the spectral radius of their iteration matrices satisfies*

$$\rho \leq \max_{1 \leq j \leq n} |A_{jj}|^{-1} \sum_{k=1 k \neq j} |A_{jk}| < 1.$$

If A is weakly diagonally dominant and irreducible, both the Jacobi and Gauss–Seidel iterations are convergent.

Proof. For the Jacobi iteration, the first statement is an immediate consequence of the Gershgorin theorem. For the second statement, we observe that all Gershgorin disks of the iteration matrix have 0 as their midpoint and a radius bounded by 1. An eigenvalue of modulus 1 must therefore belong to all disks simultaneously (because A and therefore the iteration matrix are irreducible). But this is excluded by the one strict inequality in the definition of "weakly diagonally dominant". Therefore all eigenvalues are strictly smaller than 1. The proof for the Gauss–Seidel method is similar and left as an exercise. ■

The Gauss-Seidel iteration uses more information on A than the Jacobi method and therefore seems more powerful. It is, however, important to note that it involves the solution of systems in row-echelon form, which requires serial computations. In contrast, the Jacobi method is parallelizable and does not depend on the enumeration of indices.

§3. Relaxation techniques

We can damp (or *relax*) the effect of a fixed-point method $x_{k+1} = \tilde{b} + Mx_k$ by taking a convex combination of the old and new iterates with some parameter $\omega \in (0, 1]$ on the right-hand side

$$x_{k+1} = \omega(\tilde{b} + Mx_k) + (1 - \omega)x_k = \omega\tilde{b} + M_\omega x_k \quad \text{für } M_\omega := (\omega M + (1 - \omega)I).$$

For example, for the Richardson scheme we have $M = (I - A)$ and $\tilde{b} = b$, while for Jacobi we have $\tilde{b} = D^{-1}b$ and $M = J$. Our goal is to choose ω such that $\rho(M_\omega)$ is small, which would lead to improved convergence properties.

In view of our finite difference model problem we restrict our attention, throughout this paragraph, to Hermitian positive definite Matrices $A \in \mathbb{C}^{n\times n}$. Hence, for Richardson's and Jacobi's iteration, the matrix $I - M$ will be similar to an Hermitian Matrix: while this is immediate for Richardson, we see this for the Jacobi method by considering $I - J = D^{-1}(D + L + R) = D^{-1}A$ and multiplying with $D^{\pm 1/2}$ on both sides, so that $D^{1/2}(I + J)D^{-1/2} = D^{-1/2}AD^{-1/2}$. We say that these methods are *symmetrizable* (unlike the Gauss–Seidel method). A suitable damping parameter (with optimal choice possibly larger than 1) can always enforce convergence for symmetrizable methods for positive definite problems:

Theorem 2.20. *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite and let M denote the iteration matrix of a symmetrizable iteration ($M = I - A$ for Richardson or $M = J = -D^{-1}(L + R)$ for Jacobi). For the choice*

$$\omega_* = \frac{2}{2 - \lambda_{\max}(M) - \lambda_{\min}(M)}$$

we have $\rho(M_{\omega_*}) = |1 - \omega_*(1 - \lambda_{\min}(M))| < 1$ and the relaxed iteration is convergent.

Proof. Any eigenvalue λ_j of M_ω satisfies

$$\lambda_j(M_\omega) = 1 - \omega(1 - \lambda_j(M)).$$

We compute the spectral radius

$$\rho(M_\omega) = \max_j |1 - \omega(1 - \lambda_j(M))| = \max\left\{|1 - \omega(1 - \lambda_{\min}(M))|, |1 - \omega(1 - \lambda_{\max}(M))|\right\}.$$

We want to minimize this maximum, that is we want to find a point where the two shifted and scaled absolute value functions intersect. Symmetrizability implies that the eigenvalues of $I - M$ are real and positive, whence the eigenvalues of M are strictly smaller than 1. We have that $0 < 1 - \lambda_{\max}(M) \leq 1 - \lambda_{\min}(M)$, so we seek the optimal point between the two zeros, namely between $1/(1-\lambda_{\min}(M))$ and $1/(1-\lambda_{\max}(M))$. In that interval we know the sign of the functions and obtain the condition

$$-(1 - \omega(1 - \lambda_{\min}(M))) = 1 - \omega(1 - \lambda_{\max}(M)),$$

which is solved by the value ω_* claimed above. For this value, the displayed expressions are nonnegative. Since the eigenvalues of M are smaller than 1, the value ω_* is positive, and a computation thus shows that $\rho(M_{\omega_*}) < 1$. ■

For the Gauss–Seidel method (not symmetrizable) and we will formulate a different relaxation technique, the *successive over-relaxation* (SOR). Given a parameter $0 < \omega < 2$, it corresponds to solving the following system

$$Dx_{k+1} = \omega(-Lx_{k+1} - Rx_k + b) + (1 - \omega)Dx_k.$$

The terminology "over-relaxation" refers to the fact that with the typical choice $\omega > 1$, we leave the setting of convex combinations. Such choice gives more emphasis to the new (unexplored) iterate and is made for the purpose of extrapolation or convergence acceleration. The method is called "successive" because the (over-)relaxation interacts with the process of solving the system in row-echelon form. Thus an entry of x_{k+1} depends on the entries of x_{k+1} computed earlier in the serial solution process.

Of course we can state the SOR iteration in the usual format of a stationary iteration

$$x_{k+1} = \omega(D + \omega L)^{-1}b + (D + \omega L)^{-1}(-\omega R + (1-\omega)D)x_k.$$

The iteration matrix on the right-hand side is sometimes denoted by H_ω, which is consistent with H_1 from the Gauss–Seidel method.

Theorem 2.21 (Ostrowski-Reich)**.** *Given an Hermitian positive definite matrix $A \in \mathbb{C}^{n\times n}$, the SOR method converges for any $0 < \omega < 2$.*

Proof. Without loss of generality we assume $b = 0$ because we know that convergence only depends on the spectral properties of the iteration matrix. We directly compute from its definition that the SOR iteration is equivalent to

$$(1 - \frac{\omega}{2})D(x_{k+1} - x_k) = \omega\left((-L - \frac{1}{2}D)x_{k+1} + (-L^* - \frac{1}{2}D)x_k\right)$$

where we used $R = L^*$ because A is assumed Hermitian. We multiply this from the left with $(x_{k+1} - x_k)^*$. We observe that some mixed terms cancel and compute using the relation $z^*Lz = \frac{1}{2}z^*(L + R)z$ that

$$(1 - \frac{\omega}{2})(x_{k+1} - x_k)^*D(x_{k+1} - x_k) = \frac{\omega}{2}\left(-x_{k+1}^*Ax_{k+1} + x_k^*Ax_k)\right).$$

We may exclude the pathological case $x_{k+1} = x_k$ because it would imply $Ax_k = b$. Since $0 < \omega < 2$, we then see from the foregoing formula with $k = 0$ that

$$x_1^*Ax_1 < x_0^*Ax_0.$$

From the compactness of the sphere $\{z \in \mathbb{C}^n : z^*Az = 1\}$, we then deduce the stronger property

$$\frac{x_1^*Ax_1}{x_0^*Ax_0} \leq \beta < 1$$

for some positive β and any choice of x_0 (where x_1 is understood as a function of x_0). Inductively we find

$$x_k^*Ax_k \leq \beta^k x_0^*Ax_0.$$

Since A is positive definite, this implies that x_k converges to zero as $k \to \infty$. ■

Remark 2.22. As a byproduct we note that the Gauss–Seidel method (choice $\omega = 1$) converges for any Herminian positive definite matrix. ♦

The Gauss–Seidel method is not symmetric, and our choice of solving for $D + L$ was arbitrary. One could interchange the roles of L and R, which results in a different method with analogous properties. We can symmetrize the SOR method by alternating forward and backward substitution. We count pairs of iteration steps and denote

$$\begin{aligned} Dx_{k+1/2} &= \omega(-Lx_{k+1/2} - Rx_k + b) + (1-\omega)Dx_k, \\ Dx_{k+1} &= \omega(-Lx_{k+1/2} - Rx_{k+1} + b) + (1-\omega)Dx_{k+1/2}. \end{aligned}$$

The iteration can be rewritten as

$$x_{k+1} = Mx_k + B^{-1}b$$

with the matrices

$$M := (D + \omega R)^{-1}((1 - \omega)D - \omega L)(D + \omega L)^{-1}((1 + \omega)D - \omega R)$$

and

$$B := \frac{1}{\omega(2 - \omega)}(D + \omega L)D^{-1}(D + \omega R).$$

as Exercise A.18.

§4. Orthogonal polynomials

In essence, the above basic iterative schemes consist in the evaluation of a matrix polynomial. The analysis of iterative methods therefore makes use of polynomial approximation theory.

We consider polynomials of degree n with real coefficients as functions over a nonempty interval $[a, b]$. They span an $(n + 1)$-dimensional linear space, denoted by P_n. It is a subspace of the continuous functions, $P_n \subseteq C([a, b])$. Given a positive function $\omega : [a, b] \to \mathbb{R}$ such that the integral $\int_a^b \omega(x)\,dx$ is well defined and finite, the following scalar product can be defined on $C([a, b])$,

$$\langle f, g\rangle_{L^2_\omega(a,b)} := \int_a^b f(x)g(x)\omega(x)\,dx \quad \text{for } f, g \in C([a, b]).$$

It is shown in Problem A.10 that this is indeed a scalar product. It induces a norm $\|f\|_{L^2_\omega(a,b)} = (\int_a^b |f(x)|^2\omega(x)\,dx)^{1/2}$. If we apply the Gram–Schmidt process to the monomial basis $(1, x, \ldots, x^n)$ of P_n, we obtain an orthogonal basis. We use the normalization $p_0 \equiv 1$ and $p_k = x^k + q_{k-1}(x)$ for $k = 1, \ldots, n$ and some polynomial $q_{k-1} \in P_{k-1}$. The resulting orthogonal system $(p_0, \ldots, p_n)$ consists of polynomials with leading coefficient 1.

Definition 2.23. The polynomials resulting from the procedure just described are called *orthogonal polynomials* with respect to the weight function ω. ♦

Example 2.24. For the interval $[-1, 1]$ and $\omega \equiv 1$ we obtain multiples of the *Legendre* polynomials known from Gaussian quadrature.

Example 2.25. For the interval $[-1, 1]$ and $\omega(x) = (1 - x^2)^{-1/2}$, the resulting polynomials normalized to the leading coefficient 2^{n-1} are called *Chebyshev polynomials* and denoted by $T_k(x)$.

Orthogonal polynomials always satisfy a three-term recurrence relation. In view of later applications, we first formulate a more abstract result.

Theorem 2.26. *Let V be a linear space with scalar product $\langle\cdot,\cdot\rangle$ and let $A : V \to V$ be a self-adjoint linear map. Suppose we are given a sequence $V_0 \subseteq V_1 \subseteq \cdots \subseteq V$ of subspaces with* $\dim V_k = k + 1$ *and the mapping property*

$$A(V_k) \subseteq V_{k+1} \quad \textit{and} \quad A(V_k) \not\subseteq V_k.$$

Then, for any $p_0 \in V_0$ and real numbers $\eta_0, \eta_1, \ldots, \in \mathbb{R} \setminus \{0\}$ there exists a unique extension to an orthogonal system $(p_k)_{k\geq 0}$ (with $p_k \in V_k$ for all k) satisfying the normalization

$$\langle p_k, p_k\rangle = \eta_{k-1}\langle Ap_{k-1}, p_k\rangle \quad \textit{for all } k \geq 1. \tag{2.4}$$

The p_k $(k \geq 1)$ are given through the following three-term recurrence relation

$$p_k = (\eta_{k-1}A - \delta_k)p_{k-1} - \gamma_k^2 p_{k-2} \quad \textit{für } k = 1, 2, \dots,$$

where $p_{-1} := 0$ and the coefficients read

$$\delta_k := \eta_{k-1}\frac{\langle Ap_{k-1}, p_{k-1}\rangle}{\langle p_{k-1}, p_{k-1}\rangle} \quad \textit{und} \quad \gamma_k^2 := \frac{\eta_{k-1}}{\eta_{k-2}}\frac{\langle p_{k-1}, p_{k-1}\rangle}{\langle p_{k-2}, p_{k-2}\rangle}.$$

Proof. We perform induction over $k \geq 1$. Let $(p_0, \dots, p_{k-1})$ be an orthogonal system and let $p_k \in V_k$ be orthogonal to all p_j with $j < k$. We assume the normalization (2.4) for p_k. Then $p_k - \eta_{k-1}Ap_{k-1} \in V_{k-1}$ and, hence, there are coefficients $c_0, \dots, c_{k-1}$ satisfying the expansion

$$p_k - \eta_{k-1}Ap_{k-1} = \sum_{j=0}^{k-1} c_j p_j, \quad \text{namely } c_j = \frac{\langle p_k - \eta_{k-1}Ap_{k-1}, p_j\rangle}{\langle p_j, p_j\rangle}.$$

The orthogonality property of p_k and the self-adjointness of A prove

$$\langle p_k - \eta_{k-1}Ap_{k-1}, p_j\rangle = -\langle \eta_{k-1}p_{k-1}, Ap_j\rangle,$$

whence $c_0, \dots, c_{k-3} = 0$. Furthermore

$$c_{k-2} = -\frac{\langle \eta_{k-1}Ap_{k-1}, p_{k-2}\rangle}{\langle p_{k-2}, p_{k-2}\rangle} = -\frac{\eta_{k-1}}{\eta_{k-2}}\frac{\langle p_{k-1}, p_{k-1}\rangle}{\langle p_{k-2}, p_{k-2}\rangle} \quad \text{und} \quad c_{k-1} = -\frac{\eta_{k-1}\langle Ap_{k-1}, p_{k-1}\rangle}{\langle p_{k-1}, p_{k-1}\rangle}.$$

This is the claimed recursion. The uniqueness of p_k was shown through the construction process. ■

For the application to orthogonal polynomials, we take Theorem 2.26 with $V_k = P_k$ and the inner product $\langle \cdot, \cdot \rangle_{L^2_\omega(a,b)}$. The linear map A is the multiplication by x, namely $Ap(x) = xp(x)$. The normalization (2.4) states that $p_k(x) - xp_{k-1}(x)$ belongs to P_{k-1}. From $p_0 = 1$ we deduce that p_k has 1 as its leading coefficient. We note the following consequence.

Corollary 2.27. *The orthogonal polynomials over $[a, b]$ with respect to the weight ω and leading coefficient 1 satisfy the three-term recurrence relation*

$$p_k(x) = (x - \delta_k)p_{k-1}(x) - \gamma_k^2 p_{k-2}(x) \quad \textit{for } k = 1, 2, \dots,$$

with $p_{-1} := 0$, $p_0 \equiv 1$ and the coefficients

$$\delta_k := \frac{\int_a^b x\, p_{k-1}(x)^2 \omega(x) dx}{\|p_{k-1}\|^2_{L^2_\omega(a,b)}} \quad \textit{und} \quad \gamma_k^2 := \frac{\|p_{k-1}\|^2_{L^2_\omega(a,b)}}{\|p_{k-2}\|^2_{L^2_\omega(a,b)}}.$$

■

The roots of orthogonal polynomials satisfy the following important property.

Theorem 2.28. *All roots of the orthogonal polynomial $p_k \in P_k$ over $[a, b]$ are real, simple, and lie in (a, b).*

Proof. Let $\lambda_1, \dots, \lambda_m \in (a,b)$ denote the points where p_k changes sign in (a,b). We define

$$q(x) := \prod_{j=1}^{m} (x - \lambda_j).$$

The product satisfies $p_k q \in P_{k+m}$ and has no change of sign in (a,b). Consequently,

$$\int_a^b p_k(x) q(x) \omega(x)\, dx \neq 0$$

Since p_k is orthogonal on P_{k-1}, we necessarily have $m = k$. ■

From elementary numerical analysis courses it is known that the interpolation error for standard Lagrange interpolation in the max-norm is proportional to max norm of the normalized polynomial $\prod_{j=0}^{n}(x - x_j)$ where the x_j are the interpolation points $x_0, \dots, x_{n-1}$. Consequently it is sensible to choose these points as the zeros of some $q \in P_n$ with leading coefficient 1 satisfying

$$\max_{x\in[a,b]} |q(x)| = \min \big\{ \max_{x\in[a,b]} |s(x)| : s \in P_n \text{ with leading coefficient } 1 \big\}.$$

We will achieve this via the Chebyshev polynomials T_n. In Exercise A.20, it is shown that T_n has the leading coefficient $a_n = 2^{n-1}$, satisfies the bound $|T_n(x)| \leq 1$ for any $x \in [-1,1]$ and $|T_n|$ attains the value 1 only at the *Chebyshev nodes* $z_k = \cos(k\pi/n)$ for $k = 0, \dots, n$. The sign satisfies $T_n(z_k) = (-1)^k$. The next result shows that the Chebyshev polynomials are in some sense minimal.

Theorem 2.29. *Any polynomial $p_n \in P_n$ with leading coefficient 1 attains a value with $|p_n(x)| \geq 2^{1-n}$ at some $x \in [-1,1]$.*

Proof. Assume for contradiction that $p_n \in P_n$ has leading coefficient 1 and the property $|p_n(x)| < 2^{1-n}$ for all $x \in [-1,1]$. Then, at the Chebyshev points we would have

$$2^{1-n} T_n(z_k) - p_n(z_k) \begin{cases} > 0 & \text{for } k \text{ even,} \\ < 0 & \text{for } k \text{ odd.} \end{cases}$$

By continuity, $2^{1-n}T_n - p_n$ has at least n zeros and, by construction, $2^{1-n}T_n - p_n \in P_{n-1}$. In conclusion, the difference must be zero, contradicting the strict inequalities displayed above. ■

The foregoing theorem implies that the scaled Chebyshev polynomial $2^{1-n}T_n$ solves the minmax problem stated above. A conclusion usually taught in courses of elementary numerical analysis is that the Chebyshev points are a "good" choice for interpolation points. In this lecture, the minmax property will help us to identify optimal matrix polynomials. We briefly mention how the results generalize if the interval $[-1,1]$ is transformed to $[a,b]$. The affine transformation reads

$$y(x) = 2\frac{x-a}{b-a} - 1.$$

If $p \in P_n$ is a suitable polynomial for $[-1,1]$ with leading coefficient 1, the polynomial $p(y(x))$ transformed to the interval $[a,b]$ has the leading coefficient $2^n/(b-a)^n$. The following result states a general re-scaling procedure.

Theorem 2.30. *Given an interval $[a, b]$ and some $\eta \in \mathbb{R} \setminus [a, b]$, then the polynomial*

$$\hat{T}_n(x) := \frac{T_n(y(x))}{T_n(y(\eta))}$$

minimizes $\max_{x\in[a,b]} |q(x)|$ *amongst all polynomials $q \in P_n$ with $q(\eta) = 1$.*

Proof. From Theorem 2.28 we know that all zeros of $T_n(y(x))$ lie in $[a, b]$. Therefore $c := T_n(y(\eta)) \neq 0$ and $\hat{T}_n$ is a well defined polynomial. We see that $\hat{T}_n(\eta) = 1$, and from the bound $|T_n| \leq 1$ in $[a, b]$ we obtain $|\hat{T}_n(x)| \leq 1/|c|$ for all $x \in [a, b]$. Assume for contradiction that some $q_n \in P_n$ with $q_n(\eta) = 1$ satisfies the strict inequality $|q_n| < 1/|c|$ in $[a, b]$. We first observe

$$\hat{T}_n(x) - q_n(x) = s_{n-1}(x)(x - \eta)$$

for some polynomial $s_{n-1} \in P_{n-1}$ because the difference vanishes in η. The argument from the proof of Theorem 2.29 then shows that $\hat{T}_n(x) - q_n(x)$ must have n zeros $[a, b]$, which contradicts $\eta \notin [a, b]$. ■

§5. The cg method

Throughout this paragraph, $A \in \mathbb{R}^{n\times n}$ is symmetric and positive definite (s.p.d.). Such a matrix induces a scalar product and a norm, the so-called energy norm.

Definition 2.31. For $A \in \mathbb{R}^{n\times n}$ s.p.d., the energy scalar product and norm are defined by

$$\langle y, z\rangle_A := \langle y, Az\rangle_2 \quad \text{und} \quad \|z\|_A := \sqrt{\langle z, Az\rangle_2} \quad \text{for } y, z \in \mathbb{R}^n.$$

♦

In this setting we can characterize solutions to a linear systems $Ax = b$ as minimizers of the so-called *energy functional*

$$f : \mathbb{R}^n \to \mathbb{R}, \qquad f(z) = \frac{1}{2} z^* Az - z^* b.$$

Lemma 2.32. *Let $A \in \mathbb{R}^{n\times n}$ s.p.d. and $b \in \mathbb{R}^n$. A vector $x \in \mathbb{R}^n$ solves $Ax = b$ if and only if*

$$f(x) = \min_{z\in\mathbb{R}^n} f(z).$$

Proof. Exercise A.19. ■

With this view on the linear problem, we now aim at approximating the solution x by the solution x_k of the minization problem restricted to a subspace $V_k \subseteq \mathbb{R}^n$.

Definition 2.33 (Galerkin method)**.** Given a subspace $V_k \subseteq \mathbb{R}^n$ and some $x_0 \in \mathbb{R}^n$, the *Galerkin method* is to compute $x_k \in V_k$ satisfying

$$f(x_k) = \min_{z\in x_0+V_k} f(z).$$

♦

Given some $x_k \in \mathbb{R}^n$, there gradient of f at x_k satisfies $g_k := \nabla f(x_k) = Ax_k - b$ and therefore equals the residual. As a necessary criterion for a minimum we have that $\langle g_k, v\rangle_2$ vanishes for every $v \in V_k$, written $g_k \perp_2 V_k$. As a consequence, we obtain that x_k and x_ℓ for the Galerkin solution x_ℓ from a larger space $(x_0 + V_k) \subseteq (x_0 + V_\ell)$ satisfy

$$\langle Ax_k, v\rangle_2 = \langle b, v\rangle_2 = \langle Ax_\ell, v\rangle_2 \quad \text{for all } v{\in}V_k.$$

Therefore

$$\langle A(x_\ell - x_k), v\rangle_2 = \langle x_\ell - x_k, v\rangle_A = 0 \quad \text{for all } v \in V_k,$$

or in shorthand notation $(x_\ell - x_k) \perp_A V_k$. In particular we have $(x - x_k) \perp_A V_k$. The fact that in the Galerkin method the error is A-orthogonal to the approximating subspace is referred to as *Galerkin orthogonality.*

We know that there information best available from a large sparse matrix A is the multiplication with a vector.

Definition 2.34. The spaces $V_k := \operatorname{span}\{g_0, Ag_0, \ldots, A^{k-1}g_0\}$ (for $0 \le k \le n-1$) are called *Krylov spaces.* ♦

Definition 2.35. The Galerkin method with respect to the (shifted) Krylov spaces $x_0 + V_k$ is called *cg method.* ♦

The cg therein abbreviates “conjugate gradient”, and we will later justify this name, the meaning of which is irrelevant for the moment. We observe from $x_k \in x_0 + V_k$ that the gradient $g_k = Ax_k - b$ belongs to V_k. Without having an algorithm to compute the cg approximation at the moment, we can formulate an error estimate.

Theorem 2.36. *For any starting vector $x_0 \in \mathbb{R}^n$, the cg method satisfies the error estimate*

$$\|x_k - x\|_A \le 2\left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1}\right)^k \|x_0 - x\|_A.$$

Proof. Galerkin orthogonality implies that $(x_k - x)$ is A-orthogonal to V_k. Therefore, any $y \in V_k$ satisfies

$$\|x_k - x\|_A^2 = \langle x_k - x, x_k - x\rangle_A = \langle y - x_0 - x, x_k - x\rangle_A \le \|y + x_0 - x\|_A \|x_k - x\|_A.$$

We divide by $\|x_k - x\|_A$ and take the minimum over all y. Since any $y \in x_0 + V_k$ can be written as $y = x_0 + p(A)g_0$ for some polynomial $p \in P_{k-1}$, we obtain

$$\|x_k - x\|_A = \min_{y \in x_0 + V_k} \|y - x\|_A = \min_{p \in P_{k-1}} \|x_0 + p(A)g_0 - x\|_A.$$

In particular, x_k is the best approximation from $x_0 + V_k$ to x. Since $g_0 = A(x_0 - x)$, we can write $x_0 + p(A)g_0 - x = (1 + p(A)A)(x_0 - x)$ and therefore

$$\|x_k - x\|_A = \min_{\substack{q \in P_k \\ q(0)=1}} \|q(A)(x_0 - x)\|_A \le \min_{\substack{q \in P_k \\ q(0)=1}} \|q(A)\|_A \|(x_0 - x)\|_A.$$

Since A is symmetric and positive definite, it can be diagonalized and $\sigma(A) \subseteq [a, b]$ for positive numbers a, b. We substitute $w = A^{1/2}z$ and compute

$$\|q(A)\|_A = \sup_{z \in \mathbb{R}^n \setminus \{0\}} \frac{\|q(A)z\|_A}{\|z\|_A} = \sup_{w \in \mathbb{R}^n \setminus \{0\}} \frac{\|q(A)A^{-1/2}w\|_A}{\|A^{-1/2}w\|_A} = \sup_{w \in \mathbb{R}^n \setminus \{0\}} \frac{\|q(A)w\|_2}{\|w\|_2} = \|q(A)\|_2.$$

By diagonalizing A, we see that the eigenvalues of $q(A)$ are simply given by $q(\lambda)$ for the eigenvalues λ of A. Therefore we conclude $\|q(A)\|_A = \max_{\lambda\in\sigma(A)} |q(\lambda)|$ and obtain

$$\|x_k - x\|_A \le \min_{\substack{q\in P_k\\ q(0)=1}} \max_{\lambda\in\sigma(A)} |q(\lambda)| \; \|x_0 - x\|_A.$$

Since $0 \notin [a,b]$, we can apply Theorem 2.30, which states

$$\min_{\substack{q\in P_k\\ q(0)=1}} \max_{\lambda\in[a,b]} |q(\lambda)| = \max_{\lambda\in[a,b]} \frac{|T_k(y(\lambda))|}{|T_k(y(0))|}$$

for the kth Chebyshev polynomial T_k and $y(\lambda) = 2(\lambda - a)/(b-a) - 1$. From the boundedness $|T_k| < 1$ on $[-1,1]$, its symmetry, and $y(0) = -(b+a)/(b-a)$ we infer from $b = \lambda_{\max}$ and $a = \lambda_{\min}$ that

$$\min_{\substack{q\in P_k\\ q(0)=1}} \max_{\lambda\in[a,b]} |q(\lambda)| \le |T_k(y(0))|^{-1} = \left|T_k\left(\frac{b+a}{a-b}\right)\right|^{-1} = \left|T_k\left(\frac{\kappa+1}{\kappa-1}\right)\right|^{-1}$$

for $\kappa := \kappa_2(A) = \lambda_{\max}/\lambda_{\min}$. We know from Exercise A.20 that

$$T_k(z) = \frac{1}{2}\left((z + \sqrt{z^2-1})^k + (z - \sqrt{z^2-1})^k\right)$$

and thus verify for $z = (\kappa+1)/(\kappa-1)$ by a direct computation that

$$T_k\left(\frac{\kappa+1}{\kappa-1}\right) = \frac{1}{2}\left(\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\right) \ge \frac{1}{2}\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k.$$

The assertion follows from combining the foregoing estimates. ■

From Galerkin orthogonality we deduce the relation $x_{k+1} = x_k + \alpha_k d_k$ for some $d_k \in V_{k+1}$ which is orthogonal to V_k and some scaling $\alpha_k \in \mathbb{R}$. For the gradient this implies

$$g_{k+1} = g_k + \alpha_k A d_k.$$

Before computing a representation for d_k, we deduce a formula for α_k. Since every $y_{k+1} \in V_{k+1}$ can be written as $y_{k+1} = y_k + c d_k$ for some $y_k \in V_k$ and some $c \in \mathbb{R}$, we have

$$0 = \langle g_{k+1}, y_{k+1}\rangle_2 = \langle g_k + \alpha_k A d_k, y_k + c d_k\rangle_2 = \langle g_k + \alpha_k A d_k, c d_k\rangle_2 = c(\langle g_k, d_k\rangle_2 + \alpha_k \|d_k\|_A^2).$$

where the terms involving y_k vanish due to the orthogonality of $Ax_k - b$ and d_k to V_k. With elementary manipulations we thus obtain

$$\alpha_k = -\frac{\langle g_k, d_k\rangle_2}{\langle d_k, d_k\rangle_A}.$$

It is not difficult to verify (Exercise A.24) that $\dim V_k = k$ as long as the gradient g_k does not vanish, $g_k \neq 0$ (otherwise $x_k = x$ is the global solution). Thus we are in the setting of Theorem 2.26.

Lemma 2.37. *In the cg method with $d_0 = -g_0$ we have*

$$\alpha_k = \frac{\|g_k\|_2^2}{\|d_k\|_A^2}.$$

Furthermore, the sequence

$$d_{k+1} = -g_{k+1} + \beta_k d_k \quad \textit{where } \beta_k := \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2},$$

satisfies $d_{k+1} \in V_{k+1}$ and $d_{k+1} \perp_A V_k$ as long as $g_k \neq 0$.

Proof. The proof is by induction. The assertion is obviously true for $k = 0$. Assume the result is true for $(k-1) \geq 0$. We then have from the induction hypothesis

$$-\langle g_k, d_k\rangle_2 = -\langle g_k, -g_k + \beta d_{k-1}\rangle_2 = \langle g_k, g_k\rangle_2 = \|g_k\|_2^2$$

because g_k is 2-orthogonal on V_k. This proves the representation for α_k.

From Theorem 2.26 we know that d_{k+1} is determined by a three-term recurrence relation. For $\eta_j := -\alpha_j$ we obtain

$$d_{k+1} = -\alpha_k\left(A - \frac{\langle Ad_k, d_k\rangle_A}{\|d_k\|_A^2}\right)d_k - \frac{\alpha_k}{\alpha_{k-1}}\frac{\|d_k\|_A^2}{\|d_{k-1}\|_A^2}d_{k-1}$$

with the desired orthogonality properties. With the formula for α_k, α_{k-1} that we just proved, a straightforward computation shows that the last term (including the sign) on the right-hand side equals $-\beta_{k-1}d_{k-1}$, which, thanks to the induction hypothesis, equals $-(d_k + g_k)$. Using this and $\alpha_k Ad_k = g_{k+1} - g_k$ we can simplify the three-term recurrence relation as follows

$$d_{k+1} = -g_{k+1} + g_k + \alpha_k\frac{\langle Ad_k, d_k\rangle_A}{\|d_k\|_A^2}d_k - d_k - g_k = -g_{k+1} + \left(\alpha_k\frac{\langle Ad_k, d_k\rangle_A}{\|d_k\|_A^2} - 1\right)d_k.$$

We now compute with $\alpha_k Ad_k = g_{k+1} - g_k$ and the 2-orthogonality of g_k and g_{k+1} (because $g_k \in V_k$) that

$$\alpha_k\frac{\langle Ad_k, d_k\rangle_A}{\|d_k\|_A^2} = \frac{1}{\alpha_k}\frac{\langle \alpha_k Ad_k, \alpha_k Ad_k\rangle_2}{\|d_k\|_A^2} = \frac{1}{\alpha_k}\frac{\|g_k\|_2^2 + \|g_{k+1}\|_2^2}{\|d_k\|_A^2} = \beta_k + 1.$$

Combining the last two displayed expressions yields the assertion. ■

The resulting algorithm is as follows.

Algorithm 2.38 (cg method)**.** Start: x_0 and $d_0 = -g_0 = Ax_0 - b$ For $k = 0, 1, 2, \ldots$ (as long as $g_k \neq 0$) iterate

- $\alpha_k := \frac{g_k^* g_k}{d_k^* A d_k}$
- $x_{k+1} := x_k + \alpha_k d_k$
- $g_{k+1} := g_k + \alpha_k A d_k$
- $\beta_k := \frac{g_{k+1}^* g_{k+1}}{g_k^* g_k}$
- $d_{k+1} := -g_{k+1} + \beta_k d_k$

♦

Remark 2.39. After n steps (in exact arithmetic), the Krylov space equals $\mathbb{R}^n$ and $x_n = x$ is the exact solution. In principle, we therefore have a direct method. In practice, the cg method is viewed as an iterative method and one observes good approximations for small $k \ll n$. ♦

§6. Other descent methods

The cg method can be viewed as an instance of descent methods for minimizing the functional f over $\mathbb{R}^n$. The general procedure is as follows.

Definition 2.40 (descent method)**.** We call any algorithm of the following form a *descent method*: Choose a starting vector x_0 and iterate for $k = 0, 1, 2, \ldots$:

- choose $d_k \in \mathbb{R}^n$ *(descent direction)*
- minimize the functional $f(x_k + td_k)$ with respect to t *(line search)* and denote the minimizer by $\alpha_k \in \mathbb{R}$
- set $x_{k+1} := x_k + \alpha_k d_k$

♦

Remark 2.41. The value for α_k can be explicitly computed. Indeed, the function

$$h(t) := f(x_k + td_k) = \frac{1}{2}(x_k + td_k)^* A(x_k + td_k) - (x_k + td_k)^* b$$

is quadratic in t with derivative

$$h(t)' = td_k^* Ad_k + d_k^* Ax_k - d_k^* b_k = td_k^* Ad_k + d_k^* r_k$$

with the *residual* $r_k := Ax_k - b_k$. Since h is a convex parabola, the condition $h(t)' = 0$ is necessary and sufficient of a minimum. This means

$$\alpha_k = -\frac{d_k^* r_k}{d_k^* Ad_k}.$$

♦

Example 2.42. The cg method is the descent method with $d_0 = -g_0$ and mutually A-orthogonal descent directions.

Example 2.43 (gradient descent)**.** Since the negative gradient points in the direction of the steepest descent, it is reasonable to choose $d_k = -g_k = -\nabla f(x_k)$ as descent direction. This is the *gradient method.* It is easy to verify that the iterates belong to the same affine Krylov spaces as in the cg case

$$x_k \in x_0 + \operatorname{span}\{g_0, Ag_0, \ldots, A^{k-1} g_0\}$$

and that two consecutive descent directions are 2-orthogonal, $\langle g_{k+1}, g_k \rangle_2 = 0$ (because the partial derivative $\langle \nabla f(x_{k+1}), g_k \rangle$ is zero at the minimum in the line search). However, the descent directions are not all pairwise orthogonal. It can be proven (Exercise A.23) that the gradient method converges with the following upper bound

$$\|x_k - x\|_A \le \left(\frac{\kappa_2(A) - 1}{\kappa_2(A) + 1} \right)^k \|x_0 - x\|_A.$$

This is much slower than for the cg method; the reason being that for large condition numbers of A, the directions can be “almost parallel” and the sequence x_k has an oscillatory behaviour.

The *conjugate gradient* (cg) method can therefore be seen as an improvement of the gradient method where A-orthogonality of the directions d_k is enforced. An alternative wording for “orthogonal” is “conjugate”, hence the nomenclature.

§7. Preconditioning

The convergence speed of the cg method heavily depends on the spectral condition number $\kappa_2(A)$ of the matrix A. In order to improve the convergence of the iteration, we want to transform A in such a way that the condition number is moderate. Assume we know a linear map, denoted by a matrix $C \in \mathbb{R}^{n\times n}$, which leads to moderate values of $\kappa_2(CA)$. We then transform our system to the modified system $CAx = Cb$ and apply the cg method to it. We assume C to be symmetric and positive definite. Indeed, if C has the factorization $C = EE^*$, this is equivalent to minimizing

$$\tilde f(y) = \frac{1}{2} y^* \tilde A y - \tilde b^* y \quad \text{with } \tilde A = E^* A E \text{ and } \tilde b = E^* b.$$

The minimizer is $y = E^{-1}x$ and can be theoretically computed with the cg method. In practice we do not want (and do not need) to use the factorization of C. With the above substitutions we we have indeed

$$\tilde g_k := \nabla f(y_k) = \tilde A y_k - \tilde b \quad \text{so that} \quad E\tilde g_k = CAx - Cb = Cg_k.$$

The descent directions $\tilde d_k$ is substituted as follows $E\tilde d_k = d_k$. With these substitutions we obtain a practical algorithm. (on the left we display the theoretical version for the purpose of derivation).

Algorithm 2.44 (pcg method)**.** (the actual algorithm is that on the right-hand side)

```
Theoretical derivation:
Start: $y_0$ and $\tilde d_0 = -\tilde g_0 = \tilde A x_0 - \tilde b$
For $k = 0, 1, 2, \ldots$:
    • $\alpha_k := \frac{\tilde g_k^* \tilde g_k}{\tilde d_k^* \tilde A \tilde d_k}$
    • $y_{k+1} := y_k + \alpha_k \tilde d_k$
    • $\tilde g_{k+1} := \tilde g_k + \alpha_k \tilde A \tilde d_k$
    • $\beta_k := \frac{\tilde g_{k+1}^* \tilde g_{k+1}}{\tilde g_k^* \tilde g_k}$
    • $\tilde d_{k+1} := -\tilde g_{k+1} + \beta_k \tilde d_k$
```

```
Practical version:
Start: $x_0$ and $d_0 = -Cg_0 = -C(Ax_0 - b)$
For $k = 0, 1, 2, \ldots$:
    • $\alpha_k := \frac{g_k^* C g_k}{d_k^* A d_k}$
    • $x_{k+1} := x_k + \alpha_k d_k$    ♦
    • $g_{k+1} := g_k + \alpha_k A d_k$
    • $\beta_k := \frac{g_{k+1}^* C g_{k+1}}{g_k^* C g_k}$
    • $d_{k+1} := -Cg_{k+1} + \beta_k d_k$
```

Of course, in practice the vectors Cg_k are stored as temporary variables. The matrix C is referred to as *preconditioner* and the resulting method is the *pcg method* (preconditioned cg). In practice, C need not be represented by a matrix. Instead it can be given by a function or an algorithm.

From the theoretical version of the algorithm we immediately deduce from Theorem 2.36 the error estimate

$$\|y_k - y\|_{\tilde A} \le 2 \left(\frac{\sqrt{\kappa_2(CA)} - 1}{\sqrt{\kappa_2(CA)} + 1} \right)^k \|y_0 - y\|_{\tilde A}$$

because $\tilde A$ and CA have the same eigenvalues. With $Ey = x$ we finally obtain the following.

Corollary 2.45. *For any starting vector $x_0 \in \mathbb{R}^n$, the pcg method with positive definite preconditioner C satisfies the error estimate*

$$\|x_k - x\|_A \le 2 \left(\frac{\sqrt{\kappa_2(CA)} - 1}{\sqrt{\kappa_2(CA)} + 1} \right)^k \|x_0 - x\|_A.$$

We thus have seen that finding a good preconditioner can substantially speed up the performance of the cg method.

§8. The SSOR preconditioner

We study one basic preconditioner for the cg method applied to FDM. If we perform one step of SSOR with $x_0 = 0$, we obtain $x_1 = Cb$ with

$$C := \omega(2-\omega)(D+\omega R)^{-1}D(D+\omega L)^{-1}.$$

This is our choice of an "approximate inverse" of A and thus as a preconditioner. Note that the symmetry of A implies $R = L^*$ and thus C is symmetric and positive definite.

Let us define the following quantities:

$$\mu := \max_{z\neq 0} \frac{z^* D z}{z^* A z}, \qquad \delta := \max_{z \neq 0} \frac{z^*(LD^{-1}L^* - \frac{1}{4}D)z}{z^* A z}$$

as well as

$$G(\omega) := (2-\omega)^{-1}\left(1 + \mu\frac{(2-\omega)^2}{4\omega} + \omega\delta\right) \quad \text{for } 0 < \omega < 2.$$

Theorem 2.46. *With the above definitions we have $\kappa_2(CA) \leq G(\omega)$ for any $\omega \in (0,2)$.*

Proof. The matrix CA is positive definite and its smallest resp. largest eigenvalue satisfies

$$\tilde{\lambda}_{\min} = \min_{z\neq 0} \frac{z^* A z}{z^* C^{-1} z} \quad \text{resp.} \quad \tilde{\lambda}_{\max} = \max_{z\neq 0} \frac{z^* A z}{z^* C^{-1} z}.$$

Details will be worked out as Exercise A.17. In the proof, we will show that $\tilde{\lambda}_{\max} \leq 1$ and that $\tilde{\lambda}_{\min} \geq 1/G(\omega)$, which then implies the assertion.

With the matrices

$$V := \frac{1}{\omega}((\omega-1)D + \omega L) \quad \text{and} \quad W := \frac{1}{\omega}(2-\omega)D$$

we straighforwardly compute

$$C^{-1} = (V+W)W^{-1}(V+W)^* = A + VW^{-1}V^*.$$

Since both terms on the right-hand side describe positive definite matrices, we see, after multiplying with any vector z from both sides, that $\tilde{\lambda}_{\max} \leq 1$. For bounding the lowest eigenvalue, we use the representation

$$C^{-1} = (2-\omega)^{-1}\left(A + \frac{1}{4\omega}(2-\omega)^2 D + \omega\left(LD^{-1}L^* - \frac{1}{4}D\right)\right),$$

which will be proved in Exercise A.28. The numbers μ, δ satisfy

$$z^* D z \leq \mu z^* A z \quad \text{and} \quad z^*(LD^{-1}L^* - \frac{1}{4}D)z \leq \delta z^* A z.$$

This and the above representation of C^{-1} show

$$z^* C^{-1} z \leq G(\omega) z^* A z$$

whence $\tilde{\lambda}_{\min} \geq 1/G(\omega)$. ■

To make the condition number of CA as small as possible, we can minimize the function G over the interval $(0,2)$. The result is as follows.

Lemma 2.47. *The function G attains its minimum over $(0,2)$ at the point*

$$\omega_* = \frac{2}{1+\frac{2}{\sqrt{\mu}}\sqrt{\frac{1}{2}+\delta}} \quad \textit{with the value} \quad G(\omega_*) = \sqrt{(\frac{1}{2}+\delta)\mu}+\frac{1}{2}.$$

Proof. Exercise A.29. ■

Testing the max in the definition of δ with any Cartesian unit vector reveals $\delta \geq -1/4$. If, in addition, we have $\|D^{-1/2}LD^{-1/2}\|_\infty \leq 1/2$ and $\|D^{-1/2}L^*D^{-1/2}\|_\infty \leq 1/2$, then the numerator in the definition of δ satisfies, after the substitutions $y := D^{1/2}z$ and $\tilde{L} := D^{-1/2}LD^{-1/2}$ that

$$z^*(LD^{-1}L^*)z - \frac{1}{4}z^*Dz = y^*\tilde{L}\tilde{L}^*y - \frac{1}{4}y^*y \leq 0$$

and therefore $-1/4 \leq \delta \leq 0$. We will now see that for this case, the SSOR preconditioning improves the spectral condition number to the square root of the original $\kappa_2(A)$.

Lemma 2.48. *Let A satisfy $\|D^{-1/2}LD^{-1/2}\|_\infty \leq 1/2$ and $\|D^{-1/2}L^*D^{-1/2}\|_\infty \leq 1/2$. With the optimal value ω_* we have*

$$\kappa_2(CA) \leq \sqrt{\frac{1}{2}\kappa_2(A)}+\frac{1}{2}.$$

Proof. We combine Theorem 2.46 and Lemma 2.47. We have already seen that $\delta \leq 0$ so that we are left with showing that $\mu \leq \kappa_2(A)$. We write

$$\mu = \max_{z\neq 0} \frac{z^*Dz/(z^*z)}{z^*Az/(z^*z)}.$$

From the Rayleigh quotient we see that the denominator is bounded from below by the smallest eigenvalue $\lambda_{\min}$ of A. The numerator is bounded by the largest diagonal entry d of D, which is a lower bound to the largest eigenvalue $\lambda_{\max}$ of A (because testing with an appropriate unit vector z shows $d = z^*Az \leq \lambda_{\max}$). Therefore $\mu \leq \lambda_{\max}/\lambda_{\min} = \kappa_2(A)$. ■

We conclude with the principal result of this section. Recall from Exercise A.27 that the spectral condition of the finite difference system matrix A scales like $\kappa_2(A) = \mathcal{O}(h^{-2})$.

Theorem 2.49. *The SSOR preconditioning with optimal parameter ω_* reduces the condition number of the FDM system matrix from $\kappa_2(A) = O(h^{-2})$ to $\kappa_2(CA) = O(h^{-1})$.*

Proof. We explicitly verify $\|D^{-1/2}LD^{-1/2}\|_\infty \leq 1/2$ and $\|D^{-1/2}L^*D^{-1/2}\|_\infty \leq 1/2$. Indeed, from the structure of the FDM matrix we see that in each row and column $D^{-1/2}LD^{-1/2}$ we have twice the value $-1/4$ and use the foregoing lemma. ■

In practice, it is difficult to determine the precise value of ω_*. But it turns out that the performance of the preconditioner is quite insensitive to deviations from the optimal value, see the programming exercises.

§9. The smoothing property of iterative methods

Both experience and theoretical considerations show that the classical iterative solvers behave poorly for the finite difference system because of the conditioning $\kappa_2(A) = O(h^{-2})$. We will now take closer look at the spectral decomposition of the error $x - x_k$ and deduce a favourable property of the relaxed iterations that was not quantified in our asymptotic error estimates. In the exercises it was shown that the eigenvalues of the FDM system (with homogeneous Dirichlet boundary conditions) are given by

$$\lambda_{j,k} = 4(\sin^2(\frac{1}{2}j\pi h) + \sin^2(\frac{1}{2}k\pi h))$$

for any $j, k = 1, 2, \dots, J-1$. The eigenvalues of the iteration matrix given by the relaxed Richardson method are therefore

$$\lambda_{j,k}(I - \omega A) = 1 - 4\omega(\sin^2(\frac{1}{2}j\pi h) + \sin^2(\frac{1}{2}k\pi h)).$$

If we choose $\omega = 1/4$, which corresponds to the ordinary Jacobi iteration, we see that for small values of j, k the modulus of the eigenvalues $|\lambda_{j,k}(I - \omega A)|$ is close to 1. The contribution of the error that belongs to the corresponding eigenvectors is therefore reduced very slowly, and the same is true if j and k are close to $(J-1)$. Values of j or k near $J/2$ instead lead to contributions that are rapidly damped. The choice $\omega = 1/8$, the terms related to values of j or k close to $(J-1)$ will also experience an adequate damping so that after a few iterations the error can be expected to be essentially described by linear combinations of eigenvectors for "small" j and k. In view of the eigenvectors (or eigen-mesh-functions) $U_{m,n} = \sin(j\pi m h)\sin(k\pi n h)$ we deduce that some "oscillatory" part of the error is rapidly damped so that we are left with a "smooth" part.

We want to formalize these observations. In general we call a function *oscillatory* if it is (suitably) bounded but exhibits significant variations on small scales, so it has large gradients. In one dimension the prototype is $\sin(k\pi x)$ on $[0, 1]$ with L^2 norm uniformly bounded by 1 while its derivative has an L^2 norm that grows linearly in k. Analogous properties are true in higher dimensions where we say a function u is oscillatory if

$$\|u\|_{L^2(\Omega)} \ll \|\nabla u\|_{L^2(\Omega)}.$$

For a quantification of this behaviour, we can consider the eigenvalues of the Dirichlet Laplacian. If we have a function with zero boundary conditions satisfying $-\Delta u = \lambda u$, from integration by parts we obtain

$$\|\nabla u\|^2_{L^2(\Omega)} = \int_\Omega |\nabla u|_2^2 \, dx = -\int_\Omega u \Delta u \, dx = \lambda \int_\Omega u^2 \, dx = \lambda \|u\|^2_{L^2(\Omega)}$$

so that large eigenvalues λ lead to highly oscillatory eigenfunctions. For mesh functions, we replace Δ by Δ_h and the eigenpairs of the Laplacian by the eigenpairs of the discrete Laplacian. The L^2 norm of the gradient is replaced by $x^* A x$ and therefore we see that the norm $\|x\|_A$ measures the growth of the mesh function U associated to the coefficient vector x.

Remark 2.50. We must not neglect the scaling with respect to the mesh size h. If we compare the A-norm with the Euclidean norm we obtain $\|x\|_A^2 = h^2 \lambda_{j,k} \|x\|_2^2$ for the coefficients x of the discrete eigenfunction $U_{j,k}$. ⧫

Since A is Hermitian positive definite, we can define arbitrary real powers A^s of A and define the following norm.

Definition 2.51. Given A is Hermitian positive definite and $s \in \mathbb{R}$, we define the norm

$$\|y\|_{A,s} := \sqrt{y^* A^s y}.$$

for any $y \in \mathbb{C}^n$. ♦

For $s = 0$ we have the Euclidean norm, whereas for $s = 1$ we recover the energy norm. In general we think of s as describing a scale of smoothness. We will prove a basic smoothing property for the relaxed Richardson iteration. We begin with a technical lemma.

Lemma 2.52. *Let $B \in \mathbb{C}^{n\times n}$ be Hermitian and positive definite with the bound $\|B\|_2 \leq 1$. Then $\|B(I-B)^m\|_2 \leq (m \exp(1))^{-1}$ for all $m \geq 1$.*

Proof. We see that B and $B(I-B)^m$ have the same eigenvectors and deduce from the assumed spectral bound that

$$\|B(I-B)^m x\|_2 \leq \sup_{0\leq z\leq 1} z(1-z)^m.$$

The function $z(1-z)^m$ has its only critical point at $z_\star = 1/(m+1)$ where the function attains its maximum with the value

$$\frac{1}{(m+1)}\left(\frac{m}{m+1}\right)^m = \frac{1}{m}\frac{1}{(1+\frac{1}{m})^{m+1}} \leq \frac{1}{m\exp(1)}$$

where we used the classical bound $\exp(1) < (1+1/m)^{m+1}$. ■

Given a matrix $A \in \mathbb{C}^{n\times n}$, we denote by

$$R = R(\omega, A) := I - \omega A$$

the *relaxation operator.* This is the iteration of the relaxed Richardson scheme. It is immediate to see that the error $e_k := x - x_k$ of the iteration is propagated as $e_k = R^k e_0$.

Theorem 2.53 (smoothing property)**.** *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite. The relaxed Richardson iteration with parameter $\omega \leq 1/\lambda_{\max}(A)$ satisfies*

$$\|R^k y\|_{A,s+t} \leq C(t,\omega) k^{-t/2} \|y\|_{A,s} \quad \textit{for all } y \in \mathbb{C}^n \textit{ and all } s \in \mathbb{R}, t > 0$$

with $C(t,\omega) = \left(\frac{t}{2\omega\exp(1)}\right)^{t/2}$.

Proof. We see that $A^{(s+t)/2}R^k = A^{t/2}R^k A^{s/2}$ and compute

$$\|R^k y\|_{A,s+t} = \|A^{t/2}R^k A^{s/2} y\|_2 \leq \|A^{t/2}R^k\|_2 \|y\|_{A,s}.$$

We write $A^{t/2}R^k = (AR^{2k/t})^{t/2} = \omega^{-t/2}(\omega A R^{2k/t})^{t/2}$. Since all eigenvalues of ωA lie between 0 and 1, and can apply the technical lemma to $B := \omega A$ and $m := 2k/t$ and obtain

$$\|A^{t/2}R^k\|_2 \leq \omega^{-t/2}\|\omega A R^{2k/t}\|_2^{t/2} \leq \omega^{-t/2}\left(\frac{t}{2k\exp(1)}\right)^{t/2}.$$

This concludes the proof. ■

The result is referred to as "smoothing property" based on the following reasoning. If the error y has high oscillations, we expect the estimate

$$\|y\|_{A,s+t} \lesssim \omega^{-t/2}\|y\|_{A,s}$$

to be sharp (recall that $\omega = O(1/\lambda_{\max})$). That is, for some constant $c > 0$, the converse estimate

$$\omega^{-t/2}\|y\|_{A,s} \leq c\|y\|_{A,s+t}$$

holds. In this case, the theorem states

$$\|R^k y\|_{A,s+t} \leq c\left(\frac{t}{2\exp(1)}\right)^{t/2} k^{-t/2}\|y\|_{A,s+t}$$

This means that oscillations of y are smoothed out rapidly for the first few k. If for example $t = 1$, quotients of two consecutive prefactors on the right-hand side is $\sqrt{1/2}$, $\sqrt{2/3}$, $\sqrt{3/4}$,..., which corresponds to good contraction rates for *small* k.

Example 2.54. The above spectral analysis and the theoretical smoothing result suggest that the relaxed Jacobi iteration with $\omega = 1/2$ should satisfy the smoothing property. Indeed, by the Gershgorin theorem we know that twice the diagonal entry of the finite difference matrix is an upper bound for its largest eigenvalue. For the ordinary Jacobi iteration this is not guaranteed, which corresponds to the above observation that the high frequencies are not damped. Figure 2.1 compares the iterations in a practical experiment. We recall that for the FDM, the relaxed Jacobi iteration with relaxation parapeter 1/2 corresponds to the relaxed Richardson method with $\omega = 1/4$ in one dimension and $\omega = 1/8$ in two dimensions.

§10. The two-grid algorithm

After a few smoothing steps with the Richardson relaxation, the high frequency modes in the error are basically damped out and the error can be represented with respect to a coarser grid of mesh size, say, $2h$ without essential loss of information. We therefore may project (or interpolate) the smoothed error to that coarser grid. If this is a sufficiently coarse grid, we solve the resulting finite difference system with a direct solver, which is then a feasible task. We illustrate this two-grid method in detail.

As we consider different finite different systems with respect to different grid sizes, we will indicate the current grid with the index h. We denote by $\mathcal{R}_h$ an application of the relaxed Richardson iteration with right-hand side b_h, that is

$$\mathcal{R}_h y = y - \omega(A_h y - b_h).$$

Having started with an initial guess $x_h^{(0)}$ for the solution x_h to $A_h x_h = b_h$, of course we do not know the error $e_\nu^h := x_h - \mathcal{R}_h^\nu x_h^{(0)}$ after ν relaxation steps. (otherwise we could directly compute x from it). But we know the *residual* $r_\nu^h := b_h - A_h \mathcal{R}_h^\nu x_h^{(0)}$, which corresponds to the transformed error $A_h e_\nu^h$. These quantities are thus connected via $e_k^h = A_h^{-1} r_k^h$. After a few smoothing steps we expect that the interpolated residual $J(r_k^h)$ is a good approximation to r_k^h. Here J some suitable operator restricting mesh functions from the scale h to mesh functions on the scale $2h$. We thus compute $A_{2h} z = J r_k^h$ and expect accordingly that the coarse mesh function z is a good approximation to e_ν^h. We then embed (or prolongate) z to the fine grid by some operator $I_{2h\to h}$

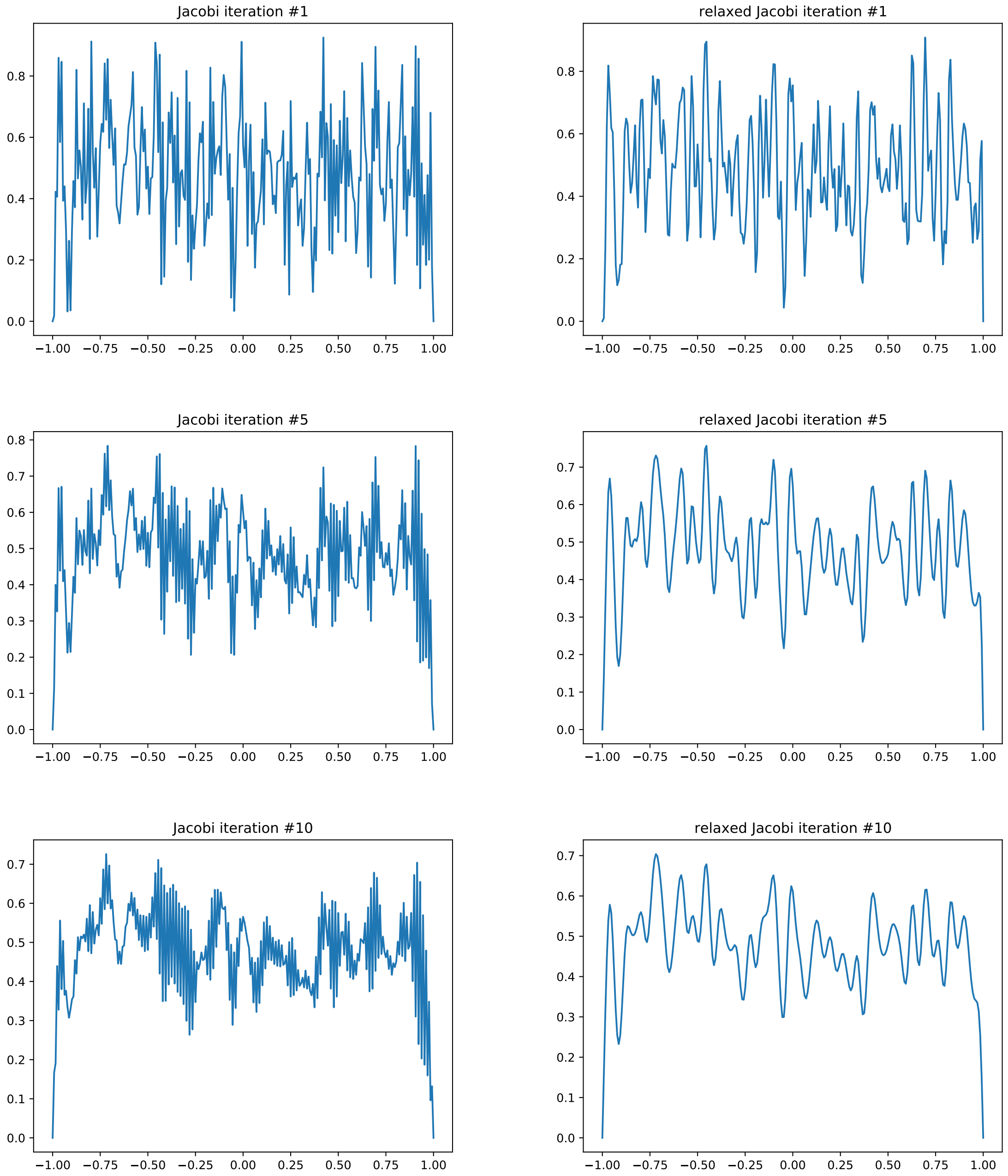


Figure 2.1.: Finite difference approximations to the one-dimensional equation $u'' = 0$ with homogeneous boundary conditions on a grid with 257 nodes. Left: Jacobi iterations 1, 5, 10; right: relaxed Jacobi ($\omega = 1/2$) iteration 1, 5, 10. The initial guess x_0 is a uniformly distributed random vector with values between 0 and 1.

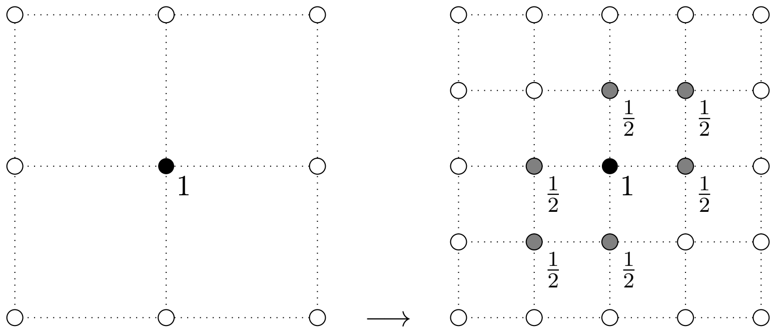


Figure 2.2.: Prolongation of a mesh function from the level $2h$ to the level h by the operator $I_{2h\to h}$.

and set $x_h^{(1)} = \mathcal{R}_h^\nu x_h^{(0)} + I_{2h\to h} z$. If $I_{2h\to h} z$ is close to e_ν^h, then $x_h^{(1)}$ is close to x_h. We can iterate this procedure with y replaced by $x_h^{(1)}$ and obtain a sequence $x_h^{(j)}$ that we expect to converge to x_h.

In order to formalize this idea, we need a definition of the two inter-grid transfer operators: the approximation of fine mesh functions by coarse mesh functions and the embedding of coarse mesh functions into the fine-mesh functions.

Definition 2.55 (prolongation operator)**.** Let $n \in \mathbb{N}$ and $h = 1/(2n)$ and $x_0 = (2j_0h, 2k_0h)$ with $0 \le j_0, k_0 \le n+1$ be a grid point in the mesh of grid size $2h$. Consider the mesh functions

$$U_{2h}(y) = \begin{cases} 1 & \text{if } y = x_0 \\ 0 & \text{else} \end{cases} \qquad \text{for all } y = (2jh, 2kh) \text{ with } 0 \le j,k \le n+1$$

$$U_h(z) = \begin{cases} 1 & \text{if } z = x_0 \\ 1/2 & \text{if } z = x_0 \pm (h,0) \text{ or } z = x_0 \pm (0,h) \text{ or } z = x_0 \pm 2^{-1/2}(h,h) \\ 0 & \text{else} \end{cases}$$

for all $z = (jh, kh)$ with $0 \le j,k \le 2n+1$. The linear operator $I_{2h\to h}$ between mesh functions on the grid of size $2h$ and the grid of size h with the property $U_h = I_{2h\to h} U_{2h}$ is called the *prolongation operator*. Figure 2.2 displays an illustration. ♦

One may wonder why we took this seemingly arbitrary choice for the prolongation. Indeed, there are many possibilities of interpreting which values would be sensible evaluations of mesh functions between two grid points. The choice we made corresponds to extending a mesh function to a function over $\bar{\Omega}$ by dividing every square of the grid in two triangles by a line segment parallel to $(1,1)$ and taking the unique continuous function that is affine when restricted to any of these triangles and coincides with the original mesh function on the grid points. It is easy to check that the resulting function has precisely the values that are prescribed by the prolongation operator. With this choice we have the property $A_{2h} = I^*_{2h\to h} A_h I_{2h\to h}$, that is, the prolongation is consistent with the action of the discrete Laplacian, see Exercise A.30. The prolongation operator is easy to realize as a sparse matrix, see the Python routine of Figure 2.3.

With this embedding, we can give a meaning to the inner product between mesh functions on different level by

$$(x_h, y_{2h})_{\mathrm{mf}} := (x_h, I_{2h\to h} y_{2h})_2.$$

```
def prolongation(n):
  N=2*n
  m=(n+1)**2;
  M=(N+1)**2
  oldnode=np.zeros(m)
  for j in range(0,n+1):
      oldnode[range(j*(n+1),(j+1)*(n+1))]=np.arange(2*j*(N+1),(2*j+1)
         *(N+1),2)
  oldnode=np.reshape(np.int_(oldnode),(n+1,n+1))
  I=np.zeros((n+1,n+1,9)); J=np.zeros((n+1,n+1,9)); V=np.zeros((n+1,n
     +1,9))
  dummy=M
  val=np.asarray([.5,.5,0,.5,1,.5,0,.5,.5])
  for j in range(0,n+1):
      for k in range(0,n+1):
          J[j,k,:]=(j*(n+1)+k)*np.ones(9)
          V[j,k,:]=val
          p=oldnode[j,k]
          nw=p+N;   north=p+N+1; ne=p+N+2
          west=p-1;              east=p+1
          sw=p-N-2; south=p-N-1; se=p-N
          if j==0: sw=dummy; south=dummy; se=dummy
          if j==n: nw=dummy; north=dummy; ne=dummy
          if k==0: nw=dummy; west=dummy; sw=dummy
          if k==n: ne=dummy; east=dummy; se=dummy
          I[j,k,:]=np.asarray([sw,south,se,west,p,east,nw,north,ne])
  I=np.reshape(I,(9*m,1)).T
  J=np.reshape(J,(9*m,1)).T
  V=np.reshape(V,(9*m,1)).T
  P=csr_matrix((V[0,:],(I[0,:],J[0,:])),shape = (M+1,m))
  R=csr_matrix((np.ones(M),(np.arange(0,M),np.arange(0,M))),shape = (M
     ,M+1))
  P=R*P
  return P
```

Figure 2.3.: The prolongation operator for an $(n+1) \times (n+1)$ grid as a sparse matrix in Python.

We see that the adjoint $I^*_{2h\to h}$ with respect to the Euclidean scalar product is the orthogonal projection to the mesh functions on the level $2h$ in $(\cdot,\cdot)_h$ inner product, indeed

$$(x_h, y_{2h})_{\mathrm{mf}} = (x_h, I_{2h\to h} y_{2h})_2 = (I^*_{2h\to h} x_h, y_{2h})_2 = (I^*_{2h\to h} x_h, y_{2h})_{\mathrm{mf}}.$$

We thus choose $I^*_{2h\to h}$ as the projection operator that approximates fine mesh functions by coarse mesh functions.

Based on the above derivation we can now formulate the two-grid algorithm.

Algorithm 2.56 (two-grid iteration for FDM)**.** We are given the right-hand side b_h of the FDM.

- Input: meshes of size $2h$ and h; number of relaxations ν; initial guess $x_h^{(0)}$
- For $k = 1, 2, \dots$

$$x_h^{(k)} = \mathcal{R}_h^\nu x_h^{(k-1)} + I_{2h\to h} A_{2h}^{-1} I^*_{2h\to h}(b_h - A_h \mathcal{R}_h^\nu x_h^{(k-1)})$$

♦

§11. The multigrid algorithm

The two-grid method was introduced above for the purpose of illustration. The symbolic notation A_{2h}^{-1} in Algorithm 2.56 indicates that a finite difference system on the scale $2h$ is solved. With respect to the coarser scale $2h$ the smoothed residual from the scale h again oscillatory and we repeat the procedure of smoothing and projecting to a coarser grid. That is, we approximate the action of A_{2h}^{-1} by another two-grid iteration with the mesh of scale $4h$. This can be repeated recursively until a sufficiently coarse mesh is reached where direct solvers are cheap.

Algorithm 2.57 (multigrid iteration for FDM)**.** We are given as input:

- a mesh hierarchy on levels $2^m h, 2^{m-1}h, \dots, 2h, h$ with some $m \in \mathbb{N}$ and $H := 2^m h$
- the right-hand side b_h of the FDM
- an initial guess $x_h^{(0)}$
- the number ν of desired smoothing steps
- a parameter $\mu \in \mathbb{N}$

For $k = 1, 2, \dots,$ the iterate

$$x_h^{(k)} = MG(h, b_h, x_h^{(k-1)}, \nu)$$

is recursively defined by

- $r_h := (b_h - A_h \mathcal{R}_h^\nu x_h^{(k-1)})$ (pre-smoothing)
- $\left\{\begin{array}{l}\textbf{if } 2h = H, \textbf{ then } z_h = A_{2h}^{-1} I^*_{2h\to h} r_h \\ \textbf{else} \text{ set } z_h^0 = 0 \text{ and } \textbf{for } j = 1,\dots,\mu \text{ do: } z_h^j := MG(2h, I^*_{2h\to h} r_h, z_h^{j-1}, \nu)\text{; set } z_h := z_h^\mu\end{array}\right.$
  (coarse-grid correction)
- $x_h^{(k)} := \mathcal{R}_h^\nu(\mathcal{R}_h^\nu x_h^{(k-1)} + I_{2h\to h} z_h)$ (post-smoothing)

♦

Of course one could choose two different values ν_1 resp. ν_2 for the pre- resp. post-smoothing, but we disregard this possibility here for a better readability of the algorithm. The parameter μ is chosen as $\mu = 1$ or $\mu = 2$ in practice. For $\mu = 1$ we call each iteration a V-cycle and for $\mu = 2$ we call it a W-cycle. Indeed, for $\mu = 1$ the iteration goes straight down to the coarsest grid and back until the finest grid is reached. This resembles the shape of the letter "V". For $\mu = 2$, each mesh is visited more often by the algorithm and the mnemonic shape for the pattern is the letter "W". More elaborate graphical illustrations can be found in textbooks or on the web. Heuristically, the W-cycle invests visits the coarse meshes more often than the V-cycle and is thus expected to handle the low-frequency part of the error more accurately.

Later in the lecture we will prove convergence properties and complexity estimates for the multigrid iteration. With the modest tools that we used in our FDM error analysis, such an analysis turns out quite difficult while the analysis of the finite element method will offer more powerful tools for accomplishing this. We confine ourselves to estimating the computational cost. A fixed number ν of relaxation steps on a grid of size $h = 1/N$ with $n = (N+1)^2$ grid points requires $\mathcal{O}(n)$ operations because our matrices are sparse. The same applies to the multiplication with the prolongation matrix or its transpose. The number of iterations for one evaluation of $MG(h, b_h, \nu)$ is therefore

$$\mathcal{O}\left(\sum_{j=0}^{m} 2^{-2j} n\right) = \mathcal{O}(n)$$

because each coarsening reduces the number of grid points by a factor 4. It is possible to prove that a fixed choice of ν and fixed maximal value k of iterations is sufficient for obtaining an error $x_h^{(k)} - x_h$ that is of the order of magnitude of the original error estimate satisfied by x_h. Therefore, multigrid is a solver of linear complexity. We will prove this in the forthcoming chapters.

In Figure 2.4, a possible implementation of the W-cycle in Python is sketched (where standard FDM routines are not further displayed). Also, the recursive implementation may not be optimal with respect to memory consumption and speed.

Example 2.58. In this example we approximate Poisson's equation on the unit square with right-hand side $f(x) = 2\pi \sin(\pi x_1) \sin(\pi x_2)$ with the multigrid method. The exact solution $u(x) = \sin(\pi x_1) \sin(\pi x_2)$ is explicitly known and used for comparison. We use $k = 3$ iterations of the W-cycle with $\nu = 2$ for both the pre- and the post-smoothing. The coarsest mesh is of size 3×3 and we compute the discrete solution on 12 further refinements of that mesh. The convergence history plot of Figure 2.5 compares the multigrid solution with the exact solution and shows that the multigrid approximation has an error comparable to the exact discrete solution u_h. Table 2.1 displays the numbers of nodes, the mesh sizes, the errors in the max norm and the computing times of a simple Laptop with Intel Core i5-6300U CPU with 4 × 2.4 GHz with 7.6 GiB RAM plus 22.8 GiB swap memory. On the finest mesh with about 60 million degrees of freedom the computation took 8 271 seconds, which is about 2 hours, which is quite long and mainly due to the computations with the swap memory. Note that the storage of A, b, $x_h^{(k)}$, and projection matrices almost fills the complete RAM on such a small machine.

```
import numpy as np
from numpy import matlib
import scipy.sparse
import scipy.sparse.linalg
from scipy.sparse import csr_matrix
from scipy.sparse import spdiags

def f_fun(x,y):  [...]   return val # right-hand side function f
def FDM_coords(n):    [...]    return coord_x, coord_y
def FDM_data(n,f):    [...]    return A, b, h
def restrict2dof(n):    [...]    return R
def prolongation(n):    [...]    return P

def relax(A, b, u,n_smooth):
    for _ in range(n_smooth):
        u = u - 1/8 * (A*u-b)
    return u

def Wcycle(coarse,fine,A,b,x,n_smooth):
    S=restrict2dof(2**fine)
    P=prolongation(2**(fine-1))
    A_inner=(S.transpose()@A)@S
    b_inner=S.transpose()@b
    if coarse==fine:
        x=S*scipy.sparse.linalg.spsolve(A_inner,b_inner)
    else:
        for _ in range(0,2):
            x=S*relax(A_inner, b_inner, S.transpose()*x,n_smooth)
            r=b-A*x;
            q=Wcycle(coarse,fine-1,P.transpose()@(A@P),P.transpose()*r
                ,0*P.transpose()*r,n_smooth)
            x=x+P*q
            x=S*relax(A_inner, b_inner, S.transpose()*x,n_smooth)
    return x

def FDM_mg(coarse,fine,f,n_iter,n_smooth):
    n=2**fine
    A, b, h = FDM_data(n,f)
    x=np.zeros((n+1)**2)
    for m in range(0,n_iter):
        x = Wcycle(coarse,fine,A,b,x,n_smooth)
    return x, h

#-------- the numerical experiment -------------------------
f=np.vectorize(f_fun)
coarse=1 #2^coarse intervals per axis
fine=8
n_iter=3
n_smooth=2
x, h=FDM_mg(coarse,fine,f,n_iter,n_smooth)
```

Figure 2.4.: Possible implementation of the W-cycle.

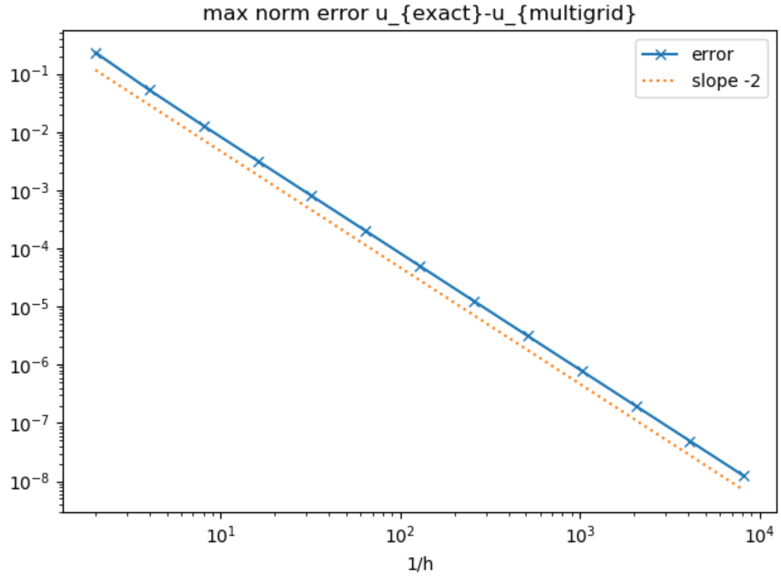


Figure 2.5.: Convergence history of $\|u - x_h^{(k)}\|_{\infty,\bar{\Omega}}$ with $k = 3$, $\nu = 2$ in Example 2.58.

mesh	nodes	h	$\|u - x_h^{(k)}\|_{\infty,\bar{\Omega}}$	time (seconds)
1	9.000E+00	5.000E-01	2.337E-01	2.717E-02
2	2.500E+01	2.500E-01	5.290E-02	2.324E-02
3	8.100E+01	1.250E-01	1.287E-02	5.230E-02
4	2.890E+02	6.250E-02	3.213E-03	1.120E-01
5	1.089E+03	3.125E-02	8.032E-04	2.275E-01
6	4.225E+03	1.562E-02	2.008E-04	5.213E-01
7	1.664E+04	7.812E-03	5.019E-05	1.257E+00
8	6.605E+04	3.906E-03	1.255E-05	4.109E+00
9	2.632E+05	1.953E-03	3.137E-06	1.083E+01
10	1.051E+06	9.766E-04	7.844E-07	3.721E+01
11	4.198E+06	4.883E-04	1.961E-07	1.369E+02
12	1.679E+07	2.441E-04	4.902E-08	5.199E+02
13	6.713E+07	1.221E-04	1.226E-08	8.271E+03

Table 2.1.: Results for $k = 3$, $\nu = 2$ in Example 2.58.

3. The finite element method

§1. Elementary Hilbert space theory

Hilbert spaces are taught in detail in any class on linear functional analysis. Here we only focus on some very basic properties.

Let X be a (real) linear space. Given a symmetric and positive definite bilinear form $(\cdot, \cdot)_X$, we define $\|x\|_X = \sqrt{(x,x)_X}$ for any $x \in X$. It is elementary to establish the Cauchy–Schwarz inequality

$$(x,y)_X \leq \|x\|_X \, \|y\|_X \quad \text{for any } x, y \in X.$$

It can be shown that $\| \cdot \|_X$ defines a norm on X (thereby justifying the notation). The proofs are the same as in the case of Euclidean vector spaces and left as an exercise. The difference to Euclidean spaces is that X can be infinite-dimensional and therefore need not be complete. If it is complete, we call X a Hilbert space.

Definition 3.1 (Hilbert space)**.** A linear space X (over $\mathbb{R}$) equipped with a symmertic and positive definite bilinear form $(\cdot, \cdot)_X$ is called *Hilbert space* if it is complete with respect to the norm $\| \cdot \|_X := \sqrt{(\cdot, \cdot)_X}$. ♦

Basically, Hilbert spaces are Banach spaces with an Euclidean structure.

Lemma 3.2 (parallelogram law)**.** *In a Hilbert space X, every $(a, b) \in X^2$ satisfies*

$$\|a - b\|_X^2 + \|a + b\|_X^2 = 2(\|a\|_X^2 + \|b\|_X^2).$$

Proof. If we expand both terms on the left-hand side with the binomial identity, we see that the mixed terms cancel. What remains are the terms on the right-hand side. ■

Theorem 3.3 (projection on complete subspaces)**.** *Let X be a Hilbert space with inner product $(\cdot, \cdot)_X$ and let $Y \subseteq X$ be a complete linear subspace. Given $x \in X$, there exists a unique element $Px \in Y$ with the property*

$$\|x - Px\|_X = \inf_{y \in Y} \|x - y\|_X.$$

The element Px is unique and characterized by the property

$$(x - Px, y)_X = 0 \quad \text{for all } y \in Y.$$

Proof. We abbreviate $\delta := \inf_{y \in Y} \|x - y\|_X$ Let $(y_k)_k$ be a sequence in Y with $\|x - y_k\|_X \to \delta$ as $k \to \infty$. To prove that the sequence is Cauchy, we let $m, n \geq 0$ and choose $a = x - y_m$, $b = x - y_n$ in the parallelogram law, which results in

$$\|y_m - y_n\|_X^2 + 4\|x - \frac{1}{2}(y_m + y_n)\|_X^2 = 2(\|x - y_m\|_X^2 + \|x - y_n\|_X^2).$$

Since y_m, y_n are from the linear space Y, their average lies in Y, and the second term on the left-hand side is bounded from below by 4δ. Since the right-hand side converges to the same value, we deduce $\|y_m - y_n\|_X \to 0$ as $m, n \to \infty$ so that $(y_k)_k$ is a Cauchy sequence. Since Y is complete, the sequence has a limit, denoted by y, which lies in Y and satisfies $\|x - y\| = \delta$. For proving uniqueness, we assume that there are $y, y' \in Y$ realizing the infimum. The above argument with the parallelogram law applied to y, y' instead of y_m, y_n shows that $y = y'$, which proves uniqueness. We thus denote $Px := y$.

For an arbitrary $z \in Y$ and $\varepsilon \in [0, 1]$, the convex combination $(1 - \varepsilon)Px + \varepsilon z$ belongs to Y, so that we infer with elementary manipluations

$$\|x - Px\|_X^2 = \delta^2 \leq \|x - (1 - \varepsilon)Px - \varepsilon z\|_X^2 = \|(x - Px) - \varepsilon(z - Px)\|_X^2.$$

Expanding the right-hand side results in the estimate

$$\|x - Px\|_X^2 \leq \|x - Px\|_X^2 + \varepsilon^2\|z - Px\|_X^2 + 2\varepsilon(x - Px, z - Px)_X.$$

Simplifying, dividing by ε, and letting $\varepsilon \to 0$, we see with the substitution $y := z - Px$ that $0 \leq (x - Px, y)_X$ for all $y \in Y$. Since this must be true for $\pm y$, the bilinearity proves the asserted variational identity. ■

Definition 3.4. The map $P : X \to Y$ from Theorem 3.3 is called *orthogonal projection* to Y. ⧫

It is easy to see that the orthogonal projection P to a subspace Y is linear and nonexpansive, that is $\|P\|_{L(X,Y)} \leq 1$, see the exercises.

We recall the dual space $X^* := L(X, \mathbb{R})$, which is the space of continuous linear functionals over X. The Riesz representation theorem states that there exists an isometric isomorphism between X and X^*. The proof is taught in every course on linear functional analysis and we will briefly discuss the proof in what follows.

Theorem 3.5 (Riesz representation theorem)**.** *Let X be a Hilbert space with inner product $(\cdot, \cdot)_X$ and let $F \in X^*$ be a continuous linear functional. Then there exists a unique element $x \in X$ with the property*

$$(y, x)_X = F(y) \quad \textit{for all } y \in X.$$

The dependence of x on F is linear and the element x satisfies $\|x\|_X = \|F\|_{X^*}$.

Proof. We consider the map $J : X \to X^*$ defined by

$$x \mapsto J(x) = [y \mapsto (y, x)_X]$$

or $J(x) = (\cdot, x)_X$ for short. It is direct to check that J is linear and satisfies $\|x\|_X \leq \|J(x)\|_{X^*} \leq \|x\|_X$ so that it is injective and an isometry. We are left with showing that J is surjective. Given some nonzero $F \in X^*$, we denote by P the orthogonal projection to the kernel $\ker(F)$ (which is closed and thus complete, see the exercises). We choose $b \in X$ with $F(b) = 1$ and set $y := b - Pb$. By scaling the element y, we can now decompose any $z \in X$ in a part in the kernel of F and a multiple of y, namely

$$z = (z - F(z)y) + F(z)y.$$

By construction, y is orthogonal to any element of $\ker(F)$, so that we compute

$$(z, y)_X = (z - F(z)y, y)_X + (F(z)y, y)_X = (F(z)y, y)_X = F(z)\|y\|_X^2.$$

Rearranging this formula reveals

$$F(z) = \|y\|_X^{-2}(z, y)_X = J(\|y\|_X^{-2}y)(z) \quad \text{for all } z \in X$$

(note that y is nonzero because $F(y) = 1$). We thus have shown $F = J(x)$ for $x = \|y\|_X^{-2}y$, whence J is surjective.

■

We now use Hilbert space methods to show well-posedness of our variational formulation.

§2. The Dirichlet problem in Sobolev spaces

Throughout this course, $\Omega \subseteq \mathbb{R}^2$ will be an open convex polygon. It is known that such domains posses, almost everywhere on the boundary, a well-defined outer unit normal vector ν. The divergence theorem teaches us the following: For a bounded Lipschitz domain Ω and a vector field $v \in C^1(\Omega; \mathbb{R}^2)$ we have

$$\int_{\partial\Omega} v \cdot \nu \, ds = \int_\Omega \operatorname{div} v \, dx.$$

Here (and throughout this text) we denote integration with respect to the n-dimensional Lebesgue measure by the symbol "dx" while integration with respect to the 1-dimensional surface measure is indicated by "ds". The divergence theorem implies the formula of integration by parts: For two differentiable functions u and v we have

$$\int_\Omega (u \, \partial_j v + v \, \partial_j u) dx = \int_{\partial\Omega} uv \, \nu_j ds$$

for any $j \in \{1, \dots, n\}$, where ν_j is the jth component of the outer unit normal. A variant thereof is called *Green's formula*

$$\int_\Omega (u\Delta v + \nabla u \cdot \nabla v) dx = \int_{\partial\Omega} u \frac{\partial v}{\partial \nu} ds,$$

where $v \in C^2(\Omega) \cap C^1(\bar{\Omega})$ is assumed.

Definition 3.6 (weak derivative)**.** Let $\Omega \subseteq \mathbb{R}^2$ be open. Let $v \in L^1_{\text{loc}}(\Omega)$ and $j \in \{1, 2\}$. If there exists a function $g \in L^1_{\text{loc}}(\Omega)$ with the property

$$\int_\Omega v \partial_j \psi \, dx = -\int_\Omega g\psi \, dx \quad \text{for all } \psi \in C_c^\infty(\Omega),$$

then this function g is called the *weak partial derivative* of v with respect to the direction j, and it is denoted by $\partial_j v$. The vector of all partial derivatives is denoted (provided it exists) by ∇v. ⧫

Remark 3.7. The weak derivative is unique (see problems). ⧫

The idea behind this definition is to extend the common notion of differentiability. If v is differentiable, then the weak and the classical derivatives coincide. There are, however, functions that are not differentiable in the classical sense, but possess a weak derivative.

Example 3.8. The absolute value function $v(x) = |x|$ on $\Omega = (-1, 1)$ is not differentiable on $(-1, 1)$. Yet, its weak derivative is given by

$$v'(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0. \end{cases} \tag{3.1}$$

Note that we can modify elements of $L^1_{\text{loc}}(\Omega)$ at $x = 0$ to any value.

From the example we see that functions with certain kinks can be weakly differentiable.

Example 3.9. We subdivide the interval $(-1,1)$ into finitely many sub-intervals $[x_j, x_{j+1}]$, where

$$-1 = x_1 < \cdots < x_N = 1 \quad \text{and} \ \cup_{j=1}^{N-1} [x_j, x_{j+1}] = \bar{\Omega} = [-1,1],$$

and consider the globally continuous functions that are affine when restricted to any of the sub-intervals $[x_j, x_{j+1}]$. Any such function is weakly differentiable.

We introduce spaces of functions that posses appropriate weak derivatives. It will turn out that these are suited for a sound theory of Poisson's equation (and similar problems). We shall prove many, but not all of the stated results.

Definition 3.10 (Sobolev spaces)**.** Let $\Omega \subseteq \mathbb{R}^2$ be bounded and open. Define

$$H^1(\Omega) := \{v \in L^2(\Omega) : \forall j \in \{1,2\}\ \partial_j v \in L^2(\Omega)\}.$$

That is, the functions from $H^1(\Omega)$ belong to $L^2(\Omega)$; their first weak derivatives exist and belong to $L^2(\Omega)$ as well. ♦

Sobolev functions have far more structure than generic L^2 functions. Recall that elements from $L^2(\Omega)$ are equivalence classes (up to equality almost everywhere) and that point evaluations are not well defined. This is generally the case for Sobolev function, too. Yet, we will see that such functions possess boundary values in some generalized sense. We first study an important property, namely that $H^1(\Omega)$ can equivalently be defined by a completion process. Let us define the following norm on $H^1(\Omega)$,

$$\|v\|_{H^1(\Omega)} := \sqrt{\|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)}}.$$

We use the convention that $\|\nabla v\|^2_{L^2(\Omega)} = \int_\Omega |\nabla v|^2 \, dx$ for the Euclidean norm $|\cdot|$.

Theorem 3.11. *Let $\Omega \subseteq \mathbb{R}^2$ be an open bounded polygon. The space $H^1(\Omega)$ is complete with respect to the norm $\|\cdot\|_{H^1(\Omega)}$, i.e. a Banach space. The space $C^\infty(\bar{\Omega})$ is dense in $H^1(\Omega)$.*

Theorem 3.12. *Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded domain with polygonal Lipschitz boundary. Then, there exists a unique continuous and linear map $S : H^1(\Omega) \to L^2(\partial\Omega)$ with the property*

$$Sv = v|_{\partial\Omega} \quad \textit{for all } v \in H^1(\Omega) \cap C^1(\bar{\Omega}).$$

Remark 3.13. A linear map $T : H^1(\Omega) \to L^2(\partial\Omega)$ is said to be continuous if there exists a constant $C_T < \infty$ such that

$$\|Tv\|_{L^2(\partial\Omega)} \le C_T \|v\|_{H^1(\Omega)} \quad \text{for all } v \in H^1(\Omega).$$

The theorem states the following. The operation of taking boundary values, which is well defined for functions from $C^1(\bar{\Omega})$, has a unique continuation to functions from $H^1(\Omega)$. Taking such generalized boundary values still leads to functions in $L^2(\partial\Omega)$, and we interpret these as boundary values of functions from $H^1(\Omega)$. This concept turns out important if we wish to pose the Dirichlet problem in Sobolev spaces. The operator S is called *trace operator*, and Sv is called the trace of v on $\partial\Omega$. ♦

We outline a proof of the trace theorem, where for simplicity of notation we assume that $\bar{\Omega} = T$ is a triangle. By density, it will be enough to show that for any edge F of T and any $v \in C^1(T)$, the following bound holds $\|v\|_{L^2(F)} \le C\|v\|_{H^1(T)}$. The constant C depends on the domain Ω but not on v. This will be proven in the following theorem.

Theorem 3.14 (trace identity and trace inequality for triangles)**.** *Let $T \subseteq \mathbb{R}^2$ be a triangle with some edge $F \subseteq T$ and opposite vertex $P \in T$. Any function $v \in C^1(T)$ then satisfies*

$$\frac{|T|}{|F|}\int_F v\,ds = \int_T v\,dx + \frac{1}{2}\int_T (\bullet - P)\cdot\nabla v\,dx$$

and

$$\|v\|^2_{L^2(F)} \leq \frac{3|F|}{2|T|}\|v\|^2_{L^2(T)} + \frac{|F|}{2|T|}\operatorname{diam}(T)^2\|\nabla v\|^2_{L^2(T)}.$$

Here, $|T|$ denotes the area of T and $|F|$ denotes the length of F.

Proof. We have $\operatorname{div}(\bullet - P) = 2$ (in two space dimensions). Integration by parts therefore reveals

$$\int_T v\,dx + \frac{1}{2}\int_T (\bullet - P)\cdot\nabla v\,dx = \int_{\partial T} v\,(\bullet - P)\cdot\nu\,ds,$$

where ν is the outer unit normal of T. We observe that, on the two edges of T different from F, the vector $(\bullet - P)$ is tangential to ∂T and, thus, its product with ν equals zero. Hence,

$$\int_{\partial T} v\,(\bullet - P)\cdot\nu\,ds = \int_F v\,(\bullet - P)\cdot\nu\,ds.$$

Since furthermore ν is constant along F, the quantity $(\bullet - P)\cdot\nu$ is constant on F as well, and its value corresponds to the orthogonal projection of $(\bullet - P)$ in direction of ν. This is precisely the length of the height on F, which by elementary geometry takes the value $2|T|/|F|$. This proves the first assertion.

In order to show the second claimed property, we apply the trace identity to v^2. Note that $\nabla(v^2) = 2v\nabla v$. We thus infer

$$\frac{|T|}{|F|}\int_F v^2\,ds = \int_T v^2\,dx + \int_T (\bullet - P)\cdot v\nabla v\,dx \leq \int_T v^2\,dx + \operatorname{diam}(T)\int_T |v|\,|\nabla v|\,dx,$$

where in the second step we have estimated the length of $(\bullet - P)$ by the diameter of T. After rearranging the identity we obtain

$$\|v\|^2_{L^2(F)} \leq \frac{|F|}{|T|}\|v\|^2_{L^2(T)} + \frac{|F|}{|T|}\operatorname{diam}(T)\int_T |v|\,|\nabla v|\,dx.$$

We use the Cauchy-Schwarz inequality and the Young inequality $2ab \leq a^2 + b^2$ to estimate the second integral as follows

$$\begin{aligned}\operatorname{diam}(T)\frac{|F|}{|T|}\int_T |v|\,|\nabla v|\,dx &= \frac{|F|}{|T|}\int_T |v|\,\big(\operatorname{diam}(T)|\nabla v|\big)\,dx \\ &\leq \frac{|F|}{2|T|}(\|v\|^2_{L^2(T)} + \operatorname{diam}(T)^2\|\nabla v\|^2_{L^2(T)}).\end{aligned}$$

This implies the second assertion. ■

As a consequence from the trace theorem, it makes sense to impose boundary values on functions from $H^1(\Omega)$. We will usually write $u|_{\partial\Omega}$ instead of Su etc., but we need to be aware that this function is only of class L^2 on $\partial\Omega$. For the Dirichlet problem, it is reasonable to consider the following subspace

$$H^1_0(\Omega) := \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\},$$

i.e., the space of Sobolev functions with zero boundary values.

Theorem 3.15. *Let $\Omega \subseteq \mathbb{R}^2$ be an open bounded polygon. The space $H_0^1(\Omega)$ is the completion $C_c^\infty(\bar{\Omega})$ with respect to the norm $\|\cdot\|_{H^1(\Omega)}$.*

For functions from $H_0^1(\Omega)$, the L^2 norm can be controlled by the L^2 norm of the gradient. This result is called Friedrichs' inequality (sometimes Poincaré–Friedrichs inequality).

Theorem 3.16 (Friedrichs' inequality). *Let Ω be an open, bounded, and connected Lipschitz domain. Then there exists a constant $C > 0$ such that*

$$\|v\|_{L^2(\Omega)} \leq C\|\nabla v\|_{L^2(\Omega)} \quad \textit{for all } v \in H_0^1(\Omega).$$

The constant is C proportional to the diameter of Ω.

Proof. The proof is left as an exercise. We sketch the basic idea. In view of Theorem 3.15, it is enough to consider $v \in C_c^\infty(\Omega)$ and then argue by density. We extend v by zero to some larger rectangular box containing Ω. After shifting coordinates, we may assume that $\Omega \subseteq (0, L)^2$, $L > 0$. Then, v is of class $C_c^\infty((0, L)^2)$ with respect to this box. For any $x \in \Omega$, we can integrate

$$v(x) = v(x_1, x_2) = v(0, x_2) + \int_0^{x_1} \partial_1 v(t, x_2)\, dt.$$

We observe that the boundary term is zero. For the remaining term, we use the Cauchy-Schwarz/Hölder inequality and obtain

$$|v(x)|^2 \leq L \int_0^L |\partial_1 v(t, x_2)|^2\, dt.$$

We now intergrate with respect to x_1

$$\int_0^L |v(x)|^2 dx_1 \leq L^2 \int_0^L |\partial_1 v(t, x_2)|^2\, dt.$$

and thereafter integrate with respect to x_2

$$\int_0^L \int_0^L |v(x)|^2\, dx_1 dx_2 \leq L^2 \int_0^L \int_0^L |\partial_1 v(t, x_2)|^2\, dt dx_2.$$

Since the support of v lies inside Ω, this implies the asserted estimate for v. By a density argument, it is true for all functions from $H_0^1(\Omega)$. ■

The most important implication of Friedrichs' inequality is that $\|\nabla \cdot\|_{L^2(\Omega)}$ defines a norm on $H_0^1(\Omega)$. (Convince yourself that this cannot be a norm on the larger space $H^1(\Omega)$ by considering constant functions.) Denoting the constant from Friedrichs' inequality by C_{F}, we indeed have the equivalence of norms

$$\|v\|_{H^1(\Omega)}^2 = \|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2 \leq (1 + C_{\mathrm{F}}^2)\|\nabla v\|_{L^2(\Omega)}^2 \leq (1 + C_{\mathrm{F}}^2)\|v\|_{H^1(\Omega)}^2. \tag{3.2}$$

We use the notation $|v|_1 = \|\nabla v\|_{L^2(\Omega)}$.

We are now in the position to formulate the Dirichlet problem in Sobolev spaces. From the above formulae we see that $-\Delta u = f$ and $u|_{\partial\Omega} = 0$ implies

$$\int_\Omega \nabla u \cdot \nabla v\, dx = \int_\Omega f v\, dx \quad \text{for all } v \in C_c^\infty(\Omega).$$

This formulation requires less derivative information on u.

Definition 3.17. Let $\Omega \subseteq \mathbb{R}^2$ be an open and bounded Lipschitz domain. Given $f \in L^2(\Omega)$, the variational (or weak) formulation of the Dirichlet problem for Poisson's equation seeks $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \text{for all } v \in H_0^1(\Omega).$$

♦

This generalizes Poisson's equation in the sense that every classical solution will also be a solution to the variational formulation (see exercises).

Lemma 3.18. *Let $\Omega \subseteq \mathbb{R}^2$ be an open and bounded Lipschitz domain. The space $H_0^1(\Omega)$ equipped with the bilinear form*

$$\int_\Omega \nabla v \cdot \nabla w \, dx$$

is a Hilbert space.

Proof. Friedrichs' inequality shows that the symmetric bilinear form is positive definite. The completeness with respect to $|\cdot|_1$ is a consequence of the equivalence of norms (3.2) and the fact that $H_0^1(\Omega)$ is a closed subspace of $H^1(\Omega)$. ■

Theorem 3.19. *Let $\Omega \subseteq \mathbb{R}^2$ be an open and bounded Lipschitz domain and let $f \in L^2(\Omega)$. The variational formulation of the Dirichlet problem of Poisson's equation has a unique solution $u \in H_0^1(\Omega)$.*

Proof. We check that

$$v \mapsto \int_\Omega f v \, dx$$

is a continuous linear functional on the Hilbert space $H_0^1(\Omega)$. This follows from the Cauchy and the Friedrichs inequality

$$\int_\Omega f v \, dx \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} C_{\mathrm{F}} |v|_1.$$

Hence, we are in the setting of the Riesz representation theorem, which states that there is a unique element $u \in H_0^1(\Omega)$ satisfying

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \text{for all } v \in H_0^1(\Omega).$$

■

By using elementary Hilbert space theory we could establish existence and uniqueness to the Dirichlet problem for any right-hand side $f \in L^2(\Omega)$.

§3. Discrete functions

The functions from the foregoing example allow for a very simple representation, and so they are generally suited for numerical computations. It is easy to verify that any such function can be characterized by the vector $(v(x_j))_{j=1}^N$ of its values at the points x_j. Between these nodal points, the values are interpolated by straight lines.

It is possible to generalize this construction to higher space dimensions. We only consider the case $n = 2$ in this lecture in order to minimize the technical efforts. Let the domain $\bar{\Omega}$ be subdivided in triangles. We consider the space of functions that are globally continuous and that are affine when restricted to any of the triangles. In order to define such spaces, we introduce a suitable class of triangular partitions.

Definition 3.20 (triangle)**.** A subset $T \subseteq \mathbb{R}^2$ is called *triangle* if there exists $(z_1, z_2, z_3) \in (\mathbb{R}^2)^3$ such that T is the convex hull of z_1, z_2, z_3 and these three points do not belong on one straight line. The points z_1, z_2, z_3 are called *vertices.* The line segments between z_j, z_k for $j \neq k$ are called *edges.* ♦

Definition 3.21 (regular triangulation)**.** Let $\mathcal{T} \subset 2^{\bar{\Omega}}$ be a finite set of triangles in $\bar{\Omega}$ ($2^{\bar{\Omega}}$ denotes the power set). The set $\mathcal{T}$ is called a *regular triangulation* of Ω if die the triangles cover the domain $\bar{\Omega}$, i.e., $\bigcup_{T\in\mathcal{T}} = \bar{\Omega}$, and if any pair $(T, K) \in \mathcal{T}^2$ satisfies one of the following relations:

(i) $T \cap K = \emptyset$

(ii) $T \cap K$ is a common vertex

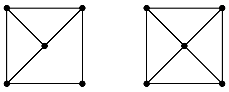
(iii) $T \cap K$ is a common edge

(iv) $T = K$.

♦

This means that the elements of a regular triangulation may only meet under certain rules.

Example 3.22. A non-regular and a regular triangulation of the square:



In what follows, $\mathcal{T}$ will always denote a regular triangulation of Ω. Let $T \in \mathcal{T}$ be a triangle. The affine functions over T are denoted by

$$P_1(T) := \{v \in L^\infty(T) : \exists(a, b, c) \in \mathbb{R}^3 \forall x \in T,\ v(x) = a + bx_1 + cx_2\}.$$

The functions that are piecewise affine with respect to $\mathcal{T}$ (but possibly globally discontinuous) are denoted by

$$P_1(\mathcal{T}) := \{v \in L^\infty(\Omega) : \forall T \in \mathcal{T}, v|_T \in P_1(T)\}.$$

Finally, the continuous and piecewise affine functions are denoted by

$$S^1(\mathcal{T}) := C^0(\Omega) \cap P_1(\mathcal{T})$$

and the subspace with zero boundary conditions reads

$$S^1_0(\mathcal{T}) := \{v \in S^1(\mathcal{T}) : v|_{\partial\Omega} = 0\}.$$

The letter S shall remind us of *splines*; a notion that is possibly known from one-dimensional interpolation.

The following property is very important, and its proof is discussed in the problems below.

Lemma 3.23. *The elements of $S^1(\mathcal{T})$ are weakly differentiable.*

Proof. Exercise. ■

The set of vertices (or *nodes*) of a triangle is denoted by $\mathcal{N}(T)$ and the set of all vertices is

$$\mathcal{N} := \{z \in \bar{\Omega} : \text{there exists } T \in \mathcal{T} \text{ having } z \text{ as a vertex}\} = \bigcup_{T\in\mathcal{T}} \mathcal{N}(T).$$

The basis we choose for $S^1(\mathcal{T})$ or $S^1_0(\mathcal{T})$ is the *nodal basis.* First, we define the nodal basis of $S^1(\mathcal{T})$ (no boundary conditions) by $(\varphi_z)_{z\in\mathcal{N}}$, where for any $z \in \mathcal{N}$ the function $\varphi_z \in S^1(\mathcal{T})$ is defined by the property

$$\varphi_z(y) = \delta_{yz} = \begin{cases} 1 & \text{if } y = z \\ 0 & \text{if } y \in \mathcal{N} \setminus \{z\}. \end{cases} \tag{3.3}$$

These functions are usually referred to as "hat functions". It will be shown in the exercises that these function indeed form a basis.

In order to define a nodal basis of $S^1_0(\mathcal{T})$, one omits the hat functions belonging to boundary vertices. To this end, we define the boundary vertices by $\mathcal{N}(\partial\Omega) := \partial\Omega \cap \mathcal{N}$ and the inner vertices by $\mathcal{N}(\Omega) := \mathcal{N} \setminus \mathcal{N}(\partial\Omega)$. The nodal basis of $S^1_0(\mathcal{T})$ then reads

$$(\varphi_z : z \in \mathcal{N}(\Omega)).$$

As in classical Lagrange interpolation, the coefficients with respect to the nodal basis are given by the nodal values. This means that any function $v_h \in S^1(\mathcal{T})$ can be expanded as follows

$$v_h = \sum_{z\in\mathcal{N}} v_h(z)\varphi_z.$$

The spaces $S^1(\mathcal{T})$ and $S^1_0(\mathcal{T})$ are called *finite element spaces.* Any continuous function $v \in C(\bar{\Omega})$ can be approximated by its interpolation $Iv \in S^1(\mathcal{T})$ as follows

$$Iv := \sum_{z\in\mathcal{N}} v(z)\varphi_z.$$

The map $I : C(\bar{\Omega}) \to S^1(\mathcal{T})$ is called *interpolation operator.* For the case of zero boundary conditions, the definition is analogous.

§4. The finite element method

The finite element method is to compute the orthogonal projection u_h of the solution $u \in H^1_0(\Omega)$ of the Dirichlet problem onto the space $S^1_0(\mathcal{T})$. It is therefore uniquely defined by the condition

$$\int_\Omega \nabla u_h \cdot \nabla v_h dx = \int_\Omega f v_h dx \quad \text{for all } v_h \in S^1_0(\mathcal{T}), \tag{3.4}$$

which describes a positive definite linear problem in a finite-dimensional space. Our Hilbert space theory teaches us that it is indeed the best approximation.

Theorem 3.24. *Let $\Omega \subset \mathbb{R}^2$ be an open, bounded, connected Lipschitz polygon with a triangulation $\mathcal{T}$. Given $f \in L^2(\Omega)$, the error between the solution $u \in H^1_0(\Omega)$ to the variational form of Poisson's equation and the finite element solution $u_h \in S^1_0(\mathcal{T})$ satisfies*

$$|u - u_h|_1 = \inf_{v_h\in S^1_0(\mathcal{T})} |u - v_h|_1.$$

We have seen that the finite element method is, in some sense, optimal. The result should illustrate the basic idea of the error analysis. It is possible to generalize the theory to more general operators (not just the Laplacian), but this is not in the focus of this lecture.

We would like to quantify the right-hand side of the best-approximation result in terms of the mesh-size (maximum diameter of the triangles in $\mathcal{T}$). The idea is to plug in a suitable approximation in the infimum for which we then derive quantified bounds. To achieve this, we will use the finite element interpolation. It is, however, not a well defined on $H^1(\Omega)$ because it takes point evaluations, which need not exist without further assumptions (see Problem A.34). This means that the interpolation operator, denoted by I_h, assigning the finite element interpolation $I_h v$ to any suitable (say continuous) function v, is not well defined on $H^1(\Omega)$. It can, however, be shown that point evaluations are well-defined in the space

$$H^2(\Omega) = \{v \in L^2(\Omega) : \text{all weak derivatives of } v \text{ up to order 2 exist as functions of } L^2(\Omega)\}$$

with norm

$$\|v\|_{H^2(\Omega)} = \sqrt{\sum_{|\alpha|\leq 2} \|\partial^\alpha v\|^2_{L^2(\Omega)}}.$$

Theorem 3.25. *Let $\Omega \subseteq \mathbb{R}^2$ be an open and bounded Lipschitz polygon. Then, we have the continuous embedding $H^2(\Omega) \hookrightarrow C(\Omega)$ and there exists a constant $C > 0$ such that*

$$\|v\|_{L^\infty(\Omega)} \leq C\|v\|_{H^2(\Omega)} \quad \textit{for any } v \in H^2(\Omega).$$

Proof. The result will be proven for triangles in Exercise A.45. For polygons, the result then follows by covering the domain with triangles and using the bound available for these. ■

We have seen that we can apply the finite element interpolation I_h under the assumption that our solution u satisfies the stronger property $u \in H^1_0(\Omega) \cap H^2(\Omega)$. For a derivation of a quantitative bound on the interpolation under this assumption, we will use —without proof— an interpolation error estimate on a reference triangle. We shall then carefully consider a transformation to arbitrary triangles and see how the estimate depends on the geometry of these.

Lemma 3.26 (interpolation on a reference triangle)**.** *Let $\hat{T}$ be the convex combination of the points $(0,0)$, $(1,0)$, $(0,1)$. There exists a constant $\hat{C} > 0$ such that for any $\hat{v} \in H^2(\hat{T})$ we have*

$$\|\nabla(\hat{v} - I_h\hat{v})\|_{L^2(\hat{T})} \leq \hat{C}\|D^2\hat{v}\|_{L^2(\hat{T})}.$$

Proof. The proof is worked out in basic finite element courses. ■

Here and throughout these notes, we use the notation $\|D^2 v\|_{L^2(T)} = \sqrt{\int_T \sum_{j,k=1}^2 |\partial_{jk} v|^2\,dx}$.

We introduce a parameter measuring the mesh quality.

Definition 3.27. Let $T \subseteq \mathbb{R}^2$ be a triangle. Let h_T denote its diameter and let ρ_T denote the diameter of the largest ball inscribed to T. The quantity h_T/ρ_T is called the *aspect ratio* of T. ⧫

Lemma 3.28. *Let $\Phi(\hat{x}) = B\hat{x} + c$ denote the affine map from a triangle $\hat{T}$ to the triangle T. Then, the spectral norm $\|\cdot\|$ of B and B^{-1} satisfies*

$$\|B\| \leq \frac{h_T}{\rho_{\hat{T}}} \quad \textit{and} \quad \|B^{-1}\| \leq \frac{h_{\hat{T}}}{\rho_T}.$$

Proof. Given any vector $\xi \in \mathbb{R}^2$ of length $|\xi| = \rho_{\hat{T}}$, there exists pair of points $\hat{x}$, $\hat{y}$ inside $\hat{T}$ with $\hat{x} - \hat{y} = \xi$ because the full ball of diameter $\rho_{\hat{T}}$ is contained in $\hat{T}$. Since $\Phi(\hat{x})$ and $\Phi(\hat{y})$ belong to T, the image under B satisfies $B\xi = B(\hat{x} - \hat{y}) = \Phi(\hat{x}) - \Phi(\hat{y})$ and its length is bounded by the diameter h_T. We thus compute

$$\|B\| = \sup_{\xi \in \mathbb{R}^2, |\xi|=1} |B\xi| = \sup_{\xi \in \mathbb{R}^2, |\xi|=\rho_{\hat{T}}} \frac{1}{\rho_{\hat{T}}} |B\xi| \leq \frac{h_T}{\rho_{\hat{T}}}.$$

The second asserted estimate follows from interchanging the roles of T and $\hat{T}$. ■

We now prove the interpolation error estimate.

Theorem 3.29 (interpolation on an arbitrary triangle)**.** *There exists a constant $C > 0$ such that for any triangle T and any $v \in H^2(T)$ we have*

$$\|\nabla(v - I_h v)\|_{L^2(T)} \leq C \frac{h_T}{\rho_T} h_T \|D^2 v\|_{L^2(T)}.$$

Proof. We consider the affine transformation

$$\Phi : \hat{T} \to T$$

from the reference triangle to T. We denote by $e := v - I_h v$ the interpolation error and observe from the change-of-variables formula that

$$\|\nabla e\|^2_{L^2(T)} = \int_T |\nabla e|^2 \, dx = \int_{\hat{T}} |(\nabla e) \cdot \Phi|^2 |\det D\Phi| \, dx$$

We use notation $\hat{v} := v \circ \Phi$ and $\hat{e} := e \circ \Phi$. The chain rule reveals for any $\hat{x} \in \hat{T}$ that

$$\nabla \hat{e}(\hat{x}) = D\Phi(\hat{x})^\top \nabla e|_{\Phi(\hat{x})}.$$

Multiplying with the inverse of $D\Phi^\top$ and taking squares thus leads to

$$|(\nabla e) \circ \Phi|^2 = |(D\Phi^\top)^{-1} \nabla \hat{e}|^2 \leq \|D\Phi^{-1}\|^2 |\nabla \hat{e}|^2$$

where $\|\cdot\|$ denotes the (pointwise) spectral matrix norm.

We observe that $D\Phi$ is constant on $\hat{T}$ (because Φ is affine). We thus obtain

$$\|\nabla e\|^2_{L^2(T)} \leq \|D\Phi^{-1}\|^2 |\det D\Phi| \|\nabla \hat{e}\|^2_{L^2(\hat{T})}.$$

By Lemma 3.26 there exists a constant $\hat{C}$, depending on $\hat{T}$, such that

$$\|\nabla \hat{e}\|^2_{L^2(\hat{T})} \leq \hat{C}^2 \|D^2 \hat{v}\|^2_{L^2(\hat{T})}.$$

Here, we have used that $\hat{e}$ is the interpolation error if $\hat{v}$. So far we have shown

$$\|\nabla e\|^2_{L^2(T)} \leq \hat{C}^2 \|D\Phi^{-1}\|^2 \int_{\hat{T}} |D^2 \hat{v}|^2 |\det D\Phi| \, dx.$$

The chain rule shows

$$D^2 \hat{v}(\hat{x}) = D\Phi(\hat{x})^\top D^2 v|_{\Phi(\hat{x})} D\Phi(\hat{x}).$$

We thus find

$$|D^2\hat{v}|^2 \leq \|D\Phi(\hat{x})\|^4 |(D^2 v)\circ\Phi|^2.$$

After transforming back to T we thus obtain

$$\|\nabla e\|_{L^2(T)}^2 \leq \hat{C}^2 \|D\Phi^{-1}\|^2 \|D\Phi\|^4 \|D^2\hat{v}\|_{L^2(T)}^2.$$

The norms of $D\Phi$ and its inverse can be estimated with Lemma 3.28 as follows

$$\|D\Phi^{-1}\|^2\|D\Phi\|^4 \leq \frac{h_{\hat{T}}^2}{\rho_T^2}\frac{h_T^4}{\rho_{\hat{T}}^4} = \frac{h_{\hat{T}}^2}{\rho_{\hat{T}}^4}\frac{h_T^4}{\rho_T^2}.$$

The terms related to $\hat{T}$ are independent of T and can be estimated by some universal constant. We thus obtain (after taking squareroots) the asserted bound on the norm of the gradient. ■

We see from the interpolation error estimate of that the interpolation error is proportional to h_T provided the aspect ratio of the triangle is bounded. We say that a family of triangulations with bounded aspect ratio is *shape-regular*. The approximation of an H^2 function is then determined by the mesh-size h_T and thus improved under mesh-refinement. We obtain:

Corollary 3.30 (global interpolation error estimate)**.** *Let $\Omega \subseteq \mathbb{R}^2$ be an open and bounded polygon. Let $\{\mathcal{T}_h\}_h$ be a shape-regular family of triangulations. Then, there is a constant $C > 0$ such that for any $v \in H^2(\Omega)$ the finite element interpolation I_h with respect to a mesh $\mathcal{T}_h$ satisfies*

$$\|\nabla(v - I_h v)\|_{L^2(\Omega)} \leq Ch\|D^2 v\|_{L^2(\Omega)}$$

for the maximal mesh-size $h = \max T \in \mathcal{T}_h h_T$.

We have seen that any $v \in H^2(\Omega)$ is approximated with order h by the finite element interpolation the H^1 norm. For convex domains, the assumption that the solution to Poisson's equation is H^2 regular, can be proven:

Theorem 3.31 (regularity on convex domains)**.** *Let $\Omega \subset \mathbb{R}^2$ be an open convex domain. Given any $f \in L^2(\Omega)$, the solution to the Dirichlet problem of the Laplacian (Poisson's equation) satisfies $u \in H_0^1(\Omega) \cap H^2(\Omega)$ with the bound*

$$\|D^2 u\|_{L^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}.$$

Proof. See the PDE literature. ■

Remark 3.32. When the domain is nonconvex, the solution may fail to belong to $H^2(\Omega)$. This is for instance the case in Exercise A.12. This is why we assume convexity throughout this lecture. ♦

Finally, we can quantify the approximation error of the finite element method.

Corollary 3.33. *Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded, convex polygon. Then, the error between u and the finite element approximation u_h with respect to a triangulation $\mathcal{T}_h$ from a shape-regular family satisfies*

$$\|\nabla(u - u_h)\|_{L^2(\Omega)} \leq h\|D^2 u\|_{L^2(\Omega)}.$$

Proof. Since $u \in H^2(\Omega)$, the interpolation $I_h u$ is well-defined. We know from the best-approximation property that

$$\|\nabla(u - u_h)\|_{L^2(\Omega)} \leq \|\nabla(u - I_h u)\|_{L^2(\Omega)}.$$

The assertion then follows from the interpolation error estimate of Corollary 3.30. ■

We can prove an improved bound for the error in the L^2 norm.

Theorem 3.34 (L^2 error bound)**.** *Let $\Omega \subset \mathbb{R}^2$ be an open convex domain. Given any $f \in L^2(\Omega)$, the solution to the Dirichlet problem of the Laplacian (Poisson's equation) and its finite element approximation satisfy*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch\|\nabla(u - u_h)\|_{L^2(\Omega)} \leq C'h^2\|D^2u\|_{L^2(\Omega)} \leq C''h^2\|f\|_{L^2(\Omega)}.$$

Proof. The technique employed in the proof is known as the *Aubin-Nitsche duality trick.* The idea is to solve for a solution $z \in H^1_0(\Omega)$ an auxiliary problem whose right-hand side is given by the error $e := u - u_h$. Let z solve

$$\int_\Omega \nabla z \cdot \nabla v \, dx = \int_\Omega ev \, dx \quad \text{for all } v \in H^1_0(\Omega).$$

We test the equation with $v := e$ and obtain

$$\|e\|^2_{L^2(\Omega)} = \int_\Omega e \, e \, dx \quad \text{for all } v \in H^1_0(\Omega) = \int_\Omega \nabla e \cdot \nabla z \, dx.$$

We now use the Galerkin orthogonality and plug in the finite element approximation z_h to z,

$$\int_\Omega \nabla e \cdot \nabla z \, dx = \int_\Omega \nabla(u - u_h) \cdot \nabla z \, dx = \int_\Omega \nabla(u - u_h) \cdot \nabla(z - z_h) \, dx.$$

Corollary 3.33 implies for the finite element errors that

$$\begin{aligned} &\|\nabla(u - u_h)\|_{L^2(\Omega)} \leq C\|D^2u_h\|_{L^2(\Omega)} \\ \text{and} \qquad &\|\nabla(z - z_h)\|_{L^2(\Omega)} \leq C\|D^2z_h\|_{L^2(\Omega)} \leq C\|e\|_{L^2(\Omega)}. \end{aligned}$$

We now combine the above formulas and divide by the norm of e to arrive at the first asserted estimate. The second one follows from Theorem 3.31. ■

§5. Mesh refinement

Given a triangulation $\mathcal{T}_H$, we can generate a finer triangulation $\mathcal{T}_h$ by subdividing every triangle in four sub-triangles by connecting the three midpoints of its edges. This procedure is sometimes referred to as *red refinement.* We note that

- the mesh size h is half the mesh size H, that is $h = H/2$,
- all new triangles generated from a coarse triangle T are congruent to T,
- mesh families generated by red refinement are shape regular,
- all vertices from $\mathcal{T}_H$ are also vertices of $\mathcal{T}_h$.

This is partly shown in the exercises. Of course, any piecewise affine function with respect to $\mathcal{T}_H$ is piecewise affine with respect to $\mathcal{T}_h$ as well. We therefore deduce the nestedness property $S^1(\mathcal{T}_H) \subseteq S^1(\mathcal{T}_h)$. Such an embedding of coarse-grid functions to fine-grid functions was already needed for FDM multigrid. Now that we are operating with functions, we can give a sound meaning to the embedding by stating an inclusion of vector spaces. In view of Exercise A.42, the finite difference embedding of Figure 2.2 now becomes transparent: it is the embedding of the corresponding finite element space if the grid is interpreted as the set of vertices of the triangulation from Figure A.1. The above embedding is linear and can be represented by a (rectangular) matrix, the *prolongation matrix*.

Let us now briefly discuss how to operate with triangulations and finite element functions on a computer (using Python). We describe a triangulation by prescribing a list of nodes and a list of triangles. The nodes are put in a list `coord` $\in \mathbb{R}^{N\times 2}$. The x and y coordinate of the jth node are written to the jth row. In the example of Figure 3.1 this means

```
coord = np.asarray([[0,0],
                    [1,0],
                    [1,1],
                    [0,1],
                    [.5,.5]])
```

for the unit square $(0,1)^2$. Here, we use the library `numpy`:

```
import numpy as np
```

Now we form triangles out of the node numbers. We use convention that the numbering is counterclockwise. The list `triangles` $\in \mathbb{R}^{N\times 3}$ contains in its jth row the three node numbers of triangle number j. In the example from Figure 3.1 this reads

```
triangles = np.asarray([[0,1,4],
                        [1,2,4],
                        [2,3,4],
                        [3,0,4]])
```

We finally save the node pairs of the boundary edges on the Dirichlet boundary

```
dirichlet= np.array([[0,1],
                     [1,2],
                     [2,3],
                     [3,0]])
```

We will comment on (and make use of) this later. In Python we can now plot our triangulation by:

```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.tri as mtri

plt.triplot(mtri.Triangulation(coord[:,0], coord[:, 1], triangles))
plt.show()
```

If we want to generate a surface plot of a piecewise affine function from $S^1(\mathcal{T})$, we can use `trisurf`. Figure 3.2 shows a complete example.

The triangulation in the above example is very coarse. Finer triangulations can be obtained red refinement. We provide a routine `red_refine.py` on the lecture webpage. We do not care about the actual code, but we just use it. It can be used as follows

```
neumann=np.zeros([0, 2])
coord, triangles, dirichlet,_,_,P = \
           red_refine(coord, triangles, dirichlet, neumann)
```

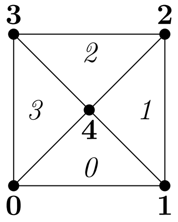


Figure 3.1.: Triangulation of the square $(0, 1)^2$ in four triangles. The bold numbers indicate the node numbers wile the numbers of the triangles are displayed in italic.

```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.tri as mtri
from mpl_toolkits import mplot3d
from mpl_toolkits.mplot3d import Axes3D

coord = np.asarray([[0,0],[1,0],[1,1],[0,1],[.5,.5]])
triangles = np.asarray([[0,1,4],[1,2,4],[2,3,4],[3,0,4]])
dirichlet= np.array([[0,1],[1,2],[2,3],[3,0]])
# show triangulation
plt.triplot(mtri.Triangulation(coord[:,0], coord[:, 1], triangles))
plt.show()
# plot the interpolation of the function x+y
func = lambda x, y:  x + y
func2=np.vectorize(func)
z=func2(coord[:,0],coord[:,1])
fig = plt.figure(figsize =(14, 9))
ax = plt.axes(projection ='3d')
trisurf = ax.plot_trisurf(coord[:,0],coord[:,1],z,
                          triangles = triangles,
                          cmap =plt.get_cmap('summer'),
                          edgecolor='Gray');
plt.show()
```

Figure 3.2.: Sample use of `triplot` and `trisurf`

```
import numpy as np
import math
import pylab
from red_refine import red_refine #our refinement routine
import scipy.sparse
import scipy.sparse.linalg
from scipy.sparse import csr_matrix
```

Figure 3.3.: The required packages for the FEM in Python

Here, `neumann` is just an empty list that, at this stage, has no importance. Later in the lecture we will also consider problems with a second type of boundary condition (so-called Neumann boundary condition), but for the moment we can ignore it; we also do not care about the two ignored output arguments of the function. The output `P` is the prolongation matrix, mapping the coefficient vector of a finite element function to its representation on the refined grid. It will be of importance in the multigrid method.

§6. Implementation of the FEM

Let us now describe how to implement the FEM on the computer. Full Python routines can be found on our lecture webpage, so that here we will focus on the important (mathematical) aspects of the implementation. In order to discretize Poisson's equation (subject to homogeneous Dirichlet boundary conditions) with the FEM, we need

- a triangulation $\mathcal{T}$, described through the data structures `coord`, `triangles`, `dirichlet`,
- the right-hand side f, e.g. given through values at certain points or as function,
- a vector b representing the linear functional $\int_\Omega f \bullet \, dx$ with respect to the nodal basis of $S^1(\mathcal{T})$,
- the so-called *stiffness matrix* A, i.e., the matrix representing the bilinear form from Poisson's equation with respect to the nodal basis of $S^1(\mathcal{T})$.

With these objects at hand, we can solve for the coefficient vector of the FEM solution u_h. It is important to restrict the matrices to the *degrees of freedom*. In our case, these correspond to the inner nodes (as the values for the boundary nodes are already fixed by the value 0). The list of degrees of freedom is usually given the variable name `dof`.

We start by specifying some required packages for (sparse) linear algebra, see Figure 3.3. The structure of the program is displayed in Figure 3.4.

It remains to describe the routines for assembling the stiffness matrix A and the right-hand side vector b. We start with A. First, we build up *local* stiffness matrices for each triangle T

$$A_T^{loc} := \Big(\int_T \nabla\varphi_j \cdot \nabla\varphi_k \, dx\Big)_{j,k=1,2,3}.$$

Here, the vertices of T are locally numbered by $1, 2, 3$. Since the φ_j are affine functions, their gradients are constant so that we arrive at the formula

$$A_T^{loc} = \operatorname{area}(T) \begin{bmatrix} \nabla\varphi_1^\top \\ \nabla\varphi_2^\top \\ \nabla\varphi_3^\top \end{bmatrix} \begin{bmatrix} \nabla\varphi_1 \; \nabla\varphi_2 \; \nabla\varphi_3 \end{bmatrix}.$$

```
def FEM(coord,triangles,dirichlet,f):
    nnodes=np.size(coord,0)
    A=stiffness_matrix(coord,triangles)
    b=RHS_vector(coord,triangles,f)
    dbnodes=np.unique(dirichlet)
    dof=np.setdiff1d(range(0,nnodes),dbnodes)
    ndof=np.size(dof)
    R=restrict2dof(dof,nnodes)
    A_inner=(R.transpose()@A)@R
    b_inner=R.transpose()@b
    x=np.zeros(nnodes)
    x[dof]=scipy.sparse.linalg.spsolve(A_inner,b_inner)
    return x, ndof
```

Figure 3.4.: The basic FEM routine.

The area is easily computed as follows. With the three vertices $z_1, z_2, z_3 \in \mathbb{R}^2$ of T, we have that

$$\operatorname{area}(T) = \frac{1}{2}\det[z_2 - z_1, z_3 - z_1].$$

For the computation of $\nabla\varphi_j$ we observe that the basis functions (or barycentric coordinates) satisfy the system

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 \\ z_1 & z_2 & z_3 \end{bmatrix}}_{\in\mathbb{R}^{3\times 3}} \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \varphi_3(x) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ x \end{bmatrix}}_{\in\mathbb{R}^{3\times 1}}$$

for any T. If we take derivatives (w.r.t. x) on both sides, we arrive at

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 \\ z_1 & z_2 & z_3 \end{bmatrix}}_{\in\mathbb{R}^{3\times 3}} \underbrace{\begin{bmatrix} \nabla\varphi_1^\top \\ \nabla\varphi_2^\top \\ \nabla\varphi_3^\top \end{bmatrix}}_{\in\mathbb{R}^{3\times 2}} = \underbrace{\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\in\mathbb{R}^{3\times 2}}.$$

Therefore

$$\begin{bmatrix} \nabla\varphi_1^\top \\ \nabla\varphi_2^\top \\ \nabla\varphi_3^\top \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ z_1 & z_2 & z_3 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

We compute all local stiffness matrices in a loop

```
    nelems=np.size(triangles,0)
    Alocal=np.zeros((nelems,3,3))

    for j in range(0,nelems):
        nodes_loc=triangles[j,:]
        coord_loc=coord[nodes_loc,:]
        T=np.array([coord_loc[1,:]-coord_loc[0,:] ,
                coord_loc[2,:]-coord_loc[0,:] ])
        area = 0.5 * ( T[0,0]*T[1,1] - T[0,1]*T[1,0] )
        T= np.concatenate((np.array([[1,1,1]]), coord_loc.T),axis=0)
        T1= np.array([[0,0],[1,0],[0,1]])
        grads = np.linalg.solve(T,T1)
        Alocal[j,:,:]=area* np.matmul(grads,grads.T)
```

Now we need to assemble the local stiffness matrices into the global stiffness matrix. The entry A_{jk} of the global stiffness matrix is given by

$$A_{jk} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_k \, dx = \sum_{T\in\mathfrak{T}} \int_T \nabla\varphi_j \cdot \nabla\varphi_k \, dx = \sum_{\substack{T\in\mathfrak{T} \\ \text{nodes } j,k \\ \text{belong to } K}} \int_T \nabla\varphi_j \cdot \nabla\varphi_k \, dx.$$

This means that, for any triangle, we save the index pairs (j, k in the above sum) assigning the global node numbers to the entries of the local stiffness matrix. We write these indices into the arrays I1, I2. We then build up a sparse matrix based on these indices (note that repeated indices imply summation).

```
nelems=np.size(triangles,0)
nnodes=np.size(coord,0)
I1=np.zeros((nelems,3,3))
I2=np.zeros((nelems,3,3))

for j in range(0,nelems):
    nodes_loc=triangles[j,:]
    I1[j,:,:] = np.concatenate((np.array([nodes_loc]),\
          np.array([nodes_loc]),np.array([nodes_loc])),axis=0)
    I2[j,:,:] = np.concatenate((np.array([nodes_loc]).T,\
            np.array([nodes_loc]).T,np.array([nodes_loc]).T),axis=1)

Alocal=np.reshape(Alocal,(9*nelems,1)).T
I1=np.reshape(I1,(9*nelems,1)).T
I2=np.reshape(I2,(9*nelems,1)).T
A=csr_matrix((Alocal[0,:],(I1[0,:],I2[0,:])),shape = (nnodes,nnodes))
```

The full routine for the stiffness matrix can be found in Figure 3.5. We now proceed with the assembling of the right-hand side. We again run a loop over all elements. Since b is not sparse, we can just update the vector in each loop iteration. For approximating the integral, we use the midpoint rule

$$\int_T f\varphi_j \, dx \approx \operatorname{area}(T) f(m)\varphi_j(m),$$

where $m = \frac{1}{3}(z_1 + z_2 + z_3)$ is the midpoint (barycentre) of T. Since φ_j is affine, we can easily compute $\varphi_j(m) = 1/3$. This results in the routine of Figure 3.6.

For testing the FEM code, we use the above data for the unit square. In the code they will be loaded by a function geom_square. We use the following right-hand side for validation

$$f(x) = 2(x_1(1 - x_1) + x_2(1 - x_2)).$$

The exact solution reads

$$u(x) = x_1(x_1 - 1)x_2(x_2 - 1).$$

For a very basic convergence test for the L^∞ norm we now execute the code from Figure 3.7.

```
def stiffness_matrix(coord,triangles):
    nelems=np.size(triangles,0)
    nnodes=np.size(coord,0)
    Alocal=np.zeros((nelems,3,3))
    I1=np.zeros((nelems,3,3))
    I2=np.zeros((nelems,3,3))

    for j in range(0,nelems):
        nodes_loc=triangles[j,:]
        coord_loc=coord[nodes_loc,:]
        T=np.array([coord_loc[1,:]-coord_loc[0,:] ,
               coord_loc[2,:]-coord_loc[0,:] ])
        area = 0.5 * ( T[0,0]*T[1,1] - T[0,1]*T[1,0] )
        tmp1= np.concatenate((np.array([[1,1,1]]), coord_loc.T),axis=0)
        tmp2= np.array([[0,0],[1,0],[0,1]])
        grads = np.linalg.solve(tmp1,tmp2)
        Alocal[j,:,:]=area* np.matmul(grads,grads.T)
        I1[j,:,:] = np.concatenate((np.array([nodes_loc]),np.array([
           nodes_loc]),np.array([nodes_loc])),axis=0)
        I2[j,:,:] = np.concatenate((np.array([nodes_loc]).T,np.array([
           nodes_loc]).T,np.array([nodes_loc]).T),axis=1)

    Alocal=np.reshape(Alocal,(9*nelems,1)).T
    I1=np.reshape(I1,(9*nelems,1)).T
    I2=np.reshape(I2,(9*nelems,1)).T
    A=csr_matrix((Alocal[0,:],(I1[0,:],I2[0,:])),shape = (nnodes,nnodes)
       )
    return A
```

Figure 3.5.: Routine for the stiffness matrix.

```
def RHS_vector(coord,triangles,f):
    nelems=np.size(triangles,0)
    nnodes=np.size(coord,0)
    b=np.zeros(nnodes)
    for j in range(0,nelems):
        nodes_loc=triangles[j,:]
        coord_loc=coord[nodes_loc,:]
        tmp=np.array([coord_loc[1,:]-coord_loc[0,:] ,
               coord_loc[2,:]-coord_loc[0,:] ])
        area = 0.5 * ( tmp[0,0]*tmp[1,1] - tmp[0,1]*tmp[1,0] )
        mid=1/3*(coord_loc[0,:]+coord_loc[1,:]+coord_loc[2,:])
        b[nodes_loc]=b[nodes_loc]+area/3*f(mid[0],mid[1])
    return b
```

Figure 3.6.: Routine for the right-hand side vector.

```
fun = lambda x, y:  (x-x**2)*(y- y**2)
u_exact=np.vectorize(fun)
f = lambda x, y:  2* ((x-x**2)+(y- y**2) )
coord, triangles, dirichlet, neumann = get_geom()
max_err=np.zeros(5)
for j in range(0,5):
    coord, triangles, dirichlet,_,_,_ = \
          red_refine(coord, triangles, dirichlet, neumann)
    x=FEM(coord, triangles, dirichlet,f)
    u_at_nodes=u_exact(coord[:,0],coord[:,1])
    max_err[j]=np.max(np.abs(u_at_nodes-x))
print(max_err)
```

Figure 3.7.: Testing the FEM code.

§7. Discrete inequalities and norms

When considering discrete functions, for example functions from our finite element spaces, we might have control (in terms of norms) over certain properties that are due the finite dimension. For example, we know that an estimate of the form $\|\nabla v\|_{L^2(\Omega)} \leq C\|v\|_{L^2(\Omega)}$ (which would some reverse Friedrichs estimate) cannot hold for all $v \in H^1(\Omega)$ with a universal constant C. As a counterexample, take for example $v(x) = \sin(kx_1)$ for an integer k. The L^2 norm is uniformly bounded, while the norm of the gradient grows linearly with k. For discrete functions, however, we can state a bound of this type, albeit with constants that degenerate with the mesh size.

Theorem 3.35 (inverse estimate)**.** *There exists a constant $C > 0$ such that any triangle $T \subseteq \mathbb{R}^2$ and any affine function $v_h \in P_1(T)$ satisfy*

$$\|\nabla v_h\|_{L^2(T)} \leq C\rho_T^{-1}\|v_h\|_{L^2(T)}.$$

If $\mathcal{T}$ is a triangulation of Ω from a shape-regular family, then, with some constant $C > 0$ that only depends on the shape-regularity, we have

$$\|\nabla v_h\|_{L^2(\Omega)} \leq Ch_{\min}^{-1}\|v_h\|_{L^2(\Omega)} \quad \textit{for all } v_h \in S^1(\mathcal{T})$$

where $h_{\min}$ is the minimal diameter from the triangles of $\mathcal{T}$.

Proof. As in previous proofs, we consider the affine transformation $\Phi : \hat{T} \to T$ from a fixed reference triangle to T and use the change-of-variables formula and the chain rule to infer

$$\|\nabla v_h\|_{L^2(T)}^2 = \int_T |\nabla v_h|^2\,dx = \int_{\hat{T}} |(\nabla v_h)\cdot\Phi|^2 |\det D\Phi|\,dx = |\det D\Phi| \int_{\hat{T}} |D\Phi^{-\top}\nabla\hat{v}_h|^2\,dx$$

where we used the notation $\hat{v} := v \circ \Phi$. Denoting by $\|\cdot\|$ the (pointwise) spectral matrix norm we obtain

$$\|\nabla v_h\|_{L^2(T)}^2 \leq |\det D\Phi| \|D\Phi^{-1}\|^2 \|\nabla\hat{v}_h\|_{L^2(\hat{T})}^2.$$

We now argue by referring only to the element $\hat{T}$, where $P_1(\hat{T})$ is a three-dimensional vector space with the seminorm $\|\nabla\cdot\|_{L^2(\hat{T})}$ and the norm $\|\cdot\|_{L^2(\hat{T})}$. We know that $P_1(\hat{T})$ is isomorphic to $\mathbb{R}^3$ where all norms are equivalent. Thus, the unit sphere of $P_1(\hat{T})$ with respect to the $L^2(\hat{T})$ norm is compact so that, by a standard argument, there exists a constant $\hat{C}$ such that

$$\|\nabla\hat{w}_h\|_{L^2(\hat{T})} \leq \hat{C}\|\nabla\hat{w}_h\|_{L^2(\hat{T})} \quad \text{for all } \hat{w}_h \in P_1(\hat{T}).$$

The dependence of $\hat{C}$ on $\hat{T}$ is very critical and it is therefore important that we use the equivalence-of-norms argument on only one fixed reference triangle. If we used a different triangle for reference, the constant could be completely different, which is the reason why we argue by change of variables. Summarizing the above, we have shown that

$$\|\nabla v_h\|_{L^2(T)}^2 \leq \hat{C}^2 \|D\Phi^{-1}\|^2 \int_{\hat{T}} |v_h \circ \Phi|^2 |\det D\Phi|\, dx.$$

From Lemma 3.28 we know that $\|D\Phi^{-1}\| \leq h_{\hat{T}}/\rho_T$. Transforming back to T thus yields the first asserted result. The second one follows by splitting the integral into triangle contributions, using the proven local bound, and estimating any ρ_T^{-1} by h_T^{-1} through the aspect ratio. ■

We write $a \lesssim b$ if an inequality $a \leq Cb$ holds for a constant $C > 0$ that may depend on the aspect ratio of the triangulation under consideration or on properties of the domain Ω, but not on the mesh size.

In what follows, we consider a sequence of uniformly (red) refined triangulations $\mathcal{T}_k$ starting from an initial triangulation $\mathcal{T}_0$, where $\mathcal{T}_{k+1}$ is the uniform refinement of $\mathcal{T}_k$. The resulting family of meshes $(\mathcal{T}_k)_k$ is shape regular and *quasi-uniform*, which means that minimal of maximal mesh size of a given mesh are comparable, i.e.,

$$h_{\max}(\mathcal{T}_k) \lesssim h_{\min}(\mathcal{T}_k) \quad \text{for any } k \geq 0.$$

We then simply write h_k for the maximal mesh size of the mesh $\mathcal{T}_k$.

Definition 3.36 (discrete norm)**.** Let $\mathcal{T}_k$ be a triangulation from the above family and $v_k \in S_0^1(\mathcal{T}_k)$ with coefficient vector y with respect to the nodal basis over $\mathcal{T}_k$. We define the norm

$$\|v_k\|_{s,k} := h_k^{1-s} \sqrt{y^* A_k^s y}.$$

♦

We have already encountered such discrete norms in the context of smoothing properties for finite differences. Indeed, in the notation of Section §9, we could write

$$\|y\|_{h_k^{(1-s)/s} A_k, s} = \|v_k\|_{s,k}.$$

In the finite element context, we now have a sound interpretation of the norm $\|v_k\|_{A_k,1}$, namely then

$$\|\nabla v_k\|_{L^2(\Omega)} = \|v_k\|_{A_k,1}.$$

For the discrete norms, we have the following generalized Cauchy–Schwarz inequality.

Lemma 3.37. *Any $v_k, w_k \in S_0^1(\mathcal{T}_k)$ satisfy.*

$$\int_\Omega \nabla v_k \cdot \nabla w_k \, dx \leq \|v_k\|_{1+t,k} \|w_k\|_{1-t,k} \quad \textit{for any } t \in \mathbb{R}.$$

Proof. Given y, z as the coefficient vectors to the functions v_k, w_k with respect to the nodal basis, we compute

$$\langle y, A_k z\rangle_2 = \langle A_k^{(1+t)/2} y, A_k^{(1-t)/2} z\rangle_2 \leq \|y\|_{A_k,1+t} \|z\|_{A_k,1-t}.$$

This implies the assertion. ■

We shall now prove that $\|v_k\|_{A_k,0}$ is equivalent to the L^2 norm of v_k.

Lemma 3.38. *Any $v_k \in S^1_0(\mathcal{T}_k)$ satisfies*

$$\|v_k\|_{L^2(\Omega)} \lesssim \|v_k\|_{0,k} \lesssim \|v_k\|_{L^2(\Omega)}.$$

Proof. We consider a single triangle $T \in \mathcal{T}_k$, where $v_k|_T$ is expanded in terms of the nodal basis of T as $v_k|_T = \sum_{j=1}^3 y_j \varphi_j$. As in previous proofs, we use the affine diffeomorphism $\Phi : \hat{T} \to T$ and write $\hat{v} = v \circ \Phi$. We transform back to a reference element and use equivalence of norms,

$$\|v_k\|^2_{L^2(T)} = |\det D\Phi| \|\hat{v}_k\|^2_{L^2(\hat{T})} \lesssim |\det D\Phi| \|y\|_2^2$$

where $\|y\|_2$ is the Euclidean norm of y. Due to the shape regularity, $|\det D\Phi| \lesssim h_k^2$ because $|\det D\Phi|$ is twice the area of T. Using again equivalence of norms, the shape regularity, and transforming back to T yields

$$h_k^2 \|y\|_2^2 \lesssim |\det D\Phi| \|\hat{v}_k\|^2_{L^2(\hat{T})} \lesssim \|v_k\|^2_{L^2(T)}.$$

We have thus shown $\|v_k\|_{L^2(T)} \lesssim h_k \|y\|_2 \lesssim \|v_k\|_{L^2(T)}$ for a single triangle T. The proof for the norm over Ω follows from summing over all triangles and the fact that each vertex belongs to a finite number of triangles whose number is uniformly bounded by the shape regularity. This means that each coefficient y_j is counted at most C many times for some $C \lesssim 1$. ■

With the tools developed in this section, we obtain a bound on the condition number of the FEM system matrix.

Lemma 3.39 (conditioning of the FEM system)**.** *The corresponding finite element system matrix A_k with respect to the triangulation $\mathcal{T}_k$ satisfies $\kappa_2(A_k) \lesssim h_k^{-2}$.*

Proof. Let $y \in \mathbb{R}^{N_k}$ where N_k is the number of interior vertices of $\mathcal{T}_k$. This defines a function $v_k \in S^1_0(\mathcal{T})$ by the expansion $v_k = \sum_j y_j \varphi_j$ in the nodal basis over $\mathcal{T}_k$. We combine Friedrichs' inequality with the inverse estimate and obtain

$$\|v_k\|_{L^2(\Omega)} \lesssim \|\nabla v_k\|_{L^2(\Omega)} \lesssim h_k^{-1} \|v_k\|_{L^2(\Omega)}.$$

Using discrete norms and the bound of Lemma 3.38 this equivalently reads

$$h_k \|y\|_2 \lesssim \|A^{1/2} y\|_2 \lesssim \|y\|_2^2$$

for the Euclidean norm. This implies that, for nonzero y, the Rayleigh quotient satisfies the bounds

$$h_k^2 \lesssim \frac{\|A^{1/2} y\|_2^2}{\|y\|_2^2} \lesssim 1$$

so that any eigenvalue λ of A satisfies $h_k^2 \lesssim \lambda \lesssim 1$. This implies the bound for the spectral condition number. ■

We have seen that the discrete norms $\|v_k\|_{1,k}$ and $\|v_k\|_{0,k}$ represent the L^2 norm ∇v_k and (up to scaling) v_k, respectively. We will use $\|\cdot\|_{2,k}$ as a surrogate for the L^2 norm of the Hessian, which of course is not well defined for a typical finite element function v_k because the gradient will be piecewise constant and discontinuous. Having the application in multigrid methods in mind, we want to quantify the best approximation of v_k by functions v_{k-1} on a coarser mesh. We will denote the orthogonal projection (with respect to the gradient inner product of $H_0^1(\Omega)$) to the space $S_0^1(\mathcal{T}_k)$ by P_k. For the approximation of the solution u to the boundary value problem in a convex domain, we proved Corollary 3.33 where we exploited that u possesses a bounded H^2 norm. This cannot be true for a general discrete function v_k, which is the reason why we resort to its discrete norm $\|\cdot\|_{2,k}$ instead.

Lemma 3.40 (approximation property)**.** *Let $\Omega \subseteq \mathbb{R}^2$ be an open convex polygon. Let $k \geq 1$ and $v_k \in S_0^1(\mathcal{T}_k)$. The best-approximation $P_{k-1}v_k \in S_0^1(\mathcal{T}_k)$ defined via*

$$\int_\Omega \nabla(v_k - P_{k-1}v_k) \cdot \nabla w_k \, dx \quad \textit{for all } w_{k-1} \in S_0^1(\mathcal{T}_{k-1})$$

satisfies

$$\|v_k - P_{k-1}v_k\|_{L^2(\Omega)} \lesssim h_k \|\nabla(v_k - P_{k-1}v_k)\|_{L^2(\Omega)} \lesssim h_k^2 \|v_k\|_{2,k}.$$

Proof. We write $e_k := v_k - P_{k-1}v_k$. As in the proof of the L^2 error estimate, we consider the solution $z \in H^1(\Omega)$ to the problem

$$\int_\Omega \nabla z \cdot \nabla w \, dx = \int_\Omega e_k w \, dx \quad \text{for all } w \in H_0^1(\Omega).$$

Since Ω is convex, we know that $z \in H_0^1(\Omega) \cap H^2(\Omega)$. We argue as in Theorem 3.34 and deduce with the FEM solution $z_{k-1} \in S_0^1(\mathcal{T}_{k-1})$ to z and Galerkin orthogonality that

$$\|e_k\|_{L^2(\Omega)}^2 = \int_\Omega \nabla e_k \cdot \nabla(z - z_{k-1}) \, dx \lesssim h_{k-1} \|\nabla e_k\|_{L^2(\Omega)} \|D^2 z\|_{L^2(\Omega)} \lesssim h_{k-1} \|\nabla e_k\|_{L^2(\Omega)} \|e_k\|_{L^2(\Omega)}$$

where we have used the interpolation estimate from Corollary 3.30 and the regularity result Theorem 3.31 for convex domains. To prove the second claimed inequality, we use Galerkin orthogonality and Lemma 3.37

$$\|\nabla(v_k - P_{k-1}v_k)\|_{L^2(\Omega)}^2 = \int_\Omega \nabla(v_k - P_{k-1}v_k) \nabla v_k \, dx \leq \|v_k - P_{k-1}v_k\|_{0,k} \, \|v_k\|_{2,k}.$$

Since $\|v_k - P_{k-1}v_k\|_{0,k} \lesssim \|v_k - P_{k-1}v_k\|_{L^2(\Omega)}$, we can use the L^2 bound that we just proved and deduce the second claimed estimate. ■

4. Iterative finite element solvers

§1. Multigrid method

We consider the situation that we are given a finite element triangulation $\mathcal{T}_L$ of the open and convex bounded polygon $\Omega \subseteq \mathbb{R}^2$ that arises from a hierarchy $\mathcal{T}_0, \mathcal{T}_1, \dots, \mathcal{T}_L$ of successively uniformly refined meshes, starting from a coarse triangulation $\mathcal{T}_0$. The meshes $(\mathcal{T}_\ell)_\ell$ from a quasi-uniform and shape-regular family. The letter ℓ shall remind us of the fact that we are dealing with different mesh levels. The finite element spaces are denoted by $V_\ell := S_0^1(\mathcal{T}_\ell)$.

For the ease of notation, we define a mesh dependent inner product $(\cdot,\cdot)_\ell$ on V_ℓ by

$$(v_\ell, w_\ell)_\ell := h_\ell^2 \sum_{z \in \mathcal{N}(\Omega)} v_h(z) w_h(z).$$

Up to scaling by h_ℓ^2, this is just the Euclidean inner product of the coefficient vectors with respect to the nodal basis. We have seen in Lemma 3.38 that $(\cdot,\cdot)_\ell$ is equivalent to the L^2 inner product on V_ℓ in the sense that it leads to an equivalent norm. By $\mathcal{A}_\ell : V_\ell \to V_\ell$ we denote the linear operator defined by

$$(\mathcal{A}_\ell v_\ell, w_\ell)_\ell = \int_\Omega \nabla v_\ell \cdot \nabla w_\ell \, dx.$$

Note that after fixing the nodal basis, the operator $\mathcal{A}_\ell$ is represented by the (scaled) stiffness matrix A_ℓ. But $\mathcal{A}_\ell$ operates on finite element functions and not on the corresponding coefficient vectors. This simplifies notation and we can write the finite element equations on the level ℓ as

$$\mathcal{A}_\ell u_\ell = f_\ell$$

for the function $f_\ell \in V_\ell$ defined by

$$(f_\ell, v_\ell)_\ell = \int_\Omega f v_\ell \, dx.$$

The mesh dependent norm from the previous section is easily identified to satisfy

$$\|v_\ell\|_{s,\ell} = \sqrt{(\mathcal{A}_\ell^s v_\ell, v_\ell)_\ell} = h^{1-s} \sqrt{y^* A_\ell y}$$

for the coefficient vector y of v_ℓ. So far, this is just mathematical morphology. As in the finite difference case, we consider the *relaxation operator*

$$R_\ell = R_\ell(\omega_\ell, \mathcal{A}_\ell) := I - \omega_\ell \mathcal{A}_\ell$$

where we will assume throughout this chapter that ω_ℓ^{-1} is an upper bound of the spectral radius of A_ℓ, which satisfies the scaling

$$\rho(A_\ell) \leq \omega_\ell^{-1} \lesssim h_\ell^{-2}.$$

We recall that the relaxation operator is the iteration operator of the relaxed Richardson iteration

$$\mathcal{R}_\ell v_\ell = v_\ell - \omega_\ell(\mathcal{A}_\ell v_\ell - f_\ell)$$

and note that $\mathcal{R}_\ell$ of course depends on f_ℓ. We will tacitly assume that the f_ℓ given on the level ℓ is chosen which is clear from the context.

In Theorem 2.53 we have already seen that the relaxation operator has a smoothing property. In the finite element setting the result reads as follows.

Theorem 4.1 (smoothing property)**.** *The relaxed Richardson iteration with parameter $\omega \leq 1/\lambda_{\max}(\mathcal{A}_\ell)$ satisfies*

$$\|R^k v_\ell\|_{s+t,\ell} \leq C(t, \omega_\ell) h_\ell^{-t} k^{-t/2} \|v_\ell\|_{s,\ell} \quad \textit{for all } v_\ell \in V_\ell, s \in \mathbb{R}, t > 0$$

with $C(t, \omega_\ell) = \left(\frac{t}{2\omega_\ell \exp(1)}\right)^{t/2}$.

Proof. The proof of Theorem 2.53 applies directly to the stiffness matrix A_ℓ. Our discrete norm incorporates a h_ℓ-dependent scaling, which leads to the additional factor h_ℓ^{-t} on the right-hand side. ∎

The smoothing property motivates a multigrid algorithm as in the case of FDM. In the context of FEM, we now have a clear idea, what the *prolongation operator* is, namely the natural injection

$$\iota_{\ell \to \ell+1} : V_\ell \to V_{\ell+1}.$$

Again, its adjoint with respect to the discrete inner product, denoted as

$$\iota^*_{\ell \to \ell+1} : V_{\ell+1} \to V_\ell$$

serves as the *restriction operator*. The definitions imply

$$(\iota^*_{\ell \to \ell+1} v_{\ell+1}, v_\ell)_\ell = (v_{\ell+1}, \iota_{\ell \to \ell+1} v_\ell)_{\ell+1} = (v_{\ell+1}, v_\ell)_{\ell+1} \quad \text{for any } v_{\ell+1} \in V_{\ell+1}, v_\ell \in V_\ell$$

so that $\iota^*_{\ell \to \ell+1}$ can be viewed as an adequate replacement of the L^2 projection from $V_{\ell+1}$ to V_ℓ.

The finite element multigrid algorithm is as follows.

Algorithm 4.2 (multigrid iteration for FEM)**.** We are given as input:

- a mesh hierarchy $(\mathcal{T}_\ell)_\ell$ on levels $\ell = 0, 1, \ldots, L$ with some $L \in \mathbb{N}$ • the right-hand side f_L • an initial guess $u_L^{(0)}$ • the number ν of desired smoothing steps • a parameter $\mu \in \mathbb{N}$ • number k_0 of multigrid iterations

For $k = 1, 2, \ldots, k_0$, the iterate

$$u_L^{(k)} = MG(L, f_L, u_L^{(k-1)}, \nu)$$

is recursively defined by

- $r_L := f_L - \mathcal{A}_L \mathcal{R}_L^\nu u_L^{(k-1)}$ (pre-smoothing)
- $\begin{cases} \textbf{if } L-1=0, \textbf{ then } z_L = \mathcal{A}_0^{-1} \iota^*_{L-1 \to L} r_L \\ \textbf{else} \text{ set } z_L^0 = 0 \text{ and } \textbf{for } j = 1, \ldots, \mu \text{ do: } z_L^j := MG(L-1, \iota^*_{L-1 \to L} r_L, z_L^{j-1}, \nu) \\ \qquad \text{set } z_L := z_L^\mu \end{cases}$
  (coarse-grid correction)
- $u_L^{(k)} := \mathcal{R}_L^\nu (\mathcal{R}_L^\nu u_L^{(k-1)} + \iota_{L-1 \to L} z_L)$ (post-smoothing)

♦

Instead of $I_{L-1\to L}z_L$ we could have simply written z_L, and this notation should rather resemble the structure of an implementation where the embedding is a nontrivial operation because the coefficient vector will change if the same function is represented with respect to two different meshes. Again, we disregard the possibility of choosing two different values ν_1 resp. ν_2 for the pre- resp. post-smoothing. As in the finite difference case, for $\mu = 1$ we call each iteration a V-cycle¸ and for $\mu = 2$ we call it a W-cycle.

An obvious choice for the termination of a good initial guess $u_L^{(0)} \in V_L$ is to take the multigrid solution from a coarser level.

Algorithm 4.3 (initial values by nested iteration)**.** We are given the input from Algorithm 4.2. For $\ell = 1,\dots,L$, compute

$$u_\ell^{(0)} = MG(\ell-1, f_{\ell-1}, u_{\ell-1}^{(k_0)}, \nu)$$

(on level $\ell - 1 = 0$ this means direct solution).
Output: initial value $u_L^{(0)}$ ♦

§2. Analysis of the W-cycle

For the error analysis of the W-cycle, it is convenient to consider the two-grid iteration as an auxiliary tool. The two-grid iteration with meshes $\mathcal{T}_{\ell-1}$ and $\mathcal{T}_\ell$ corresponds to multigrid, where the coarse-grid correction on the level $L-1$ is carried out exactly. The scheme then simplifies to:

Algorithm 4.4 (two-grid iteration for FEM)**.** We are given as input:

- mesh $\mathcal{T}_{L-1}$ and its refinement $\mathcal{T}_L$ • f_L, $u_L^{(0)}$, ν as in Algorithm 4.2

For $k = 1, 2, \dots,$ do

- $r_L := f_L - \mathcal{A}_L\mathcal{R}_L^\nu u_L^{(k-1)}$ (smoothing)
- $u_L^{(k)} = \mathcal{R}_L^\nu u_L^{(k-1)} + \mathcal{A}_{L-1}^{-1}\iota_{L-1\to L}^* r_L$ (coarse-grid correction)

♦

We will now prove that for sufficiently large ν, the two-grid algorithm is contractive.

Lemma 4.5 (two-grid contraction)**.** *Denote by $u_L \in V_L$ the exact finite element solution and abbreviate the error by $e^{(k)} := u_L - u_L^{(k)}$. In the two-grid method with ν smoothing steps we have*

$$\|\nabla e^{(k)}\|_{L^2(\Omega)} \lesssim \nu^{-1/2}\|\nabla e^{(k-1)}\|_{L^2(\Omega)} \quad \textit{for any } k \geq 1.$$

The constant hidden in $\lesssim$ is independent of ν.

Proof. We denote the coarse-grid correction by $z_L^{\mathrm{ex}} := \mathcal{A}_{L-1}^{-1}\iota_{L-1\to L}^* r_L$. With obvious computations, the definition of r_L and $f_L = \mathcal{A}_L u_L$ that

$$(\mathcal{A}_{L-1}z_L^{\mathrm{ex}}, w_{L-1})_{L-1} = (r_L, w_{L-1})_L = (\mathcal{A}_L(u_L - \mathcal{R}_L^\nu u_L^{(k-1)}), w_{L-1})_L$$

for any test function $w_{L-1} \in V_{L-1}$. We have therefore shown that the coarse-grid correction satisfies

$$z_L^{\mathrm{ex}} = P_{L-1}(u_L - \mathcal{R}_L^\nu u_L^{(k-1)}). \tag{4.1}$$

The definition of $u_L^{(k)}$ from the two-grid method therefore implies

$$e^{(k)} = (1 - P_{L-1})(u_L - \mathcal{R}_L^\nu u_L^{(k-1)}).$$

We have represented the error as the best-approximation error of a discrete function of the level L by functions from the level $L-1$. The approximation property from Lemma 3.40 and the representation of the Richardson error through the relaxation operator lead to

$$\|\nabla e^{(k)}\|_{L^2(\Omega)} \lesssim h_L \|u_L - \mathcal{R}_L^\nu u_L^{(k-1)}\|_{2,L} = h_L \|R_L^\nu e^{(k-1)}\|_{2,L}.$$

We combine this estimate with the smoothing property from Theorem 4.1 (with $s = t = 1$) and obtain

$$\|\nabla e^{(k)}\|_{L^2(\Omega)} \lesssim \nu^{-1/2} \|e^{(k-1)}\|_{1,L} = \nu^{-1/2} \|\nabla e^{(k-1)}\|_{L^2(\Omega)}.$$

■

We will now prove contraction of the W-cycle. We will use that the relaxation operator is nonexpansive with respect to the energy norm

$$\|\nabla R_\ell v_\ell\|_{L^2(\Omega)} \le \|\nabla v_\ell\|_{L^2(\Omega)}.$$

Indeed, we recall $R_\ell = I - \omega_\ell \mathcal{A}_\ell$ and note that the eigenvalues of R_ℓ lie between 0 and 1 for our choice of ω_ℓ. We write

$$\|\nabla R_\ell v_\ell\|_{L^2(\Omega)} = \|\mathcal{A}_\ell^{1/2} R_\ell v_\ell\|_{0,\ell} = \|R_\ell \mathcal{A}_\ell^{1/2} v_\ell\|_{0,\ell} \tag{4.2}$$

and obtain the claimed estimate from Theorem 2.5.

Theorem 4.6 (W-cycle contraction)**.** *For any $\gamma \in (0,1)$ there exists a number $\nu > 0$ of relaxation steps such that the W-cycle multigrid solution $u_L^{(k)} \in V_L$ satisfies*

$$\|\nabla(u_L - u_L^{(k)})\|_{L^2(\Omega)} \le \gamma \|\nabla(u_L - u_L^{(k-1)})\|_{L^2(\Omega)} \le \gamma^k \|\nabla(u_L - u_L^{(0)})\|_{L^2(\Omega)}.$$

Proof. We denote by C the constant from Lemma 4.5 such that $\|\nabla e^{(k)}\|_{L^2(\Omega)} \le C\nu^{-1/2} \|\nabla e^{(k-1)}\|_{L^2(\Omega)}$ for the two-grid iteration. We choose ν as $\nu \ge (C/(\gamma - \gamma^2))^2$ and prove the result by induction over the levels. For $L = 1$ we have the two-grid method where the claim is satisfied for $\nu \ge (C/\gamma)^2$, which is satisfied for our choice of ν. Let now the claimed estimate hold for some $L - 1 \ge 1$. On the level L we assume in view of (4.2) without loss of generality that no post-smoothing is performed. We use the auxiliary two-grid solution

$$\tilde{u}_L^{(k)} = \mathcal{R}_L^\nu u_L^{(k-1)} + z_L^{\mathrm{ex}} \quad \text{with} \quad z_L^{\mathrm{ex}} := \mathcal{A}_{L-1}^{-1} \iota_{L-1\to L}^* r_L$$

for comparison and obtain from the triangle inequality

$$\|\nabla(u_L - u_L^{(k)})\|_{L^2(\Omega)} \le \|\nabla(u_L - \tilde{u}_L^{(k)})\|_{L^2(\Omega)} + \|\nabla(\tilde{u}_L^{(k)} - u_L^{(k)})\|_{L^2(\Omega)}.$$

For the first term on the right-hand side we use Lemma 4.5, which implies

$$\|\nabla(u_L - \tilde{u}_L^{(k)})\|_{L^2(\Omega)} \le C\nu^{-1/2} \|\nabla(u_L - u_L^{(k-1)})\|_{L^2(\Omega)}.$$

The remaining term is the approximation error of the coarse-grid correction, which can be written as

$$\|\nabla(\tilde{u}_L^{(k)} - u_L^{(k)})\|_{L^2(\Omega)} = \|\nabla(z_L^{\mathrm{ex}} - z_L^2)\|_{L^2(\Omega)}$$

where $z_L^2 \in V_{L-1}$ is the multigrid approximation of z_L^{ex} on the level $L-1$ with the W-cycle. From the induction hypothesis we obtain

$$\|\nabla(z_L^{\mathrm{ex}} - z_L^2)\|_{L^2(\Omega)} \leq \gamma^2 \|\nabla z_L^{\mathrm{ex}}\|_{L^2(\Omega)}$$

(recall that the initial guess z_L^0 was chosen to be zero). The γ^2 stems from the fact that the multigrid is run twice in the W-cycle. Since the exact coarse-grid correction is the projected error, as shown in (4.1), and the orthogonal projection is nonexpansive, we have fomr (4.2) that

$$\|\nabla z_L^{\mathrm{ex}}\|_{L^2(\Omega)} = \|\nabla P_{L-1} R^\nu (u_L - u_L^{(k-1)})\|_{L^2(\Omega)} \leq \|\nabla(u_L - u_L^{(k-1)})\|_{L^2(\Omega)}.$$

We combine the above estimates and obtain

$$\|\nabla(u_L - u_L^{(k)})\|_{L^2(\Omega)} \leq (C\nu^{-1/2} + \gamma^2)\|\nabla(u_L - u_L^{(k-1)})\|_{L^2(\Omega)}.$$

Our choice of ν then proves the assertion. ■

Theorem 4.7 (W-cycle error estimate)**.** *We consider multigrid with the W-cycle and initial values computed from nested iteration. If $\nu > 0$ is chosen such that the W-cycle is contractive with $\gamma < 1$ in the sense of Theorem 4.6, then there exists $k_0 > 0$ such that the multigrid solution $u_L^{(k)}$ for $k \geq k_0$ satisfies*

$$\|\nabla(u_L - u_L^{(k)})\|_{L^2(\Omega)} \lesssim h_L \|f\|_{L^2(\Omega)}.$$

The constant hidden in the notation $\lesssim$ is independent of L and f.

Proof. On any level $\ell = 0, \ldots, L$, we denote $e_\ell := u_\ell - u_\ell^{(k)}$ with $e_0 := 0$. From the contraction property from Theorem 4.6 and the choice of the initial value, we obtain

$$\|\nabla e_\ell\|_{L^2(\Omega)} \leq \gamma^k \|\nabla(u_\ell - u_{\ell-1}^{(k)})\|_{L^2(\Omega)}.$$

For the right-hand side, we use the triangle inequality and the a priori error estimate from Corollary 3.33 and deduce

$$\begin{aligned} \|\nabla(u_\ell - u_{\ell-1}^{(k)})\|_{L^2(\Omega)} &\leq \|\nabla(u - u_\ell)\|_{L^2(\Omega)} + \|\nabla(u - u_{\ell-1})\|_{L^2(\Omega)} + \|\nabla e_{\ell-1}\|_{L^2(\Omega)} \\ &\leq C h_\ell \|f\|_{L^2(\Omega)} + \|\nabla e_{\ell-1}\|_{L^2(\Omega)}. \end{aligned}$$

Combining the foregoing bounds yields

$$\|\nabla e_\ell\|_{L^2(\Omega)} \leq \gamma^k (C h_\ell \|f\|_{L^2(\Omega)} + \|\nabla e_{\ell-1}\|_{L^2(\Omega)}).$$

By an induction argument, we therefore infer with $2h_\ell = h_{\ell-1}$ (and thus $h_\ell = 2^{L-\ell} h_L$) that

$$\|\nabla e_L\|_{L^2(\Omega)} \leq C\|f\|_{L^2(\Omega)} \sum_{\ell=1}^{L} h_\ell \gamma^{(1+L-\ell)k} = 2^{-1} C \|f\|_{L^2(\Omega)} h_L \sum_{\ell=1}^{L} (2\gamma^k)^{(1+L-\ell)}.$$

If k_0 is large enough such that $2\gamma^{k_0} < 1$, then the geometric sum is uniformly bounded and the assertion follows. ■

On the level ℓ, the number of degrees of freedom n_ℓ (size of the FEM system to be solved) is proportional to 4^ℓ. On simple triangulations like the FDM grid, this is clear. Generally, it follows from Euler's formula, cf. Exercise A.50. Matrix-vector multiplication with the sparse matrix A_ℓ requires Cn_ℓ operations for some constant C. The parameters k_0, ν in Theorem 4.7 are uniformly bounded. With those numbers being fixed, we therefore have that the number of operations $\mathrm{NOP}(\ell)$ satisfies

$$\mathrm{NOP}(L) \leq Cn_L + 2\,\mathrm{NOP}(\ell-1) \leq C\sum_{\ell=1}^{L} 2^{L-\ell} n_\ell$$

the factor 2 is due to the choice of the W-cycle. Since $n_\ell \lesssim 4^{-(L-\ell)} n_L$, we obtain with a geometric series argument that

$$\mathrm{NOP}(L) \lesssim n_L,$$

i.e., the multigrid method provides an error of the order h_L with linear computational cost.

§3. Discrete Sobolev inequality

We know that taking point evaluations is not bounded with respect to the H^1 norm. But when working with discrete functions, we can quantify the deterioration of the norm in terms of the mesh size.

Theorem 4.8 (discrete Sobolev inequality)**.** *Let T_H be a triangle from $\mathcal{T}_H$ of mesh size H and $v_h \in S^1(\mathcal{T}_h)$ with respect to a refinement $\mathcal{T}_h$ of mesh size h. Then*

$$\|v_h\|_{L^\infty(T_H)} \lesssim H^{-1}\|v_h\|_{L^2(T_H)} + (1+\log(h/H))^{1/2})\|\nabla v_h\|_{H^1(T_H)}.$$

Proof. The proof very closely follows [BS08, §4.9]. We first note that the case of h close to H follows from a simple scaling argument, whence we may focus on the case $h < H/2 =: R_0$. An elementary consideration shows that there is a sector of angle ω and length $R > 0$ with $R \leq H \lesssim R$ denoted in polar coordinates by $K_R = \{(r,\varphi) : 0 < r < R, 0 < \varphi < \omega\}$ such that any $x \in T_H$ satisfies

$$x + QK_H \subseteq T_H$$

for some planar isometry Q. We consider $T_h \in \mathcal{T}_h$ where the maximum of $|v_h|$ is achieved and assume without loss of generality that the barycentre of T_h is the origin 0 and the corresponding Q is the identity. We start by bounding the value $v_h(0)$. There is a positive number $0 < s < 1$ (only depending on the shape regularity) such that the scaled sector K_{sh} is contained in T_h. Integration until any $0 < r < R$ yields

$$v_h(r,\varphi) - v_h(0) = \int_0^r \partial_\rho v_h(\rho,\varphi)d\rho.$$

We use Young's inequality and find

$$\frac{1}{2}|v_h(0)|^2 \leq |v_h(r,\varphi)|^2 + \left|\int_0^r \partial_\rho v_h(\rho,\varphi)d\rho\right|^2.$$

Given any $R_0 < r < R$, we now split the integral into a part that lies inside T_h and a remainder, and thereafter use Hölder's inequality

$$\begin{aligned}\left|\int_0^r \partial_\rho v_h(\rho,\varphi)d\rho\right| &\leq \left|\int_0^{sh} \partial_\rho v_h(\rho,\varphi)d\rho\right| + \left|\int_{sh}^r \partial_\rho v_h(\rho,\varphi)d\rho\right| \\ &\leq sh\|\nabla v_h\|_{L^\infty(T_h)} + \left|\int_{sh}^r (\partial_\rho v_h(\rho,\varphi))^2\rho d\rho\right|^{1/2}\left|\int_{sh}^r \rho^{-1}d\rho\right|^{1/2}.\end{aligned}$$

The last appearing expression equals $\sqrt{\log(r/(sh))} \leq \sqrt{\log(R/(sh))} \lesssim \sqrt{1+\log(H/h)}$. Furthermore, from the shape regularity we obtain $sh\|\nabla v_h\|_{L^\infty(T_h)} \lesssim \|\nabla v_h\|_{L^2(T_h)}$. Therefore combining the above estimates yields with Young's inequality that

$$|v_h(0)|^2 \lesssim |v_h(r,\varphi)|^2 + \|\nabla v_h\|^2_{L^\infty(T_h)} + (1+\log(H/h)) \left| \int_{sh}^R (\partial_\rho v_h(\rho,\varphi))^2 \rho d\rho \right|.$$

We note $H^2 \lesssim \int_0^\omega \int_{R_0}^R r dr d\varphi \lesssim H^2$. Multiplying the displayed formula by r and integrating therefore results in

$$H^2|v_h(0)|^2 \lesssim \|v_h\|^2_{L^2(T_H)} + H^2\|\nabla v_h\|^2_{L^2(T_H)} + (1+\log(H/h))H^2\|\nabla v_h\|^2_{L^2(T_H)}.$$

We have therefore shown

$$|v_h(0)|^2 \lesssim H^{-2}\|v_h\|^2_{L^2(T_H)} + (1+\log(H/h))\|\nabla v_h\|^2_{L^2(T_H)}.$$

Since a scaling argument with the function $\tilde{v}_h(y) = v_h(y) - v_h(0)$ shows $\|\tilde{v}_h\|_{L^\infty(T_h)} \lesssim \|\nabla v_h\|_{L^2(T_h)}$, the triangle inequality shows the assertion of the theorem. ■

§4. Hierarchical multilevel decomposition

We consider a mesh hierarchy $\mathcal{T}_0, \dots, \mathcal{T}_L$ with finite element spaces $V_0, \dots, V_L$ as in ´previous sections. The nodal finite element interpolation to V_ℓ is denoted by $I_\ell : C(\bar{\Omega}) \to V_\ell$. Any function $v_L \in V_L$ is then decomposed as

$$v_L = \sum_{\ell=0}^{L} (I_\ell - I_{\ell-1}) v_L$$

where we set $I_{-1} := 0$. This corresponds to a multilevel decomposition

$$V_L = \bigoplus_{\ell=0}^{L} W_\ell$$

with

$$W_\ell := \{v_\ell - I_{\ell-1}v_\ell : v_\ell \in V_\ell\}.$$

The space W_ℓ is spanned by the hat functions that correspond to interior vertices of $\mathcal{T}_\ell$ that were not in $\mathcal{T}_{\ell-1}$.

We recall the scaling argument from the inverse inequality. By transforming to a reference element, evoking equivalence of norms and using the shape-regularity, we can prove:

Lemma 4.9 (interpolation estimate for 1-level difference)**.** *For $\ell \in \{1, \dots, L\}$, any $v_\ell \in V_\ell$, and any $T \in \mathcal{T}_{\ell-1}$ we have*

$$\|v_\ell - I_{\ell-1}v_\ell\|_{L^2(T)} \lesssim h_\ell \|\nabla v_\ell\|_{L^2(T)}.$$

We prove an L^2 estimate for the single contributions of the above decomposition of v_L.

Lemma 4.10. *Any $v_L \in V_L$ and any $\ell \in \{0, \dots, L\}$ satisfy*

$$\|(I_\ell - I_{\ell-1})v_L\|_{L^2(\Omega)} \lesssim h_\ell(1+\sqrt{L-\ell})\|\nabla v_L\|_{L^2(\Omega)}.$$

Proof. In the case $\ell = 0$, the result is immediately implied by the discrete Sobolev embedding, a scaling argument, and Friedrichs' inequality. We therefore consider the case $\ell > 0$. Let $T \in \mathcal{T}_\ell$ and $c \in \mathbb{R}$ be arbitrary. We denote $v_\ell := (I_\ell v_L - c) \in P_1(T)$. From a scaling argument we obtain

$$\|(I_\ell - I_{\ell-1})v_L\|_{L^2(T)} = \|v_\ell - I_{\ell-1}v_\ell\|_{L^2(T)} \leq h_\ell\|v_\ell\|_{L^\infty(T)} \leq h_\ell\|v_L - c\|_{L^\infty(T)}$$

where we have used in the last step that the maximum of $|v_\ell|$ corresponds to one of the nodal values of $|v_L - c|$ on T . Here, we have abused notation and denoted by $I_{\ell-1}$ the local interpolation on T (note that v_ℓ is only locally defined). The discrete Sobolev inequality from Theorem 4.8 leads to

$$h_\ell\|v_L - c\|_{L^\infty(T)} \lesssim \|v_L - c\|_{L^2(T)} + h_\ell(1 + |\log(h_\ell/h_L)|)^{1/2}\|\nabla v_L\|_{L^2(T)}.$$

For the choice $c = \int_T v_L\,dx/\,|T|$ as the integral mean, the Poincaré inequality (which is proven in any basic course on Sobolev spaces or finite elements) states

$$\|v_L - c\|_{L^2(T)} \lesssim h_\ell\|\nabla v_L\|_{L^2(T)}.$$

Combining the above estimates, summing over all $T \in \mathcal{T}_\ell$ and taking logarithms in the relation $h_L = 2^{L-\ell}h_\ell$ then leads to the assertion.

■

Lemma 4.11 (strengthened Cauchy–Schwarz inequality)**.** *Let* $0 \leq j \leq k \leq L$. *Any* $w_j \in W_j$ *and* $w_k \in W_k$ *satisfy*

$$\int_\Omega \nabla w_j \cdot \nabla w_k\,dx \lesssim 2^{(j-k)/2}\|\nabla w_j\|_{L^2(\Omega)}\|\nabla w_k\|_{L^2(\Omega)}.$$

Proof. We consider a coarse triangle $T \in \mathcal{T}_j$, use integration by parts, and the shape regularity (note that ∇v_j is constant over T) to infer

$$\int_T \nabla w_j \cdot \nabla w_k\,dx = \int_{\partial T}(\partial w_j/\partial\nu_T)w_k\,ds \lesssim h_j^{-1}\|\nabla w_j\|_{L^2(T)}\int_{\partial T}|w_k|\,ds.$$

The boundary integral is an integral of a piecewise affine function. With the shape regularity we obtain

$$\int_{\partial T}|w_k|\,ds \lesssim h_k \sum_{z\in\mathcal{N}_k\cap\partial T}|w_k(z)| \lesssim h_k(h_j/h_k)^{1/2}\left(\sum_{z\in\mathcal{N}_k\cap\partial T}|w_k(z)|^2\right)^{1/2}$$

where we have used the Cauchy–Schwarz inequality for vectors in the last step. We know from Lemma 3.38 that

$$\sqrt{\sum_{z\in\mathcal{N}_k\cap\partial T}|w_k(z)|^2} \lesssim h_k^{-1}\|w_k\|_{L^2(T)}$$

so that, after combining the above estimates, we conclude

$$\int_T \nabla w_j \cdot \nabla w_k\,dx \lesssim h_j^{-1}(h_j/h_k)^{1/2}\|\nabla w_j\|_{L^2(T)}\|w_k\|_{L^2(T)}.$$

We now use that $w_k = w_k - I_{k-1}w_k$ and exclude the trivial case $k = j$. Thus $k \geq 1$ and Lemma 4.9 leads to

$$\int_T \nabla w_j \cdot \nabla w_k\,dx \lesssim (h_k/h_j)^{1/2}\|\nabla w_j\|_{L^2(T)}\|\nabla w_k\|_{L^2(T)}.$$

The relation $h_j = 2^{k-j}h_k$ then leads to the asserted estimate for T. The global estimate follows from splitting the integral in local contributions, using the proven bound, and applying the Cauchy–Schwarz inequality in $\mathbb{R}^{\operatorname{card}\mathcal{T}_j}$. ■

§5. Hierarchical basis preconditioner

We will define a preconditioner for the FEM stiffness matrix A_L with respect to the triangulation $\mathcal{T}_L$. We begin by working with functions and therefore again represent the FEM system by an operator

$$\langle \mathcal{A}_\ell v_\ell, w_\ell \rangle = \int_\Omega \nabla v_\ell \cdot \nabla w_\ell \, dx$$

where the angle brackets represent the duality pairing between V_ℓ and V_ℓ^*. By $\iota_\ell : W_\ell \to V_L$ we denote the embedding. The dual operator $\iota_\ell^* : V_L^* \to W_\ell^*$ is then defined as usual by

$$\langle \iota_\ell^* f, w_\ell \rangle = \langle f, \iota_\ell w_\ell \rangle \quad \text{for any } f \in V_L^*, w_\ell \in W_\ell.$$

On W_ℓ we define the operator $\mathcal{B}_\ell : W_\ell \to W_\ell^*$ by

$$\langle \mathcal{B}_\ell w_\ell, \tilde{w}_\ell \rangle := \sum_{z \in \mathcal{N}_\ell(\Omega) \setminus \mathcal{N}_{\ell-1}(\Omega)} w_\ell(z) \tilde{w}_\ell(z).$$

If a nodal basis is introduced in W_ℓ, this corresponds to the Euclidean inner product of the coefficient vectors. Clearly, $\mathcal{B}_\ell$ is invertible.

Definition 4.12 (hierarchical basis preconditioner)**.** The hierarchical basis preconditioner is defined by

$$\mathcal{B} := \sum_{\ell=0}^{L} \iota_\ell \mathcal{B}_\ell^{-1} \iota_\ell^*.$$

♦

Lemma 4.13. *The operator $\mathcal{B} : V_L^* \to V_L$ is symmetric and positive definite.*

Proof. The operators $\mathcal{B}_\ell$ are symmetric and therefore $\mathcal{B}_\ell^{-1}$ and $\iota_\ell \mathcal{B}_\ell^{-1} \iota_\ell^*$ have the same property, and so has $\mathcal{B}$. Since the $\mathcal{B}_\ell$ are positive definite and the multilevel decomposition of V_L is a direct sum, the operator $\mathcal{B}$ is positive definite as well. ■

We recall Young's inequality for convolutions (using the counting measure). For given sequences $a = (a_j)_j$ and $b = (b_k)_k$ of nonnegative real numbers, it states

$$\|a * b\|_{\ell^2} \le \|a\|_{\ell^2} \|b\|_{\ell^1}.$$

The proof departs from Hölder's inequality (with $p = q = 1/2$)

$$\sum_j \left(\sum_k a_{j-k} b_k \right)^2 = \sum_j \left(\sum_k (a_{j-k} b_k^{1/2}) b_k^{1/2} \right)^2 \le \|b\|_{\ell^1} \sum_j \sum_k (a_{j-k}^2 b_k).$$

Fubini's theorem shows that the sum equals $\|a\|_{\ell^2}^2 \|b\|_{\ell^1}$.

Lemma 4.14. *Let $v_L \in V_L$ have the multilevel decomposition $v_L = \sum_{\ell=0}^L w_\ell$ with $w_\ell \in W_\ell$. Then*

$$\langle \mathcal{A}_L v_L, v_L \rangle \lesssim \sum_{\ell=0}^{L} \langle \mathcal{B}_\ell w_\ell, w_\ell \rangle \lesssim (1 + |\log h_L|^2) \langle \mathcal{A}_L v_L, v_L \rangle.$$

Proof. With the decomposition of v_L and the strengthened Cauchy–Schwarz estimate we deduce

$$\langle \mathcal{A}_L v_L, v_L \rangle \lesssim \sum_j \sum_k 2^{-|j-k|/2} \|\nabla w_k\|_{L^2(\Omega)} \|\nabla w_j\|_{L^2(\Omega)}.$$

We apply the Cauchy–Schwarz inequality for vectors (with respect to j) and obtain

$$\langle \mathcal{A}_L v_L, v_L \rangle \lesssim \left(\sum_j \left(\sum_k 2^{-|j-k|/2} \|\nabla w_k\|_{L^2(\Omega)} \right)^2 \right)^{1/2} \left(\sum_j \|\nabla w_j\|^2_{L^2(\Omega)} \right)^{1/2}.$$

For the term inside the first square-root we use Young's inequality for convolutions and finally obtain with the geometric series (for $1/\sqrt{2}$) that

$$\langle \mathcal{A}_L v_L, v_L \rangle \lesssim \sum_{\ell=0}^{L} \|\nabla w_\ell\|^2_{L^2(\Omega)}.$$

The combination of the norm equivalence from Lemma 3.38 and inverse estimates then shows the first asserted estimate. In order to establish the second claimed estimate, we first combine Lemma 3.38 with Lemma 4.10 and obtain

$$\sum_{\ell=0}^{L} \langle \mathcal{B}_\ell w_\ell, w_\ell \rangle \lesssim \sum_{\ell=0}^{L} h_\ell^{-2} \|w_\ell\|^2_{L^2(\Omega)} \lesssim \sum_{\ell=0}^{L} (1 + \sqrt{L-\ell})^2 \|\nabla v_L\|^2_{L^2(\Omega)} \lesssim L^2 \|\nabla v_L\|^2_{L^2(\Omega)}.$$

Since $L \lesssim 1 + \log h_L$, we have shown

$$\sum_{\ell=0}^{L} \langle \mathcal{B}_\ell w_\ell, w_\ell \rangle \lesssim (1 + |\log h_L|^2) \langle \mathcal{A}_L v_L, v_L \rangle$$

and hence the second asserted bound. ∎

Eventually, we are interested in the eigenvalues of $\mathcal{B}\mathcal{A}_L$. We need the following tool.

Lemma 4.15. *Any* $v_L \in V_L$ *satisfies* $\langle \mathcal{B}^{-1} v_L, v_L \rangle = \sum_{\ell=0}^{L} \langle \mathcal{B}_\ell w_\ell, w_\ell \rangle$.

Proof. We note that the unique multilevel decomposition of $v_L \in V_L$ into contributions $w_\ell \in W_\ell$ satisfies

$$w_\ell = \mathcal{B}_\ell^{-1} \iota_\ell^* \mathcal{B}^{-1} v_L.$$

Indeed, we directly check that $w_\ell \in W_\ell$ and

$$\sum_{\ell=0}^{L} \mathcal{B}_\ell^{-1} \iota_\ell^* \mathcal{B}^{-1} v_L = \sum_{\ell=0}^{L} \iota_\ell \mathcal{B}_\ell^{-1} \iota_\ell^* \mathcal{B}^{-1} v_L = \mathcal{B}\mathcal{B}^{-1} v_L = v_L.$$

We now write

$$\langle \mathcal{B}^{-1} v_L, v_L \rangle = \sum_{\ell=0}^{L} \langle \mathcal{B}^{-1} v_L, w_\ell \rangle.$$

Each contribution of the sum then satisfies

$$\langle \mathcal{B}^{-1} v_L, w_\ell \rangle = \langle \mathcal{B}_\ell \mathcal{B}_\ell^{-1} \iota_\ell^* \mathcal{B}^{-1} v_L, w_\ell \rangle = \langle \mathcal{B}_\ell w_\ell, w_\ell \rangle$$

and the assertion follows. ∎

Theorem 4.16. *The hierarchical basis preconditioner satisfies*

$$\frac{\lambda_{\max}(\mathcal{B}\mathcal{A}_L)}{\lambda_{\min}(\mathcal{B}\mathcal{A}_L)} \lesssim (1 + |\log h_L|^2).$$

Proof. Since $\mathcal{A}_L$ and $\mathcal{B}$ are symmetric and positive definite, we know that the extremal eigenvalues of $\mathcal{B}\mathcal{A}_L$ are characterized as the minimum and the maximum of the Rayleigh quotient

$$\frac{\langle \mathcal{A}_L v_L, v_L \rangle}{\langle \mathcal{B}^{-1} v_L, v_L \rangle} \quad \text{for } 0 \neq v_L \in V_L.$$

By Lemma 4.15, the denominator equals $\sum_{\ell=0}^{L} \langle \mathcal{B}_\ell w_\ell, w_\ell \rangle$ so that Lemma 4.14 implies

$$\frac{1}{1 + |\log h_L|^2} \lesssim \lambda_{\max}(\mathcal{B}\mathcal{A}_L) \leq \lambda_{\max}(\mathcal{B}\mathcal{A}_L) \lesssim 1.$$

This implies the assertion. ■

A. Problems

Exercise A.1. Prove that the Laplacian is represented in polar coordinates (r, φ) as follows

$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \varphi^2}.$$

Exercise A.2. Let the following function be given

$$\Phi(x) = \begin{cases} -\frac{1}{2\pi}\log|x| & \text{if } n = 2 \\ \frac{1}{n(n-2)\alpha(n)}\frac{1}{|x|^{n-2}} & \text{if } n \geq 2. \end{cases}$$

Here, $\alpha(n) \neq 0$ is some real number. Show that $\Delta\Phi(x) = 0$ holds for all $x \in \mathbb{R}^n \setminus \{0\}$.

Exercise A.3. Prove the approximation properties of difference quotients from Lemma 1.7.

Exercise A.4. Show that the discrete problem from Definition 1.8 and the stated matrix-vector system are equivalent.

Exercise A.5 (convergence rates in Hölder norms)**.** Let $k \in \mathbb{N}_0$ and $0 < \alpha \leq 1$ and define the following norm

$$\|v\|_{C^{k,\alpha}(\bar\Omega)} = \|v\|_{C^k(\bar\Omega)} + \max_{|\beta|=k} \sup_{\substack{x,y\in\Omega \\ x\neq y}} \frac{|\partial^\beta v(x) - \partial^\beta v(y)|}{|x-y|^\alpha}.$$

A continuous function v with finite norm $\|v\|_{C^{k,\alpha}(\bar\Omega)}$ is said to be uniformly Hölder continuous of class $C^{k,\alpha}$. Prove that the finite difference method satisfies the following convergence estimate

$$|u - U|_{\infty,\bar\Omega} \leq Ch^\alpha \max_{j=1,2} \|\partial^2_{x_j} u\|_{C^{0,\alpha}(\bar\Omega)}$$

provided $\|u\|_{C^{2,\alpha}(\bar\Omega)} < \infty$.

Hint: Use first-order Taylor expansion with Lagrange form of the remainder.

Exercise A.6. Work out the details in the Taylor expansions for the derivation of the 9-point stencil.

Exercise A.7. Prove that the 9-point stencil satisfies a discrete maximum principle and work out an error estimate for the finite difference error $|u-U|_{\infty,\Omega}$ for the Laplacian on the unit square with homogeneous Dirichlet boundary conditions.

Exercise A.8 (operator norm)**.** Let $\|\cdot\|$ be a norm on $\mathbb{K}^n$.

(a) Prove that the map $\|A\| := \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|Ax\|}{\|x\|}$ is a norm on $\mathbb{K}^{n\times n}$ which is submultiplicative and compatible with the underlying vector norm.

(b) Let $A \in \mathbb{K}^{n\times n}$. Prove that the operator norm related to the Euclidean distance $\|\cdot\|_2$ satisfies

$$\|A\|_2 = \|A^*\|_2.$$

(c) Prove that the Frobenius norm $\|A\|_{\mathrm{F}} := \left(\sum_{j,k=1}^{n} |A_{jk}|^2\right)^{1/2}$ is compatible with the Euclidean norm, but is not the operator norm if $n \geq 2$ (a 2 × 2 counterexample is sufficient).

Exercise A.9 (norms of maximal row or column sum)**.** Prove that the norm of the maximal column sum

$$\|A\|_1 := \max_{1\leq k\leq n} \sum_{j=1}^{n} |A_{jk}|$$

on $\mathbb{R}^{n\times n}$ is the operator norm corresponding to the (vector) ℓ^1 norm. Show that the operator norm corresponding to the ℓ^∞ norm is given by the maximal row sum.

Exercise A.10 (inner product)**.** ¸ We consider the space $C([a,b])$ of continuous functions over the interval $[a,b]$ and a positive real function $\omega \in C([a,b])$ mit $\omega > 0$.

(a) Show that the map $C([a,b])^2 \ni (f,g) \mapsto \langle f,g\rangle_{L^2(a,b),\omega} \in \mathbb{C}$ given by $\langle f,g\rangle_{L^2(a,b),\omega} := \int_a^b f(x)\overline{g(x)}\omega(x)\,dx$ is a scalar product on $C([a,b])$.

(b) Show that $\|f\|_{L^2(a,b),\omega} = \langle f,f\rangle_{L^2(a,b),\omega}^{1/2}$ is a norm on $C([a,b])$.

Exercise A.11. Let V, W be normed linear spaces where W is complete. Prove that the space $L(V,W)$ of bounded linear operators endowed with the operator norm is a complete normed linear space.

Exercise A.12. Let $\Omega = (-1,1)^2 \setminus ([0,1]\times[-1,0])$ be the Γ-shaped (or L-shaped) domain. Let u be given by

$$u(x,y) = (1-x^2)(1-y^2)r^{2/3}\sin\left(\frac{2\varphi}{3}\right).$$

Here, we use polar coordinates $0 < r < 1$ and $0 < \varphi < 3\pi/2$; note that $x = r\cos\varphi$ and $y = r\sin\varphi$.

(a) Prove that u satisfies $-\Delta u = f$ for some $f \in C^0(\bar{\Omega})$ and $u|_{\partial\Omega} = 0$. Compute f.

(b) Prove that u does not possess bounded derivatives and, thus, does not belong to $C^1(\bar{\Omega})$.

Exercise A.13 (condition number of s.p.d. matrices)**.** Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite. Shoow that

(a) $\lambda \in \mathbb{R}$ is an eigenvalue of A if and only if λ^2 is an eigenvalue of A^*A.

(b) $\lambda \in \mathbb{R}$ is an eigenvalue of A if and only if λ^{-1} is an eigenvalue of A^{-1}.

(c) $\|A\|_2 = \lambda_{\max}$ for the largest eigenvalue $\lambda_{\max}$ of A.

(d) $\kappa_2(A) = \lambda_{\max}/\lambda_{\min}$, where $\lambda_{\min}$ is the smallest eigenvalue of A.

Exercise A.14 (stationary iterations)**.** Decide whether the Jacobi or Gauss–Seidel method are convergent for solving $A_jx = b$ for the matrices

$$A_1 = \begin{bmatrix} 2 & -1 & 2 \\ 1 & 2 & -2 \\ 2 & 2 & 2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 5 & 5 & 0 \\ -1 & 5 & 4 \\ 2 & 3 & 8 \end{bmatrix}.$$

Exercise A.15 (convergence of iterative methods for the FDM system)**.** Prove that the Jacobi and the Gauss–Seidel method are convergent for the finite difference system. Prove furthermore that the relaxed Richardson iteration is convergent if the relaxation parameter is chosen $\omega < 1/8$.

Exercise A.16 (Richardson iteration)**.** Let $A \in \mathbb{C}^n$ be a matrix with a pair of nonzero eigenvalues $\lambda, \mu \in \mathbb{C}$ of opposite sign, i.e., $\lambda/|\lambda| = -\mu/|\mu|$. Prove that the relaxed Richardson iteration

$$x_{k+1} = \omega b + (I - \omega A)x_k$$

is divergent for any choice of the parameter ω.

Exercise A.17 (Rayleigh quotient)**.** Let $A, C \in \mathbb{R}^{n \times n}$ be symmetric positive definite matrices. Prove that the smallest resp. largest eigenvalue of CA satisfies

$$\tilde{\lambda}_{\min} = \min_{z \neq 0} \frac{z^* A z}{z^* C^{-1} z} \quad \text{resp.} \quad \tilde{\lambda}_{\max} = \max_{z \neq 0} \frac{z^* A z}{z^* C^{-1} z}.$$

Exercise A.18 (SSOR)**.** Show that the SSOR iteration can be rewritten as

$$x_{k+1} = M x_k + B^{-1} b$$

with the matrices

$$M := (D + \omega R)^{-1}((1 - \omega)D - \omega L)(D + \omega L)^{-1}((1 + \omega)D - \omega R)$$

and

$$B := \frac{1}{\omega(2 - \omega)}(D + \omega L)D^{-1}(D + \omega R).$$

Exercise A.19 (energy minimization; energy norm)**.** Let $A \in \mathbb{R}^{n \times n}$ by symmetric and positive definite and let $b \in \mathbb{R}^n$. We define the following function over $\mathbb{R}^n$

$$f(x) = \frac{1}{2}\langle x, Ax \rangle_2 - \langle x, b \rangle_2 \quad \text{for any } x \in \mathbb{R}^n.$$

(a) Compute the gradient of f.

(b) Show that $x \in \mathbb{R}^n$ satisfies $Ax = b$ if and only if $f(x) = \min_{z \in \mathbb{R}^n} f(z)$.

(c) Show that $\|x\|_A := \langle x, Ax \rangle_2^{1/2}$, $x \in \mathbb{R}^n$, is a norm on $\mathbb{R}^n$.

Exercise A.20 (Chebyshev polynomials)**.** Consider the functions $T_n(x) = \cos(n \arccos x)$ over $[-1, 1]$ for $n \in \mathbb{N}_0$ and prove:

(a) The functions T_n satisfy the recurrence relation $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$ for $n \geq 1$.

(b) The functions $(T_n)_{n \in \mathbb{N}_0}$ are an orthogonal system for the inner product $\langle \cdot, \cdot \rangle_{L^2_\omega(-1,1)}$ with the weight function $\omega(x) = (1 - x^2)^{-1/2}$.

(c) The T_n are identical to the Chebyshev polynomials.

(d) We have $|T_n(x)| \leq 1$, and $|T_n(x)| = 1$ if and only if $x = \cos(k\pi/n)$ for some $k \in \{0, \ldots, n\}$.

(e) The polynomials T_n can be represented as $T_n(x) = \frac{1}{2}\big((x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n\big)$.

Exercise A.21. Prove that a descent method with descent directions $d_0, \ldots, d_{n-1} \in \mathbb{R}^n \setminus \{0\}$ is a Galerkin method with respect to the spaces $W_k := \operatorname{span}\{d_0, \ldots, d_{k-1}\}$ if and only if all descent directions are mutually A-orthogonal.

Exercise A.22 (Kantorovich inequality). Let $A \in \mathbb{R}^{n\times n}$ by symmetric and positive definite with eigenvalues $0 < \lambda_1 \leq \dots \leq \lambda_n$ and $\kappa := \kappa_2(A)$. Prove the following:

(a) Define $\mu := \sqrt{\lambda_1 \lambda_n}$. Then we have for all $n \in \{1, \dots, n\}$ that

$$\kappa^{-1/2} \leq \lambda_j/\mu \leq \kappa^{1/2} \quad \text{und} \quad \lambda_j/\mu + \mu/\lambda_j \leq \kappa^{1/2} + \kappa^{-1/2}.$$

(*Hint:* Monotonicity properties of $z \mapsto z + z^{-1}$)

(b) The matrices $\mu^{-1}A+\mu A^{-1}$ and A have the same eigenvectors. The corresponding eigenvalues are bounded from above by $\kappa^{1/2} + \kappa^{-1/2}$.

(c) Any $x \in \mathbb{R}^n$ satisfies

$$\mu^{-1}\langle x, Ax\rangle_2 + \mu\langle x, A^{-1}x\rangle_2 \leq (\kappa^{1/2} + \kappa^{-1/2})\|x\|_2^2 \quad (\textit{Hint:} \text{ Exercise A.13}).$$

(d) Any $x \in \mathbb{R}^n \setminus \{0\}$ satisfies $\displaystyle\frac{\langle x, Ax\rangle_2 \langle x, A^{-1}x\rangle_2}{\|x\|_2^4} \leq \left(\frac{1}{2}\kappa^{1/2} + \frac{1}{2}\kappa^{-1/2}\right)^2.$

Hint: You might use that $4ab \leq (|a| + |b|)^2$ for any real a, b.

Exercise A.23 (convergence of the gradient method). Let $A \in \mathbb{R}^{n\times n}$ be s.p.d, $b \in \mathbb{R}^n$, and $f(x) = \frac{1}{2}\langle x, Ax\rangle_2 - \langle x, b\rangle_2$. Denote by $x_\star$ the solution to $Ax_\star = b$ and denote by x_k the iterates of the gradient method with $d_k = -\nabla f(x_k)$. Prove the following.

(a) Any $x \in \mathbb{R}^n$ satisfies $f(x) = f(x_\star) + \frac{1}{2}\|x - x_\star\|_A^2$.

(b) We have $f(x_{k+1}) = f(x_k) - \dfrac{1}{2}\dfrac{\|d_k\|_2^4}{\langle d_k, Ad_k\rangle_2}$.

(c) We have $d_k = -A(x_k - x_\star)$ and $\|x_k - x_\star\|_A^2 = \langle d_k, A^{-1}d_k\rangle_2$. Furthermore, the following identities hold

$$\|x_{k+1} - x_\star\|_A^2 = \|x_k - x_\star\|_A^2 - \frac{\|d_k\|_2^4}{\langle d_k, Ad_k\rangle_2}$$

$$\text{and} \quad \|x_{k+1} - x_\star\|_A^2 = \|x_k - x_\star\|_A^2 \left[1 - \frac{\|d_k\|_2^4}{\langle d_k, Ad_k\rangle_2 \langle d_k A^{-1} d_k\rangle_2}\right].$$

(d) Use the Kantorovich Lemma (Exercise A.22) to prove the error bound

$$\|x_k - x_\star\|_A \leq \left(\frac{\kappa_2(A) - 1}{\kappa_2(A) + 1}\right)^k \|x_0 - x_\star\|_A.$$

Exercise A.24. Prove that in the cg method the Krylov spaces satisfy $g_k \in V_k$ and $\dim V_k = k$ as long as $g_k \neq 0$.

Exercise A.25 (eigenvalues of the Dirichlet Laplacian). Prove that all eigenvalues λ and eigenfunctions $u \neq 0$ to the problem

$$-\Delta u = \lambda u \text{ in } \Omega \quad \text{and} \quad u|_{\partial\Omega} = 0$$

for the unit square $\Omega = (0, 1)^2$ are given by

$$u(x) = \sin(j\pi x_1)\sin(k\pi x_2) \quad \text{and} \quad \lambda = (j^2 + k^2)\pi^2 \quad \text{for any } j, k = 1, 2, \dots.$$

(*Hint:* separation of variables)

Exercise A.26 (eigenvalues of the finite difference system)**.** Show that all eigenvalues λ and eigenfunctions U of the finite difference system $Ax = \lambda h^2 x$ over the unit square are given by

$$U_{m,n} = \sin(j\pi mh)\sin(k\pi nh) \quad \text{and} \quad \lambda = 4(\sin^2(\tfrac{1}{2}j\pi h) + \sin^2(\tfrac{1}{2}k\pi h))$$

for any $j, k = 1, 2, \ldots, J-1$ and all interior grid points $x_{m,n} = (mh, nh)$ with $m, n = 1, \ldots, J-1$ and $h = 1/J$.
Hint: You may use the relation $\sin((m+1)y) = 2\cos(j\pi h)\sin(my) - \sin((m-1)y)$ for $y = jh\pi$.

Exercise A.27. Prove that the spectral condition number of the FDM system scales like $O(h^{-2})$.
Hint: Exercise A.26

Exercise A.28. Prove that the SSOR preconditioner satisfies the representation

$$C^{-1} = (2-\omega)^{-1}\left(A + \frac{1}{4\omega}(2-\omega)^2 D + \omega\left(LD^{-1}L^* - \frac{1}{4}D\right)\right).$$

Exercise A.29. Prove Lemma 2.47

Exercise A.30. Prove that the finite difference matrices for grid-size h and $2h$ are related through the prolongation operator as follows $A_{2h} = I^*_{2h\to h} A_h I_{2h\to h}$.

Exercise A.31. Prove, based on the divergence theorem, the formula of integration by parts as well as Green's formula.

Exercise A.32. Let X be a Hilbert space. (a) Prove that the kernel $\ker(F)$ of any continuous linear functional $F \in X^*$ is closed. (b) Prove that the orthogonal projection $P : X \to Y$ to a closed subspace Y is linear and nonexpansive, i.e., $\|P\|_{L(X,Y)} \leq 1$. (c) Prove that the map $J : X \to X^*$ from the Riesz representation theorem is an isometry.

Exercise A.33. Prove the *fundamental lemma of calculus of variations* stated as (a)–(b).

(a) Let the function $g \in C^0(\Omega)$ satisfy $\int_\Omega g\psi\,dx = 0$ for all $\psi \in C_c^\infty(\Omega)$. Then $g = 0$ in Ω.

(b) The assertion of (a) remains valid if $g \in L^1_{\text{loc}}(\Omega)$ (with the same conclusion a.e. in Ω).

(c) Show that the weak derivative is unique (using (b)).

Exercise A.34. Show that the function $v(x) = \log(|\log(|x|)|)$ on the disc $\Omega = \{x \in \mathbb{R}^2 : |x| < 1/\exp(1)\}$ is weakly differentiable but neither bounded nor continuous over Ω. Prove that $\|v\|_{L^2(\Omega)} < \infty$ and $\|\nabla v\|_{L^2(\Omega)} < \infty$. (*Hint:* Polar coordinates.)

Exercise A.35. Show that the notions of classical and weak derivative coincide for continuously differentiable functions.

Exercise A.36. (a) Draw a regular triangulation of the square $(0,1)^2$ with 7 triangles.

(b) Let K, T be triangles that intersect in one point $z = T \cap K$. The point z is vertex to T but not to K. Such point is called a *hanging node*. Draw a picture of this situation and convince yourself that regular triangulations cannot contain any hanging node.

Exercise A.37. (a) Prove the claims from Example 3.8 and Example 3.9.

(b) Draw plots of such piecewise affine function for some examples.

(c) Is the sign function from (3.1) weakly differentiable?

Exercise A.38. Show that the functions $(\varphi_z)_{z\in\mathcal{N}}$ are uniquely defined by (3.3) and that they form as basis of $S^1(\mathcal{T})$ and a partition of unity over $\bar{\Omega}$. Draw the graph of one of the basis functions φ_z on an example triangulation.

Exercise A.39. Let $\mathcal{T}$ be a regular triangulation of $\Omega \subseteq \mathbb{R}^2$ and let $v \in P_1(\mathcal{T})$ be a piecewise affine function. For each interior edge F with adjacent triangles T_+ and T_- (i.e., $F = T_+ \cap T_-$), the jump across F is defined by $[v]_F := v|_{T_+} - v|_{T_-}$. (a) Prove that

$$v \in H^1(\Omega) \iff [v]_F = 0 \quad \text{for all interior edges } F.$$

(b) Show that the finite element space satisfies $S^1(\mathcal{T}) \subseteq H^1(\Omega)$.

Exercise A.40. Show that $\|\cdot\|_{H^1(\Omega)}$ is a norm on $H^1(\Omega)$. Does $\|\nabla\cdot\|_{L^2(\Omega)}$ define a norm on $H^1(\Omega)$ as well?

Exercise A.41. Show that any function $u \in C^1(\bar{\Omega}) \cap C^2(\Omega)$, satisfying $-\Delta u = f$ for $f \in C^0(\bar{\Omega})$ and $u|_{\partial\Omega} = 0$, also satisfies the weak formulation of Poisson's equation.

Exercise A.42. The finite difference grid can be triangulated as displayed in Figure A.1. Prove that in this case the system matrices of FDM and FEM coincide.

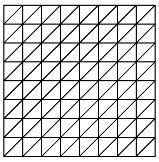


Figure A.1.: Triangulation of Exercise A.42.

Exercise A.43 (barycentric coordinates)**.** Let $T \subseteq \mathbb{R}^2$ be a triangle with vertices z_1, z_2, z_3. Show that to any point $x \in T$ there exist unique real numbers $\lambda_1(x)$, $\lambda_2(x)$, $\lambda_3(x)$ with the properties

$$x = \lambda_1(x)z_1 + \lambda_2(x)z_2 + \lambda_3(x)z_3 \qquad \text{and} \qquad \lambda_1(x) + \lambda_2(x) + \lambda_3(x) = 1.$$

The λ_j are called *barycentric coordinates.* Show furthermore that the barycentric coordinates (as functions of x) coincide with the three nodal basis functions for the vertices of T.

Exercise A.44. *(nodal interpolation not L^2 or H^1 stable)* For a triangle $T \subseteq \mathbb{R}^2$, prove that there is no constant C such that the nodal P_1 interpolation I satisfies

$$\begin{aligned} \|Iu\|_{L^2(T)} &\le C\|u\|_{L^2(T)} \text{ for all } u \in C^\infty(T) \\ \text{or} \quad \|\nabla Iu\|_{L^2(T)} &\le C\|\nabla u\|_{L^2(T)} \text{ for all } u \in C^\infty(T). \end{aligned}$$

Exercise A.45. Let $T \subseteq \mathbb{R}^2$ be a triangle and $v \in H^2(T) := \{w \in H^1(T) : \partial_j w \in H^1(T) \text{ for } j = 1, 2\}$ with norm

$$\|v\|_{H^2(T)} = \sqrt{\sum_{|\alpha|\le 2} \|\partial^\alpha v\|^2_{L^2(T)}}.$$

(a) Consider a sub-triangle $t := \operatorname{conv}\{A, B, C\}$ with $E := \operatorname{conv}\{A, B\}$ and with tangent vector τ. Apply the trace inequality to $f|_E := \nabla v \cdot \tau$ and prove that

$$|v(B) - v(A)| \leq |E|^{1/2} \varrho^{-1/2} 2 \big(1 + \operatorname{diam}(t)^2\big)^{1/2} \|v\|_{H^2(t)}$$

for $\varrho := 2|t|/|E|$.

(b) For any two points A and B in T there exists $C \in T$ such that (with $E := \operatorname{conv}\{A, B\}$ and $t := \operatorname{conv}\{A, B, C\}$), ϱ^{-1} is uniformly bounded by some constant $C(T)$ that depends only on T, but not on A, B, or t.

(c) Conclude that v is Hölder continuous with exponent 1/2.

Remark: This shows the embedding $H^2(T) \hookrightarrow C^{0,1/2}(T)$ on a triangle.

Exercise A.46. Let $f \in L^2(\Omega)$ and recall the energy functional

$$J(v) := \frac{1}{2} \|\nabla v\|^2_{L^2(\Omega)} - \int_\Omega f v \, dx \quad \text{for } v \in H^1_0(\Omega).$$

Prove that the error of the finite element method for the Poisson problem with right-hand side f satisfies

$$\|\nabla(u - u_h)\|^2_{L^2(\Omega)} = 2(J(u_h) - J(u)) = \|\nabla u\|^2_{L^2(\Omega)} - \|\nabla u_h\|^2_{L^2(\Omega)}.$$

Exercise A.47. Let $\mathcal{T}$ be a triangulation. Prove that the aspect ratio of the triangles stays bounded under iterative red refinement.

Exercise A.48. A family of triangulations satisfies the *minimal angle condition* if there is a lower bound $0 < \alpha_0$ to all interior angles of the triangles from that family. Prove that the *minimal angle condition* implies shape regularity.

Exercise A.49. Prove that there exists a constant C only dependent on the shape regularity such that any finite element function $v_h \in S^1(\mathcal{T})$ satisfies

$$\|\nabla v_h\|_{L^2(T)} \leq C h_T^{-1} \|v_h\|_{L^2(T)} \quad \text{for all } T \in \mathcal{T}.$$

This estimate is called *inverse inequality. (Hint: Use transformation to a reference element $\hat{T}$. Use equivalence-of-norms argument in the finite dimensional space $P_1(\hat{T})$ with a constant $C(\hat{T})$ only depending on $\hat{T}$. Afterwards, transform back.)*

Exercise A.50. Let $(\mathcal{T}_\ell)_\ell$ be a sequence of uniform refinements from an initial triangulation $\mathcal{T}_0$ of the convex domain $\Omega \subseteq \mathbb{R}^2$. Let $\mathcal{N}_\ell(\Omega)$ denote the set of interior vertices and let $\mathcal{E}_\ell(\Omega)$ denote the set of interior edges of $\mathcal{T}_\ell$.

(a) Prove the Euler formula $\operatorname{card}(\mathcal{N}_\ell(\Omega)) = 1 + \operatorname{card}(\mathcal{E}_\ell(\Omega)) - \operatorname{card}(\mathcal{T}_\ell)$.

(b) Prove the recurrence relations

$$\operatorname{card}(\mathcal{E}_{\ell+1}(\Omega)) = 2 \operatorname{card}(\mathcal{E}_\ell(\Omega)) + 3 \operatorname{card}(\mathcal{T}_\ell) \quad \text{and} \quad \operatorname{card}(\mathcal{T}_{\ell+1}) = 4 \operatorname{card}(\mathcal{T}_\ell)$$

(c) Combine (a) and (b) to deduce $4^\ell \lesssim \operatorname{card}(\mathcal{N}_\ell(\Omega)) \lesssim 4^\ell$ for $\ell \geq 1$.

B. Programming exercises

Exercise B.1. Install a suitable Python environment on your computer. Use the `NumPy` library to perform the elementary matrix-vector multiplications

$$\begin{bmatrix} 2 & 4 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -3 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Exercise B.2. Have a look at the `scipy.sparse` library, in particular `dia_matrix` and `linalg`. Use these tools to set up the (sparse) system matrix of the finite difference method.

Exercise B.3. Implement the finite difference method for the Poisson problem on the square domain for zero boundary conditions and the right-hand side $f(x) = -e^{x_1}x_1(x_1(x_2^2 - x_2 + 2) + 3x_2^2 - 3x_2 - 2)$. You can use the command `spsolve` for a direct solver for sparse matrices. Use different mesh sizes $h = 2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}$. Compare the computed solution with the exact solution (given by $u(x) = e^{x_1}(x_1 - x_1^2)(x_2 - x_2^2)$) at the grid-points by considering the error in the maximum-norm. Visualize the computed solutions using surface plot tools from Python (see also Problem B.4).

Exercise B.4. Inform yourself about the possibilities of creating surface plots in Python and visualize the finite difference solution from the previous exercise.

Hint: A basic example taken from https://www.geeksforgeeks.org/3d-surface-plotting-in-python-using-matplotlib/

```
# Import libraries
from mpl_toolkits import mplot3d
import numpy as np
import matplotlib.pyplot as plt
# Creating dataset
x = np.outer(np.linspace(-3, 3, 32), np.ones(32))
y = x.copy().T # transpose
z = (np.sin(x **2) + np.cos(y **2) )
# Creating figure
fig = plt.figure(figsize =(14, 9))
ax = plt.axes(projection =’3d’)
# Creating plot
ax.plot_surface(x, y, z)
# show plot
plt.show()
```

Exercise B.5. Find a way to extend the FDM to the L-shaped domain by eliminating points outside $\bar{\Omega}$ from the resulting system. Test the method for the setting from Problem A.12 where the boundary condition is given by the (known) exact solution. Which convergence properties do you observe?

Exercise B.6. Implement the 9-point stencil finite difference method for the Poisson problem on the square domain for zero boundary conditions and the right-hand side from Problem B.3 Use different mesh sizes and compare the computed solution with the exact solution at the grid-points by considering the error in the maximum-norm. Compare the convergence speed with that of the 5-point stencil.

Exercise B.7. Compare (experimentally) the performance (in terms of convergence rates) of the 5-point and the 9-point stencil for the example on the L-shaped domain (see Problems A.12 and B.5). Give a theoretical explanation of what you observe.

Exercise B.8. Implement the Jacobi and the Gauss–Seidel method for sparse matrices. Consider the Dirichlet problem from B.3 and its FDM discretization for different mesh sizes $h = 2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}$. Solve the linear systems with the Jacobi and the Gauss–Seidel method. Plot the maximum error between the reference discrete solution u_h (obtained by a direct solver) and the approximate discrete solution $u_{h,k}$ obtained after k steps of the iterative solver in a semilogarithmic diagram. Do the convergence properties depend on the discretization parameter h?

Exercise B.9. Implement the cg and the pcg method for sparse martices. Check the convergence rate (with respect to the cg iteration) for the finite difference system for different mesh size. Test whether this can be improved if one (relaxed) Jacobi or Gauss-Seidel step is used as preconditioner.

Exercise B.10. Test the convergence propertis of the pcg method with the SSOR preconditioner ($\omega = 1.3$).

Exercise B.11. Start from the example triangulation from Figure 3.1 and plot the interpolation of the function $u(x, y) = \sin(12\pi x)y^2$ on a sequence of 6 red-refined triangulations.

Exercise B.12. Do a convergence study of the FEM for the unit square the right-hand side f given in B.3 with respect to the following error norm

$$\|\nabla(u - u_h)\|_{L^2(\Omega)}.$$

For computing the gradient of u_h on a given element T, use the local representation in terms of the nodal basis. The gradients of the basis vectors were already computed in the loop for the stiffness matrix. Perform an analogous convergence study for the error in the L^2 norm and compare the convergence rates (with respect to the maximal diameter of the triangles in the triangulations, the so-called mesh size). Visualize the results in a loglog-diagram (horizontal axis: mesh size, vertical axis: error in the different norms).

Exercise B.13. (a) Write the data structures for a triangulation of the L-shaped domain $\Omega := (-1, 1)^2 \setminus ([0, 1] \times [-1, 0])$ with Dirichlet boundary $\partial\Omega$.

(b) Plot the convergence history for $-\Delta u = 1$ on the L-shaped domain (cf. Problem A.46; the exact solution satisfies $\|\nabla u\|^2 = 0.2140750232$). Compare the convergence rate with the results on the square domain.

Bibliography

[AB84] O. Axelsson and V. A. Barker. *Finite element solution of boundary value problems. Theory and computation.* Computer Science and Applied Mathematics. Academic Press, Inc., Orlando, FL, 1984.

[Bra07] D. Braess. *Finite Elements. Theory, Fast Solvers, and Applications in Elasticity Theory.* Cambridge University Press, Cambridge, third edition, 2007.

[BS08] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics.* Springer, New York, third edition, 2008.