Finite Elements II

Thomas Richter

January 20, 2020

Contents

1 Finite Elements Techniques

1.1 Finite Elements

In this section,we introduce the basic concepts for describing the finite element method. We start by introducing finite elements for the Laplace equation

$$u \in \mathcal{V} = H_0^1(\Omega): \quad (\nabla u, \nabla \phi)_\Omega = (f, \phi)_\Omega \quad \forall \phi \in \mathcal{V},$$

for a given right hand side $f \in L^2(\Omega)$. The domain $\Omega \subset \mathbb{R}^d$ is two or three dimensional. First, we partition this domain into a triangulation (or mesh) Ω_h, consisting of open elements $K \subset \mathbb{R}^d$. These elements are simple geometric structures like triangles, quadrilaterals or tetrahedra, or mixtures. The boundary of each element K has a finite number of edges $e \subset \partial K$ and nodes $x \in \partial K$. The edges $e \subset \partial K$ are not necessarily straight, see Figure 1.1 for different triangulations. We define

Definition 1 (Structural regularity)**.** *A triangulation $\Omega_h = \{K_1, \dots, K_N\}$ of the domain $\Omega \subset \mathbb{R}^d$ is called* structural regular*, if the elements cover the domain*

$$\bar{\Omega} = \bigcup_{n=1}^{N} \bar{K}_i,$$

and if different elements are disjoint

$$K_i \cap K_j = \emptyset \quad \forall i \neq j,$$

and if the intersection of the closure is either a common node, or a (complete) common edge

$$i \neq j \quad \bar{K}_i \cap \bar{K}_j = \begin{cases} \emptyset & \text{or,} \\ e \in K_i \text{ and } e \in K_j & \text{or,} \\ x \in K_i \text{ and } x \in K_j. \end{cases}$$

It is obvious that a curved domain $\Omega \subset \mathbb{R}^d$ cannot be triangulated with geometric structures like triangles. Instead we must use meshes with curved elements, as shown in Figure 1.1 (right). A common approach for curved finite element meshes is to base every element $K \in \Omega_h$ on one common reference element $\hat{K}$.

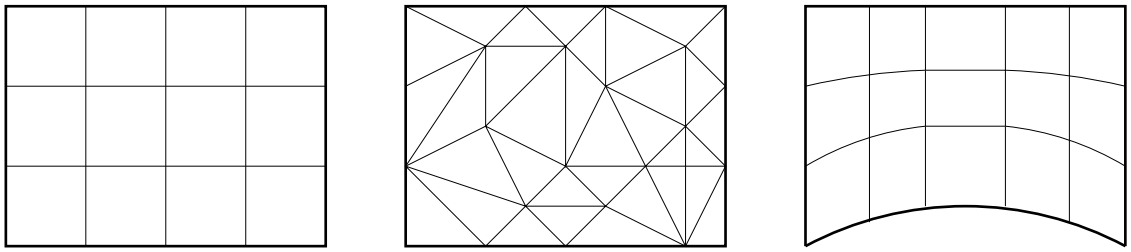


Figure 1.1: Different triangulations Ω_h for a domain $\Omega \subset \mathbb{R}^2$. From left to right: structured quadrilateral mesh, unstructured triangular mesh and structured quadrilateral mesh using curved parametric elements.

Definition 2 (Parametric triangulation)**.** *Let $\hat{K} \subset \mathbb{R}^d$ be the reference element, i.e. an element of size $|\hat{K}| = \mathcal{O}(1)$ and high regularity (e.g. the unit quad, the right-angled unit-triangle). Further, let*

$$\hat{P}(\hat{K}) \subset C(\overline{\hat{K}})^d$$

be a function space mapping to $\mathbb{R}^d$. A triangulation Ω_h is called parametric*, if every element $K \in \Omega_h$ arises from the mapping of the reference element*

$$\forall K \in \Omega_h \text{ there exists a } T_K \in \hat{P}(\hat{K}) \text{ such that } K = T_K(\hat{K}).$$

Usually, for the reference element $\hat{K}$ one chooses the unit triangle, the unit quad or the unit hex. If the space $\hat{P}(\hat{K})$ is the space of affine mappings, the parametric triangulation Ω_h will consist of standard elements with straight edges only. If we consider mappings $\hat{P}(\hat{K})$ of higher polynomial degree (or even rational functions), we can generate curved elements that can be used to approximate domains with curved boundaries. Next, we define conditions on the shape of each element.

Definition 3 (Shape regularity of triangular meshes)**.** *A family of triangular meshes Ω_h, $(h > 0)$ is is called* shape regular*, if there exists a constant $c > 0$, independent on $h > 0$, such that*

$$\max_{K \in \Omega_h} \frac{h_K}{\rho_K} \geqslant c,$$

where by $h_K := \operatorname{diam}(K)$ we denote the diameter of K and by ρ_K the radius of the largest inscribed circle.

Shape regularity of a sequence of meshes describes that all triangles have approximately the same shape. It holds

Lemma 4 (Shape regularity of triangular meshes)**.** *For a family of triangular meshes Ω_h, $(h > 0)$ the following conditions are equivalent*

1. *The family of meshes is shape regular according to Definition 3.*
2. (Minimum angle condition) *There exists a constant $c > 0$ independent of $h > 0$, such that all interior angles α are bound away from zero $\alpha \geqslant c$.*

3. (Maximum angle condition) *There exists a constant* $c > 0$ *independent of* $h > 0$*, such that all interior angles* α *are bound away from* π *by* $\alpha \leqslant \pi - c$*.*

Describing suitable shape regularity conditions for other types of finite element meshes is more complicated. Already for quadrilateral meshes one must combine minimum and maximum angle conditions to prevent that quadrilaterals can degenerate to triangles. Instead we introduce a more general concept of shape regularity that can be applied to all kinds of parametric meshes.

Definition 5 (Shape regularity of parametric meshes)**.** *A family parametric mesh* Ω_h*,* $h > 0$ *with reference element* $\hat{K}$ *is called shape regular, if there exists a constant* $c > 0$*, such that it holds*

$$\frac{1}{c}\|\nabla T_K\| \, \|\nabla T_K^{-1}\| \leqslant c \quad \forall K \in \Omega_h,$$

where $T_K : \hat{K} \to K$ *is the reference map for element* $K \in \Omega_h$ *and with a constant* $c > 0$ *that does not depend on* $h > 0$ *or* $K \in \Omega_h$*.*

This definition of shape regularity is less obvious, it however is directly usable for deriving interpolation estimates. These estimates are usually shown on fixed reference elements $\hat{K}$ and then carried over to a specific $K \in \Omega_h$ by using this reference mapping. For triangular meshes, we can show that this general definition is equivalent to those given in Lemma 4:

Lemma 6. *Let* Ω_h *be a triangular mesh. The condition of Definition 5 is equivalent to those given in Lemma 4.*

Proof. Let $\hat{K} = \{(x, y) \in \mathbb{R}^2_+, \ 0 < x + y < 1\}$ be the reference triangle. Further, Let T_K be an element map, given as (neglecting rotation, translation and isotropic scaling, as all these operations do not effect the shape)

$$T_K(x, y) = \begin{pmatrix} 1 & s \\ 0 & a \end{pmatrix},$$

where $1 : a$ with $a > 0$ indicates the anisotropic aspect ratio and s refers to the shearing. It holds

$$\|\nabla T_K\|_1 \, \|\nabla T_K^{-1}\|_1 = \max\Big\{1, |s| + |a|\Big\} \max\Big\{1, \frac{|s|}{|a|} + \frac{1}{|a|}\Big\},$$

and for this expression to be bounded independent of $h > 0$, it must holds

$$|s|, |a| \leqslant c < \infty, \quad |a| \geqslant \frac{1}{c}.$$

By these limits, the three vertices of the reference triangle get mapped to

$$(0,0) \mapsto (0,0), \quad (0,1) \mapsto (s,a), \quad (1,0) \mapsto (1,0), \quad s \in [-c, c], \ a \in [c^{-1}, c],$$

such that the regularity conditions are fulfilled. □

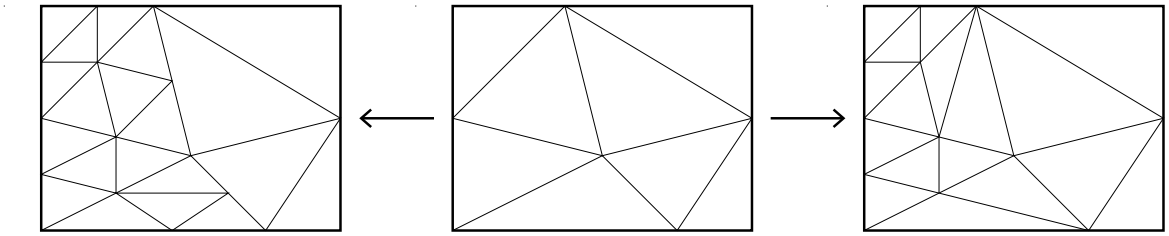


Figure 1.2: Different types of mesh-refinement. Left: refinement of a triangle into three triangles using hanging nodes. Right: using different refinement types such that no hanging nodes appear.

Remark 7 (Boundary approximation)**.** *Most commonly, one considers polynomial spaces for the element maps* T_K*. Such a mapping allows for higher order representation of curved boundaries, see Figure 1.1. A more recent approach is the* Isogeometric Analysis*: the geometrical domains in application problems are usually designed with the help of CAD-programs. These programs use splines (NURBS) for representing the domain. The Isogeometric Analysis uses these splines for defining the finite element mesh and the discrete finite element spaces. By this construction, every geometrical error is eliminated, see Bazilevs et al.* [4]*.*

The quality and the resolution of the finite element mesh will determine the accuracy of the finite element approximation. Constructing a finite element mesh of a domain $\Omega \subset \mathbb{R}^d$ can be a very difficult task. In particular for complex technical structures, a mesh can consist of millions of elements just for resolving the complex geometry. Further, it is possible that this complex mesh will not have the proper resolution at the correct spots to deliver good approximative solutions. In such cases, a mesh must be *refined*. Refining a mesh is either done by *remeshing* and constructing a completely new triangulation with better resolution in certain areas or by *mesh refinement*. Here, elements of a mesh are split into smaller elements. By mesh refinement, we usually must break the structural regularity assumption. In Figure 1.2, we show different procedures for mesh refinement. While the left sketch shows a refinement that does not satisfy the structural-regularity condition, it is simple as every triangle can be split in the same way. The right refinement type yields shape regular meshes, different types of refinement must however be used depending on the refinement topology of a triangle. If we consider quadrilateral meshes one always uses a simple refinement model, where each quad is split into four smaller quads. A shape-regular refinement is only possible, if quadrilateral elements are coupled with triangular elements.

If refinement techniques are used that generate non shape regular meshes, nodes on the middle of edges appear. These nodes will be called *hanging nodes* and must be treated with special care when constructing finite element spaces. We will allow for meshes with one hanging node on an edge only. A mesh without hanging nodes and with element that all have approximately the same size is called a *uniform mesh*. A mesh with areas of mesh refinement is called a *locally refined mesh*.

Locally refined meshes are *unstructured meshes* as apposed to *structured meshes* like meshes

of tensor product type with nodes

$$x_k = (x_r, y_s, z_t) = h \cdot (r, s, t) \quad 0 \leqslant r, s, t \leqslant M.$$

For unstructured meshes, there is no uniform mesh topology. Each node can be part of a different number of elements. Finite element libraries on unstructured meshes call for a dynamic memory layout. Compared to structured finite element approaches, this calls for a large computational overhead. Efficient use of modern hardware with efficient memory usage is troublesome. On the other hand, unstructured meshes allow for a more efficient distribution of the unknowns. The same approximation property can be reached with far smaller problem sizes. Often there is a narrow balance between both approaches. The concept of *mesh grading* sometimes allows for an efficient compromise between fully unstructured meshes and structured meshes. Instead of changing the number of unknowns locally, one only moves the nodal points attached to the degrees of freedom in order to reach adequate local accuracy, see [3].

If the boundaries Γ of the domain $\Omega \subset \mathbb{R}^d$ are not polygonal, a triangulation into simplices will not match the domain

$$\Omega \neq \bigcup_{K \in \Omega_h} K$$

and a geometric error will occur. For better approximation of *curved domains*, the element mapping T_K must not be affine, but is allowed to have a higher degree. Figure 1.1 (right) shows the approximation of a circular domain with a quadrilateral mesh using affine mappings and a quadrilateral mesh using a piecewise biquadratic mapping.

1.1.1 Finite Element Spaces

Finite element spaces V_h are defined locally on the elements $K \in \Omega_h$ of the mesh Ω_h. The most basic finite element space on a triangular mesh is the space of piecewise linear functions

$$V_h = \{\phi \in C(\bar{\Omega}) \mid \phi\big|_K \in \operatorname{span}\{1, x, y\}\ \forall K \in \Omega_h\}.$$

On every triangle $K \in \Omega_h$, the finite element space is locally constructed by three basisfunctions $\phi_K^1, \phi_K^2, \phi_K^3$ such that for the three nodes x_K^1, x_K^2, x_K^3 it holds $\phi_K^j(x_K^i) = \delta_{ij}$. These basis functions are glued together with the basis functions of neighboring elements.

Such a local construction of finite element spaces is difficult on general elements like quadrilaterals that arise from the transformation of a reference element. A more general approach uses parametric finite elements, where the basis is defined on the reference element $\hat{K}$. As reference elements, we consider the reference triangle or quad in two dimensions and the reference tetrahedra or hex in three dimensions. Let

$$\begin{aligned} \hat{P}_r &:= \operatorname{span}\{x^\alpha,\ 0 \leqslant \sum_i \alpha_i \leqslant r \text{ with } 0 \leqslant \alpha_i \leqslant r\}, \\ \hat{Q}_r &:= \operatorname{span}\{x^\alpha,\ 0 \leqslant \alpha_i \leqslant r\}, \end{aligned} \tag{1.1}$$

where $\alpha \in \mathbb{N}^d$ is a multi-index. By $\{\phi_1, \dots, \phi_n\}$ we denote a basis of these polynomial spaces. Besides the usual monomial basis functions $\phi_\alpha = x^\alpha$, we make use of nodal basis functions: let $\hat{x}_i \in \hat{K}$ be uniformly distributed piecewise distinct points in $\hat{K}$. Then, the nodal basis functions $\hat{\phi}_i \in P_r$ (or $\hat{\phi}_i \in Q_r$) are defined by the property

$$\hat{\phi}_i(\hat{x}_j) = \delta_{ij}, \quad i, j = 1, \dots, n.$$

By this notation, we can define the nodal basis functions $\phi_i,\ i = 1, \dots, n$ for every mesh-element $K \in \Omega_h$ by using the domain-map

$$\phi_i^K := \hat{\phi}_i \circ T_K^{-1},\ i = 1, \dots, n.$$

If the element map $T_K : \hat{K} \to K$ is affine, the resulting basis functions are polynomials in the same space as the reference basis. For general polynomial mapping however, the nodal basis usually consists of rational functions. In the mesh-nodes $x_i^K := T_K(\hat{x}_i^K)$ it holds $\phi_i^K(x_j) = \delta_{ij}$. The global finite element space of order r is then given by

$$V_h^r := \{\phi \in C^0(\bar{\Omega}) :\ \phi\big|_K \circ T_K \in P_r \text{ (or } Q_r)\} = \operatorname{span}\{\phi_i,\ i = 1, \dots, N\},$$

where the basis functions ϕ_i are given by gluing the local basis functions ϕ_i^K together. Finite element spaces with global continuity are H^1-*conforming* $V_h \subset \mathcal{V} = H^1(\Omega)$. For some applications, higher regularity is required. If $V_h \subset C^1(\Omega)$, the finite element space will be H^2-conforming, i.e. $V_h \subset H^2(\Omega)$.

In Figure 1.3 we show typical distributions of the nodes for different reference elements and indicate the corresponding polynomial space. The first three columns show Lagrangian elements on triangles and quadrilaterals up to degree 3. The last column shows two H^2-conforming elements, where node values and derivative values are specified.

Sometimes, we do not require H^1-conformity. Considering the Navier-Stokes equations, the pressure must only be discretized with L^2-conforming finite elements that are not necessarily continuous. Here, we introduce discontinuous finite element spaces:

$$V_h^{r,dc} := \{\phi \in L^2(\Omega) \to \mathbb{R} :\ \phi\big|_K \in P_r \text{ (or } Q_r),\ \forall K \in \Omega_h\}.$$

This space is not H^1-conforming, but at least L^2-conforming. Continuous finite element spaces using the element mapping T_K are called *parametric finite elements*. If the element map is a polynomial from the same polynomial space $[P_r]^d$ or $[Q_r]^d$ as the finite element basis, the approach is called *isoparametric* .

Interpolation with finite elements

Interpolation operators $I_h : V \to V$ are the most important tool for finite element error analysis since due to their local construction they allow for a local analysis. The nodal interpolant N_h of a function $u : K \to \mathbb{R}$ is constructed by a simple point-wise process on every element:

$$N_h u(x) = \sum_{i=1}^{L} u(x_i^K)\phi_i^K(x)$$

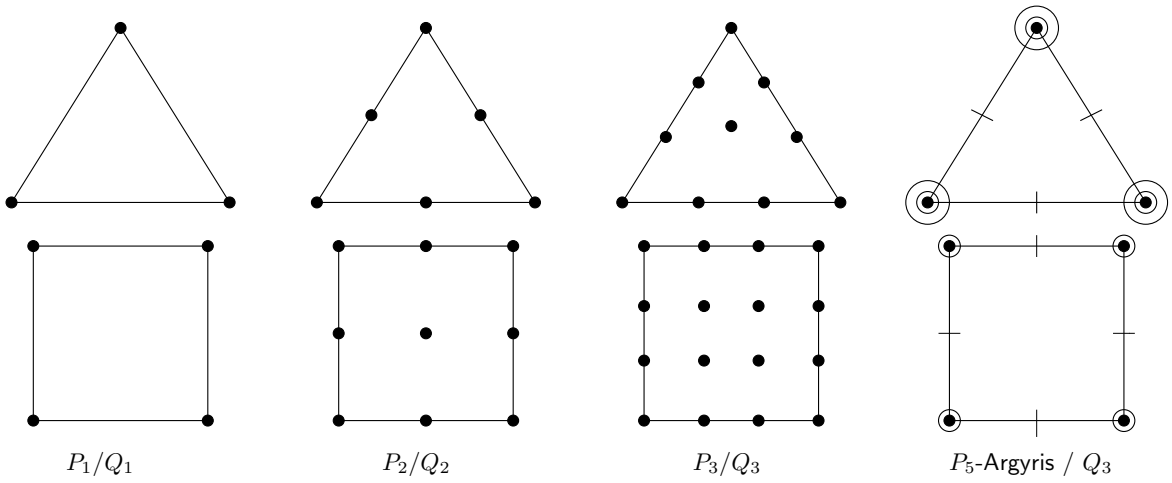


Figure 1.3: Some finite elements. The first three columns show the degrees of freedom in the classical H^1-conforming Lagrange elements. Node-values are equally distributed in the elements. The two elements on the right are H^2-conforming. Besides the node values we indicate the first derivative $\nabla\phi$ (two values each, small circle), the second derivatives $\nabla^2\phi$ (three values each, big circles) and the normal derivatives $\partial_n\phi$ (one value each) in the edge midpoints. Altogether, we specify the $3 \cdot (1+2+3)+3 = 21$ unknowns of the P_5 and $4 \cdot (1+2)+4 = 16$ unknowns of Q_3.

Nodal interpolants are completely local, only information on one element $K \in \Omega_h$ is required. They are however only well-defined for functions $u \in C(\bar{K})$. This is not the case for the Sobolev-space H^1. Here, single point-values must not be finite. If we require interpolants of functions with such minimal regularity, we must replace the point-evaluation $u(x_i^K)$ by some kind of averages.

Lemma 8 (Nodal interpolation on $K \in \Omega_h$)**.** *Let $K \in \Omega_h$ and $T_K : \hat{K} \to K$ and $u \in H^{r+1}(K)$. Then it holds for the nodal interpolation of degree $r \geqslant 1$*

$$\|\nabla^k(u - N_h u)\|_K \leqslant h^{r+1-k}\|\nabla^{r+1}u\|_K, \quad 0 \leqslant k \leqslant r,$$

and, on the boundary of K it holds

$$\|u - N_h u\|_{\partial K} \leqslant h^{r+\frac{1}{2}}\|\nabla^{r+1}u\|_K.$$

Proof. For the proof, we refer to the literature [2]. □

An important result in the context of interpolation is the following *Bramble-Hilbert-Lemma*. Proofs are given in all standard books on finite elements, e.g. [7, 8].

Lemma 9 (Bramble-Hilbert-Lemma)**.** *Let $F(\cdot) : H^m(\hat{K}) \to \mathbb{R}$ be a functional satisfying*

1. *boundness*

$$|F(v)| \leqslant c_1\|v\|_{H^m(\hat{K})} \quad \forall v \in H^m(\hat{K})$$

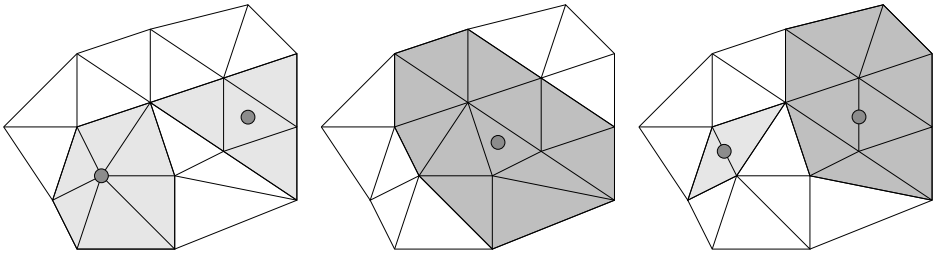


Figure 1.4: *Patches* used in defining the Clement interpolation: Left: a nodal patch P_i and the small element patch P_K. Middle: large element patch $\tilde{P}_K$ and right: small edge patch P_e and large edge patch $\tilde{P}_e$ (dark gray).

2. *sublinearity*
$$|F(u+v)| \leqslant c_2(|F(u)| + |F(v)|) \quad \forall u, v \in H^m(\hat{K})$$

3. *zero on* $P^{m-1}(\hat{K})$
$$F(q) = 0 \quad \forall q \in P^{m-1}(\hat{K})$$

Then it holds
$$|F(v)| \leqslant c(\hat{K})|v|_{H^m(\hat{K})},$$
with the H^m*-seminorm* $|\cdot|_{H^m}$.

The nodal interpolation operator $u \mapsto N_h u$ is only well-defined, if $u \in C(\bar{\Omega})$. Point values of u must be well defined. For H^1-functions, this regularity is not necessarily given, such that a nodal interpolation can never by H^1-stable. For $u \in H^1(\Omega)$, interpolation operators must be defined in terms of averages. The most famous H^1-stable interpolation operator is the *Clement-Interpolation*:

Lemma 10 (Clement-Interpolation)**.** *Let* $u \in H^1(\Omega)$ *and* $V_h \subset H^1(\Omega)$ *be a Lagrangian finite element space with basis* $\phi_i(x_j) = \delta_{ij}$ *on the triangulation* Ω_h*. Then, the Clement-Interpolation* $C_h u \in V_h$ *given as*
$$C_h u = \sum_{i=1}^{n} \chi_i(u)\phi_i, \quad \chi_i(u) := \begin{cases} \frac{1}{|P_i|}\int_{P_{x_i}} u(x)\,dx & x_i \notin \partial\Omega_h \\ 0 & x_i \in \partial\Omega_h \end{cases}$$
with the patches P_i *defined as unions of all elements that touch a node* x_i
$$P_i := \bigcup_{K \in \Omega_h,\, x_i \in \bar{K}} K$$
is H^1*-stable*
$$\|\nabla C_h u\| \leqslant c\|\nabla u\| \quad \forall u \in H^1(\Omega).$$
and satisfies
$$\|\nabla(u - C_h u)\|_K \leqslant ch\|\nabla u\|_{\tilde{P}_K},$$

where $\tilde{P}_K$ *is the larger element patch*

$$\tilde{P}_K := \bigcup_{x_i \in \bar{K}} P_i.$$

Proof. (*i*) We start by showing H^1 stability of the nodal functionals $\chi(u)$ that will replace the simple nodal value extraction $\chi_i^N(u) = u(x_i)$ used in the nodal interpolation.

Let $\hat{P}_x$ be a reference nodal patch satisfying $|\hat{P}_x| = \mathcal{O}(1)$. Each patch $P_i \in \Omega_h$ is given as piecewise affine mapping, affine on each $K \in P_i$, of $\hat{P}_x$. For $\hat{u} \in L^2(\hat{P}_x)$ it holds

$$|\hat{\chi}(\hat{u})| \leqslant \frac{1}{|\hat{P}_x|} \int_{\hat{P}_x} |\hat{u}(\hat{x})| \, d\hat{x} \leqslant \frac{1}{|\hat{P}_x|^{\frac{1}{2}}} \|\hat{u}\|_{\hat{P}_x} \leqslant c \|\hat{u}\|_{\hat{P}_x},$$

where the constant $c > 0$ does not depend on $h > 0$ since we assume shape regularity of the mesh. By T_i we denote the piecewise affine map $T_i : \hat{P}_x \to P_i$. Its gradient ∇T_i is to be understood as an elementwise constant function that satisfies

$$\| \det(\nabla T_i) \|_\infty = \mathcal{O}\left(\frac{|P_i|}{|\hat{P}_x|} \right) = \mathcal{O}\left(h^d \right),$$

where $d = 2$ in the two-dimensional case and $d = 3$ on a three-dimensional tetrahedral mesh.

Now, let $u \in H^1(P_i)$ and $\hat{u}$ be the correnspondingfunction on the reference patch $\hat{P}_x$, e.g. $\hat{u}(\hat{x}) = u(T_i(\hat{x}))$. Then it holds

$$\|u - \chi_i(u)\|^2_{L^2(P_i)} \leqslant c h^d \|\hat{u} - \hat{\chi}(\hat{u})\|^2_{L^2(\hat{P}_x)}.$$

We study $F(\hat{u}) = \hat{u} - \hat{\chi}(\hat{u})$, which is linear and vanishes on the space of constant functions. Applying the Bramble-Hilbert lemma and mapping back to P_i gives

$$\|u - \chi_i(u)\|^2_{L^2(P_i)} \leqslant c c_{bhl} h^d \|\nabla \hat{u}\|^2_{L^2(P_i)} \leqslant c c_{bhl} h^2 \|\nabla u\|^2_{L^2(P_i)}.$$

(*ii*) We define the interpolant as

$$C_h u(x) := \sum_{x_i \in \Omega_h} \chi_i(u) \phi_h^{(i)}(x) \quad \forall u \in \mathcal{V}.$$

The nodal basis functions satisfy

$$\sum_{x_i \in \bar{T}} \phi_h^{(i)}(x) \Big|_T \equiv 1,$$

if we also include the basis functions on points $x_i \in \partial\Omega_h$. This is ok, since we set the nodal functional to zero here, i.e. $\chi_i(u) = 0$ for $x_i \in \Omega_h$. Using $\|\phi_h^{(i)}\|_{L^\infty(K)} \leqslant 1$ it holds

$$\begin{aligned}
\|u - C_h u\|_K &= \left\| u \left(\sum_{x_i \in \bar{K}} \phi_h^{(i)} \right) - \sum_{x_i \in \bar{K}} \chi_i(u) \phi_h^{(i)} \right\|_{L^2(K)} \\
&\leqslant \sum_{x_i \in \bar{K}} \|(u - \chi_i(u)) \phi_h^{(i)}\|_{L^2(K)} \leqslant \sum_{x_i \in \bar{K}} \|u - \chi_i(u)\|_{L^2(P_i)} \leqslant \sum_{x_i \in \bar{K}} c_{bhl} h \|\nabla u\|_{L^2(P_i)} \\
&\leqslant c_{bhl} \sqrt{c(K)} h \|\nabla u\|_{L^2(\tilde{P}_K)},
\end{aligned}$$

where $c(K)$ is the number of elements that overlap within each patch. Given shape regularity of the mesh we estimate $c(K) \leqslant c$ uniform in $h > 0$. The corresponding estimate on the edge is derived by mapping to a reference map and applying the Bramble Hilbert Lemma here. □

Remark 11 (Interpolation on anisotropic meshes). *The Clement interpolation is an H^1-stable operator*

$$\|\nabla C_h u\|_K \leqslant c\|\nabla u\|_{P_K},$$

with a constant $c > 0$ that does not depend on $h > 0$. The Clement operator however fails, if the mesh-elements $K \in \Omega_h$ are anisotropic with $h_{min}(K) \ll h_{max}(K)$. On such elements, it only holds

$$\|\nabla C_h u\|_K \leqslant c\frac{h_{max}(K)}{h_{min}(K)}\|\nabla u\|_{P_K}.$$

An H^1-stable alternative to the Clement operator, which is also stable on anisotropic meshes, is the Scott & Zhang operator. Here, the nodal values are also defined as averages, but averaging is only applied over edges of elements. This helps to avoid mixing of mesh-sizes in different directions. See [16] for basics on the Scott & Zhang interpolation operator, and [2] for an analysis of interpolation operators on anisotropic meshes.

1.1.2 Finite Element Analysis for Elliptic Problems

Let $\mathcal{V} = H_0^1(\Omega)$ and $V_h \subset \mathcal{V}$ be a finite element subspace. By $a(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ we denote an elliptic and continuous bilinear-form, such that there exist constants $c_1, c_2 > 0$ such that

$$a(u,v) \leqslant c_1\|\nabla u\|_\Omega \|\nabla v\|_\Omega, \quad a(u,u) \geqslant c_2\|\nabla u\|^2.$$

This bilinear-form defines a the *energy norm*.

Lemma 12 (Energy norm). *Let $a(\cdot,\cdot)$ be a $\mathcal{V}$-elliptic and continuous bilinear form. Then,*

$$\|u\|_a := \sqrt{a(u,u)},$$

defines a norm that is equivalent to the $\mathcal{V}$-norm.

For $f \in L^2(\Omega)$ we denote by $u \in \mathcal{V}$ and $u_h \in V_h$ solutions to

$$a(u,\phi) = (f,\phi) \quad \forall \phi \in \mathcal{V}, \quad a(u_h,\phi_h) = (f,\phi_h) \quad \forall \phi_h \in V_h. \tag{1.2}$$

It holds

Lemma 13 (Galerkin orthogonality). *For the solution $u \in \mathcal{V}$ and the conforming Galerkin-solution $u_h \in V_h \subset \mathcal{V}$ it holds*

$$a(u - u_h, \phi_h) = 0 \quad \forall \phi_h \in V_h.$$

Proof. This follows as $V_h \subset \mathcal{V}$ allows subtract the two equations in (1.2) and choose $\phi := \phi_h \in \mathcal{V}$. □

Using Galerkin orthogonality we can directly show the following important property:

Lemma 14 (Best approximation, Cea's Lemma)**.** *The conforming finite element approximation is the best approximation in the energy norm* $\|u\|_a := \sqrt{a(u,u)}$

$$\|u - u_h\|_a \leqslant c_1 \min_{\phi_h \in V_h} \|u - \phi_h\|_a$$

and it holds

$$\|\nabla(u - u_h)\| \leqslant \frac{c_1}{c_2} \min_{\phi_h \in V_h} \|\nabla(u - \phi_h)\|.$$

Proof. This follows with of Galerkin orthogonality

$$\begin{aligned} c_2\|\nabla(u - u_h)\|^2 \leqslant \|u - u_h\|_a^2 &= a(u - u_h, u - u_h) \\ &= a(u - u_h, u - \phi_h) \leqslant c_1\|u - u_h\|_a \, \|u - \phi_h\| \end{aligned}$$

□

Using this best-approximation property, and choosing the interpolant $\phi_h := I_h u \in V_h$ we get a first error estimate.

Lemma 15 (Energy norm a priori estimates)**.** *Let* $u \in H^{r+1}(\Omega) \cap \mathcal{V}$ *be the solution to* (1.2)*, and* $u_h \in V_h^r \subset \mathcal{V}$ *be the finite element solution. It holds:*

$$\|\nabla(u - u_h)\| \leqslant c h^r \|\nabla^{r+1} u\|.$$

Proof. This result follows by combining best approximation and interpolation estimates. □

Lemma 16 (L^2-norm a priori error estimate)**.** *Let* $u \in H^{r+1}(\Omega) \cap \mathcal{V}$ *and* $u_h \in V_h^r$ *be solutions to* (1.2)*. Then, it holds*

$$\|u - u_h\| \leqslant c h^{r+1} \|\nabla^{r+1} u\|.$$

Proof. Let $z \in \mathcal{V}$ be the solution to the adjoint problem

$$a(\phi, z) = (f_z, \phi) \quad \forall \phi \in V, \quad f_z := \frac{u - u_h}{\|u - u_h\|}.$$

As the finite element space is conforming it holds that $f_z \in \mathcal{V} \hookrightarrow L^2(\Omega)$. Hence given sufficient solution of the domain elliptic regularity gives $z \in H^2(\Omega) \cap \mathcal{V}$ and

$$\|z\|_{H^2(\Omega)} \leqslant c_s \|f_z\| = c_s.$$

We choose $\phi = u - u_h$ to get by using Galerkin orthogonality:

$$\begin{aligned}\|u - u_h\| &= a(u - u_h, z) = a(u - u_h, z - I_h z)\\ &\leqslant c\|\nabla(u - u_h)\|\,\|\nabla(z - I_h z)\|\\ &\leqslant ch^r\|\nabla^{r+1}u\|\,c_I h\|\nabla^2 z\|\\ &\leqslant ch^{r+1}\|\nabla^{r+1}u\|.\end{aligned}$$

□

Here, we only report on the most basic a priori error estimates, namely the energy norm error and the L^2-error. Estimating the error in a pointwise sense, i.e. in the norm $\|u - u_h\|_{L^\infty(\Omega)}$ is more complex.

Lemma 17 (Maximum norm priori estimates)**.** *Let $u \in C^2(\bar{\Omega}) \cap \mathcal{V}$ and $u_h \in V_h^1$ be solutions to* (1.2)*. Then, it holds*

$$\max_\Omega |u - u_h| \leqslant ch^2\big\{|\ln(h) + 1|\big\} \max_\Omega |\nabla^2 u|$$

Proofs are found in [13, 15, 12].

As $|\ln(h)| \to \infty$ for $h \to \infty$, the convergence rate is slightly less than h^2. The logarithmic term is sharp and it is possible to construct meshes, where this result is numerically validated. For higher order finite elements (starting with quadratic finite elements) one observes (and one can proof) the full order of convergence

$$m \geqslant 2: \quad u_h \in V_h^m \quad \max_\Omega |u - u_h| \leqslant ch^{m+1} \max_\Omega |\nabla^{m+1} u|.$$

1.2 Finite Elements on Curved domains

The standard finite element analysis is strongly depending on the conformity of the Galerkin approach $V_h \subset \mathcal{V}$ which is essential for getting Galerkin-Orthogonality. If the domain Ω is curved and cannot be matched by the finite element mesh $\Omega_h \neq \Omega$, the finite element space will not be conforming. In this section, we shortly discuss the approximation of the Laplace problem

$$u \in H_0^1(\Omega): \quad (\nabla u, \nabla \phi)_\Omega = (f, \phi)_\Omega \quad \forall \phi \in H_0^1(\Omega), \tag{1.3}$$

on a domain $\Omega \subset \mathbb{R}^d$ that is curved and smooth, i.e., the boundary $\partial\Omega$ locally allows for a C^{r+1}-parameterization, with $r \in \mathbb{N}_+$. Finite elements on curved domains must deal with two difficulties:

1. A polygonal mesh will never exactly match the domain Ω. Hence, the discrete equation
$$u_h \in V_h: \quad (\nabla u_h, \nabla \phi_h)_{\Omega_h} = (\tilde{f}, \phi_h)_{\Omega_h} \quad \forall \phi_h \in V_h,$$
is given on a different domain. The right hand side f must not even be defined on all of Ω_h, which is the case, if the domain Ω has concave boundary parts, where Ω_h might

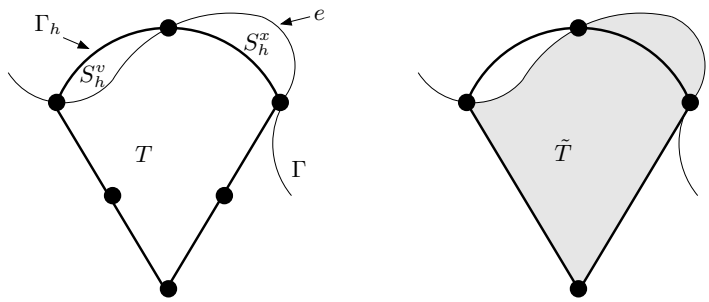


Figure 1.5: Left: geometric remainders for curved boundary approximation. Definition of the mesh snippets $S_h^v = \Omega_h \setminus \Omega$ and $S_h^x = \Omega \setminus \Omega_h$. Right: Definition of the curved extended element $\tilde{T}$ fitting the domain Ω. Exemplarily for quadratic isoparametric elements.

reach out. For this reason, we denoted a modified (extended) right hand side by $\tilde{f}$. For details, we refer to Remark 22.

2. The boundary conditions cannot be exactly satisfied. We consider homogenous Dirichlet conditions only. While $u \in H_0^1(\Omega)$ is zero on all of $\partial\Omega$, $u_h \in V_h$ is zero in the boundary nodes on $\partial\Omega$ but otherwise, it is zero on $\partial\Omega_h \neq \partial\Omega$.

Finite element analysis on curved domains is discussed in literature [7]. General proofs for isoparametric finite elements on curved domains, including optimal order a priori error bounds for the energy error are given in [11].

To cope with the two problems mentioned above, we will start by stating some definitions and lemma. Parts of the boundary can be convex or concave. We define the remainders by

$$S_h^x = \Omega \setminus \Omega_h, \quad S_h^v = \Omega_h \setminus \Omega, \quad S_h = S_h^x \cup S_h^v. \tag{1.4}$$

For a parametric triangulation Ω_h of Ω, see Definition 2, it holds

Lemma 18 (Isoparametric triangulation of Curved domains)**.** *Let $\Omega \subset \mathbb{R}^d$ be a domain with smooth boundary allowing for a C^{r+1}-parameterization with $r \geqslant 1$. Let Ω_h be an isoparametric mesh of Ω with polynomial degree r. For the area of the mesh snippets S_h^x, S_h^v, S_h it holds*

$$|S_h^x| = |S_h^v| = |S_h| = O(h^{r+1}).$$

Proof. This follows by simple geometrical arguments. Let $T \in \Omega_h$ be an element at the boundary and S be that part of S_h which is connected to the element T, see Figure 1.5. Further, let $e \in \partial T$ be the (curved) edge at the boundary Γ_h, which is a $d-1$-dimensional manifold in $\mathbb{R}^d$ with area $|e| = O(h^{d-1})$. Assume that $\psi : e \to \mathbb{R}$ is the parameterization of $\partial\Omega$ over e (see again Figure 1.5). $\psi(s)$ has $r+1$ zero's along the edge in 2d. Hence,

$$\max_{[0,h]} |\psi| \leqslant c h^{r+1} \max_{[0,h]} |\psi^{r+1}|.$$

Therefore, as $|e| = O(h^{d-1})$, it holds

$$|S| = O(h^{r+d}) \quad \Rightarrow \quad |S_h| = O(h^{r+1}).$$

□

The previous lemma shows that standard finite elements will always suffer from a geometrical error. By the use of Isogeometric analysis [10] this error could be completely avoided for domains that can be described by splines.

Another technical difficulty is given by the mismatch of Ω and Ω_h. Functions $u \in H_0^1(\Omega)$ and $u_h \in V_h$ are defined on different domains, such that the expression $u - u_h$ must be discussed. The following lemma will show a way to give $u_h \in V_h$ a meaning both on Ω_h and on Ω.

Lemma 19 (Boundary extension of discrete functions)**.** *Under the assumptions of Lemma 18, let $h \leqslant h_0 \in \mathbb{R}$ and $T \in \Omega_h$ be an element at the boundary $\partial\Omega$ with boundary edge $e \in \partial T$. By $\tilde{T}$ we denote the curved triangle fitting the domain's boundary, see Figure 1.5. For $u_h \in V_h$ we define by $\tilde{u}_h|_T$ the polynomial extension of $u_h|_T$ to $\tilde{T}$. It holds*

$$c_1 \|u_h\|_{H^s(T)} \leqslant \|\tilde{u}_h\|_{H^s(\tilde{T})} \leqslant c_2 \|u_h\|_{H^s(T)}, \ s = 0, 1, 2,$$

with two constants $c_1, c_2 > 0$ that do not depend on T or h.

Proof. This follows by considering equivalence of (discrete) norms and the negligible size of the remainders.

$$|T| = |\tilde{T}| = O(h^d), \quad |(T \setminus \tilde{T}) \cup (\tilde{T} \setminus T)| = O(h^{r+d}).$$

□

In the following, we will always use the notation u_h even on $\tilde{T}$.

While $u_h \in V_h$ is well-defined on Ω_h (including S_h^v) and can be extended to Ω including S_h^x, functions $u \in H_0^1(\Omega)$ are only well-defined on Ω including S_h^x. An extension to the concave part S_h^v might fail due to limited regularity. For the analysis, we need one further - trace inequality-like - estimate:

Lemma 20 (Geometric boundary error)**.** *Let $u \in H_0^1(\Omega)$. There exists a constant $c > 0$, such that for the convex remainder S_h^x, it holds*

$$\|u\|_{S_h^x} \leqslant c h^{\frac{r+1}{2}} \|u\|_{H^1(\Omega)}.$$

Further, let $u_h \in V_h$. It holds

$$\|u_h\|_{H^s(S_h)} \leqslant c h^{\frac{r}{2}} \|u_h\|_{H^s(\Omega)}, \ s = 0, 1.$$

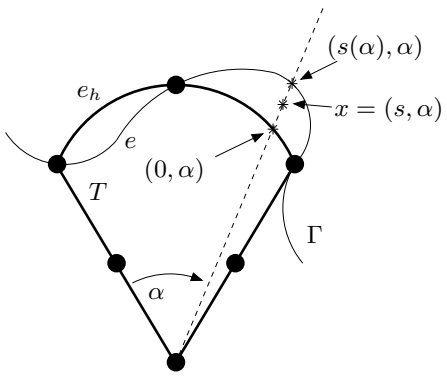


Figure 1.6: Local coordinate system on curved elements. Sketch for the proofs of Lemma 20 and 21. The boundary Γ with segment $e \subset \Gamma$ is given as parameterization of Γ_h with segments e_h, i.e. $s_T : e_h \to e$.

Proof. For the proof, we refer to Figure 1.6. Let $T \in \Omega_h$ be an element on the boundary, $e_h \in \partial T$ be the edge of the element, $\tilde{T}$ the extended element and $e \in \tilde{T}$ be the edge at the boundary $\partial\Omega$. By S we denote the remainder between T and $\tilde{T}$.

(*i*) Let $x \in S$ be given as $x = (s, \alpha)$, where α is the angle and s the radial coordinate, see Figure 1.6. The local coordinate system is such, that $(0, \alpha) \in e_h \subset \Gamma_h$ is a point on the boundary of the (curved) triangle and $(s(\alpha), \alpha)$ is the corresponding point on the domain's boundary part $e \subset \Gamma$. It holds $|s(\alpha)| = O(h^{r+1})$, compare Lemma 18. Let $u \in C^1(\bar{S})$. It holds

$$u(s,\alpha) = u(0,\alpha) + \int_0^s \partial_r u(t,\alpha)\,dt,$$

and hence

$$|u(s,\alpha)|^2 \leqslant c\left(|u(0,\alpha)|^2 + |s|\int_0^s |\partial_r u(t,\alpha)|^2\,dt\right).$$

Integration over S (in s and α) and noting that $|s| \leqslant |s(\alpha)| \leqslant ch^{r+1}$ gives

$$\|u\|_S^2 \leqslant ch^{r+1}\|u\|_e^2 + h^{2r+1}\|\nabla u\|_S^2. \tag{1.5}$$

(*ii*) To proof the first estimate, we continue with (1.5) by summing over all boundary elements, using trace inequality and Poincaré and extending S_h^x to Ω

$$\|u\|_{S_h} \leqslant ch^{\frac{r+1}{2}}\|\nabla u\|_\Omega.$$

(*iii*) For the second inequality, we apply the local trace inequality and extend from S to T

$$\|u_h\|_S^2 \leqslant ch^{r+1}\left(h^{-1}\|u_h\|_T^2 + h\|\nabla u_h\|_T^2\right) + h^{2r+1}\|\nabla u_h\|_T^2.$$

Using the inverse inequality, we get

$$\|u_h\|_S^2 \leqslant ch^r \|u_h\|_T^2,$$

such that the result follows by summing over all boundary snippets. This argumentation is also valid for ∇u_h. □

Discrete functions $\phi_h \in V_h$ are not zero on $\partial\Omega$ but zero on $\partial\Omega_h$.

Lemma 21 (Curved boundary error)**.** *Let $\phi_h \in V_h$ be arbitrary. It holds*

$$\|\phi_h\|_{\partial\Omega} \leqslant ch^{r+\frac{1}{2}} \|\nabla\phi_h\|_\Omega$$

Proof. We again refer to Figure 1.6. Let $T \in \Omega_h$ and $(s(\alpha), \alpha) \in e$ be a point on the boundary of $\partial\Omega$. By $(0, \alpha) \in e_h \subset \partial T$ we denote the corresponding point on the boundary of the triangle. It holds for $\phi_h \in V_h$

$$\phi_h(s(\alpha), \alpha) = \phi_h(0, \alpha) + \int_0^{s(\alpha)} \partial_r \phi_h(t, \alpha)\, dt,$$

and hence by squaring and integrating over α and by noting that $|s(\alpha)| = O(h^{r+1})$ we get

$$\|\phi_h\|_e^2 \leqslant \|\phi_h\|_{e_h}^2 + ch^{r+1} \|\nabla\phi_h\|_S^2. \tag{1.6}$$

With Lemma 20 and using $\phi_h = 0$ on e_h, gives

$$\|\phi_h\|_e^2 \leqslant ch^{2r+1} \|\nabla\phi_h\|_\Omega^2,$$

such that the result follows by summing over all boundary parts. □

Remark 22 (Extension of the right hand side at concave domain boundaries)**.** *As discussed in the beginning of this section, problems might already arise with the definition of the right hand side $f : \Omega \to \mathbb{R}$, which is not necessarily well-defined on the discrete domain Ω_h. This issue is easily handled by defining a projection or interpolation $f_h \in V_h$ to be used as discrete right hand side:*

$$(f_h, \phi_h)_\Omega = (f, \phi_h)_\Omega \quad \forall \phi_h \in V_h.$$

An additional error of type

$$(f - f_h, \phi)_\Omega \leqslant c\|f - f_h\|_{H^{-1}(\Omega)} \|\nabla\phi\|_\Omega,$$

will arise. By exploiting the weak norm and orthogonality of $f - f_h$ such estimates can be given with optimal order and without requiring additional regularity of $f \in H^{r-1}(\Omega)$:

$$\begin{aligned}
\|f - f_h\|_{H^{-1}(\Omega)} &= \sup_{\phi \in H_0^1(\Omega)} \frac{(f - f_h, \phi)_\Omega}{\|\nabla\phi\|} \\
&= \sup_{\phi \in H_0^1(\Omega)} \frac{(f - f_h, \phi - \bar{\phi})_\Omega}{\|\nabla\phi\|} \\
&\leqslant ch^r \|\nabla^{r-1} f\|_\Omega.
\end{aligned}$$

To shorten the proof of the following lemma we will not give details on this issue and just consider f as a well-defined right hand side function.

With these preparations, we can show the following essential theorem, that gives the a priori error estimate for the Laplace equation on smooth and curved domains:

Theorem 23 (A priori error on curved domains)**.** *Let $r \in \mathbb{N}_+$. Let Ω be a domain with boundary that allows for parametrization of degree $r+1$. Let $f \in H^{r-1}(\Omega) \cap L^2(\Omega)$. Let $u_h \in V_h$ be the isoparametric finite element solution of degree r. It holds*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant ch^r \|f\|_{H^{r-1}(\Omega)}$$

and

$$\|u - u_h\| \leqslant ch^{r+1} \|f\|_{H^{r-1}(\Omega)}.$$

Proof. (i) We start with the H^1 error estimate and derive a modified Galerkin orthogonality. For $\phi_h \in V_h$ it holds (where we use the extension $\tilde{\phi}_h \cong \phi_h$ defined by Lemma 19 without further notice)

$$(f, \phi_h)_\Omega = (-\Delta u, \phi_h)_\Omega = (\nabla u, \nabla \phi_h)_\Omega - \langle \partial_n u, \phi_h \rangle_{\partial\Omega}.$$

The discrete problem is defined on Ω_h with $\Omega_h = \Omega \cup S_h^\nu \setminus S_h^x$. It holds

$$(f, \phi_h)_\Omega + (f, \phi_h)_{S_h^\nu} - (f, \phi_h)_{S_h^x} = (\nabla u_h, \nabla \phi_h)_\Omega + (\nabla u_h, \nabla \phi_h)_{S_h^\nu} - (\nabla u_h, \nabla \phi_h)_{S_h^x}.$$

Then, for the finite element error $e_h = u - u_h$, we get the following disturbed Galerkin orthogonality:

$$\begin{aligned}(\nabla e_h, \nabla \phi_h)_\Omega = &-(f, \phi_h)_{S_h^\nu} + (f, \phi_h)_{S_h^x} \\ &+ \langle \partial_n u, \phi_h \rangle_{\partial\Omega} + (\nabla u_h, \nabla \phi_h)_{S_h^\nu} - (\nabla u_h, \nabla \phi_h)_{S_h^x}.\end{aligned} \tag{1.7}$$

(ii) Now, we can estimate the energy error by picking $\phi_h = I_h u - u_h$:

$$\begin{aligned}\|\nabla e_h\|_\Omega^2 \leqslant &\|\nabla e_h\|_\Omega \|\nabla(u - I_h u)\|_\Omega + \|f\|_S \|I_h u - u_h\|_S \\ &+ \|\nabla u_h\|_S \|\nabla(I_h u - u_h)\|_S + \|\partial_n u\|_{\partial\Omega} \|I_h u - u_h\|_{\partial\Omega},\end{aligned} \tag{1.8}$$

where we enlarged S_h^x and S_h^ν to S. The single terms can be estimates with help of Lemma 20 and 21 and the standard interpolation estimate. Exemplarily we discuss the boundary term. With Lemma 21

$$\begin{aligned}\|\partial_n u\|_{\partial\Omega} \|I_h u - u_h\|_{\partial\Omega} \leqslant &c\|u\|_{H^2(\Omega)} ch^{\frac{r+1}{2}} \|\nabla(I_h u - u_h)\|_\Omega \\ &c\|u\|_{H^2(\Omega)} h^{\frac{r+1}{2}} \left(\|\nabla(u - I_h u)\|_\Omega + \|\nabla(u - u_h)\|_\Omega\right).\end{aligned}$$

The remaining terms can be handled in a similar fashion, such that combination with Young's inequality gives the final estimate.

(iii) For estimating the L^2-error, we define the adjoint problem:

$$-\Delta z = \frac{e_h}{\|e_h\|} \text{ on } \Omega \text{ with } z = 0 \text{ on } \partial\Omega,$$

such that

$$\|z\|_{H^2(\Omega)} \leqslant c_s.$$

Multiplication with e_h and integration over Ω yields

$$\|e_h\|_\Omega = (e_h, -\Delta z)_\Omega = (\nabla e_h, \nabla z)_\Omega + \langle u_h, \partial_n z\rangle_{\partial\Omega},$$

as $u = 0$ on $\partial\Omega$. Using (1.7) with $\phi_h = I_h z$, it follows

$$\begin{aligned}\|e_h\| \leqslant \|\nabla e_h\|_\Omega \|\nabla(z - I_h z)\|_\Omega + \|u_h\|_{\partial\Omega}\|\partial_n z\|_{\partial\Omega} + \|\partial_n u\|_{\partial\Omega}\|I_h z\|_{\partial\Omega} \\ + \|f\|_S \|I_h z\|_S + \|\nabla u_h\|_S \|\nabla I_h z\|_S.\end{aligned} \tag{1.9}$$

The first term can be estimated with help of the energy estimate and the interpolation estimates, followed by the stability of the adjoint solution $\|z\|_{H^2(\Omega)} \leqslant c_s$. For the second term, we first use (1.6) and get by introducing $\pm u$

$$\begin{aligned}\|u_h\|_{\partial\Omega} \leqslant ch^{\frac{r+1}{2}} \|\nabla u_h\|_s \leqslant ch^{\frac{r+1}{2}} \left(\|\nabla e_h\|_S + \|\nabla u\|_S\right) \\ \leqslant ch^{\frac{r+1}{2}} \|\nabla e_h\| + ch^{r+1}\|u\|_{H^2(\Omega)}.\end{aligned}$$

This procedure will also be used for the third term. The right hand side part in the fourth term of (1.9) is estimated with Lemma 20

$$\|f\|_S \leqslant ch^{\frac{r+1}{2}} \|f\|_{H^1(\Omega)}.$$

For the interpolation part $\|I_h z\|$ we first use the intermediate result (1.5) from the proof of Lemma 20 and introduce $\pm z$ on the boundary to get with interpolation estimates

$$\begin{aligned}\|I_h z\|_S \leqslant ch^{\frac{r+1}{2}} \|z - I_h z\|_{\partial\Omega} + ch^{\frac{r+1}{2}} \underbrace{\|z\|_{\partial\Omega}}_{=0} + ch^{r+\frac{1}{2}} \|\nabla I_h z\|_S \\ \leqslant ch^{2+\frac{r}{2}} \|z\|_{H^2(\Omega)} + ch^{r+\frac{1}{2}} \|z\|_{H^1(\Omega)}.\end{aligned}$$

Overall, the fourth term in (1.9) is estimated as

$$\|f\|_S \|I_h z\|_S \leqslant ch^{r+\frac{3}{2}} \|f\|_{H^1(\Omega)}.$$

This trick is also used in the final term of (1.9). As $(r+1)/2 \leqslant r + 1/2$

$$\begin{aligned}\|\nabla I_h z\|_S \leqslant ch^{\frac{r+1}{2}} \|\nabla I_h z\|_{\partial\Omega} + ch^{r+\frac{1}{2}} \|\nabla^2 I_h z\|_{S_h} \\ \leqslant ch^{\frac{r+1}{2}} \left(\|\nabla(z - I_h z)\|_{\partial\Omega} + \|\nabla^2(z - I_h z)\|_\Omega + \|z\|_{H^2(\Omega)}\right)\end{aligned}$$

The same estimate is applied to ∇u_h

$$\|\nabla u_h\|_S \leqslant ch^{\frac{r+1}{2}} \left(\|\nabla(u - u_h)\|_{\partial\Omega} + \|\nabla^2(u - u_h)\|_\Omega + \|u\|_{H^2(\Omega)}\right).$$

Together with the stability estimates of the interpolation and higher order estimates of the discrete solution (that can be shown by introducing $\pm I_h u$ and applying the inverse estimate to the discrete parts) we get

$$\|\nabla u_h\|_S \|\nabla I_h z\|_S \leqslant ch^{r+1} \|f\|_{L^2(\Omega)}.$$

□

2 Solution Techniques

Das Bestimmen der Finite Elemente Approximation $u_h \in V_h$ einer linearen Randwertaufgabe erfordert das Lösen eines linearen Gleichungssystems

$$\mathbf{A}_h \mathbf{u}_h = \mathbf{b}_h.$$

In diesem Kapitel befassen wir uns mit iterativen Approximationsverfahren für große lineare Gleichungssysteme, welche von der Finite Elemente Diskretisierung einer partiellen Differentialgleichung stammen. Wir beschränken uns dabei wieder im Wesentlichen auf das Dirichlet-Problem des Laplace-Operators:

$$-\Delta u = f \quad \text{in } \Omega \subset \mathbb{R}^2, \quad u = 0 \text{ auf } \partial\Omega.$$

Bei Bedarf gehen wir auf andere, z.B. nicht-symmetrische Gleichungen, oder auch die (einfache) Erweiterung auf dreidimensionale Probleme ein.

Bei der konformen und konsistenten Finite Elemente Methode erbt die Matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ wichtige Eigenschaften des Differentialoperators, d.h., falls z.B. die Randwertaufgabe positiv definit ist, so ist es auch die Matrix, falls sie symmetrisch ist, so ist auch die Matrix symmetrisch. Üblicherweise ist $N \gg 1000$ sehr groß. Daher scheiden direkte Verfahren, wie die LR- oder Cholesky-Zerlegung als Lösungsmethoden aus. Nach Konstruktion der Finite Elemente Basis ist die Matrix aber dünn besetzt, je nach Ansatzgrad, Dimension und Gleichung (d.h. Größe des Differnetialgleichungssystems) zwischen 5 und über 100 Einträgen pro Zeile. Üblicherweise, insbesondere bei Verwendung von allgemeinen Gittern, hat die Matrix keine Bandstruktur sondern die Struktur ändert sich von Zeile zu Zeile.

Durch numerische Quadratur treten zwangsläufig Rundungsfehler auf. Beim Lösen des linearen Gleichungssystems werden diese Fehler verstärkt. Es gilt die folgende Abschätzung.

Lemma 24 (Störungssatz)**.** *Let $\delta\mathbf{A}_h$ and $\delta\mathbf{b}_h$ be distortions of system matrix and right hand side satisfying*

$$\mu := \operatorname{cond}_2(\mathbf{A}_h)\frac{\|\delta\mathbf{A}_h\|}{\|\mathbf{A}_h\|} < 1, \quad \|\delta\mathbf{A}_h\| < \|\mathbf{A}_h^{-1}\|^{-1}.$$

Then, for the distorted solution $\tilde{\mathbf{u}}_h = \mathbf{u}_h + \delta\mathbf{u}_h$ to $(\mathbf{A}_h + \delta\mathbf{A}_h)\tilde{\mathbf{u}}_h = \mathbf{b}_h + \delta\mathbf{b}_h$ it holds

$$\frac{\|\delta\mathbf{u}_h\|}{\|\mathbf{u}_h\|} \leqslant \frac{\operatorname{cond}_2(\mathbf{A}_h)}{1-\mu}\left\{\frac{\|\delta\mathbf{A}_h\|}{\|\mathbf{A}_h\|} + \frac{\|\delta\mathbf{b}_h\|}{\|\mathbf{b}_h\|}\right\}.$$

Proof. (*i*) We start with the simple case and consider an error in the right hand side only $\tilde{\mathbf{b}}_h = \mathbf{b}_h + \delta\mathbf{b}_h$. It holds

$$\mathbf{A}_h\mathbf{u}_h = \mathbf{b}_h, \quad \mathbf{A}_h\hat{\mathbf{u}}_h = \tilde{\mathbf{b}}_h \quad \Rightarrow \quad \mathbf{u}_h - \hat{\mathbf{u}}_h = \mathbf{A}_h^{-1}\delta\mathbf{b}_h$$

and hence with $\mathbf{A}\mathbf{u}_h = \mathbf{b}_h$

$$\frac{\|\mathbf{u}_h - \hat{\mathbf{u}}_h\|}{\|\mathbf{A}_h\|\,\|\mathbf{u}_h\|} \leqslant \frac{\|\mathbf{u}_h - \hat{\mathbf{u}}_h\|}{\|\mathbf{A}_h\mathbf{u}_h\|} = \frac{\|\mathbf{u}_h - \hat{\mathbf{u}}_h\|}{\|\mathbf{b}_h\|} \leqslant \|\mathbf{A}_h^{-1}\|\frac{\|\delta\mathbf{b}_h\|}{\|\mathbf{b}_h\|}.$$

Multiplication with $\|\mathbf{A}\|$ gives the first result

$$\frac{\|\mathbf{u}_h - \hat{\mathbf{u}}_h\|}{\|\mathbf{u}_h\|} \leqslant \operatorname{cond}(\mathbf{A}_h)\frac{\|\delta\mathbf{b}_h\|}{\|\mathbf{b}_h\|} \tag{2.1}$$

if we consider an error in the right hand side only.

(*ii*) Now, we consider an error in the matrix $\tilde{\mathbf{A}}_h = \mathbf{A}_h + \delta\mathbf{A}_h$. It holds

$$\mathbf{A}_h\mathbf{u}_h = \mathbf{b}_h, \quad \tilde{\mathbf{A}}_h\bar{\mathbf{u}}_h = \mathbf{b}_h \quad \Rightarrow \quad \mathbf{u}_h - \bar{\mathbf{u}}_h = \tilde{\mathbf{A}}_h^{-1}\delta\mathbf{A}_h\mathbf{u}_h.$$

Using the assumption $\|\mathbf{A}_h^{-1}\delta\mathbf{A}_h\| \leqslant \|\mathbf{A}_h^{-1}\|\,\|\delta\mathbf{A}_h\| < 1$ gives

$$\|(\mathbf{A}_h + \delta\mathbf{A}_h)^{-1}\delta\mathbf{A}_h\| = \|(I + \mathbf{A}_h^{-1}\delta\mathbf{A})^{-1}\mathbf{A}_h^{-1}\delta\mathbf{A}_h\| < \frac{\|\mathbf{A}_h^{-1}\|\,\|\delta\mathbf{A}_h\|}{1 - \|\mathbf{A}_h^{-1}\delta\mathbf{A}_h\|}.$$

Hence

$$\|\mathbf{u}_h - \bar{\mathbf{u}}_h\| \leqslant \frac{\|\mathbf{A}_h^{-1}\|\,\|\delta\mathbf{A}_h\|}{1 - \|\mathbf{A}_h^{-1}\delta\mathbf{A}_h\|}\|\mathbf{u}_h\|.$$

We define by $\|\mathbf{u}_h\|$ and extend the fraction by $\|\mathbf{A}_h\|/\|\mathbf{A}_h\|$ (twice) to get

$$\frac{\|\mathbf{u}_h - \bar{\mathbf{u}}_h\|}{\|\mathbf{u}_h\|} \leqslant \frac{\operatorname{cond}(\mathbf{A}_h)}{1 - \operatorname{cond}(\mathbf{A}_h)\frac{\|\delta\mathbf{A}_h\|}{\|\mathbf{A}_h\|}}\frac{\|\delta\mathbf{A}_h\|}{\|\mathbf{A}_h\|}$$

(*iii*) The final result follows by combining both estimates. □

Lemma 25. *Let $\|\cdot\|$ be a matrix norm, induced by a corresponding vector nom. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be such that $\|\mathbf{A}\| < 1$. Then, $I + \mathbf{A}$ is a regular matrix and it holds*

$$\|(I + \mathbf{A})^{-1}\| \leqslant \frac{1}{1 - \|\mathbf{A}\|}.$$

Proof. As $\|\cdot\|$ is an induced norm

$$\|(I + \mathbf{A})a\| \geqslant \|x\| - \|\mathbf{A}x\| \geqslant (1 - \|\mathbf{A}\|)\|x\|.$$

Therefore $1 - \|\mathbf{A}\| > 0$ shows, that $I + \mathbf{A}$ is injective and regular. Finally

$$\begin{aligned} 1 = \|I\| &= \|(I + \mathbf{A})(I + \mathbf{A})^{-1}\| + \|(I + \mathbf{A})^{-1} + \mathbf{A}(I + \mathbf{A}) \quad -1\| \\ &\geqslant \|(I + \mathbf{A})^{-1}\| - \|\mathbf{A}\|\,\|(I + \mathbf{A})^{-1}\| \\ &= \|(I + \mathbf{A})^{-1}\|(1 - \|\mathbf{A}\|) > 0. \end{aligned}$$

□

2.1 Eigenschaften der linearen Gleichungssysteme

Die bestimmende Größe für die Fehlerfortpflanzung ist also die Konditionszahl der Matrix. Wir beweisen:

Lemma 26 (Konditionierung der Steifigkeitsmatrix)**.** *Auf einer Folge von regulären Gittern Ω_h gelten für die Konditionszahlen der Steifigkeitsmatrix $\mathbf{A}_h$ (der Poisson-Gleichung) sowie für die Massenmatrix $\mathbf{M}_h$:*

$$\mathrm{cond}_2(\mathbf{A}_h) = O(h^{-2}), \quad \mathrm{cond}_2(\mathbf{M}_h) = O(1).$$

Proof. (i) Beide Matrizen sind positiv definit. Die Spektralkondition ist also gegeben durch:

$$\mathrm{cond}_2(\mathbf{A}_h) = \frac{\lambda_{\max}(\mathbf{A}_h)}{\lambda_{\min}(\mathbf{A}_h)}, \quad \mathrm{cond}_2(\mathbf{M}_h) = \frac{\lambda_{\max}(\mathbf{M}_h)}{\lambda_{\min}(\mathbf{M}_h)}.$$

Für die Eigenwerte einer positiv definiten Matrix $\mathbf{A}$ gilt

$$\lambda_{\min}(\mathbf{A}) = \min_{\mathbf{v}\in\mathbb{R}^N} \frac{\langle \mathbf{A}\mathbf{v}, \mathbf{v}\rangle}{|\mathbf{v}|^2} \leqslant \max_{\mathbf{v}\in\mathbb{R}^N} \frac{\langle \mathbf{A}\mathbf{v}, \mathbf{v}\rangle}{|\mathbf{v}|^2} = \lambda_{\max}(\mathbf{A}).$$

(ii) Wir bestimmen zunächst die Konditionszahl der Massenmatrix. Mit den Element- Massenmatrizen $\mathbf{M}_T$ und der elementweisen Einschränkung $\mathbf{v}_T = \mathbf{v}_{|T}$ gilt für einen Vektor $v_h \in V_h$ mit Koeffizienten $\mathbf{v}$:

$$\begin{aligned}\langle \mathbf{M}_h\mathbf{v}, \mathbf{v}\rangle &= \sum_{T\in\Omega_h} \frac{\langle \mathbf{M}_T\mathbf{v}_T, \mathbf{v}_T\rangle}{|\mathbf{v}_T|^2}|\mathbf{v}_T|^2 \\ &\geqslant \min_{T\in\Omega_h,\, \mathbf{v}\in\mathbb{R}^N} \frac{\langle \mathbf{M}_T\mathbf{v}_T, \mathbf{v}_T\rangle}{|\mathbf{v}_T|^2} \sum_{T\in\Omega_h} |\mathbf{v}_T|^2 \geqslant \min_{T\in\Omega_h}\{\lambda_{\min}(\mathbf{M}_T)\}|\mathbf{v}|^2,\end{aligned}$$

Entsprechend gilt für den maximalen Eigenwert:

$$\langle \mathbf{M}_h\mathbf{v}, \mathbf{v}\rangle \leqslant \max_{T\in\Omega_h,\mathbf{v}\in\mathbb{R}^N} \frac{\langle M_T\mathbf{v}_T, \mathbf{v}_T\rangle}{|\mathbf{v}_T|^2} \sum_{T\in\Omega_h} |\mathbf{v}_T|^2 \leqslant \min_{T\in\Omega_h}\{\lambda_{\max}(\mathbf{M}_T)\}d_{\max}|\mathbf{v}|^2,$$

wobei $d_{\max}$ die maximale Zahl Zellen ist, die sich in einem Knoten treffen. (Diese Konstante ist auf formregulären Gittern gleichmäßig in $h > 0$ beschränkt).

Für die Einträge der Massenmatrix gilt bei Transformation auf das Referenzelement:

$$\mathbf{m}_{ij} = |\det B_T|\,\hat{\mathbf{m}}_{ij},$$

und es gilt also mit $|\det B_T| = O(h_T^d)$ für die Eigenwerte von $\mathbf{M}_T$:

$$\lambda_{\max}(\mathbf{M}_T) = |\det B_T|\lambda_{\max}(M_{\hat{T}}) \leqslant ch_T^d, \quad \lambda_{\min}(\mathbf{M}_T) = |\det B_T|\lambda_{\min}(M_{\hat{T}}) \geqslant ch_T^d.$$

Die Matrix $\mathbf{M}_{\hat{T}}$ ist fest, die Eigenwerte können durch Konstanten abgeschätzt werden. Es folgt:

$$\lambda_{\min}(\mathbf{M}_h) \geqslant ch^d, \quad \lambda_{\max}(\mathbf{M}_h) \leqslant ch^d \quad \Rightarrow \quad \mathrm{cond}_2(\mathbf{M}_h) = O(1).$$

(*iii*) Die Eigenwerte der Steifigkeitsmatrix $\mathbf{A}_h$ wollen wir auf die Eigenwerte der Massenmatrix $\mathbf{M}_h$ zurückführen. Es gilt:

$$\lambda_{\min}(\mathbf{A}_h) \geqslant \min_{\mathbf{v}\in\mathbb{R}^N} \frac{\langle \mathbf{A}_h\mathbf{v},\mathbf{v}\rangle}{\langle \mathbf{M}_h\mathbf{v},\mathbf{v}\rangle} \min_{\mathbf{v}\in\mathbb{R}^N} \frac{\langle \mathbf{M}_h\mathbf{v},\mathbf{v}\rangle}{|\mathbf{v}|^2} = \min_{v_h\in V_h} \frac{a(v_h,v_h)}{\|v_h\|^2}\lambda_{\min}(\mathbf{M}_h),$$
$$\lambda_{\max}(\mathbf{A}_h) \leqslant \max_{\mathbf{v}\in\mathbb{R}^N} \frac{\langle \mathbf{A}_h\mathbf{v},\mathbf{v}\rangle}{\langle \mathbf{M}_h\mathbf{v},\mathbf{v}\rangle} \max_{\mathbf{v}\in\mathbb{R}^N} \frac{\langle \mathbf{M}_h\mathbf{v},\mathbf{v}\rangle}{|\mathbf{v}|^2} = \max_{v_h\in V_h} \frac{a(v_h,v_h)}{\|v_h\|^2}\lambda_{\max}(\mathbf{M}_h).$$

Weiter gilt wegen $V_h \subset V := H_0^1(\Omega)$:

$$\min_{v_h\in V_h} \frac{a(v_h,v_h)}{\|v_h\|^2} \geqslant \inf_{v\in H_0^1(\Omega)} \frac{a(v,v)}{\|v\|^2} =: \lambda_{\min}(\Delta),$$

mit dem kleinsten Eigenwert des Laplace-Operators auf Ω. Mit der inversen Beziehung, Satz **??** gilt ferner:

$$a(v_h,v_h) \leqslant \sum_{T\in\Omega_h} \|\nabla v_h\|_T^2 \leqslant c \sum_{T\in\Omega_h} h_T^{-2}\|v_h\|_T^2 \leqslant c \max_{T\in\Omega_h} h_T^{-2}\|v_h\|^2,$$

Insgesamt gilt also:

$$\lambda_{\min}(\Delta)\lambda_{\min}(\mathbf{M}_h) \leqslant \lambda_{\min}(\mathbf{A}_h) \leqslant \lambda_{\max}(\mathbf{A}_h) \leqslant c \max_{T\in\Omega_h} h_T^{-2}\lambda_{\max}(\mathbf{M}_h).$$

Mit $\lambda_{\min}(\Delta) = c_0 > 0$ und den Eigenwerten der Massenmatrix folgt die Behauptung. □

It is important to note, that the h-dependency of the condition number

$$\mathrm{cond}_2(\mathbf{A}_h) = \mathcal{O}\left(h^{-2}\right)$$

comes from the degree of the differential operator. $-\Delta$ is a second order differential operator. The mass matrix corresponds to the identity operator id and here, the condition behaves like $\mathcal{O}(1)$. Finite element approximations of the operator $(-\Delta)^2$, where

$$a(u,\phi) = (\Delta u, \Delta\phi)$$

have a system matrix with a condition number that scales as $\mathcal{O}\left(h^{-4}\right)$. The condition number does not depend on the polynomial degree of V_h and it also does not depend on the dimension d of $\Omega \subset \mathbb{R}^d$.

Aufgrund des Konstruktionsprinzip von Finite Elemente Ansätzen sind die Matrizen dünn besetzt, denn z.B. für Lagrange-Ansätze gilt:

$$\mathrm{supp}\,(\phi_h^{(i)}) \cap \mathrm{supp}\,(\phi_h^{(j)}) \neq \emptyset \quad \Leftrightarrow \quad \exists T \in \Omega_h,\ x_i, x_j \in \bar{T}.$$

Die genaue Anzahl von Matrix-Einträgen pro Matrixzeile hängt jedoch stark vom jeweiligen Finite Elemente Ansatz und auch von der zugrundeliegenden Triangulierung ab.

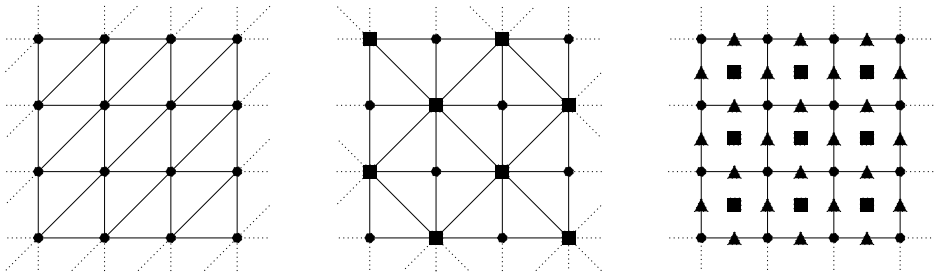


Figure 2.1: Ausschnitt von zwei uniformen Dreiecksgittern und jeweilige Knotenfunktionale der linearen Finiten Elemente Diskretisierung, sowie uniformes Viereckgitter mit den Knotenfunktionalen der biquadratischen Finiten Elemente.

Example 27 (Matrix-Struktur)**.** *Zunächst betrachten wir die Diskretisierung der Poisson-Gleichung mit linearen Finiten Elementen auf einem gleichmäßigen Dreiecksgitter wie in Abbildung 2.1 links. Auf jeder Zelle* T *liegen drei Knotenpunkte* $\mathbf{x}_i$*, das heißt pro Zelle überschneiden sich drei Basisfunktionen. In jedem Knoten* $\mathbf{x}_i$ *kommen auf diesem Gitter sechs Eckpunkte zusammen. Jede Matrix-Zeile (die* j*-te Zeile beinhaltet jeweils die Kopplungen der Testfunktion* $\phi_h^{(j)}$ *zu allen anderen Testfunktionen) hat neben dem Diagonaleintrag noch sechs Nebendiagonaleinträge.*

In der mittleren Abbildung ist ein weiteres, auch sehr regelmäßiges Dreiecksgitter gezeigt. Hier gibt es Knoten $\mathbf{x}_i$*, welche Eckpunkte (die rund markierten) von vier Dreiecken sind, sowie Knoten (die eckigen), in welchen 8 Dreiecke zusammenkommen. Neben der Nebendiagonale gibt es also noch vier oder noch 8 weitere Matrixeinträge pro Zeile.*

In der Abbildung rechts wird ein uniformes Gitter mit den Knotenpunkten $\mathbf{x}_i$ *der biquadratischen Finiten Elemente gezeigt. Hier muss die Analyse der Matrixeinträge je nach Typ der Knotenfunktionale erfolgen, ob die Punkte im Innern (Viereck), in den Eckpunkten (Kreis) oder auf den Kanten (Dreieck) liegen. In jedem Elemente* T *überschneiden sich die Träger von 9 Basisfunktionen. Die Matrixzeile, welche zu einem Knoten im Innern eines Elementes gehört hat demnach inklusive Diagonalelement 9 Einträge, da für die zugehörige Basisfunktion gilt supp* $(\phi_h^{(i)}) = T$*. Basisfunktionen zu Punkten auf den Kanten haben einen Träger auf genau zwei Vierecken und erzeugen somit 15 Matrixeinträge. Für Knotenpunkte auf einer Ecke entstehen 25 Matrixeinträge.*

Die betrachteten Beispiele zeigen, dass auch bei einfachen Ansätzen und gleichmäßigen Gittern keine einheitliche Matrixstruktur erwartet werden kann. Vielmehr muss davon ausgegangen sein, dass die Matrix generell unstrukturiert ist und sich von Zeile zu Zeile ändern kann. Dies ist insbesondere der Fall, wenn allgemeine Gitter zugelassen werden, in denen sich pro Eckknoten unterschiedlich viele Elemente treffen.

Das "Aussehen" der Matrix $\mathbf{A}$ hängt zusätzlich noch von der gewählten Nummerierung der Freiheitsgrade im Gitter ab. Angenommen das Gitter habe $N = M^2$ Knoten, mit M Knoten in jeder Raumrichtung. Werden z.B. die Freiheitsgrade im Gitter links von Abbildung 2.1

lexikographisch, also von unten links, nach oben rechts nummeriert ergibt sich eine Block-Bandmatrix mit der Struktur:

$$\mathbf{A} = \left.\begin{pmatrix} B & -I & & & & \\ -I & B & -I & & & \\ & -I & B & -I & & \\ & & \ddots & \ddots & \ddots & \\ & & & -I & B & -I \\ & & & & -I & B \end{pmatrix}\right\} M, \quad B = \left.\begin{pmatrix} 4 & -1 & & & & \\ -1 & 4 & -1 & & & \\ & -1 & 4 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 4 & -1 \\ & & & & -1 & 4 \end{pmatrix}\right\} M,$$

Man beachte, dass jede Matrixzeile nur 5 Einträge ungleich Null hat, obwohl nach obiger Diskussion 7 Einträge entstehen sollten. Die beiden Einträge zur jeweils Diagonalen Kopplung verschwinden jedoch auf diesem gleichmäßigen Gitter aus Symmetriegründen und sind im Allgemeinen vorhanden. Auf allgemeinen Gittern, oder bei anderer Nummerierung der Freiheitsgrade im Gitter ist üblicherweise keine Bandstruktur gegeben. Zum effizienten Speichern dieser dünn besetzten Matrizen sind daher spezielle Speicherstrukturen notwendig.

2.2 Krylow-Raum-Methoden

Einfache Fixpunktiterationen lassen sich in folgender Form schreiben:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{d}^k, \quad k = 1, 2, \ldots,$$

wobei $\mathbf{d}^k$ in jedem Schritt die Richtung angibt, in der die Lösung verbessert wird. Beim Jacobi-Verfahren bestimmt sich diese Richtung z.B. als $\mathbf{d}^k = D^{-1}(\mathbf{b} - \mathbf{A}\mathbf{x}^k)$, beim Gauß-Seidel-Verfahren als $\mathbf{d}^k = (D+L)^{-1}(\mathbf{b}-\mathbf{A}\mathbf{x}^k)$. Um diese allgemeine Iteration zu verbessern, setzen wir an zwei Punkten an: Zunächst fügen wir in jedem Schritt der Iteration einen Relaxationsparameter ω^k ein:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \omega^k \mathbf{d}^k, \quad k = 1, 2, \ldots,$$

welchen wir Schritt für Schritt optimal bestimmen werden. Anschließend versuchen wir neue Suchrichtungen $\mathbf{d}^k$ auf eine systematische Art und Weise zu entwickeln. Das heißt, wir suchen eine Richtung $\mathbf{d}^k$, in der die größte Fehlerreduktion zu erwarten ist. Hier wird der Begriff des *Krylow-Raums* zur Geltung kommen.

2.2.1 Abstiegs- und Gradientenverfahren

In diesem Abschnitt beschränken wir uns auf symmetrisch positiv definite Matrizen $\mathbf{A} \in \mathbb{R}^{n\times n}$. Zentral für das gesamte Kapitel ist die folgende Charakterisierung zur Lösung eines linearen Gleichungssystems mit symmetrisch positiv definiter Matrix:

Lemma 28 (Lineares Gleichungssystem und Minimierung)**.** *Es sei* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *eine symmetrisch positiv definite Matrix,* $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$*. Die folgenden Bedingungen sind äquivalent:*

$$\begin{aligned} &(i) && \mathbf{Ax} = \mathbf{b}, \\ &(ii) && Q(\mathbf{x}) \leqslant Q(\mathbf{y}) \quad \forall \mathbf{y} \in \mathbb{R}^n, \quad Q(\mathbf{y}) = \frac{1}{2}(\mathbf{Ay}, \mathbf{y})_2 - (\mathbf{b}, \mathbf{y})_2. \end{aligned}$$

Proof. $(i) \Rightarrow (ii)$ $\mathbf{x}$ sei Lösung des linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$. Dann gilt mit beliebigem $\mathbf{y} \in \mathbb{R}^n$:

$$\begin{aligned} 2Q(\mathbf{y}) - 2Q(\mathbf{x}) &= (\mathbf{Ay}, \mathbf{y})_2 - 2(\mathbf{b}, \mathbf{y})_2 - (\mathbf{Ax}, \mathbf{x})_2 + 2(\mathbf{b}, \mathbf{x}) \\ &= (\mathbf{Ay}, \mathbf{y})_2 - 2(\mathbf{Ax}, \mathbf{y})_2 + (\mathbf{Ax}, \mathbf{x})_2 \\ &= (\mathbf{A}(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x})_2 \geqslant 0, \end{aligned}$$

d.h. $Q(\mathbf{y}) \geqslant Q(\mathbf{x})$.

$(ii) \Rightarrow (i)$ Umgekehrt sei $Q(\mathbf{x})$ nun Minimum. Das heißt, $\mathbf{x} \in \mathbb{R}^n$ ist stationärer Punkt der quadratischen Form $Q(\mathbf{x})$, also

$$0 \overset{!}{=} \frac{\partial}{\partial \mathbf{x}_i} Q(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}_i} \Big\{ \frac{1}{2}(\mathbf{Ax}, \mathbf{x})_2 - (\mathbf{b}, \mathbf{x})_2 \Big\} = (\mathbf{Ax})_i - \mathbf{b}_i, \quad i = 1, \ldots, n.$$

Das heißt, $\mathbf{x}$ ist Lösung des linearen Gleichungssystems. □

Anstelle der Bestimmung einer Lösung eines linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$, suchen wir das Minimum des sogenannten *Energiefunktionals*

$$Q(\mathbf{x}) \to \min.$$

Dieser Zugang ist Grundlage der im Folgenden diskutieren Verfahren und auch Basis der allgemeinen Klasse von Krylow-Raum-Verfahren.

Remark 29. *Die Bezeichnung Energiefunktional ist physikalisch motiviert, da dieses Funktional die Energie eines Systems repräsentiert. Zu dieser Energie kann eine dementsprechende Norm assoziert werden, die wir in Kürze definieren.*

Wir betrachten zunächst nur symmetrisch positiv definite Matrizen, daher ist durch $\|\mathbf{x}\|_A := \sqrt{(\mathbf{Ax}, \mathbf{x})_2}$ eine Norm, die sogenannte *Energienorm*, gegeben. Die Minimierung des Energiefunktionals $Q(\cdot)$ ist auch äquivalent zur Minimierung des Fehlers $\mathbf{x}^k - \mathbf{x}$ in der zugehörigen Energienorm. Denn angenommen $\mathbf{x} \in \mathbb{R}^n$ sei die Lösung des Gleichungssystems und $y \in \mathbb{R}^n$ eine beliebige Approximation an diese Lösung. Dann gilt

$$\|\mathbf{y} - \mathbf{x}\|_{\mathbf{A}}^2 = (A(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x})_2 = (\mathbf{Ay}, \mathbf{y}) - \underbrace{2(\mathbf{Ay}, \mathbf{x})}_{=2(b,y)} + (\mathbf{Ax}, \mathbf{x}) = 2Q(\mathbf{y}) + (\mathbf{Ax}, \mathbf{x}).$$

Minimierung von $Q(\mathbf{y})$ minimiert auch den Fehler in der Energienorm.

Die Idee des *Abstiegsverfahrens* ist die sukzessive Reduktion des Energiefunktionals $Q(\cdot)$ für eine Folge von Approximationen $\mathbf{x}^k$. Dabei wird in jedem Schritt des Verfahrens zunächst eine sogenannte Abstiegsrichtung $\mathbf{d}^k$ gewählt, im Anschluss wird die Approximation in dieser Richtung verbessert: $\mathbf{x}^{k+1} = \mathbf{x}^k + \omega^k \mathbf{d}^k$. Die Schrittweite $\omega^k \in \mathbb{R}$ wird dabei so gewählt, dass der resultierende Wert des Energiefunktionals minimal wird. Wir fassen zusammen:

Listing 2.1: Abstiegsverfahren

Es sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit, $\mathbf{x}^0, b \in \mathbb{R}^n$.

```
Fuer k von 0,1,2,...
  Waehle Abstiegsrichtung d^k ∈ R^n
  Bestimme ω^k als Minimum von ω^k = arg min_{ω^k ∈ R} Q(x^k + ω^k d^k)
  Update x^{k+1} = x^k + ω^k d^k
```

Wir gehen davon aus, dass die Suchrichtungen gegeben seien. Es bleibt, die optimale Schrittweite zu berechnen. Hier handelt es sich um ein einfaches skalares Minimierungsproblem. Wir bestimmen

$$0 \stackrel{!}{=} \frac{\partial}{\partial \omega^k} Q(\mathbf{x}^k + \omega^k \mathbf{d}^k) = \omega^k (\mathbf{A}\mathbf{d}^k, \mathbf{d}^k) + (\mathbf{A}\mathbf{x}^k, \mathbf{d}^k) - (b, \mathbf{d}^k),$$

also

$$\omega^k = \frac{(\mathbf{b} - \mathbf{A}\mathbf{x}^k, \mathbf{d}^k)}{(\mathbf{A}\mathbf{d}^k, \mathbf{d}^k)}. \tag{2.2}$$

Eine Möglichkeit zur Bestimmung der Abstiegsrichtung ist eine Kombination mit Jacobi- oder Gauß-Seidel-Verfahren. Die Richtungen werden über diese Verfahren bestimmt, die Schrittweite wird optimal berechnet. Wir betrachten hierzu ein Beispiel:

Example 30 (Abstiegsverfahren, Jacobi und Gauß-Seidel)**.** *Es sei* $\mathbf{A}\mathbf{x} = b$ *mit*

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ -3 \\ 4 \end{pmatrix}, \quad x = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}.$$

Mit dem Startvektor $\mathbf{x}^0 = 0$ *führen wir jeweils zehn Schritte mit Jacobi-, Gauß-Seidel-Verfahren sowie jeweils mit den entsprechenden Kombinationen unter Verwendung des optimalen Abstiegsschritts* ω^k *durch. In Abbildung 2.2 links fassen wir für alle Verfahren die Fehler zusammen. In der unteren Abbildung stellen wir den Approximationsverlauf* $\mathbf{x}^k \in \mathbb{R}^3$ *für Jacobi sowie Jacobi-Abstiegsverfahren grafisch dar. Obwohl der Verlauf des Jacobi-Abstiegsverfahrens wesentlich "geradliniger" scheint, konvergiert dieses ebenso langsam wie das Jacobi-Verfahren selbst. Nur im Fall des Gauß-Seidel-Verfahrens wird die Konvergenz durch Wahl optimaler Schrittweite* ω^k *wesentlich beschleunigt.*

Abschließend werden wir ein erstes Verfahren entwickeln, welches die neue Suchrichtung $\mathbf{d}^k \in \mathbb{R}^n$ systematisch so bestimmt, dass das quadratische Funktional $Q(\mathbf{x})$ möglichst stark minimiert werden kann. Wir suchen also die Richtung des *stärksten Abfalls*. Zu einem Punkt

$\mathbf{x} \in \mathbb{R}^n$ ist dies gerade die Richtung $d \in \mathbb{R}^n$, die normal (sprich senkrecht) auf der Niveaumenge $N(\mathbf{x})$ steht:

$$N(\mathbf{x}) := \{\mathbf{y} \in \mathbb{R}^n : \ Q(\mathbf{y}) = Q(\mathbf{x})\}$$

In einem Punkt $\mathbf{x}$ ist die Niveaumenge aufgespannt durch alle Richtungen $\delta\mathbf{x} \in \mathbb{R}^n$ mit

$$0 \stackrel{!}{=} Q'(\mathbf{x}) \cdot \delta\mathbf{x} = (\nabla Q(\mathbf{x}), \delta\mathbf{x}) = (\mathbf{b} - \mathbf{A}\mathbf{x}, \delta\mathbf{x}).$$

Die Vektoren $\delta\mathbf{x}$, welche die Niveaumenge aufspannen, stehen orthogonal auf dem Defekt $\mathbf{b} - \mathbf{A}\mathbf{x}$, dieser zeigt daher in Richtung der stärksten Änderung von $Q(\cdot)$. Wir wählen $\mathbf{d}^k := \mathbf{b} - \mathbf{A}\mathbf{x}^k$. Die so gefundene Suchrichtung wird dann mit dem Abstiegsverfahren kombiniert, d.h., wir iterieren

$$\mathbf{d}^k := \mathbf{b} - \mathbf{A}\mathbf{x}^k, \quad \omega^k := \frac{\|\mathbf{d}^k\|_2^2}{(\mathbf{A}\mathbf{d}^k, \mathbf{d}^k)_2}, \quad \mathbf{x}^{k+1} := \mathbf{x}^k + \omega^k \mathbf{d}^k.$$

Wir definieren das *Gradientenverfahren*:

Listing 2.2: Gradientenverfahren

Es sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit, $b \in \mathbb{R}^n$. Es sei $\mathbf{x}^0 \in \mathbb{R}^n$ beliebig und $\mathbf{d}^0 := \mathbf{b} - \mathbf{A}\mathbf{x}^0$.

```
1 Fuer k von 0, 1, 2, ...
2   r^k := A d^k
3   ω^k = ||d_k||_2^2 / (r^k, d^k)_2
4   x^{k+1} = x^k + ω^k d^k
5   d^{k+1} = d^k − ω^k r^k
```

Durch Einführen eines zweiten Hilfsvektors $r^k \in \mathbb{R}^n$ kann in jeder Iteration ein Matrix-Vektor-Produkt gespart werden. Für Matrizen mit Diagonalanteil $D = \alpha I$ ist das Gradientenverfahren gerade das Jacobi-Verfahren in Verbindung mit dem Abstiegsverfahren. Daher kann für dieses Verfahren im Allgemeinen auch keine verbesserte Konvergenzaussage erreicht werden. Es stellt jedoch den Einstieg in eine ganze Klasse von komplexeren Verfahren, den Krylow-Unterraum-Verfahren, dar. Wir zeigen:

Lemma 31 (Gradientenverfahren)**.** *Es sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit. Dann konvergiert das Gradientenverfahren für jeden Startvektor $\mathbf{x}^0 \in \mathbb{R}^n$ gegen die Lösung des Gleichungssystems $\mathbf{A}\mathbf{x} = b$.*

Proof. Es sei $\mathbf{x}^k \in \mathbb{R}^n$ eine gegebene Approximation. Weiter sei $d := \mathbf{b} - \mathbf{A}\mathbf{x}^k$. Dann berechnet sich ein Schritt des Gradientenverfahrens als

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \frac{(\mathbf{d}, \mathbf{d})}{(\mathbf{A}\mathbf{d}, \mathbf{d})} d.$$

Für das Energiefunktional gilt:

$$\begin{aligned}
Q(\mathbf{x}^{k+1}) &= \frac{1}{2}(\mathbf{A}\mathbf{x}^{k+1}, \mathbf{x}^{k+1}) - (\mathbf{b}, \mathbf{x}^{k+1}) \\
&= \frac{1}{2}(\mathbf{A}\mathbf{x}^{k}, \mathbf{x}^{k}) + \frac{1}{2}\frac{(\mathbf{d},\mathbf{d})^2}{(\mathbf{A}\mathbf{d},\mathbf{d})^2}(\mathbf{A}\mathbf{d},\mathbf{d}) + \frac{(\mathbf{d},\mathbf{d})}{(\mathbf{A}\mathbf{d},\mathbf{d})}(\mathbf{A}\mathbf{x}^{k},\mathbf{d}) \\
&\quad - (b, \mathbf{x}^{k}) - \frac{(\mathbf{d},\mathbf{d})}{(\mathbf{A}\mathbf{d},\mathbf{d})}(\mathbf{b},\mathbf{d}) \\
&= Q(\mathbf{x}^{k}) + \frac{(\mathbf{d},\mathbf{d})}{(\mathbf{A}\mathbf{d},\mathbf{d})}\left\{\frac{1}{2}(\mathbf{d},\mathbf{d}) + (\mathbf{A}\mathbf{x}^{k},\mathbf{d}) - (\mathbf{b},\mathbf{d})\right\} \\
&= Q(\mathbf{x}^{k}) + \frac{(\mathbf{d},\mathbf{d})}{(\mathbf{A}\mathbf{d},\mathbf{d})}\left\{\frac{1}{2}(\mathbf{d},\mathbf{d}) + (\underbrace{\mathbf{A}\mathbf{x}^{k} - \mathbf{b}}_{=-\mathbf{d}},\mathbf{d})\right\}
\end{aligned}$$

Also folgt

$$Q(\mathbf{x}^{k+1}) = Q(\mathbf{x}^{k}) - \frac{(\mathbf{d},\mathbf{d})^2}{2(\mathbf{A}\mathbf{d},\mathbf{d})}.$$

Wegen der positiven Definitheit von $\mathbf{A}$ gilt

$$\lambda_{\min}(\mathbf{A})(\mathbf{d},\mathbf{d}) \leqslant (\mathbf{A}\mathbf{d},\mathbf{d}) \leqslant \lambda_{\max}(\mathbf{A})(\mathbf{d},\mathbf{d})$$

und schließlich ist mit

$$Q(\mathbf{x}^{k+1}) \leqslant Q(\mathbf{x}^{k}) - \underbrace{\frac{(\mathbf{d},\mathbf{d})}{2\lambda_{\max}}}_{>0}$$

die Folge $Q(\mathbf{x}^k)$ monoton fallend. Solange $d = \mathbf{b} - \mathbf{A}\mathbf{x} \neq 0$, fällt die Folge streng monoton. Weiter ist $Q(\mathbf{x}^k)$ nach unten durch $Q(\mathbf{x})$ beschränkt. Also konvergiert die Folge $Q(\mathbf{x}^k) \to c \in \mathbb{R}^n$. Im Grenzwert muss $0 = (\mathbf{d},\mathbf{d}) = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2$ gelten, sprich $\mathbf{A}\mathbf{x} = b$. □

Schließlich zeigen wir noch eine (suboptimale) Abschätzung der Konvergenzgeschwindigkeit des Gradientenverfahrens:

Lemma 32 (Konvergenz des Gradientenverfahrens (vereinfacht))**.** *Es sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine symmetrisch positiv definite Matrix. Dann gilt für das Gradientenverfahren zur Lösung von $\mathbf{A}\mathbf{x} = b$ die Fehlerabschätzung*

$$\|\mathbf{x}^k - \mathbf{x}\|_A \leqslant \left(1 - \frac{1}{\kappa}\right)^k, \quad \kappa := \operatorname{cond}_2(\mathbf{A}) = \frac{\lambda_{max}(\mathbf{A})}{\lambda_{min}(\mathbf{A})}.$$

Proof. Die Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ ist symmetrisch positiv definit. Es gibt also ein System aus n Eigenwerten

$$0 < \lambda_{min} =: \lambda_1 \leqslant \cdots \leqslant \lambda_n =: \lambda_{max}$$

und orthonormalen Eigenvektoren $\mathbf{w}_1, \ldots, \mathbf{w}_n \in \mathbb{R}^n$. Es sei

$$\mathbf{e}^k = \mathbf{x}^k - \mathbf{x} = \sum_{i=1}^{n} \mathbf{e}_i^k \mathbf{w}_i \tag{2.3}$$

eine Entwicklung des Fehlers in den Eigenvektoren. Für einen Iterationsschritt des Gradientenverfahrens gilt mit

$$\mathbf{d}^k = \mathbf{b} - \mathbf{A}\mathbf{x}^k = -\mathbf{A}\mathbf{e}^k$$

die Fehlerfortpflanzung

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \frac{(\mathbf{d}^k, \mathbf{d}^k)}{(\mathbf{A}\mathbf{d}^k, \mathbf{d}^k)}\mathbf{d}^k \quad \Rightarrow \quad \mathbf{e}^{k+1} = \mathbf{e}^k - \frac{(\mathbf{A}\mathbf{e}^k, \mathbf{A}\mathbf{e}^k)}{(\mathbf{A}^2\mathbf{e}^k, \mathbf{A}\mathbf{e}^k)}\mathbf{A}\mathbf{e}^k. \tag{2.4}$$

Mit der Darstellung (2.3) und mit $A\mathbf{w}_i = \lambda_i \mathbf{w}_i$ folgt

$$\mathbf{e}^{k+1} = \sum_{i=1}^{N} \left(1 - \frac{(\mathbf{A}\mathbf{e}^k, \mathbf{A}\mathbf{e}^k)}{(\mathbf{A}^2\mathbf{e}^k, \mathbf{A}\mathbf{e}^k)}\lambda_i\right) \mathbf{e}_i^k \mathbf{w}_i = \sum_{i=1}^{N} (1 - \mu_i)\mathbf{e}_i^k \mathbf{w}_i \tag{2.5}$$

mit

$$\mu_i = \lambda_i \frac{(\mathbf{A}\mathbf{e}^k, \mathbf{A}\mathbf{e}^k)}{(\mathbf{A}^2\mathbf{e}^k, \mathbf{A}\mathbf{e}^k)}.$$

Wegen der Orthonormalität der Eigenvektoren und wegen $\lambda_i > 0$ folgt

$$\begin{aligned} \mu_i &= \lambda_i \frac{\left(\sum_{j=1}^n \lambda_j \mathbf{e}_j^k \mathbf{w}_j, \sum_{j=1}^n \lambda_j \mathbf{e}_j^k \mathbf{w}_j\right)}{\left(\sum_{j=1}^n \lambda_j^2 \mathbf{e}_j^k \mathbf{w}_j, \sum_{j=1}^n \lambda_j \mathbf{e}_j^k \mathbf{w}_j\right)} \\ &= \lambda_i \frac{\sum_{j=1}^n \lambda_j^2 (\mathbf{e}_j^k)^2}{\sum_{j=1}^n \lambda_j^3 (\mathbf{e}_j^k)^2} \\ &\geqslant \lambda_{min} \frac{\sum_{j=1}^n \lambda_j^2 (\mathbf{e}_j^k)^2}{\lambda_{max} \sum_{j=1}^n \lambda_j^2 (\mathbf{e}_j^k)^2} = \frac{\lambda_{max}}{\lambda_{min}}. \end{aligned} \tag{2.6}$$

Wir machen nun mit (2.5) weiter und erhalten in der $\mathbf{A}$-Norm

$$\begin{aligned} \|\mathbf{e}^{k+1}\|_{\mathbf{A}}^2 = (\mathbf{A}\mathbf{e}^{k+1}, \mathbf{e}^{k+1}) &= \left(\mathbf{A}\sum_{i=1}^{N}(1-\mu_i)\mathbf{e}_i^k\mathbf{w}_i, \sum_{i=1}^{N}(1-\mu_i)\mathbf{e}_i^k\mathbf{w}_i\right) \\ &= \sum_{i=1}^{N}(1-\mu_i)^2\lambda_i(\mathbf{e}_i^k)^2. \end{aligned}$$

Und mit (2.6) und dem Zusammenhang

$$\|\mathbf{e}^k\|_{\mathbf{A}}^2 = (\mathbf{A}\mathbf{e}^k, \mathbf{e}^k) = \sum_{i=1}^{N} \lambda_i (\mathbf{e}_i^k)^2$$

folgt dann

$$\|\mathbf{e}^{k+1}\|_{\mathbf{A}}^2 \leqslant \left(1 - \frac{1}{\kappa}\right)^2 \|\mathbf{e}^k\|_{\mathbf{A}}^2, \quad \kappa := \mathrm{cond}_2(\mathbf{A}) = \frac{\lambda_{max}}{\lambda_{min}}.$$

Wiederholte Abschätzung ergibt schließlich

$$\|\mathbf{e}^k\|_A \leqslant \left(1 - \frac{1}{\kappa}\right)^k \|\mathbf{x}^0 - \mathbf{x}\|_A.$$

□

Remark 33 (Optimale Fehlerabschätzung für das Gradientenverfahren)**.** *Mit größerem Aufwand lässt sich für das Gradientenverfahren unter denselben Voraussetzungen die optimale Fehlerabschätzung*

$$\|\mathbf{x}^k - \mathbf{x}\|_A \leqslant \left(\frac{1-1/\kappa}{1+1/\kappa}\right)^k, \quad \kappa := \mathrm{cond}_2(\mathbf{A}),$$

herleiten. Es gilt asymptotisch

$$\frac{1-\frac{1}{\kappa}}{1+\frac{1}{\kappa}} = 1 - \frac{2}{\kappa} + O\left(\frac{1}{\kappa^2}\right),$$

d.h., es liegt doppelt so schnelle Konvergenz vor. Schlüssel für diesen optimalen Beweis ist der Satz von Kantorovich, eine Abschätzung für die Eigenwerte einer positiv definiten Matrix. Für den Beweis verweisen wir auf die Literatur [**?**]*.*

Die asymptotische Konvergenzrate des Gradientenverfahrens wird durch die Kondition der Matrix bestimmt. Für die Modellmatrix gilt $\kappa = O(n^2)$, siehe Beispiel **??**. Also gilt

$$\rho = \frac{1-\frac{1}{n^2}}{1+\frac{1}{n^2}} = 1 - \frac{2}{n^2} + O\left(\frac{1}{n^4}\right).$$

Die Konvergenz ist demnach ebenso langsam wie die des Jacobi-Verfahrens (wir haben bereits diskutiert, dass es für die Modellmatrix mit dem Jacobi-Abstiegsverfahren übereinstimmt). Für das Gradientenverfahren gilt jedoch der folgende Zusammenhang, der Basis des CG-Verfahrens ist:

Lemma 34 (Abstiegsrichtungen im Gradientenverfahren)**.** *Es sei* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *symmetrisch positiv definit. Dann stehen je zwei aufeinanderfolgende Abstiegsrichtungen* $\mathbf{d}^k$ *und* $\mathbf{d}^{k+1}$ *des Gradientenverfahrens orthogonal aufeinander, also* $(\mathbf{d}^k, \mathbf{d}^{k+1}) = 0$*.*

Proof. Zum Beweis siehe Algorithmus 2.2. Es gilt

$$\mathbf{d}^{k+1} = \mathbf{d}^k - \omega^k \mathbf{r}^k = \mathbf{d}^k - \frac{(\mathbf{d}^k, \mathbf{d}^k)}{(\mathbf{A}\mathbf{d}^k, \mathbf{d}^k)}\mathbf{A}\mathbf{d}^k.$$

Also gilt

$$(\mathbf{d}^{k+1}, \mathbf{d}^k) = (\mathbf{d}^k, \mathbf{d}^k) - \frac{(\mathbf{d}^k, \mathbf{d}^k)}{(\mathbf{A}\mathbf{d}^k, \mathbf{d}^k)}(\mathbf{A}\mathbf{d}^k, \mathbf{d}^k) = (\mathbf{d}^k, \mathbf{d}^k) - (\mathbf{d}^k, \mathbf{d}^k) = 0.$$

□

2.2.2 Das CG-Verfahren

Der Zusammenhang aus Satz 34 gilt nur für jeweils aufeinanderfolgende Abstiegsrichtungen, im Allgemeinen gilt jedoch $\mathbf{d}^k \not\perp \mathbf{d}^{k+2}$. In Abbildung 2.2 rechts ist der Approximationsverlauf des Jacobi-Abstiegsverfahrens, welches hier mit dem Gradientenverfahren übereinstimmt, dargestellt. Zwei aufeinanderfolgende Richtungen sind zwar je orthogonal,

die dritte Richtung steht jedoch wieder nahezu parallel auf der ersten. Dies führt dazu, dass das Gradientenverfahren im Allgemeinen sehr langsam konvergiert. Das CG-Verfahren, auch "Verfahren der konjugierten Gradienten" genannt, entwickelt diesen Ansatz weiter und wählt Suchrichtungen $\{\mathbf{d}^0, \dots, \mathbf{d}^{k-1}\}$, die paarweise orthogonal sind. Orthogonalität wird dabei im **A**-Skalarprodukt (wir betrachten weiter nur symmetrisch positiv definite Matrizen) erreicht:

$$(\mathbf{A}\mathbf{d}^r, \mathbf{d}^s) = 0 \quad \forall r \neq s$$

Im k-ten Schritt wird die Approximation $\mathbf{x}^k = \mathbf{x}^0 + \sum_{i=0}^{k-1} \alpha_i \mathbf{d}^i$ als das Minimum über alle $\alpha = (\alpha_0, \dots, \alpha_{k-1})$ bezüglich $Q(\mathbf{x}^k)$ gesucht:

$$\begin{aligned}\min_{\alpha \in \mathbb{R}^k} Q\left(\mathbf{x}^0 + \sum_{i=0}^{k-1} \alpha_i \mathbf{d}^i\right) \\ = \min_{\alpha \in \mathbb{R}^k} \left\{ \frac{1}{2} \left(\mathbf{A}\mathbf{x}^0 + \sum_{i=0}^{k-1} \alpha_i A\mathbf{d}^i, \mathbf{x}^0 + \sum_{i=0}^{k-1} \alpha_i \mathbf{d}^i \right) - \left(b, \mathbf{x}^0 + \sum_{i=0}^{k-1} \alpha_i \mathbf{d}^i \right) \right\}\end{aligned}$$

Der stationäre Punkt ist bestimmt durch

$$\begin{aligned}0 \stackrel{!}{=} \frac{\partial}{\partial \alpha_j} Q(\mathbf{x}^k) = \left(\mathbf{A}\mathbf{x}^0 + \sum_{i=0}^{k-1} \alpha_i \mathbf{A}\mathbf{d}^i, \mathbf{d}^j \right) - (b, \mathbf{d}^j) \\ = -\left(\mathbf{b} - \mathbf{A}\mathbf{x}^k, \mathbf{d}^j\right), \quad j = 0, \dots, k-1.\end{aligned}$$

Das heißt, das neue Residuum $\mathbf{b} - \mathbf{A}\mathbf{x}^k$ steht orthogonal auf allen Suchrichtungen $\mathbf{d}^j$ für $j = 0, \dots, k-1$. Dieses Gleichungssystem

$$(\mathbf{b} - \mathbf{A}\mathbf{x}^k, \mathbf{d}^j) = 0 \quad \forall j = 0, \dots, k-1 \tag{2.7}$$

wird *Galerkin-Gleichung* genannt. Beim Entwurf des CG-Verfahrens ist es nun wichtig, dass die neu gewählte Suchrichtung $\mathbf{d}^k$ nicht im Erzeugnis der bisherigen Suchrichtungen $\operatorname{span}\{\mathbf{d}^0, \dots, \mathbf{d}^{k-1}\}$ enthalten ist. Denn in diesem Fall wird der Suchraum nicht größer und die Approximation kann nicht verbessert werden. Daher wählt man für das CG-Verfahren ausgehend von einer Startapproximation $\mathbf{x}^0 \in \mathbb{R}^n$ mit $\mathbf{d}^0 := \mathbf{b} - \mathbf{A}\mathbf{x}^0$ den *Krylow-Raum* $K_k(\mathbf{d}^0, A)$:

$$K_k(\mathbf{d}^0, A) := \operatorname{span}\{\mathbf{d}^0, \mathbf{A}\mathbf{d}^0, \dots, \mathbf{A}^{k-1}\mathbf{d}^0\}$$

Es gilt:

Lemma 35. *Angenommen, es gilt $\mathbf{A}^k\mathbf{d}^0 \in K_k$. Dann liegt die Lösung $\mathbf{x} \in \mathbb{R}^n$ von $\mathbf{A}\mathbf{x} = b$ im k-ten Krylow-Raum $K_k(d_0, A)$.*

Proof. Es sei K_k gegeben und $\mathbf{x}^k \in \mathbf{x}_0 + K_k$ die beste Approximation, welche die Galerkin-Gleichung (2.7) erfüllt. Es sei $r^k := \mathbf{b} - \mathbf{A}\mathbf{x}^k$. Wegen

$$r^k = \mathbf{b} - \mathbf{A}\mathbf{x}^k = \underbrace{\mathbf{b} - \mathbf{A}\mathbf{x}^0}_{=\mathbf{d}^0} + \mathbf{A}\underbrace{(\mathbf{x}^0 - \mathbf{x}^k)}_{\in K_k} \in \mathbf{d}^0 + AK_k$$

gilt $r^k \in K_{k+1}$. Angenommen, nun, $K_{k+1} \subset K_k$. Dann gilt $r^k \in K_k$. Die Galerkin-Gleichung besagt $r^k \perp K_k$, d.h., es gilt zwingend $r^k = 0$ und $\mathbf{A}\mathbf{x}^k = b$. □

Falls das CG-Verfahren abbricht, weil keine neuen Suchrichtungen hinzukommen, so ist die Lösung gefunden. Angenommen, die **A**-orthogonalen Suchrichtungen $\{\mathbf{d}^0, \mathbf{d}^1, \ldots, \mathbf{d}^{k-1}\}$ liegen vor, so kann die CG-Approximation durch Ausnutzen der Basisdarstellung $\mathbf{x}^k = \mathbf{x}^0 + \sum \alpha_i \mathbf{d}^i$ aus der Galerkin-Gleichung berechnet werden:

$$\left(\mathbf{b} - \mathbf{A}\mathbf{x}^0 - \sum_{i=0}^{k-1} \alpha_i \mathbf{A}\mathbf{d}^i, d_j\right) = 0 \quad \Rightarrow \quad (\mathbf{b} - \mathbf{A}\mathbf{x}^0, \mathbf{d}^j) = \alpha_j (\mathbf{A}\mathbf{d}^j, \mathbf{d}^j)$$

$$\Rightarrow \quad \alpha_j = \frac{(\mathbf{d}^0, \mathbf{d}^j)}{(\mathbf{A}\mathbf{d}^j, \mathbf{d}^j)}$$

Die **A**-orthogonale Basis $\{\mathbf{d}^0, \ldots, \mathbf{d}^{k-1}\}$ des Krylow-Raums $K_k(\mathbf{d}^0, A)$ kann z.B. mit dem Gram-Schmidt-Verfahren berechnet werden. Der Nachteil dieser Methode ist der hohe Aufwand des Gram-Schmidt-Verfahrens. Zur Orthogonalisierung eines Vektors bzgl. einer bereits vorhandenen Basis $\{\mathbf{d}^0, \ldots, \mathbf{d}^{k-1}\}$ sind k Skalarprodukte erforderlich. Seine Leistungsfähigkeit erlangt das CG-Verfahren durch Ausnutzen einer zweistufigen Rekursionsformel, welche die **A**-orthogonale Basis effizient und stabil berechnet:

Lemma 36 (Zweistufige Rekursionsformel zur Orthogonalisierung)**.** *Es sei* $\mathbf{A} \in \mathbb{R}^{n \times n}$ *symmetrisch positiv definit sowie* $\mathbf{x}^0 \in \mathbb{R}^n$ *und* $\mathbf{d}^0 := \mathbf{b} - \mathbf{A}\mathbf{x}^0$*. Dann wird durch die Iteration für* $k = 1, 2, \ldots$

$$r^k := \mathbf{b} - \mathbf{A}\mathbf{x}^k, \quad \beta_{k-1} := -\frac{(r^k, \mathbf{A}\mathbf{d}^{k-1})}{(\mathbf{d}^{k-1}, \mathbf{A}\mathbf{d}^{k-1})}, \quad \mathbf{d}^k := r^k - \beta_{k-1}\mathbf{d}^{k-1}$$

eine **A***-orthogonale Basis mit* $(\mathbf{A}\mathbf{d}^r, \mathbf{d}^s) = 0$ *für* $r \neq s$ *erzeugt. Dabei ist* $\mathbf{x}^k$ *in Schritt* k *definiert als die Galerkin-Lösung* $(\mathbf{b} - \mathbf{A}\mathbf{x}^k, \mathbf{d}^j) = 0$ *für* $j = 0, \ldots, k-1$*.*

Proof. Es sei durch $\{\mathbf{d}^0, \ldots, \mathbf{d}^{k-1}\}$ eine **A**-orthogonale Basis des $K_k(\mathbf{d}^0, A)$ gegeben. Weiter sei $\mathbf{x}^k \in \mathbf{x}^0 + K_k(\mathbf{d}^0, A)$ die Galerkin-Lösung zu (2.7). Es sei $r^k := \mathbf{b} - \mathbf{A}\mathbf{x}^k \in K_{k+1}$ und wir fordern, dass $r^k \notin K_k(\mathbf{d}^0, A)$. Ansonsten bricht die Iteration nach Hilfssatz 35 ab. Wir bestimmen $\mathbf{d}^k$ mit dem Ansatz

$$\mathbf{d}^k = r^k - \sum_{j=0}^{k-1} \beta_j^{k-1} \mathbf{d}^j. \tag{2.8}$$

Die Orthogonalitätsbedingung besagt:

$$0 \stackrel{!}{=} (\mathbf{d}^k, \mathbf{A}\mathbf{d}^i) = (r^k, \mathbf{A}\mathbf{d}^i) + \sum_{j=0}^{k-1} \beta_j^{k-1} (\mathbf{d}^j, \mathbf{A}\mathbf{d}^i)$$

$$= (r^k, \mathbf{A}\mathbf{d}^i) + \beta_i^{k-1} (\mathbf{d}^i, \mathbf{A}\mathbf{d}^i), \quad i = 0, \ldots, k-1$$

Es gilt $(r^k, \mathbf{A}\mathbf{d}^i) = (\mathbf{b} - \mathbf{A}\mathbf{x}^k, \mathbf{A}\mathbf{d}^i) = 0$ für $i = 0, \ldots, k-2$, da $Ar^k \perp K_{k-1}$. Hieraus folgt $\beta_i^{k-1} = 0$ für $i = 0, 1, \ldots, k-2$. Für $i = k-1$ gilt

$$\beta_{k-1} := \beta_{k-1}^{k-1} = -\frac{(r^k, \mathbf{A}\mathbf{d}^{k-1})}{(\mathbf{d}^{k-1}, \mathbf{A}\mathbf{d}^{k-1})}.$$

Schließlich gilt mit (2.8) $\mathbf{d}^k = r^k - \beta_{k-1}\mathbf{d}^{k-1}$. □

Mit diesen Vorarbeiten können wir alle Bestandteile des CG-Verfahrens zusammensetzen. Es sei also mit $\mathbf{x}^0$ eine Startlösung und mit $\mathbf{d}^0 := \mathbf{b} - \mathbf{A}\mathbf{x}^0$ der Startdefekt gegeben. Angenommen, $K_k := \operatorname{span}\{\mathbf{d}^0, \ldots, \mathbf{d}^{k-1}\}$ sowie $\mathbf{x}^k \in \mathbf{x}^0 + K_k$ und der Defekt $r^k = \mathbf{b} - \mathbf{A}\mathbf{x}^k$ liegen vor. Dann berechnet sich $\mathbf{d}^k$ gemäß Hilfssatz 36 als

$$\beta_{k-1} = -\frac{(r^k, \mathbf{A}\mathbf{d}^{k-1})}{(\mathbf{d}^{k-1}, \mathbf{A}\mathbf{d}^{k-1})}, \quad \mathbf{d}^k = r^k - \beta_{k-1}\mathbf{d}^{k-1}. \tag{2.9}$$

Für den neuen Entwicklungskoeffizienten α_k in $\mathbf{x}^{k+1} = \mathbf{x}^0 + \sum_{i=0}^{k} \alpha_i \mathbf{d}^i$ gilt durch Testen der Galerkin-Gleichung (2.7) mit $\mathbf{d}^k$

$$\begin{aligned}\left(\underbrace{\mathbf{b} - \mathbf{A}\mathbf{x}^0}_{=\mathbf{d}^0} - \sum_{i=0}^{k} \alpha_i \mathbf{A}\mathbf{d}^i, \mathbf{d}^k\right) &= (\mathbf{b} - \mathbf{A}\mathbf{x}^0, \mathbf{d}^k) - \alpha_k(\mathbf{A}\mathbf{d}^k, \mathbf{d}^k) \\ &= (\mathbf{b} - \mathbf{A}\mathbf{x}^0 + \underbrace{\mathbf{A}(\mathbf{x}^0 - \mathbf{x}^k)}_{\in K_k}, \mathbf{d}^k) - \alpha_k(\mathbf{A}\mathbf{d}^k, \mathbf{d}^k).\end{aligned}$$

Also

$$\alpha_k = \frac{(r^k, \mathbf{d}^k)}{(\mathbf{A}\mathbf{d}^k, \mathbf{d}^k)}, \quad \mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{d}^k. \tag{2.10}$$

Hieraus lässt sich auch unmittelbar der neue Defekt r^{k+1} bestimmen:

$$r^{k+1} = \mathbf{b} - \mathbf{A}\mathbf{x}^{k+1} = \mathbf{b} - \mathbf{A}\mathbf{x}^k - \alpha_k \mathbf{A}\mathbf{d}^k = r^k - \alpha_k \mathbf{A}\mathbf{d}^k \tag{2.11}$$

Wir fassen $(2.9 - 2.11)$ zusammen und formulieren das klassische CG-Verfahren:

Listing 2.3: CG-Verfahren

Es sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit, $\mathbf{x}^0 \in \mathbb{R}^n$ und $r^0 = \mathbf{d}^0 = \mathbf{b} - \mathbf{A}\mathbf{x}^0$ gegeben. Iteriere für $k = 0, 1, \ldots$

1. $\alpha_k = \frac{(r^k, \mathbf{d}^k)}{(\mathbf{A}\mathbf{d}^k, \mathbf{d}^k)}$
2. $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{d}^k$
3. $r^{k+1} = r^k - \alpha_k \mathbf{A}\mathbf{d}^k$
4. $\beta_k = \frac{(r^{k+1}, \mathbf{A}\mathbf{d}^k)}{(\mathbf{d}^k, \mathbf{A}\mathbf{d}^k)}$
5. $\mathbf{d}^{k+1} = r^{k+1} - \beta_k \mathbf{d}^k$

Bei exakter Arithmetik liefert das CG-Verfahren für ein n-dimensionales Problem eine Lösung nach (höchstens) n Schritten und kann daher prinzipiell als direktes Verfahren betrachtet werden.

Lemma 37 (CG als direkte Methode)**.** *Das CG-Verfahren bricht für jeden Startvektor $\mathbf{x}^0 \in \mathbb{R}^n$ bei rundungsfreier Rechnung nach spätestens n Schritten mit $\mathbf{x}^n = \mathbf{x}$ ab. In jedem Schritt gilt*

$$Q(\mathbf{x}^k) = \min_{\alpha\in\mathbb{R}} Q(\mathbf{x}^{k-1} + \alpha\mathbf{d}^{k-1}) = \min_{y\in\mathbf{x}^0+K_k} Q(\mathbf{y})$$

bzw.

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}^k\|_{\mathbf{A}^{-1}} = \min_{y\in\mathbf{x}^0+K_k} \|\mathbf{b} - \mathbf{A}\mathbf{y}\|_{\mathbf{A}^{-1}}$$

in der Norm

$$\|\mathbf{x}\|_{\mathbf{A}^{-1}} = (\mathbf{A}^{-1}\mathbf{x}, \mathbf{x})_2^{\frac{1}{2}}.$$

Proof. Die Eigenschaft, ein direktes Verfahren zu sein, folgt unmittelbar aus Hilfssatz 35.

Die Iterierte des CG-Verfahrens ist zunächst bestimmt als

$$Q(\mathbf{x}^k) = \min_{y \in \mathbf{x}^0 + K_k} Q(\mathbf{y}),$$

was gleichbedeutend ist mit (2.7). Mit dem Ansatz

$$\mathbf{x}^k = \mathbf{x}^0 + \sum_{k=0}^{k-1} \alpha_k \mathbf{d}^{k-1} = \mathbf{x}^0 + \underbrace{y^{k-1}}_{\in K_{t-1}} + \alpha_{t-1}\mathbf{d}^{k-1}$$

folgt

$$(\mathbf{b} - \mathbf{A}\mathbf{x}^k, \mathbf{d}^j) = (\mathbf{b} - \mathbf{A}\mathbf{y}^{k-1}, d_j) - \alpha_{t-1}(\mathbf{A}\mathbf{d}^{k-1}, \mathbf{d}^j) = 0 \quad \forall j = 0, \ldots, t-1$$

also

$$(\mathbf{b} - \mathbf{A}\mathbf{y}^{k-1}, d_j) = 0 \; \forall j = 0, \ldots, t-2,$$

und somit $y^{k-1} = \mathbf{x}^{k-1}$ sowie

$$Q(\mathbf{x}^k) = \min_{\alpha \in \mathbb{R}} Q(\mathbf{x}^{k-1} + \alpha \mathbf{d}^{k-1}).$$

Schließlich gilt mithilfe der Symmetrie $A = \mathbf{A}^T$

$$\begin{aligned}\|\mathbf{b} - \mathbf{A}\mathbf{y}\|_{\mathbf{A}^{-1}}^2 = (\mathbf{A}^{-1}[\mathbf{b} - \mathbf{A}\mathbf{y}], \mathbf{b} - \mathbf{A}\mathbf{y})_2 &= (\mathbf{A}\mathbf{y}, \mathbf{y})_2 - (\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}\mathbf{y})_2 - (\mathbf{y}, \mathbf{b})_2 \\ &= (\mathbf{A}\mathbf{y}, \mathbf{y})_2 - 2(\mathbf{b}, \mathbf{y})_2,\end{aligned}$$

also die Beziehung $\|\mathbf{b} - \mathbf{A}\mathbf{y}\|_{\mathbf{A}^{-1}}^2 = 2Q(\mathbf{y})$. □

Obwohl das CG-Verfahren prinzipiell eine direkte Methode ist, wird es in der Praxis als approximative, iterative Methode eingesetzt. Durch Rundungsfehler werden die Suchrichtungen $\{\mathbf{d}^0, \ldots, \mathbf{d}^{k-1}\}$ nie wirklich orthogonal sein. Die Konvergenzanalyse des CG-Verfahrens erweist sich als sehr aufwendig. Schlüssel ist die folgende Charakterisierung einer Iteration $\mathbf{x}^k = \mathbf{x}^0 + K_k$ als

$$\mathbf{x}^k = \mathbf{x}^0 + p_{k-1}(\mathbf{A})\mathbf{d}^0,$$

wobei $p_{k-1} \in P_{k-1}$ ein Polynom in $\mathbf{A}$ ist:

$$p_{k-1}(\mathbf{A}) = \sum_{i=0}^{k-1} \alpha_i \mathbf{A}^i$$

Hiermit kann die Minimierungseigenschaft aus Satz 37 geschrieben werden als

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}^k\|_{\mathbf{A}^{-1}} = \min_{y \in \mathbf{x}^0 + K_k} \|\mathbf{b} - \mathbf{A}\mathbf{y}\|_{\mathbf{A}^{-1}} = \min_{q \in P_{k-1}} \|\mathbf{b} - \mathbf{A}\mathbf{x}^0 - Aq(\mathbf{A})\mathbf{d}^0\|_{\mathbf{A}^{-1}}.$$

Wenn wir zur $\|\cdot\|_A$-Norm übergehen, folgt mit $\mathbf{d}^0 = \mathbf{b} - \mathbf{A}\mathbf{x}^0 = A(\mathbf{x} - \mathbf{x}^0)$

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}^k\|_{\mathbf{A}^{-1}} = \|\mathbf{x} - \mathbf{x}^k\|_A = \min_{q \in P_{k-1}} \|(\mathbf{x} - \mathbf{x}^0) - q(\mathbf{A})A(\mathbf{x} - \mathbf{x}^0)\|_A,$$

also

$$\|\mathbf{x} - \mathbf{x}^k\|_A = \min_{q \in P_{k-1}} \|[I - q(\mathbf{A})A](\mathbf{x} - \mathbf{x}^0)\|_A.$$

Im Sinne einer Bestapproximation können wir diese Aufgabe schreiben als

$$p \in P_{k-1}: \quad \|[I - p(\mathbf{A})A](\mathbf{x} - \mathbf{x}^0)\|_A = \min_{q \in P_{k-1}} \|[I + q(\mathbf{A})A](\mathbf{x} - \mathbf{x}^0)\|_A. \tag{2.12}$$

Gesucht ist eine Bestapproximation. Der Konvergenzbeweis zum CG-Verfahren baut daher auf dem entsprechenden Abschnitt **??** und insbesondere Abschnitt **??** auf. Es ist $q(\mathbf{A})\mathbf{A} \in P_k(\mathbf{A})$, also suchen wir ein Polynom $q \in P_k$ mit der Eigenschaft $q(0) = 1$, sodass

$$\|\mathbf{x}^k - \mathbf{x}\|_A \leqslant \min_{q \in P_k,\, q(0)=1} \|q(\mathbf{A})\|_A \|\mathbf{x} - \mathbf{x}^0\|_A. \tag{2.13}$$

Die Konvergenz des CG-Verfahrens hängt davon ab, ob es uns gelingt, ein Polynom $q \in P_k$ mit der Eigenschaft $p(0) = 1$ zu finden, mit möglichst kleiner Norm in $\mathbf{A}$. Wir zeigen:

Lemma 38 (Schranke für Matrixpolynome)**.** *Es sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit mit Eigenwerten $0 < \lambda_1 \leqslant \cdots \leqslant \lambda_n$, $p \in P_k$ ein Polynom mit $p(0) = 1$:*

$$\|p(\mathbf{A})\|_A \leqslant M, \quad M := \min_{p \in P_k,\, p(0)=1} \sup_{\lambda \in [\lambda_1, \lambda_n]} |p(\lambda)|.$$

Proof. Es sei $\{q_1, \ldots, q_n\}$ eine Orthogonalbasis aus Eigenvektoren. Ein beliebiges $y \in \mathbb{R}^n$ hat die Darstellung

$$y = \sum_{i=1}^{n} \gamma_i q_i.$$

Mit $Q = [q_1, \ldots, q_n]$ gilt

$$A = Q^T D Q, \quad D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

sodass für Polynome $p \in P_{k-1}$ folgt:

$$p(\mathbf{A}) = \sum_{i=0}^{k-1} \alpha_i \mathbf{A}^i = \sum_{i=0}^{k-1} \alpha_i (Q^T D Q)^i = Q^T \sum_{i=0}^{k-1} \alpha_i D^i Q = Q^T p(D) Q$$

Dann ist

$$\|p(\mathbf{A})y\|_{\mathbf{A}}^2 = \sum_{i=1}^{n} \lambda_i p(\lambda_i)^2 \gamma_i^2 \leqslant \underbrace{\sup_{\lambda \in [\lambda_1, \lambda_n]} |p(\lambda)|^2}_{=:M^2} \sum_{i=1}^{n} \lambda_i \gamma_i^2 =: M^2 \|\mathbf{y}\|_{\mathbf{A}}^2.$$

Und schließlich folgt

$$\|p(\mathbf{A})\|_A = \sup_{y \in \mathbb{R}^n,\, y \neq 0} \frac{\|p(\mathbf{A})y\|_A}{\|\mathbf{y}\|_A} = M.$$

□

Mit diesem Resultat und der Fehlerabschätzung (2.13) können wir nun eine Konvergenzabschätzung für das CG-Verfahren herleiten.

Lemma 39 (Konvergenz des CG-Verfahrens)**.** *Es sei* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *symmetrisch positiv definit, durch* $b \in \mathbb{R}^n$ *die rechte Seite und durch* $\mathbf{x}^0 \in \mathbb{R}^n$ *ein beliebiger Startwert gegeben. Dann gilt*

$$\|\mathbf{x}^k - \mathbf{x}\|_A \leqslant 2\left(\frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}}\right)^k \|\mathbf{x}^0 - \mathbf{x}\|_A, \quad k \geqslant 0,$$

mit der Spektralkondition $\kappa = \operatorname{cond}_2(\mathbf{A})$ *der Matrix* $\mathbf{A}$*.*

Proof. Aus dem Hilfssatz und der Abschätzung (2.13) folgt

$$\|\mathbf{x}^k - \mathbf{x}\|_A \leqslant M\|\mathbf{x}^0 - \mathbf{x}\|_A$$

mit

$$M = \min_{q\in P_k,\, q(0)=1} \max_{\lambda\in[\lambda_1,\lambda_n]} |q(\lambda)|.$$

Es gilt, eine möglichst scharfe Abschätzung für die Größe M zu finden. Gesucht ist ein Polynom $q \in P_k$, welches am Nullpunkt den Wert eins annimmt, $q(0) = 1$, und auf dem Bereich $[\lambda_1, \lambda_n]$ möglichst nahe (in der Maximumsnorm) an 0 liegt.

Hierzu verwenden wir die Tschebyscheff-Approximation. Wir suchen die beste Approximation $p \in P_k$ zur Nullfunktion auf $[\lambda_1, \lambda_n]$. Dieses Polynom soll zusätzlich die Normierungseigenschaft $p(0) = 1$ besitzen. Daher scheidet die triviale Lösung $p = 0$ aus. Das Tschebyscheff-Polynom (siehe Abschnitt **??** und Satz **??**)

$$T_k = \cos\big(k \arccos(x)\big)$$

hat die Eigenschaft

$$2^{-k-1} \max_{[-1,1]} |T_k(x)| = \min_{\alpha_0,\ldots,\alpha_{k-1}} \max_{[-1,1]} |x^k + \sum_{i=0}^{k-1} \alpha_i x^k|,$$

ist also bei Normierung des größten Monoms das Polynom, dessen Maximum auf $[-1, 1]$ minimal ist. Wir wählen nun die Transformation

$$x \mapsto \frac{\lambda_n + \lambda_1 - 2t}{\lambda_n - \lambda_1}$$

und erhalten mit

$$p(t) = T_k\left(\frac{\lambda_n + \lambda_1 - 2t}{\lambda_n - \lambda_1}\right) T_k\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)^{-1}$$

das Polynom von Grad k, welches auf $[\lambda_1, \lambda_n]$ minimal ist und der Normierung

$$p(0) = 1$$

genügt. Es gilt

$$\sup_{t\in[\lambda_1,\lambda_n]} |p(t)| = T_k\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right)^{-1} = T_k\left(\frac{\kappa+1}{\kappa-1}\right)^{-1} \tag{2.14}$$

mit der Spektralkondition

$$\kappa := \frac{\lambda_n}{\lambda_1}.$$

Wir nutzen die Darstellung der Tschebyscheff-Polynome außerhalb von $[-1,1]$ aus Satz **??**:

$$T_n(x) = \frac{1}{2}\Big[(x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^n\Big]$$

Für $x = \frac{\kappa+1}{\kappa-1}$ gilt

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

und entsprechend

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

Hiermit kann (2.14) abgeschätzt werden:

$$T_k\left(\frac{\kappa+1}{\kappa-1}\right) = \frac{1}{2}\Big[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\Big] \geqslant \frac{1}{2}\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k$$

Also folgt

$$\sup_{t\in[\lambda_1,\lambda_n]} T_k\left(\frac{\kappa+1}{\kappa-1}\right)^{-1} \leqslant 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k = 2\left(\frac{1-\frac{1}{\sqrt{\kappa}}}{1+\frac{1}{\sqrt{\kappa}}}\right)^k.$$

□

In Abbildung 2.3 zeigen wir für die Situation $\lambda_1 = 1$ und $\lambda_n = 4$ die entsprechenden Polynome für $n = 2, 3, 4$.

Die Konvergenz des CG-Verfahrens ist gerade doppelt so schnell wie die des Gradientenverfahrens. Es gilt

$$\rho := \frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}} = 1 - 2\sqrt{\kappa} + O(\kappa).$$

Für die Modellmatrix folgt $\rho = 1-\frac{1}{n}$. Das CG-Verfahren ist für dünn besetzte symmetrische Gleichungssysteme eines der effizientesten Iterationsverfahren. Die Konvergenz hängt wesentlich von der Kondition $\mathrm{cond}_2(\mathbf{A})$ der Matrix $\mathbf{A}$ ab. Für $\mathrm{cond}_2(\mathbf{A}) \approx 1$ ist das Verfahren optimal: Zur Reduktion des Fehlers um einen gegebenen Faktor ϵ ist eine feste Anzahl von Schritten notwendig.

Abschließend diskutieren wir anhand der Modellmatrix die verschiedenen Verfahren im Vergleich:

Example 40 (LGS mit der Modellmatrix)**.** *Es sei* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *mit* $n = m^2$ *die Modellmatrix*

$$A = \begin{pmatrix} B & -I & & \\ -I & B & -I & \\ & -I & B & -I \\ & & -I & B \end{pmatrix}, \quad B = \begin{pmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & -1 & 4 \end{pmatrix}, \quad I = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$$

mit $B, I \in \mathbb{R}^{m\times m}$*. Die Matrix* $\mathbf{A}$ *ist eine Bandmatrix mit Bandbreite* $2m$*. Weiter ist die Matrix* $\mathbf{A}$ *symmetrisch positiv definit. Sie ist irreduzibel und diagonaldominant und erfüllt in den Rändern der Blöcke das starke Zeilensummenkriterium. Alle bisher diskutieren Verfahren können auf die Matrix* $\mathbf{A}$ *angewendet werden. Größter sowie kleinster Eigenwert und Spektralkondition von* $\mathbf{A}$ *berechnen sich zu*

$$\lambda_{min} = \frac{2\pi^2}{n} + O(n^{-2}), \quad \lambda_{max} = 8 - \frac{2\pi^2}{n} + O(n^{-2}), \quad \kappa \approx \frac{4}{\pi^2} n \approx n.$$

Direkte Verfahren

Die LR-Zerlegung ist (da $\mathbf{A}$ *positiv definit) ohne Permutierung durchzuführen. Gemäß Satz* **??** *beträgt der Aufwand hierzu* $N_{LR} = nm^2 = 4n^2$ *Operationen. Die Lösung ist dann bis auf Rundungsfehlereinflüsse exakt gegeben. Alternativ kann die Cholesky-Zerlegung von* $\mathbf{A}$ *erstellt werden. Hierzu sind* $N_{LL} = 2n^2$ *Operationen notwendig. LR- bzw. Cholesky-Zerlegung sind dicht besetzte Bandmatrizen. Das anschließende Lösen mit Vorwärts- und Rückwärtselimination benötigt weitere* $2nm$ *Operationen. Wir fassen die notwendigen Operationen in folgender Tabelle zusammen:*

$n = m^2$	N_{LR}	N_{LL}
100	$5 \cdot 10^4$	$2 \cdot 10^4$
10000	$5 \cdot 10^9$	$2 \cdot 10^9$
1000000	$5 \cdot 10^{12}$	$2 \cdot 10^{12}$

Bei effizienter Implementierung auf moderner Hardware ist das Gleichungssystem mit $10\,000$ *Unbekannten in wenigen Sekunden, das größte Gleichungssystem in wenigen Stunden lösbar. Grundlegend für die effiziente Anwendung direkter Verfahren ist eine Vorsortierung der dünn besetzten Systeme, siehe Abschnitt* **??***.*

Einfache Fixpunktiterationen

Wir schätzen zunächst für Jacobi- sowie Gauß-Seidel-Verfahren die Spektralradien ab. Mit einem Eigenwert λ *und Eigenvektor* w *von* $\mathbf{A}$ *gilt*

$$\begin{aligned} Aw = \lambda w \quad &\Rightarrow \quad Dw + (L + R)w = \lambda w \\ &\Rightarrow \quad -D^{-1}(L + R)w = -D^{-1}(\lambda I - D)w. \end{aligned}$$

Das heißt, wegen $D_{ii} = 4$ *gilt*

$$Jw = \frac{\lambda - 4}{4} w$$

und die Eigenwerte von J *liegen zwischen*

$$\lambda_{min}(J) = \frac{1}{4}(\lambda_{min}(\mathbf{A}) - 4) \approx -1 + \frac{\pi^2}{2n},$$
$$\lambda_{max}(J) = \frac{1}{4}(\lambda_{max}(\mathbf{A}) - 4) \approx 1 - \frac{\pi^2}{2n}.$$

Minimaler und maximaler Eigenwert liegen jeweils nahe an -1 *bzw. an* 1*. Für die Konvergenzrate gilt*

$$\rho_J := \mathrm{spr}(J) = 1 - \frac{\pi^2}{2n}$$

und mit (**??**) *folgt für das Gauß-Seidel-Verfahren*

$$\rho_H := \rho_J^2 \approx 1 - \frac{\pi^2}{n}.$$

Zur Reduktion des Fehlers um einen gegebenen Faktor ϵ *sind* t *Schritte erforderlich:*

$$\rho_J^{t_J} = \epsilon \quad \Rightarrow \quad t_J = \frac{\log(\epsilon)}{\log(\rho)} \approx \frac{2}{\pi^2}\log(\epsilon)n, \quad t_H \approx \frac{1}{\pi^2}\log(\epsilon)n.$$

Jeder Schritt hat den Aufwand eines Matrix-Vektor-Produkts, d.h. im gegebenen Fall $5n$*. Schließlich bestimmen wir gemäß* (**??**) *den optimalen SOR-Parameter zu*

$$\omega_{opt} \approx 2 - 2\frac{\pi}{\sqrt{n}}.$$

Dann gilt

$$\rho_\omega = \omega_{opt} - 1 \approx 1 - 2\frac{\pi}{\sqrt{n}}.$$

Hieraus erhalten wir

$$t_\omega \approx \frac{\log(\epsilon)}{2\pi}\sqrt{n}.$$

Der Aufwand des SOR-Verfahrens entspricht dem des Gauß-Seidel-Verfahrens mit einer zusätzlichen Relaxation, d.h. $6n$ *Operationen pro Schritt. Wir fassen für die drei Verfahren Konvergenzrate, Anzahl der notwendigen Schritte und Gesamtaufwand zusammen. Dabei ist stets* $\epsilon = 10^{-4}$ *gewählt:*

	Jacobi		*Gauß-Seidel*		*SOR*		
$n = m^2$	t_J	N_j	t_H	N_H	ω_{opt}	t_J	N_J
100	*180*	10^5	*90*	$5 \cdot 10^4$	*1.53*	*9*	10^5
10000	*18500*	10^9	*9300*	$5 \cdot 10^8$	*1.94*	*142*	10^7
1000000	*1866600*	10^{13}	*933000*	$5 \cdot 10^{12}$	*1.99*	*1460*	10^9

Während das größte Gleichungssystem mit dem optimalen SOR-Verfahren in wenigen Sekunden gelöst werden kann, benötigen Jacobi- und Gauß-Seidel-Verfahren etliche Stunden.

Abstiegsverfahren

Schließlich betrachten wir Gradienten- und CG-Verfahren. Es gilt für die Kondition

$$\kappa(\mathbf{A}) = \frac{\lambda_{max}(\mathbf{A})}{\lambda_{min}(\mathbf{A})} = \frac{4n}{\pi^2} - 1,$$

also folgt für die Konvergenzraten von Gradienten und CG-Verfahren

$$\rho_{GR} \approx 1 - \frac{2}{\kappa} \approx 1 - \frac{\pi^2}{2n}, \quad \rho_{CG} \approx 1 - \frac{\pi}{\sqrt{n}}.$$

Bei effizienter Implementierung benötigt das Gradientenverfahren pro Iterationsschritt eine Matrix-Vektor-Multiplikation ($5n$ *Operationen*)*, zwei Skalarprodukte* ($2n$ *Operationen*) *sowie zwei Vektor-Additionen* ($2n$ *Operationen*)*, insgesamt somit* $9n$ *Operationen. Der Aufwand des CG-Verfahrens ist mit* $15n$ *Operationen etwas größer. Die Anzahl der Schritte bestimmt sich zu*

$$\rho_{GR}^t = 10^{-4} \quad \Rightarrow \quad t_{GR} \approx \frac{n}{\pi^2}, \quad t_{CG} \approx \frac{\sqrt{n}}{\pi}.$$

Wir fassen zusammen:

	Gradienten		*CG*	
$n = m^2$	t_{GR}	N_{GR}	t_{CG}	N_{CG}
100	*18*	10^4	*9*	10^4
10000	*2 330*	10^8	*140*	10^7
1000000	*233 300*	10^{12}	*1400*	10^{10}

Wie bereits bekannt, ist das Gradientenverfahren ebenso ineffektiv wie das Jacobi-Verfahren. Das CG-Verfahren erreicht etwa die gleiche Effizienz wie das SOR-Verfahren bei optimaler Wahl des Relaxationsparameters. Dieses Ergebnis darf nicht falsch interpretiert werden: Im Allgemeinen ist dieser Relaxationsparameter nicht verfügbar und muss grob approximiert werden. Das heißt, in der praktischen Anwendung wird das SOR-Verfahren sehr viel schlechter konvergieren und das CG-Verfahren ist im Allgemeinen überlegen.

Vorkonditionierung

Die Konvergenzrate des CG-Verfahrens hängt von der Konditionszahl der Systemmatrix ab. Es gilt

$$\rho_{CG} = \frac{1 - \frac{1}{\sqrt{\kappa}}}{1 + \frac{1}{\sqrt{\kappa}}} = 1 - \frac{2}{\sqrt{\kappa}} + O\left(\frac{1}{\kappa}\right).$$

Example 41. *Für Diskretisierungen von elliptischen partiellen Differentialgleichungen in zwei Dimensionen müssen wir z.B.* $\kappa = O(N)$ *erwarten, also*

$$\rho_{CG} \sim 1 - \frac{1}{\sqrt{N}},$$

sodass die Konvergenzrate mit der Problemgröße immer schlechter wird.

Die Idee der Vorkonditionierung ist eine Reformulierung des linearen Gleichungssystems. Hierzu sei $P \in \mathbb{P}^{n\times n}$ eine Matrix, welche die Schreibweise

$$P = KK^T$$

erlaubt. Dann gilt

$$\mathbf{A}x = b \quad \Leftrightarrow \quad \underbrace{K^{-1}A(K^T)^{-1}}_{=:\tilde{A}} \underbrace{K^T x}_{=:\tilde{x}} = \underbrace{K^{-1}b}_{=:\tilde{b}},$$

also

$$\tilde{\mathbf{A}}\tilde{x} = \tilde{b}.$$

Im Fall

$$\mathrm{cond}_2(\tilde{A}) \ll \mathrm{cond}_2(\mathbf{A})$$

und falls die Anwendung von K^{-1} "preiswert" ist, so kann durch Betrachtung des *vorkonditionierten Systems* $\tilde{\mathbf{A}}\tilde{x} = \tilde{b}$ eine wesentliche Beschleunigung erreicht werden. Die Bedingung $P = KK^T$ ist notwendig, damit die Matrix $\tilde{A}$ wieder symmetrisch ist.

Das CG-Verfahren mit Vorkonditionierung kann folgendermaßen formuliert werden:

Listing 2.4: CG-Verfahren mit Vorkonditionierung
Es sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit, $P = KK^T$ ein symmetrischer Vorkonditionierer.

```
1  Waehle Startwert x^0 ∈ R^n
2  r^0 = b − Ax^0
3  Pp^0 = r^0
4  d^0 = p^0
5  Fuer k von 0, 1, ...
6     α_k = (r^k, d^k) / (Ad^k, d^k)
7     x^{k+1} = x^k + α_k d^k
8     r^{k+1} = r^k − α_k A d^k
9     Pp^{k+1} = r^{k+1}
10    β_k = (r^{k+1}, p^{k+1}) / (r^k, g^k)
11    d^{k+1} = p^{k+1} + β_k d^k
```

In jedem Schritt der Verfahrens kommt als zusätzlicher Aufwand die Anwendung des Vorkonditionierers P hinzu. Bei der Auswahl des Vorkonditionierers ist darauf zu achten, dass er die Schreibweise $P = KK^T$ erlaubt - auch wenn die Teile K und K^T im Verfahren nicht genutzt

werden. Für die Wahl des Vorkonditionierers gelten ähnliche Bedingungen wie für die Wahl der Iterationsmatrix einfacher Fixpunktverfahren, die Bedingung

$$P \approx \mathbf{A}^{-1}$$

ist somit optimal für eine möglichst gute Kondition des vorkonditionierten Problems, wobei die Bedingung

$$P \approx I,$$

wesentlich für die effiziente Anwendung des Vorkonditionierers ist. Dementsprechend kommen als Vorkonditionierer auch die üblichen Approximationsverfahren zum Einsatz:

- *Jacobi-Vorkonditionierung*
  Wir wählen $P \approx D^{-1}$, wobei D der Diagonalanteil der Matrix $\mathbf{A}$ ist. Hier gilt

  $$D = D^{\frac{1}{2}}(D^{\frac{1}{2}})^T,$$

  das heißt, bei $D_{ii} > 0$ ist dieser Vorkonditionierer zulässig. Für die vorkonditionierte Matrix gilt

  $$\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \quad \Rightarrow \quad \tilde{a}_{ii} = 1$$

  und die Vorkonditionierung bewirkt eine Skalierung der Matrixeinträge. Diese Art der Vorkonditionierung kann die Kondition verbessern.

- *SSOR-Vorkonditionierung*
  Das SSOR-Verfahren ist eine symmetrische Variante des SOR-Verfahrens und basiert auf der Zerlegung

  $$P = (D + \omega L)D^{-1}(D + \omega R) = \underbrace{(D^{\frac{1}{2}} + \omega L D^{-\frac{1}{2}})}_{K}\underbrace{(D^{\frac{1}{2}} + \omega D^{-\frac{1}{2}} R)}_{=K^T},$$

  welche im Fall symmetrischer Matrizen $L = R^T$ die notwendige Bedingung an einen CG-Vorkonditionierer erfüllt.

  Für die Modellmatrix der Diskretisierung der Laplace-Gleichung kann bei optimaler Wahl von ω (welches eine nicht triviale Aufgabe ist) die Beziehung

  $$\mathrm{cond}_2(\tilde{A}) = \sqrt{\mathrm{cond}_2(\mathbf{A})}$$

  gezeigt werden. Die Konvergenzrate verbessert sich somit erheblich. Die Anzahl der notwendigen Schritte zum Erreichen einer Fehlerreduktion um den festen Faktor ϵ verbessert sich zu

  $$t_{CG}(\epsilon) = \frac{\log(\epsilon)}{\log(1 - \kappa^{-\frac{1}{2}})} \approx -\frac{\log(\epsilon)}{\sqrt{\kappa}}, \quad \tilde{t}_{CG}(\epsilon) = \frac{\log(\epsilon)}{\log(1 - \kappa^{-\frac{1}{4}})} \approx \frac{\log(\epsilon)}{\sqrt[4]{\kappa}}.$$

  Anstelle von z.B. 100 werden nur zehn Schritte benötigt. Die Bestimmung eines optimalen Parameters ω ist im Allgemeinen jedoch sehr schwer.

- *Unvollständige Cholesky-Zerlegung*
  Schließlich betrachten wir die Vorkonditionierung durch unvollständige Cholesky-Zerlegung. Zur symmetrisch positiv definiten Matrix $\mathbf{A}$ erstellen wir die approximative Zerlegung

  $$\mathbf{A} \approx C^T C,$$

  wobei die Zerlegung in C^T und C die gleiche Besetzungsstruktur der Matrix $\mathbf{A}$ verwendet. Das heißt, $C_{ij} \neq 0$ nur dann, wenn auch $A_{ij} \neq 0$. Es gilt

  $$A = C^T C + N,$$

  wobei die Größe (also die Zahl der Nicht-Nullen) und Bedeutung von N ganz wesentlich von der Sortierung der Matrix $\mathbf{A}$ abhängt (siehe Abschnitt **??**). Eine Analyse dieser Vorkonditionierung ist schwierig. In der Praxis zeigt sich jedoch, dass dieses Verfahren den anderen Methoden weit überlegen ist. Insbesondere kommt es ohne die Bestimmung des Parameters ω aus. Die Anwendung ist natürlich weitaus "teurer" als die Anwendung von Jacobi- oder SSOR-Vorkonditionierung.

Krylow-Teilraum-Verfahren für allgemeine Matrizen

Die CG-Iteration beruht stark auf der Symmetrie und positiven Definitheit der Matrix $\mathbf{A}$. Wir wollen hier zum Abschluss kurz Methoden vorstellen, die auch bei allgemeinen Matrizen verwendet werden können.

Es sei also $\mathbf{A} \in \mathbb{R}^{n \times n}$ eine reguläre, sonst aber allgemeine Matrix. Eine Symmetrisierung des Problems

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

kann einfach durch Multiplikation mit $\mathbf{A}^T$, also durch

$$\mathbf{A}^T \mathbf{A} x = \mathbf{A}^T b$$

erreicht werden. Die Matrix $B = \mathbf{A}^T \mathbf{A}$ ist positiv definit, da

$$(Bx, x)_2 = (\mathbf{A}^T \mathbf{A} x, x)_2 = (\mathbf{A}x, \mathbf{A}x)_2 = \|\mathbf{A}x\|_2,$$

und prinzipiell könnte das CG-Verfahren auf $\mathbf{A}^T \mathbf{A}$ angewendet werden. Anstelle einer sind dann zwei Matrix-Vektor-Multiplikationen pro Schritt durchzuführen. Gravierender ist die Auswirkung auf die Konvergenzrate, da

$$\kappa(B) = \mathrm{cond}_2(\mathbf{A}^T \mathbf{A}) = \mathrm{cond}_2(\mathbf{A})^2$$

gilt. Die Konvergenz verschlechtert sich somit wesentlich.

Die GMRES-Methode, *generalized minimal residual* überträgt die Idee des CG-Verfahrens auf allgemeine Matrizen. Aufbauend auf dem Krylow-Raum

$$K_k(\mathbf{d}^0, A) = \{\mathbf{d}^0, \mathbf{A}\mathbf{d}^0, \ldots, \mathbf{A}^{k-1}\mathbf{d}^0\}$$

wird zunächst eine Orthonormalbasis erstellt. Beim GMRES-Verfahren wird die Orthogonalität bzgl. des euklidischen Skalarprodukts erreicht:

$$(\mathbf{d}^i, \mathbf{d}^j)_2 = \delta_{ij}, \quad i, j = 0, \ldots, k-1$$

Die Näherung $\mathbf{x}^k \in \mathbf{x}^0 + K_k$ wird dann mithilfe der Galerkin-Gleichung

$$(\mathbf{b} - \mathbf{A}\mathbf{x}^k, \mathbf{A}\mathbf{d}^j)_2 = 0, \quad j = 0, \ldots, k-1 \tag{2.15}$$

berechnet. Bei allgemeinen Matrizen kann kein zweistufiges Orthogonalisierungsverfahren hergeleitet werden. Daher hat das GMRES-Verfahren einen weit höheren Aufwand als die CG-Iteration. Die Orthogonalbasis wird mit dem sogenannten *Arnoldi-Verfahren* entwickelt. Dieses erlaubt die Wiederverwendung der Faktoren, welche im Zuge der Orthonormalisierung berechnet wurden, zur Lösung der Galerkin-Gleichung (2.15). Die Orthogonalisierung geschieht üblicherweise mit Givens-Rotationen oder auf Basis von Householder-Transformationen.

Die Schwachstelle des GMRES-Verfahrens ist der wachsende Aufwand mit der Anzahl der Schritte, da die Orthogonalisierung stets bis zum Anfang zurückgeführt werden muss. In der Praxis wird das GMRES-Verfahren daher meistens mit *Restart* durchgeführt: Zur Orthogonalisierung wird stets nur eine feste Anzahl von Suchrichtungen N_o gespeichert. Nach jeweils N_o Schritten wird ein neuer Krylow-Raum aufgebaut und die Orthogonalisierung erneut gestartet.

Das GMRES-Verfahren ist die Standardmethode zur iterativen Lösung von großen linearen Gleichungssystemen mit dünn besetzter Matrix. Es lässt sich leicht mit Vorkonditionierung verbinden. Da die Matrix $\mathbf{A}$ sowieso nicht symmetrisch ist, erlaubt das GMRES-Verfahren eine sehr flexible Wahl von Vorkonditionierern.

Alternativ zum GMRES-Verfahren ist die *biconjugate gradient stabilized-Methode* (BiCGStab) die zweite gebräuchliche Iteration. Der Ansatz des BiCGStab-Verfahrens ist, Fehler in der Orthogonalität bewusst hinzunehmen. Das Verfahren basiert auf zwei kurzen Rekursionen für gestörte orthogonale Vektoren. Eine Analyse des Verfahrens ist schwierig. In der praktischen Anwendung erweist es sich als ähnlich effizient wie die GMRES-Iteration. Die einzelnen Schritte sind schneller, da kurze Rekursionsformeln genutzt werden können. Dafür ist die Konvergenzrate meist schlechter. Auch das BiCGStab-Verfahren kann mit Vorkonditionierern beschleunigt werden.

Verlässt man das Gebiet der positiv definiten Matrizen, so wird die Analyse von iterativen Lösern für lineare Gleichungssyteme schnell sehr aufwendig. Die Anzahl der Verfahren ist sehr vielfältig. Wir verweisen auf die Literatur. Eine algorithmische Darstellung der verschiedenen Verfahren findet sich in [**?**], eine umfangreiche Einführung und Analyse in [14] oder [9].

2.2.3 Verfahren für nicht-symmetrische Gleichungssysteme

Satz **??** ist Grundlage des CG-Verfahrens, da er die Äquivalenz zwischen linearem Gleichungssystem und Minimierung der quadratischen Form $Q(\cdot)$ herstellt. Dieser Satz gilt nur für symmetrische Matrizen. Allgemeiner erhalten wir:

Lemma 42 (Minimierung des Residuums)**.** *Es sei $\mathbf{A} \in \mathbb{R}^{N\times N}$ eine reguläre Matrix. Dann ist das lineare Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ äquivalent zur Minimierung des Residuums*

$$\|\mathbf{b} - \mathbf{Ax}\| = \min_{\mathbf{y}\in\mathbb{R}^N} \|\mathbf{b} - \mathbf{Ay}\|.$$

Proof. Sei $\mathbf{x}$ das Minimum des Residuums in der l_2-Norm. Dann gilt:

$$0 = \frac{d}{ds}\|\mathbf{b} - \mathbf{A}(\mathbf{x} + s\mathbf{y})\|^2\Big|_{s=0} \quad \forall \mathbf{y} \in \mathbb{R}^N,$$

und also

$$0 = \frac{d}{ds}\|\mathbf{b} - \mathbf{Ax} - s\mathbf{Ay}\|^2\Big|_{s=0} = 2\langle \mathbf{b} - \mathbf{Ax}, \mathbf{Ay}\rangle \quad \forall \mathbf{y} \in \mathbb{R}^N.$$

Da $\mathbf{A}$ regulär ist folgt $\langle \mathbf{b} - \mathbf{Ax}, \mathbf{y}\rangle = 0$ für alle $\mathbf{y} \in \mathbb{R}^N$ und schließlich $\mathbf{Ax} = \mathbf{b}$.

Umgekehrt sei nun $\mathbf{x}$ die Lösung des linearen Gleichungssystems. Dann gilt für $\mathbf{y}$ beliebig:

$$\begin{aligned}\|\mathbf{b} - \mathbf{Ay}\|^2 - \|\mathbf{b} - \mathbf{Ax}\|^2 &= \langle \mathbf{Ay}, \mathbf{Ay}\rangle - \langle \mathbf{Ax}, \mathbf{Ax}\rangle - 2\langle \mathbf{b}, \mathbf{Ay}\rangle + 2\langle \mathbf{b}, \mathbf{Ax}\rangle \\ &= \langle \mathbf{Ay}, \mathbf{Ay}\rangle + \langle \mathbf{Ax}, \mathbf{b}\rangle - 2\langle \mathbf{Ax}, \mathbf{Ay}\rangle = \|\mathbf{Ax} - \mathbf{Ay}\|^2 > 0.\end{aligned}$$

□

Anstelle der quadratischen Form soll jetzt direkt das Residuum der Gleichung minimiert werden. Mit dem Krylow-Raum

$$K_t := K_t(\mathbf{r}^{(0)}; \mathbf{A}) = \operatorname{span}\{\mathbf{r}^{(0)}, \mathbf{Ar}^{(0)}, \ldots, \mathbf{A}^{t-1}\mathbf{r}^{(0)}\},$$

wird ein $\mathbf{x}^{(t)} \in \mathbf{x}^{(0)} + K_t$ gesucht, so dass gilt:

$$\|\mathbf{b} - \mathbf{Ax}^{(t)}\|^2 = \min_{\mathbf{y}\in K_t} \|\mathbf{b} - \mathbf{Ay}\|^2. \tag{2.16}$$

Für dieses Minimum gilt wieder eine Orthogonalitätsbeziehung:

Lemma 43 (Galerkin-Gleichung)**.** *Die Lösung $\mathbf{x}^{(t)} \in \mathbf{x}^{(0)} + K_t$ der Minimierungsaufgabe (2.16) ist eindeutig durch die Galerkin-Gleichung beschrieben:*

$$\langle \mathbf{b} - \mathbf{Ax}^{(t)}, \mathbf{Ay}\rangle = 0 \quad \forall \mathbf{y} \in K_t. \tag{2.17}$$

Proof. Übung. □

Zum Ansatz sei zunächst eine orthogonale Basis $\{\mathbf{d}^{(0)}, \ldots, \mathbf{d}^{(t-1)}\}$ des K_t gegeben. Durch

$$\mathbf{Q}_t := [\mathbf{d}^{(0)}, \ldots, \mathbf{d}^{(t-1)}] \in \mathbb{R}^{N \times t},$$

ist eine Matrix mit $\mathbf{Q}^T\mathbf{Q} = I$ gegeben. Jede Lösung $\mathbf{x}^{(t)} \in \mathbf{x}^{(0)} + K_t$ kann dann in der Form

$$\mathbf{x}^{(t)} = \mathbf{x}^{(0)} + \mathbf{Q}_t\mathbf{y}^{(t)}, \quad \mathbf{y}^{(t)} \in \mathbb{R}^t$$

geschrieben werden. Wir setzen diesen Ansatz in die Galerkin-Gleichung (2.17) ein und erhalten:

$$\langle \mathbf{r}^{(0)} - \mathbf{A}\mathbf{Q}_t\mathbf{y}^{(t)}, \mathbf{A}\mathbf{Q}\mathbf{z}^{(i)} \rangle = 0 \quad \forall \mathbf{z}^{(i)} \in \mathbb{R}^t.$$

Um die gesuchte Approximation $\mathbf{x}^{(t)} = \mathbf{Q}_t\mathbf{y}^{(t)}$ zu finden, muss also nur ein lineares Gleichungssystem mit t Unbekannten und Gleichungen gelöst werden.

Diese Idee ist Grundlage des *Generalized Minimal Residual Verfahrens* (GMRES). In einem ersten Schritt wird eine Orthonormalbasis des K_t erstellt. Anschließend wird das lineare Gleichungssystem im $\mathbb{R}^t$ gelöst um schließlich mittels $\mathbf{x}^{(t)} = \mathbf{Q}_t\mathbf{y}^{(t)}$ die Approximation zu erhalten.

Für allgemeine Matrizen $\mathbf{A}$ sind Konvergenzaussagen schwer zu treffen. Wegen der Herleitung über die Minimierung des Residuums konvergiert dieses aber monoton, wie bei dem CG-Verfahren. Im Fall fehlerfreier Arithmetik ist das GMRES-Verfahren ein direktes Lösungsverfahren. In der Anwendung werden jedoch stets nur wenige Schritte durchgeführt. Kern des Verfahrens ist die Orthogonalisierung des Krylow-Raums K_t. Je größer der Raum K_t umso aufwändiger ist auch die Orthogonalisierung. Aus diesem Grund wird üblicherweise nur eine feste Zahl m von Schritten des GMRES-Verfahrens ausgeführt und anschließend mit einer besseren Startlösung $\mathbf{x}^{(0)'} = \mathbf{x}^{(m)}$ neu gestartet. (GMRES-Verfahren mit *Restart*).

2.2.4 Vorkonditionierung

Satz **??** sagt, dass die Konvergenzgeschwindigkeit des CG-Verfahrens im Wesentlichen von der Spektralkondition der Matrix $\mathbf{A}$ abhängt. Die Idee der *Vorkonditionierung* ist es, das Gleichungssystem durch Multiplikation einer Matrix $\mathbf{P}^{-1}$ derart zu ändern, so dass für die Matrix $\mathbf{P}^{-1}\mathbf{A}$ gilt:

$$\mathrm{cond}_2(\mathbf{P}^{-1}\mathbf{A}) \ll \mathrm{cond}_s(\mathbf{A}).$$

Dann wird anstelle von $\mathbf{A}\mathbf{x} = \mathbf{b}$ das System

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{x} = \mathbf{P}\mathbf{b},$$

gelöst, welches wegen der weitaus besseren Spektralkondition von $\mathbf{P}^{-1}\mathbf{A}$ schneller konvergiert. Wir gehen zunächst davon aus, dass die Matrix $\mathbf{A}$ symmetrisch positiv definit ist und mit dem CG-Verfahren gelöst werden soll. Das vorkonditionierte System muss dann auch symmetrisch positiv definit sein. Der symmetrisch positiv definite Vorkonditionierer $\mathbf{P}$ liege also in der Form:

$$\mathbf{P} = \mathbf{K}\mathbf{K}^T,$$

vor. Dann lösen wir:

$$\mathbf{K}^{-T}\mathbf{K}^{-1}\mathbf{A}(K^{-T}K^{T})\mathbf{x} = \mathbf{K}^{-T}\mathbf{K}^{-1}\mathbf{b} \quad \Rightarrow \quad \underbrace{\mathbf{K}^{-1}\mathbf{A}K^{-T}}_{=:\tilde{\mathbf{A}}}\underbrace{K^{T}\mathbf{x}}_{=\tilde{\mathbf{x}}} = \underbrace{\mathbf{K}^{-1}\mathbf{b}}_{=\tilde{\mathbf{b}}}.$$

Für die Matrix $\tilde{A}$ gilt:

$$\mathbf{K}^{-T}\tilde{\mathbf{A}}\mathbf{K}^{T} = (\mathbf{K}^{-T}\mathbf{K}^{-1})\mathbf{A}(\mathbf{K}^{-T}\mathbf{K}^{T}) = \mathbf{P}^{-1}\mathbf{A}.$$

Die Matrix $\tilde{\mathbf{A}}$ ist also ähnlich zur Matrix $\mathbf{P}^{-1}\mathbf{A}$. Ähnliche Matrizen haben die gleichen Eigenwerte. Im Fall $\mathbf{P} = \mathbf{A}$ ist die Matrix $\tilde{\mathbf{A}}$ also ähnlich zur Einheitsmatrix mit der Kondition $\mathrm{cond}_2(I) = 1$. In diesem Fall wäre das Konvergenzverhalten des CG-Verfahrens optimal.

Im Folgenden formulieren wir das sogenannte *Vorkonditionierte CG-Verfahren* oder *Preconditioned Conjugate Gradient* (PCG):

Algorithm 44 (PCG-Verfahren)**.** *Sei* $\mathbf{P} = \mathbf{K}\mathbf{K}^{T}$ *ein Vorkonditionierer sowie* $\mathbf{x}^{(0)} \in \mathbb{R}^{N}$ *ein Startwert. Setze* $\mathbf{d}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$ *sowie* $\mathbf{z}^{(0)} = \mathbf{P}^{-1}\mathbf{d}^{(0)}$*. Iteriere*

$$\begin{aligned}
&(i) && \alpha_t = \frac{\langle \mathbf{r}^{(t)}, \mathbf{z}^{(t)} \rangle}{\langle \mathbf{A}\mathbf{d}^{(t)}, \mathbf{d}^{(t)} \rangle} \\
&(ii) && \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \alpha_t \mathbf{d}^{(t)} \\
&(iii) && \mathbf{r}^{(t+1)} = \mathbf{r}^{(t)} - \alpha_t \mathbf{A}\mathbf{d}^{(t)} \\
&(iv) && \mathbf{z}^{(t+1)} = \mathbf{P}^{-1}\mathbf{r}^{(t+1)} \\
&(v) && \beta_t = \frac{\langle \mathbf{r}^{(t+1)}, \mathbf{z}^{(t+1)} \rangle}{\langle \mathbf{r}^{(t)}, \mathbf{z}^{(t)} \rangle} \\
&(vi) && \mathbf{d}^{(t+1)} = \mathbf{r}^{(t+1)} + \beta_t \mathbf{d}^{(t)}.
\end{aligned}$$

Im Vergleich zum einfachen CG-Verfahren fällt als zusätzlicher Aufwand insbesondere das Lösen des Vorkonditionierersystems $\mathbf{P}\mathbf{z}^{(t+1)} = \mathbf{r}^{(t+1)}$ an. Hierzu kann die Zerlegung $\mathbf{P} = \mathbf{K}\mathbf{K}^{T}$ genutzt werden. Der Vorkonditionierer sollte also einerseit möglichst einfach zu invertieren sein, auf der anderen Seite sollte $\mathbf{P} \approx \mathbf{A}$, insbesondere sollten das Eigenwerte von $\mathbf{P}^{-1}\mathbf{A}$ möglichst nahe beieinander liegen.

Jacobi-Vorkonditionierung Für $\mathbf{P}$ eignet sich z.B. das Jacobi-Verfahren mit

$$\mathbf{P}_J = \mathrm{diag}(\mathbf{A}) =: \mathbf{D}, \quad \mathbf{P}_J = \mathbf{D}^{\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}.$$

Dieser Vorkonditionierer ist sehr "billig" anzuwenden und sorgt dafür, dass die Einträge der Matrix $\mathbf{A}$ skaliert werden. Insbesondere gilt $\tilde{a}_{ii} = 1$ für $i = 1, \ldots, N$. Dies kann zu einer Reduktion der Kondition führen. Mit Hilfe der Gerschgorin-Kreise kann eine einfache Abschätzung für die Eigenwerte gefunden werden. Bei $\tilde{A}$ liegt der Mittelpunkt stets bei 1.

SOR-Vorkonditionierung Das Gauß-Seidel Verfahren kann nicht in der Form $\mathbf{P} = \mathbf{K}\mathbf{K}^T$ faktorisiert werden, es ist nicht symmetrisch. Die Matrix $\mathbf{A}$ sei additiv zerlegt in $\mathbf{A} = \mathbb{L} + \mathbb{D} + \mathbb{R} = \mathbb{L} + \mathbb{D} + \mathbb{L}^T$, da $\mathbf{A}$ symmetrisch. Das SOR-Verfahren hat die Iterationsmatrix:

$$P_{SOR} = (\mathbf{D} + \omega\mathbf{L})\mathbb{D}^{-1}(\mathbf{D} + \omega\mathbf{L}^T) = \underbrace{(\mathbf{D}^{\frac{1}{2}} + \omega\mathbf{L}\mathbf{D}^{-\frac{1}{2}})}_{K}\underbrace{(\mathbf{D}^{\frac{1}{2}} + \omega\mathbf{D}^{-\frac{1}{2}}\mathbf{L}^T)}_{K^T}$$

Bei optimaler Wahl der Relaxationsparameters ω (dies ist im Allgemeinen jedoch nicht möglich) wird eine wesentliche Reduktion der Kondition erreicht:

$$\mathrm{cond}_2(\tilde{\mathbf{A}}) = \sqrt{\mathrm{cond}_2(\mathbf{A})}.$$

Unvollständige Cholesky-Vorkonditionierung Abschließend stellen wir noch ein schwer zu analysierendes, jedoch höchst erfolgreiches Verfahren vor. Durch

$$\tilde{\mathbf{A}} \approx \tilde{\mathbf{C}}\tilde{\mathbf{C}^T},$$

sei die *unvollständige Cholesky-Zerlegung* von $\mathbf{A}$ gegeben. Die Cholesky-Zerlegung als Spezialfall der LR-Zerlegung ist üblicherweise voll besetzt. Mit $\tilde{\mathbf{C}}\tilde{\mathbf{C}^T}$ bezeichnen wir die Approximative Zerlegung der Matrix $\mathbf{A}$. Elemente c_{ij} von $\tilde{\mathbf{C}}$ werden künstlich auf Null gesetzt, wenn für den Matrixeintrag $a_{ij} = 0$ gilt.

Durch gute Vorkonditionierung, etwa mit dem SOR-Verfahren oder der Cholesky-Zerlegung kann die Konditionierung des vorkonditionierten Systems (bei der Poisson-Gleichung) auf $\mathrm{cond}(\tilde{\mathbf{A}}) = O(h^{-1})$ verbessert werden. Hierdurch wird die Konvergenzrate des CG-Verfahrens von $1 - O(h)$ auf $1 - O(\sqrt{h})$ verbessert. Dies führt zu einem Gesamtaufwand von $O(N^{\frac{5}{4}})$.

2.3 Mehrgitterverfahren

Alle bisher betrachteten iterativen Lösungsverfahren hängen von der Kondition der Matrix und damit bei der Behandlung der FE Diskretisierung der Poisson-Gleichung vom Gitter.

Im Folgenden betrachten wir die eindimensionale Diskretisierung mit linearen Finiten Elementen auf einem uniformen Gitter mit N Elementen. Als prototypisches Iterationsverfahren werden wir die gedämpfte Richardson-Iteration genauer untersuchen. Es ist:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \theta(\mathbf{b} - \mathbf{A}\mathbf{x}^{(t)}) = \underbrace{(I - \theta\mathbf{A})}_{=:\mathbf{B}_\theta}\mathbf{x}^{(t)} + \theta\mathbf{b},$$

mit der Iterationsmatrix $\mathbf{B}_\theta$.

Die positiv definite Matrix $\mathbf{A}$ hat in *Stencil-Notation* die Form:

$$\mathbf{A} = \begin{bmatrix} -1 & 2 & -1 \end{bmatrix}.$$

Diese Matrix hat die Eigenvektoren $\omega_k \in \mathbb{R}^{N+1}$ mit Eigenwerten λ_k:

$$\begin{aligned} \lambda_k &= 2\left(1 - \cos\left(\frac{k}{N}\pi\right)\right) \\ (\omega_k)_i &= \sin\left(\frac{ki}{N}\pi\right), \quad i = 0, \ldots, N. \end{aligned} \tag{2.18}$$

Die Konvergenzordnung der Iterativen Verfahren ist im Allgemeinen sehr langsam, sie hängt vom Spektralradius der Iterationsmatrix $\mathbf{B} := I - \theta\mathbf{A}$ ab. Es gilt wieder in Stencil-Notation

$$\mathbf{B}_\theta = \begin{bmatrix} \theta & (1-2\theta) & \theta \end{bmatrix}$$

Wir wählen nun $\theta = 4^{-1} \approx \lambda_{max}(\mathbf{A}_h)^{-1}$. Dann ist

$$\mathbf{B} = \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$$

Diese Matrix $\mathbf{B} := \mathbf{B}_{\frac{1}{4}}$ hat wieder die gleichen Eigenvektoren (2.18) diesmal mit den Eigenwerten:

$$\lambda_k^B = \frac{1}{2}\left(1 + \cos\left(\frac{k}{N}\pi\right)\right)$$

Die Eigenwerte liegen alle im Intervall $\lambda_k^B \in (0, 1)$, der größte Eigenwert von $\mathbf{B}$ verhält sich wie $\lambda_{max}(\mathbf{B}) = 1 - O(h^2)$.

Wir untersuchen die Konvergenz des Richardson-Verfahrens für die Poisson-Gleichung im Detail. Es sei also $\mathbf{x}^{(t)}$ die letzte Approximation mit Fehler $\mathbf{e}^{(t)} := \mathbf{x} - \mathbf{x}^{(t)}$. Für den Fehler in der nächsten Iteration gilt:

$$\mathbf{e}^{(t+1)} = \mathbf{B}\mathbf{e}^{(t)}.$$

Wir schreiben den Vektor $\mathbf{e}^{(t)}$ in einer Entwicklung in Eigenwerten

$$\mathbf{e}^{(t)} = \sum_{k=1}^{N-1} \mathbf{e}_k^{(t)} \omega_k.$$

Dabei sind die $\mathbf{e}_k^{(t)} \in \mathbb{R}$ die Entwicklungskoeffizienten in der Eigenvektor-Darstellung und nicht die kartesischen Koeffizienten. Wir nennen (wegen der Sinus-Form der Eigenvektoren) die einzelnen Komponenten ω_k die *Frequenzen* des Fehlers. Wir können nun die Wirkung des Richardson-Verfahrens für jede einzelne Frequenz ω_k verfolgen. Es gilt

$$\mathbf{e}_k^{(t+1)} = \lambda_k^B \mathbf{e}_k^{(t)},$$

das heißt, die k-te Komponente des Fehlers wird genau um den Eigenwert λ^B gedämpft. In Abbildung 2.4 zeigen wir den Verlauf der Eigenwerte λ_k^B. Fehlerkomponenten, welche zu niedrigen Frequenzen gehören werden nur sehr langsam reduziert, während alle hochfrequenten Anteile schnell reduziert werden. Wir definieren:

Definition 45 (Fehlerfrequenzen)**.** *Komponenten welche zu Eigenwerten λ_k für $k > \frac{N}{2}$ gehören heißen* hochfrequente Anteile*, alle anderen Komponenten heißen* niederfrequent*. Hochfrequente Anteile sind gerade die, welche auf gröberen Gittern nicht dargestellt werden können.*

Das Richardson-Verfahren ist zwar ein schlechter Löser für die Poisson-Gleichung, es *glättet* hochfrequente Fehleranteile allerding sehr schnell aus. In Abbildung 2.5 zeigen wir den Fehler der Richardson-Iteration über einige Schritte. Der Gesamtfehler wird nach zehn Schritten nicht wesentlich kleiner, allerdings ist der Fehler bereits nach zwei Schritten stark geglättet. Das Richardson-Verfahren scheint also in der Lage zu sein, lokale Fehleranteile sehr schnell zu *glätten*.

Das Mehrgitterverfahren beruht nun auf der Idee, dass die Frage, ob ein Fehleranteil hochfrequent ist oder nicht vom zugrundeliegenden Gitter abhängt. Auf einem Gitter mit $N = 10$ Elementen ist die Frequenz

$$(\omega_4)_i = \sin\left(\frac{4i}{10}\pi\right), \quad i = 0, \ldots, 10,$$

niederfrequent, auf einem gröberen Gittern mit $N' = 5$ Elementen ist die gleiche Schwingung jedoch hochfrequent:

$$(\omega_4)'_i = \sin\left(\frac{4i}{5}\pi\right), \quad i = 0, \ldots, 5.$$

In Abbildung 2.6 ist diese Frequenz auf beiden Gittern aufgetragen. Wir fassen zusammen:

- Einfache Iterationsverfahren wie die Richardson-Iteration sind schlechte Löser aber gute Glätter für hochfrequente Fehleranteiler.
- Niederfrequente Anteile auf einem Gitter Ω_h sind hochfrequente Anteile auf einem gröberen Gitter Ω_{2h}.
- Und ganz trivial: je gröber das Gitter, umso geringer der Aufwand.

Beim Mehrgitter-Verfahren sollen diese Leitsätze kombiniert werden. Auf dem feinsten Gitter Ω_h wird das Gleichungssystem nicht gelöst, es werden zunächst nur hochfrequente Fehleranteile ausgeglättet und es bleiben nur niederfrequente Fehler $\mathbf{e}_h \to \mathbf{e}_h^{nf}$. Diese werden auf ein gröberes Gitter Ω_H transferiert $\mathbf{e}_h^{nf} \to \mathbf{e}_H^{hf}$, wo sie wieder hochfrequent sind. Auf diesem Grobgitter ist das verbleibende Problem sehr viel kleiner und kann einfacher gelöst werden.

2.3.1 Hierarchische Finite Elemente Ansätze

Wir müssen zur Notation zunächst einige Begriffe einführen. Durch

$$\Omega_H = \Omega_0, \Omega_1, \ldots, \Omega_L = \Omega_h,$$

sei eine Familie von Gittern des Gebiets Ω gegeben. Wir definieren:

Definition 46 (Geschachtelte Gitter). *Seien Ω_H, Ω_h zwei Triangulierungen des Gebiets Ω. Die Gitter heißen geschachtelt $\Omega_H \Subset \Omega_h$, falls für jeden Knoten $x_i \in \Omega_H$ gilt $x_i \in \Omega_h$ und jedes Element $K \in \Omega_h$ durch Verfeinerung eines Elementes $K' \in \Omega_H$ entstanden ist.*

In Abbildung 2.7 zeigen einfache Beispiele von Gittern die geschachtelt sind und Gittern, die nicht geschachtelt sind. Auf jedem Gitter Ω_l sei durch V_l ein Finite Elemente Ansatzraum definiert und wir definieren die lokalen Probleme:

$$u_l \in V_l : \quad a(u_k, \phi_l) = (f, \phi_l) \quad \forall \phi_l \in V_l$$

Als kompakte Schreibweise definieren wir einen Operator $\mathcal{A}_l : V_l \to V_l$ mittels

$$(\mathcal{A}_l u_l, v_l) = a(u_l, v_l) \quad \forall u_l, v_l \in V_l.$$

Mit der Vektordarstellung

$$u_l = \sum_{i=1}^{N_l} \mathbf{u}_l^i \phi_h^{(i)},$$

und der Steifigkeitsmatrix $\mathbf{A}_l$ ist dies Problem äquivalent zu dem linearen Gleichungssystem

$$\mathbf{A}_l \mathbf{u}_l = \mathbf{b}_l, \quad (\mathbf{A}_l)_{ij} = a(\phi_h^{(j)}, \phi_h^{(i)}), \quad (\mathbf{b}_l)_i = (f, \phi_h^{(i)}).$$

Es gilt:

Lemma 47 (Geschachtelte Finite Elemente Räume). *Es seien $\Omega_H \Subset \Omega_h$ geschachtelte Gitter und V_H sowie V_h isoparametrische Finite Elemente Räume mit dem gleichen Finite Elemente Ansatz $\{P(\hat{T}), \chi(\hat{T})\}$ auf den Gittern Ω_H bzw. Ω_h. Es gilt:*

$$V_H \subset V_h.$$

Proof. Übung. □

Definition 48 (Gittertransfer). *Es seien $V_{l-1} \subset V_l$ zwei geschachtelte Finite-Elemente Räume. Für eine Funktion $v_h \in V_h$ definieren wir den* Restriktionsoperator $\mathcal{R}_{l-1} : V_l \to V_{l-1}$ *durch*

$$(\mathcal{R}_{l-1} v_h, \phi_H) = (v_h, \phi_H) \quad \forall \phi_H \in V_H,$$

sowie für eine Funktion $v_H \in V_H$ den Prolongationsoperator $\mathcal{P}_l : V_{l-1} \to V_l$ *durch*

$$\mathcal{P}_l v_H = v_H.$$

Remark 49 (Eigenschaften der Restriktion). *Seien $\Omega_H \Subset \Omega_h$ zwei geschachtelte Gitter. Die Restriktion $\mathcal{R}_H : V_h \to V_H$ ist gerade die L^2-Projektion von V_h in den gröberen Raum V_H. Die L^2-Projektion ist eine globale Operation, um sie zu berechnen muss ein lineares Gleichungssystem mit der Massenmatrix $\mathbf{M}_H$ gelöst werden. Obwohl ein lineares Gleichungssystem mit der Massenmatrix relativ einfach zu lösen ist (da $cond_2(\mathbf{M}_H) = O(1)$) sollte dies aus Effizienzgründen vermieden werden.*

In der Anwendung des Mehrgitterverfahrens werden wir allerdings gar nicht die Finite Elemente Funktion $\mathcal{R}_H v_h \in V_H$ benötigen, sondern nur den Vektor $\mathbf{x}_H \in \mathbb{R}^{N_H}$ mit

$$(\mathbf{x}_H)_i := (\mathcal{R}_H v_h, \phi_H^{(i)}) = (v_h, \phi_H^{(i)}), \quad i = 1, \dots, N_H. \tag{2.19}$$

Zu der beliebigen Funktion $v_h \in V_h$ definieren wir auf dem feinen Gitter Ω_h den Vektor $\mathbf{x}_h \in \mathbb{R}^{N_h}$ als

$$(\mathbf{x}_h)_i = (v_h, \phi_h^{(i)}).$$

Die Räume $V_H \subset V_h$ sind geschachtelt, d.h., jede Knotenbasisfunktion $\phi_H^{(i)} \in V_H$ ist auch im feinen Raum $\phi_H^{(i)} \in V_h$ und hat dort die Basisdarstellung

$$\phi_H^{(i)} = \sum_{j=1}^{N_h} \mu_{ij} \phi_h^{(j)}, \tag{2.20}$$

mit einer Koeffizientenmatrix $\mathbf{R}_H \in \mathbb{R}^{N_H \times N_h}$ mit $(\mathbf{R}_H)_{ij} = \mu_{ij}$ welche sehr dünn besetzt ist, also mit $\mu_{ij} \neq 0$ nur für sehr wenige (insbesondere unabhängig von h und H) Einträge. In Abbildung 2.8 zeigen wir die Kombination einer Basisfunktion $\phi_H^{(i)}$ durch drei Funktionen des feinen Gitters.

Die Einträge des gesuchten Vektors $\mathbf{x}_H$ aus (2.19) sind dann bestimmt durch die Gleichungen:

$$(\mathbf{x}_H)_i = (v_h, \phi_H^{(i)}) = \sum_{j=1}^{N_h} \mu_{ij} \underbrace{(v_h, \phi_h^{(j)})}_{=:(\mathbf{x}_h)_j} = (\mathbf{R}_H \mathbf{x}_h)_i.$$

Das heißt, wenn nicht die Funktion $\mathcal{R}_H v_h$ von Interesse ist, sondern nur die Funktion r_h getestet mit den Basisfunktionen $\phi_h^{(i)}$, also ein Vektor $\mathbf{x}_h = ((r_h, \phi_h^{(i)}))_{i=1}^{N_h}$, dann besteht der einfache Zusammenhang

$$\mathbf{x}_H = \mathcal{R}_H \mathbf{x}_h.$$

Ebenso betrachten wir nun

Remark 50 (Eigenschaften der Prolongation)**.** *Die Prolongation $\mathcal{P}_h : V_H \to V_h$ ist die Identität auf V_H. Mit der Darstellung (2.20) erhalten wir für jeden Vektor $v_H \in V_H$ unmittelbar*

$$\begin{aligned} v_H(x) &= \sum_{i=1}^{N_H} (\mathbf{v}_H)_i \phi_H^{(i)}(x) = \sum_{i=1}^{N_H} \sum_{j=1}^{N_h} \mu_{ij} (\mathbf{v}_H)_i \phi_h^{(j)}(x) \\ &= \sum_{j=1}^{N_h} \underbrace{\left(\sum_{i=1}^{N_H} \mu_{ij} (\mathbf{v}_H)_i \right)}_{=:(\mathbf{v}_h)_j} \phi_h^{(j)}(x) \end{aligned}$$

Für den Vektor $\mathbf{v}_h \in \mathbb{R}^{N_h}$ gilt also

$$\mathbf{v}_h = \mathbf{R}_H^T \mathbf{v}_H.$$

Die Wirkung der Prolongation auf einen Knotenvektor ist also gerade durch die transponierte Matrix der Restriktion eines integrierten Vektors beschrieben. Man beachte, dass die Hintereinanderausführung von Restriktion und Prolongation nicht die Identität ergibt!

Der enge Zusammenhang zwischen Formulierungen in den Funktionenräumen V_h und V_H sowie zwischen den Vektorräumen $\mathbb{R}^{N_H}$ sowie $\mathbb{R}^{N_h}$ ist typisch für die Analyse des Mehrgitterverfahrens.

2.3.2 Das Zweigitter-Verfahren

Als Vorstufe zum Mehrgitterverfahren beschreiben wir zunächst die *Zweigitteriteration*. Diese formulieren wir im Finite-Elemente Kontext:

Algorithm 51 (Zweigitteriteration in Finite-Elemente Räumen)**.** *Seien $V_H \subset V_h$ zwei geschachtelte FE-Räume sowie $u_h^{(0)} \in V_h$ eine Approximation der Lösung $u_h \in V_h$. Weiter sei $\omega \in (0,1]$ ein Dämpfungsparameter und durch $\mathcal{G}_h : V_h \to V_h$ eine Glättungsiteration gegeben. Iteriere:*

$$u_h^{(t+1)} = \mathcal{ZG}(u_h^{(t)}, f_h),$$

mit der Zweigitter-Iteration:

(i)	*Vorglätten:*	$\bar{u}_h = \mathcal{G}_h^{\nu_1}(u_h^{(t)}) := \mathcal{G}_h^{\nu_1}(u_h^{(t)}, a(\cdot,\cdot), f)$
(ii)	*Grobgitterproblem:*	$w_H \in V_H : \quad a(\bar{u}_h + w_H, \phi_H) = (f, \phi_H) \quad \forall \phi_H \in V_H$
(iii)	*Nachglätten:*	$u_h^{(t+1)} = \mathcal{G}_h^{\nu_2}(\bar{u}_h + \omega w_H)$

Das Verfahren besteht aus zwei Schritten: zunächst werden im feinen Finite Elemente Raum V_h die hochfrequenten Fehleranteile geglättet. Im Anschluss wird durch die Grobgitter-Korrektur das Problem im groben Raum $V_H \subset V_h$ approximiert. Schließlich kann ein weiteres Mal geglättet werden. Die Approximation wird in zwei Stufen aufgeteilt: die hohen Fehlerfrequenzen werden mit einigen wenigen (üblicherweise $\nu \sim 3$) Schritten eines einfachen Iterationsverfahrens reduziert, alle niedrig frequenten Fehleranteile werden auf dem Grobgitter behandelt.

Remark 52 (Nachglätten)**.** *Die Zweigitteriteration (und später auch die Mehrgitteriteration) konvergieren auch ohne Nachglättung nur mit Vorglättung. Auch für die Beweise ist die Nachglättung nicht wesentlich. Wir setzen daher im Folgenden ohne Einschränkung $\nu_2 = 0$.*

Im Folgenden beschreiben wir die einzelnen Schritte der Zweigitter-Iteration genauer und leiten darüber hinaus eine Vektor-Schreibweise des Zweigitter-Verfahrens her.

Die Vorglättung Mit $\bar{u}_h = \mathcal{G}_h^\nu(u_h^{(0)})$ ist die ν-fache Ausführung der Glättungsoperation bezeichnet. Allgemein sei $u_h^{(t)} = \mathcal{G}_h(u_h^{(t-1)}) := \mathcal{G}_h u_h^{(t-1)} + g_h$ affin linear und eine Fixpunkt-Iteration mit $\mathcal{G}_h(u_h) = u_h$. Für den Fehler $\bar{e}_h := u_h - \bar{u}_h = u_h - u_h^{(\nu)}$ gilt dann

$$\bar{e}_h = u_h - u_h^{(\nu)} = \mathcal{G}_h(u_h) - \mathcal{G}_h(u_h^{(\nu-1)}) = \mathcal{G}_h e_h^{(\nu-1)} = \mathcal{G}_h^\nu e_h^{(0)} = \mathcal{G}_h^\nu(u_h - u_h^{(0)}). \quad (2.21)$$

Wir betrachten als Beispiel die gedämpfte Richardson-Iteration. In Finite Elemente Schreibweise suchen wir $u_h^{(t)} \in V_h$, so dass

$$(u_h^{(t)}, \phi_h) = (\mathcal{G}_h(u_h^{(t-1)}), \phi) := (u_h^{(t-1)}, \phi_h) + \theta\big((f_h, \phi_h) - a(u_h^{(t-1)}, \phi_h)\big) \quad \forall \phi_h \in V_h,$$

also

$$u_h^{(t+1)} = (\mathcal{I}_h - \theta \mathcal{A}_h) u_h^{(t)} + \theta f_h.$$

Formuliert mit Vektoren gilt die Vorschrift:

$$\mathbf{M}_h \mathbf{u}_h^{(t)} = \mathbf{M}_h \mathbf{u}_h^{(t-1)} + \theta(\mathbf{b}_h - \mathbf{A}_h \mathbf{u}_h^{(t)}),$$

oder mit einer Verfahrensmatrix:

$$\mathbf{u}_h^{(t)} = \mathbf{G}_h(\mathbf{u}_h^{(t-1)}) := \mathbf{G}_h \mathbf{u}_h^{(t-1)} + \mathbf{g}_H, \quad \mathbf{G}_h := I - \theta \mathbf{M}_h^{-1} \mathbf{A}_h, \quad \mathbf{g}_h := \theta \mathbf{M}_h^{-1} \mathbf{b}_h.$$

Dann erhalten wir die Fehlerfortpflanzungen in Finite Elemente und Vektorschreibweise:

$$\bar{\mathbf{e}}_h = \mathbf{u}_h - \bar{\mathbf{u}}_h = \mathbf{G}_h^\nu \mathbf{e}_h^{(0)}, \quad \bar{e}_h = u_h - \bar{u}_h = \mathcal{G}_h^\nu e_h^{(0)}. \quad (2.22)$$

Die Grobgitterkorrektur Wir suchen die Lösung $w_H \in V_H$ von

$$a(w_H, \phi_H) = (f, \phi_H) - a(\bar{u}_h, \phi_H) \quad \forall \phi_H \in V_H. \quad (2.23)$$

Die rechte Seite dieses Problems ist das Residuum zur vorgeglätteten Approximation $\bar{u}_h \in V_h$ bezüglich der groben Basisfunktionen $\phi_H \in V_H$. Wir betrachten das Residuum $\bar{r}_h \in V_h$ als

$$(\bar{r}_h, \phi_h) := (f, \phi_h) - a(\bar{u}_h, \phi_h) \quad \forall \phi_h \in V_h.$$

Dann ist die rechte Seite $\bar{r}_H \in V_H$ von (2.23) gegeben als

$$\bar{r}_H = \mathcal{R}_H \bar{r}_h. \quad (2.24)$$

Mit dem Operator $\mathcal{A}_H$ schreiben wir (2.23) in der Form

$$w_H = \mathcal{A}_H^{-1} \bar{r}_H, \quad (2.25)$$

In Vektor-Schreibweise definieren wir zunächst

$$\mathbf{r}_h := \mathbf{b}_h - \mathbf{A}_h \bar{u}_h,$$

für welchen gilt:

$$(\mathbf{r}_h)_i = (r_h, \phi_h^{(i)}).$$

Also, mit Remark 49 ist

$$\mathbf{r}_H = \mathbf{R}_H \mathbf{r}_h, \quad (\mathbf{r}_H)_i = (r_H, \phi_h^{(i)}),$$

und die Grobgitterlösung $\mathbf{w}_H \in \mathbb{R}^{N_H}$ berechnet sich als

$$\mathbf{w}_H = \mathbf{A}_H^{-1} \bar{\mathbf{r}}_H, \tag{2.26}$$

oder eben

$$\mathbf{w}_H = \mathbf{A}_H^{-1} \mathbf{R}_H (\mathbf{b}_h - \mathbf{A}_h \bar{\mathbf{u}}_h). \tag{2.27}$$

Update Schließlich ergibt sich die neue Iteration durch:

$$u_h^{(1)} = \bar{u}_h + \omega \mathcal{P}_h w_H, \tag{2.28}$$

bzw. in Vektorschreibweise

$$\mathbf{u}_h^{(1)} = \bar{\mathbf{u}}_h + \omega \mathbf{R}_H^T \mathbf{W}_H. \tag{2.29}$$

Wir fassen zusammen:

Algorithm 53 (Zweigitteriteration)**.** *Seien $\Omega_H \Subset \Omega_h$ zwei geschachtelte Triangulierungen von Ω und $V_H \subset V_h$ zwei geschachtelte Finite Elemente Räume mit $\dim(V_H) = N_H$ und $\dim(V_h) = N_h$. Sei $\mathbf{u}_h^{(0)} \in \mathbb{R}^{N_h}$ ein Startvektor. $\nu > 0$ sei die Anzahl an Glättungsschritten, $\omega \in (0, 1]$ ein Dämpfungsparameter. Iteriere*

$$u_h^{(t+1)} = \mathcal{ZG}(u_h^{(t)}, f_h),$$

mit der Zweigitter-Iteration $\mathcal{ZG}(u_h^{(t)}, f_h)$:

(i)	*Vorglätten:*	$\bar{\mathbf{u}}_h = \mathbf{G}_h^{\nu_1}(\mathbf{u}_h^{(t)})$	$\bar{u}_h = \mathcal{G}_h^{\nu_1}(u_h^{(t)})$
(ii)	*Residuum:*	$\mathbf{r}_h = \mathbf{b}_h - \mathbf{A}_h \bar{\mathbf{u}}_h$	$r_h = f_h - \mathcal{A}_h \bar{u}_h$
(iii)	*Restriktion:*	$\mathbf{r}_H = \mathbf{R}_H \mathbf{r}_h$	$r_H = \mathcal{R}_H r_H$
(iv)	*Grobgitterproblem:*	$\mathbf{w}_H = \mathbf{A}_H^{-1} \mathbf{r}_H$	$w_H = \mathcal{A}_H^{-1} r_H$
(v)	*Prolongation:*	$\mathbf{u}_h^{(t+1)} = \bar{\mathbf{u}}_h + \omega \mathbf{R}_H^T \mathbf{w}_H$	$u_h^{(t+1)} = \bar{u}_h + \omega \mathcal{P}_h w_H$

Fehlerfortpflanzung Zur Analyse der Zweigitter-Konvergenz müssen wir nun eine Fehlerfortpflanzung herleiten, also einen Zusammenhang zwischen $e_h^{(t+1)}$ nach einem Schritt und $e_h^{(t)}$ vor dem Schritt. Wir entwickeln diese Fehlerdarstellung wieder simultan in Funktionenschreibweise und für Vektoren. Für den neuen Fehler gilt:

$$e_h^{(t+1)} = u_h - u_h^{(t+1)}, \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = \mathbf{u}_h - \mathbf{u}_h^{(t+1)}.$$

Mit (2.28) bzw. (2.29) erhalten wir

$$e_h^{(t+1)} = \bar{e}_h - \mathcal{P}_h w_H, \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = \bar{\mathbf{e}}_h - \mathbf{R}_H^T \mathbf{w}_H$$

Weiter, mit (2.25) bzw. (2.26)

$$e_h^{(t+1)} = \bar{e}_h - \mathcal{P}_h \mathcal{A}_H^{-1} \bar{r}_H, \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = \bar{\mathbf{e}}_h - \mathbf{R}_H^T \mathbf{A}_H^{-1} \bar{\mathbf{r}}_H.$$

Die rechte Seite des Grobgitterproblems ist durch (2.24) bzw. durch (2.27) in Vektorschreibweise gegeben:

$$e_h^{(t+1)} = \bar{e}_h - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H (f - \mathcal{A}_h \bar{u}_h), \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = \bar{\mathbf{e}}_h - \mathbf{R}_H^T \mathbf{A}_H^{-1} \mathbf{R}_H (\mathbf{b}_h - \mathbf{A}_h \bar{\mathbf{u}}_h).$$

Wir nutzen nun für die exakte Lösung $\mathcal{A}_h u_h = f$ bzw. $\mathbf{A}_h \mathbf{u}_h = \mathbf{b}_h$ und erhalten einen Bezug zum Fehler nach der Vorglättung

$$e_h^{(t+1)} = \bar{e}_h - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H \mathcal{A}_h (u_h - \bar{u}_h) \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = \bar{\mathbf{e}}_h - \mathbf{R}_H^T \mathbf{A}_H^{-1} \mathbf{R}_H \mathbf{A}_h (\mathbf{u}_h - \bar{\mathbf{u}}_h),$$

also:

$$e_h^{(t+1)} = (\mathcal{I}_h - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H \mathcal{A}_h) \bar{e}_h, \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = [I_h - \mathbf{R}_H^T \mathbf{A}_H^{-1} \mathbf{R}_H \mathbf{A}_h] \bar{\mathbf{e}}_h.$$

Für diesen gilt mit Darstellung (2.21) und (2.22)

$$e_h^{(t+1)} = (\mathcal{I}_h - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H \mathcal{A}_h) \mathcal{G}_h^{\nu_1} e_h^{(t)} \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = [I_h - \mathbf{R}_H^T \mathbf{A}_H^{-1} \mathbf{R}_H \mathbf{A}_h] \mathbf{G}_h^{\nu_1} \mathbf{e}_h^{(t)}.$$

Zusammengefasst erhalten wir den *Zweigitter-Operator* $\mathcal{B}_{ZG}(\nu)$ und die *Zweigitter-Matrix* $\mathbf{B}_{ZG}(\nu)$:

$$\mathcal{B}_{ZG}(\nu) = (\mathcal{I}_h - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H \mathcal{A}_h) \mathcal{G}_h^{\nu}, \quad \mathbf{B}_{ZG}(\nu) = [I_h - \mathbf{R}_H^T \mathbf{A}_H^{-1} \mathbf{R}_H \mathbf{A}_h] \mathbf{G}_h^{\nu}$$

Die Zweigitter-Iteration spaltet sich in zwei Bestandteile: die Glättung und die Grobgitter-Korrektur. Bei der mathematischen Konvergenzanalyse des Verfahrens werden diese beiden Anteile getrennt. Wir beweisen getrennt:

Lemma 54 (Glättungseigenschaft)**.** *Die gedämpfte Richardson-Iteration mit* $\theta = \lambda_{max}(A_h)^{-1}$ *erfüllt die Glättungseigenschaft*

$$\|\mathcal{A}_h \mathcal{G}_h^{\nu} \nu_h\|_{L^2(\Omega)} \leqslant \frac{c_G}{\nu h^2} \|\nu_h\|_{L^2(\Omega)} \quad \forall \nu_h \in V_h.$$

sowie

Lemma 55 (Approximationseigenschaft)**.** *Sei* $\Omega_H \Subset \Omega_h$ *mit* $H \leqslant cH$ *und* $c > 0$ *unabhängig von* h *und* H *zwei geschachtelte Gitter. Für die Grobgitterkorrektur gilt die Approximationseigenschaft*

$$\|(\mathcal{A}_h^{-1} - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H) \nu_h\|_{L^2(\Omega)} \leqslant c_A h^2 \|\nu_h\|_{L^2(\Omega)} \quad \forall \nu_h \in V_h.$$

Mit diesen beiden Sätzen folgt unmittelbar das Konvergenzresultat für die Zweigitter-Iteration:

Lemma 56 (Konvergenz der Zweigitter-Iteration)**.** *Es sei $\mathcal{G}$ ein Glättungsoperator mit der Eigenschaft von Satz 54. Weiter sei $H \leqslant ch$. Dann gilt für hinreichend viele Glättungsschritte $\nu_1 > \nu$ mit ν unabhängig von h*

$$\|\mathcal{B}_{ZG}(\nu)\| \leqslant \rho_{ZG}(\nu) = \frac{c}{\nu} < 1.$$

Proof. Mit Satz 54 und Satz 55 gilt:

$$\|\mathcal{B}_{ZG}(\nu)v_h\| \leqslant \frac{c_A c_G}{\nu}\|v_h\| \quad \forall v_h \in V_h.$$

Mit $\nu > c_A c_G$ folgt die Aussage des Satzes. □

Wir beweisen zunächst die Glättungseigenschaft:

Proof. (Beweis von Satz 54) Der Operator $\mathcal{A}_h : V_h \to V_h$ mit $(\mathcal{A}_h u_h, v_h) = (u_h, \mathcal{A}_h v_h)$ ist selbstadjungiert, und hat positive reelle Eigenwerte $0 < \lambda_1 \leqslant \cdots \leqslant \lambda_{N_h}$ und verfügt über ein zugehöriges System aus L^2-orthonormalen Eigenvektoren $\omega_h^{(1)}, \ldots, \omega_h^{(N_h)}$. Jede Funktion $v_h \in V_h$ schreiben wir nun in der Form

$$v_h = \sum_{i=1}^{N_h} \gamma_i \omega_h^{(i)}, \quad \gamma_i = (v_h, \omega_h^{(i)}), \quad \|v_h\|^2_{L^2(\Omega)} = \sum_{i=1}^{N_h} \gamma_i^2. \tag{2.30}$$

Mit $\theta := \lambda_{N_h}^{-1}$ ist durch die Richardson-Iteration

$$\mathcal{G}_h := \mathcal{I}_h - \frac{1}{\lambda_{N_h}}\mathcal{A}_h,$$

ein Operator $\mathcal{G}_h : V_h \to V_h$ definiert. Für ein beliebiges $v_h \in V_h$ gilt in Schreibweise (2.30)

$$\mathcal{A}_h \mathcal{G}_h^\nu v_h = \sum_{i=1}^{N_h} \gamma_i \lambda_i \left(1 - \frac{\lambda_i}{\lambda_{N_h}}\right)^\nu \omega_h^{(i)}.$$

In der Norm:

$$\begin{aligned}
\|\mathcal{A}_h \mathcal{G}_h^\nu v_h\|^2 &= \sum_{i=1}^{N_h} \gamma_i^2 \lambda_i^2 \left(1 - \frac{\lambda_i}{\lambda_{N_h}}\right)^{2\nu} \\
&\leqslant \lambda_{N_h}^2 \max_{1 \leqslant i \leqslant N_h} \left\{ \left(\frac{\lambda_i}{\lambda_{N_h}}\right)^2 \left(1 - \frac{\lambda_i}{\lambda_{N_h}}\right)^{2\nu} \right\} \sum_{i=1}^{N_h} \gamma_i^2 \\
&= \lambda_{N_h}^2 \max_{1 \leqslant i \leqslant N_h} \left\{ \left(\frac{\lambda_i}{\lambda_{N_h}}\right)^2 \left(1 - \frac{\lambda_i}{\lambda_{N_h}}\right)^{2\nu} \right\} \|v_h\|^2.
\end{aligned}$$

Es gilt $0 < \lambda_1/\lambda_{N_h} \leqslant 1$ und mit der Ungleichung

$$\max_{0 \leqslant x \leqslant 1} \{x(1-x)^\nu\} \leqslant (1+\nu)^{-1}, \quad \nu \geqslant 1$$

folgt

$$\|\mathcal{A}_h \mathcal{G}_h^\nu v_h\|^2 \leqslant \lambda_{N_h}^2 (1+\nu)^{-2} \|v_h\|^2. \tag{2.31}$$

Für den größten Eigenwert des Operators $\mathcal{A}_h$ gilt der Zusammenhang

$$\lambda_{N_h} \|\omega_h^{(N_h)}\|^2 = (\mathcal{A}_h \omega_h^{(N_h)}, \omega_h^{(N_h)}) = \langle \mathbf{A}_h \boldsymbol{\omega}_h^{(N_h)}, \boldsymbol{\omega}_h^{(N_h)} \rangle \leqslant \lambda_{\max}(\mathbf{A}_h) |\boldsymbol{\omega}_h^{(N_h)}|^2,$$

mit dem Koeffizientenvektor $\boldsymbol{\omega}_h^{(N_h)}$ von $\omega_h^{(N_h)}$ in der Knotenbasisdarstellung. Jetzt gilt

$$|\boldsymbol{\omega}_h^{(N_h)}|^2 = \langle \mathbf{M}_h^{-1} \mathbf{M}_h \boldsymbol{\omega}_h^{(N_h)}, \boldsymbol{\omega}_h^{(N_h)} \rangle \leqslant \lambda_{\min}(\mathbf{M}_h)^{-1} (\omega_h^{(N_h)}, \omega_h^{(N_h)}).$$

Mit $\lambda_{\min}(\mathbf{M}_h) = O(h^2)$ sowie $\lambda_{\max}(\mathbf{A}_h) = O(1)$ folgt $\lambda_{N_h} \leqslant c h^{-2}$ und aus (2.31) schließlich die Behauptung. □

Es bleibt, die Approximationseigenschaft zu zeigen:

Proof. (Beweis von Satz 55) Es sei $f_h \in V_h$ beliebig. Dann ist $v_h = \mathcal{A}_h^{-1} f_h$ definiert durch

$$a(v_h, \phi_h) = (f_h, \phi_h) \quad \forall \phi_h \in V_h. \tag{2.32}$$

Weiter sei $f_H \in V_H$ gegeben durch $f_H = \mathcal{R}_H v_h$, also

$$(f_H, \phi_H) = (f_h, \phi_H) \quad \forall \phi_H \in V_H.$$

Also ist $v_H = \mathcal{A}_H^{-1} f_H = \mathcal{A}_H^{-1} \mathcal{R}_H f_h$ definiert durch

$$a(v_H, \phi_H) = (f_h, \phi_H) \quad \forall \phi_H \in V_H. \tag{2.33}$$

Schließlich definieren wir eine Funktion $v \in H_0^1(\Omega) \cap H^2(\Omega)$ als Lösung der Randwertaufgabe

$$a(v, \phi) = (f_h, \phi) \quad \forall \phi \in H_0^1(\Omega). \tag{2.34}$$

Für diese Lösung gilt die a priori Abschätzung

$$\|\nabla^2 v\| \leqslant c \|f_h\|. \tag{2.35}$$

Die Lösungen $v_h \in V_h$ von (2.32) und $v_H \in V_H$ (2.33) sind gerade die Galerkin-Approximationen der Lösung $v \in H_0^1(\Omega) \cap H^2(\Omega)$ von (2.34). Es gilt demnach die L^2-Fehlerabschätzung:

$$\|v - v_h\| \leqslant c h^2 \|\nabla^2 v\|, \quad \|v - v_H\| \leqslant c H^2 \|\nabla^2 v\|.$$

Mit der Annahme $H \leqslant c h$ und der a priori Abschätzung (2.35) folgt

$$\|v_h - v_H\| \leqslant \|v - v_h\| + \|v - v_H\| \leqslant c h^2 \|\nabla^2 v\| \leqslant c h^2 \|f_h\|.$$

Also folgt für den Grobgitter-Operator

$$\|\mathcal{A}_h^{-1} f_h - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H f_h\| \leqslant c h^2 \|f_h\| \quad \forall f_h \in V_h.$$

Dies vervollständigt den Beweis. □

2.3.3 Mehrgitter-Verfahren

Beim Zweigitter-Verfahren erfolgt die Grobgitter-Korrektur durch Lösen eines Systems $\mathbf{A}_H u_H = f_H$. Dieses Problem ist zwar kleiner als das ursprüngliche, kann jedoch immer noch zu groß zum effizienten Lösen sein.

Beim Mehrgitter-Verfahren zur Lösung von $\mathbf{A}_L u_L = f_L$ im Raum $V_h = V_L$ wird zur Lösung des Grobgitter-Problems $A_{L-1} w_{L-1} = r_{L-1}$ wieder das Zweigitter-Verfahren verwendet. Dies geschieht auf rekursive Art bis wir beim Problem auf dem gröbsten Gitter $\mathbf{A}_0 u_0 = r_0$ ankommen. Dieses nun sehr kleine Gleichungssystem kann mit einem direkten Verfahren gelöst werden.

Algorithm 57 (Mehrgitter-Verfahren)**.** *Durch $\Omega_H = \Omega_0 \Subset \Omega_1 \Subset \ldots \Subset \Omega_L = \Omega_h$ sei eine Familie von geschachtelten Triangulierungen mit $h_{l-1} \leqslant c h_l$ gegeben. Sei $u_L^{(0)} := u_h^{(0)}$ ein Startwert. Durch $\nu_1, \nu_2 \geqslant 0$ sei die Anzahl der Vor- bzw. Nachglättungsschritte gegeben. Weiter sei $R \geqslant 1$ und $\omega_l \in (0, 1]$ ein Dämpfungsparameter. Iteriere für $t \geqslant 0$*

$$u_L^{(t+1)} = \mathcal{MG}(L, u_L^{(t)}, f_L),$$

mit der Mehrgitter-Iteration $\mathcal{MG}(l, u_l^{(t)}, f_l)$:

$l = 0$		*Direkter Löser:*	$u_0^{(t+1)} = \mathcal{A}_0^{-1} f_0$
$l > 0$	(i)	*Vorglätten:*	$\bar{u}_l = \mathcal{G}_l^{\nu_1}(u_l^{(t)})$
	(ii)	*Residuum:*	$r_l = f_l - \mathcal{A}_l \bar{u}_l$
	(iii)	*Restriktion:*	$r_{l-1} = \mathcal{R}_{l-1} r_l$
	(iv)	*Grobgitterproblem:*	$w_{l-1}^{(0)} = 0$
		$1 \leqslant r \leqslant R$:	$w_{l-1}^{(r)} = \mathcal{MG}(l-1, w_{l-1}^{(r-1)}, r_{l-1})$
	(v)	*Prolongation:*	$w_l = \mathcal{P}_l w_{l-1}^{(R)}$
	(vi)	*Update:*	$\bar{\bar{u}}_l = \bar{u}_l + \omega_l w_l$
	(vii)	*Nachglätten:*	$u_l^{(t+1)} = \mathcal{G}_l^{\nu+2}(\bar{\bar{u}}_l)$

Da die Grobgitter-Korrektur nun lediglich eine Approximation ist, führen wir $R \geqslant 1$ Korrekturschritte aus. R wird üblicherweise sehr klein als $R = 1$ oder als $R = 2$ gewählt. Im Fall $R = 1$ spricht man vom *V*-Zyklus, im Fall $R = 2$ vom *W*-Zyklus der Mehrgitter-Iteration, siehe Abbildung 2.9 für eine Darstellung des Iterationsschemas. Neben dem *V* und *W*-Zyklus existieren weitere Varianten wie der F-Zyklus, beim dem in einer Richtung $R = 1$ und in der anderen Richtung $R = 2$ gewählt wird. Weiter kann es zweckdienlich sein, den Mehrgitterprozess nicht auf dem feinsten Gitter Ω_h sondern auf dem gröbsten Gitter Ω_H zu starten. Iterativ werden mit dem Mehrgitter-Verfahren auf den Gittern $\Omega_0, \Omega_1, \ldots$ Lösungen mit hinreichender Genauigkeit erstellt. Dieses *geschachtelte Mehrgitter-Verfahren* erreicht die optimale Komplexität $O(N_L)$.

Wenn der V-Zyklus konvergiert, so ist er sehr effizient. Bei Problemen mit Unsymmetrien, nicht-glatten Koeffizienten, Singularitäten ist der V-Zyklus hingegen oft instabil. Hier bietet sich dann der robuste W-Zyklus an.

Wir beweisen nun:

Lemma 58 (Konvergenz des Mehrgitter-Verfahrens)**.** *Der Zweigitter-Zyklus sei auf jedem Paar der Familie* $V_0 \subset V_1 \subset \cdots \subset V_L$ *konvergent mit* $\rho_{ZG}(\nu) \to 0$ *für* $\nu \to 0$ *gleichmäßig bezüglich* l*. Dann konvergiert für* $\nu \geqslant \nu_0$ *groß genug der Mehrgitteralgorithmus im W-Zyklus mit einer von* L *(und somit* h*) unabhängigen Konvergenzrate* $\rho_{MG} < 1$ *bezüglich der* L^2*-Norm:*

$$\|u_L - \mathcal{MG}(L, u_L^{(t)}, f_L)\| \leqslant \rho_{MG}\|u_L - u_L^{(t)}\|.$$

Proof. (*i*) Wir setzen $\nu_2 = 0$. Wir führen den Beweis per Induktion nach L. Zunächst sei $\nu \geqslant \nu_0$ so gewählt, dass für die Zweigitter-Konvergenz gilt $\rho_{ZG} \leqslant \frac{1}{8}$. Wir wollen zeigen, dass dann gilt $\rho_{MG} \leqslant \frac{1}{4}$. Im Fall $L \leqslant 2$ liegt gerade das Zweigitter-Verfahren vor und die Aussage ist richtig.

(*ii*) Sei nun $L > 2$. Für das Gitterlevel $L-1$ gilt nach Voraussetzung $\rho_{MG} \leqslant \frac{1}{4}$. Sei $u_L^{(t)}$ die letzte Iteration.

Wir führen nun als Hilfsgröße die Lösung der Zweigitter-Iteration (mit exakter Lösung des Grobgitter-Problems) ein:

$$\bar{u}_L^{(t+1)} = \mathcal{ZG}(L, u_L^{(t)}, f_L).$$

Dann gilt für die Differenz zwischen dieser Lösung und zweimaliger Grobgitter-Korrektur mit dem Mehrgitter-Verfahren ($0 = \omega_{l-1}^{(0)} \xrightarrow{\mathcal{MG}(l-1)} \omega_{l-1}^{(1)} \xrightarrow{\mathcal{MG}(l-1)} \omega_{l-1}^{(2)}$)

$$u_L^{(t+1)} - \bar{u}_L^{(t+1)} = \mathcal{P}_L(w_{l-1}^{(2)} - \bar{w}_{l-1}),$$

wobei $w_{l-1}^{(2)}$ die mit Mehrgitter approximierte Grobgitterlösung ist und $\bar{w}_{l-1}$ die exakte Lösung des Grobgitterproblems. Es gilt:

$$\|w_{l-1}^{(2)} - \bar{w}_{l-1}\| \leqslant \rho_{MG}^2\|w_{l-1}^{(0)} - \bar{w}_{l-1}\| = \rho_{MG}^2\|\bar{w}_{l-1}\|, \tag{2.36}$$

da der Startwert in Schritt (*iv*) von Algorithmus 57 gerade $w_{l-1}^{(0)} = 0$ ist. Mit

$$\begin{aligned}\bar{w}_{l-1} &= \mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1}(f_l - \mathcal{A}_l\mathcal{G}_l^\nu(u_l^{(t)})) \\ &= \mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1}\mathcal{A}_l(u_l - \mathcal{G}_l^\nu(u_l^{(t)})) \\ &= \mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1}\mathcal{A}_l\mathcal{G}_l^\nu(u_l - u_l^{(t)}),\end{aligned}$$

da $\mathcal{G}_l(u_l) = u_l$ eine Fixpunktiteration ist. Also gilt mit (2.36)

$$\|w_{l-1}^{(2)} - \bar{w}_{l-1}\| \leqslant \rho_{MG}^2\|\mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1}\mathcal{A}_l\mathcal{G}_l^\nu\|\,\|u_l - u_l^{(t)}\|.$$

Wir schätzen nun die Norm des Grobgitter-Operators mit Hilfe des Zweigitter-Operators ab:

$$\mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1}\mathcal{A}_l\mathcal{G}_l^\nu = \mathcal{G}_l^\nu - (\mathcal{A}_l^{-1} - \mathcal{P}_l\mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1})\mathcal{A}_l\mathcal{G}_l^\nu = \mathcal{G}_l^\nu - \mathcal{Z}\mathcal{G}_l$$

Hieraus folgt:

$$\|\mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1}\mathcal{A}_l\mathcal{G}_l^\nu\| \leqslant \|\mathcal{G}_l^\nu\| + \|\mathcal{Z}\mathcal{G}\| \leqslant 1 + \rho_{ZG} \leqslant 2.$$

Insgesamt erhalten wir:

$$\|u_l - u_l^{(t+1)}\| \leqslant (\rho_{ZG} + 2\rho_{MG}^2)\|u_l - u_l^{(t)}\|.$$

Mit $\rho_{ZG} \leqslant \frac{1}{8}$ und $\rho_{MG} \leqslant \frac{1}{4}$ aus der Induktionsannahme folgt die Aussage des Satzes. □

Im Abschluss betrachten wir nun den numerischen Aufwand in jedem Mehrgitterschritt. Hierzu benötigen wir einige Hilfsgrößen, welche auf jedem Gitter Ω_l den Aufwand der einzelnen Mehrgitterkomponenten bezüglich der Anzahl der Freiheitsgrade messen. Es sei

$$C_0(N_0) := OP(\mathcal{A}_0^{-1})/N_0,$$

der Aufwand zur Grobgitterlösung pro Grobgitter-Freiheitsgrad. Man beachte, dass C_0 von N_0 anhängt, da die Grobgitterlösung im Allgemeinen nicht mit linearer Laufzeit erfolgen kann. Dennoch, da N_0 fest ist, kann $C_0(N_0)$ in diesem Zusammenhang als konstant angesehen werden. Weiter sei $C_s := OP(\mathcal{S}_l)/N_l$ der Aufwand pro Glättungs-Iteration pro Knoten auf dem Gitterlevel Ω_l sowie $C_T := OP(\mathcal{R}_l)/N_l$ der Aufwand zur Restriktion und auch Prolongation pro Freiheitsgrad und $C_r := OP(r_l)/N_l$ der numerische Aufwand zur Berechnung des Residuums. Dann gilt:

Lemma 59 (Komplexität des Mehrgitterverfahrens)**.** *Es sei* $\kappa := \max(N_{l-1}/N_l) < 1$, $R = 1$ *oder* $R = 2$ *die Anzahl der Grobgitteriterationen und* ν_1, ν_2 *die Anzahl der Vor- bzw. Nachglättungsschritte. Dann gilt im Fall* $q := \kappa R < 1$ *für den numerischen Aufwand der Mehrgitteriteration*

$$OP(\mathcal{MG}(L)) = \frac{1}{1-q}\left((\nu_1+\nu_2)C_s + C_r + 2C_T\right)N_L + C_0(N_0)q^L N_L$$

Der Gesamtaufwand zur Reduktion des L^2*-Fehlers bezüglich der* L^2*-Norm auf die Diskretisierungsgenauigkeit* $O(h^2)$ *beträgt daher* $O(N_L \ln(N_L))$.

Proof. Wir betrachten zunächst eine Gitterebene l mit N_l Freiheitsgraden und zählen die auf dieser Ebene notwendigen arithmetischen Operation mit Ausnahme der Grobgitterkorrektur. Mit der Vorbereitung gilt hier:

$$OP(\mathcal{G}_l) = \tilde{C}N_l, \quad \tilde{C} := (\nu_1+\nu_2)C_s + C_r + 2C_T \tag{2.37}$$

D.h., auf der feinsten Gitterebene gilt rekursiv:

$$\begin{aligned} OP(\mathcal{MG}_L) &= \tilde{C}N_L + R \cdot OP(\mathcal{MG}_{L-1}) \\ &= \tilde{C}N_L + R\tilde{C}N_{L-1} + R^2 \cdot OP(\mathcal{MG}_{L-2}) \\ &= \tilde{C}N_L + R\tilde{C}N_{L-1} + R^2CN_{L-2} + \cdots + R^{L-1}\tilde{C}N_1 + R^L C_0(N_0)N_0, \end{aligned}$$

wobei $C_0(N_0)N_0$ der Aufwand zum Lösen des Grobgitterproblems ist. Mit $N_{l-1} \leqslant \kappa N_l$ und $q := \kappa R < 1$ folgt:

$$\begin{aligned} OP(\mathcal{MG}_L) &\leqslant \tilde{C} \sum_{l=0}^{L-1} (\kappa R)^l N_L + C_0(N_0)(\kappa R)^L N_L \\ &= \tilde{C} \frac{1-q^L}{1-q} N_L + C_0(N_0) q^L N_L \\ &\leqslant \left(\frac{\tilde{C}}{1-q} + C_0(N_0) q^L \right) N_L. \end{aligned}$$

Mit (2.37) folgt die Komplexitätsabschätzung. Zur Abschätzung der Gesamtkomplexität machen wir den Ansatz:

$$\rho_{MG}^t = h_L^2 \sim N_L^{-\frac{2}{d}} \quad \rightarrow \quad t \sim -\frac{\ln(N_L)}{\ln(\rho_{MG})}.$$

□

Die Mehrgitter-Iteration erreicht also fast die optimale Komplexität $O(N_L)$ zum Lösen des Gleichungssystems mit hinreichender Genauigkeit. Der Logarithmische Term ist hierbei in der Anwendung nicht wesentlich. Für den Beweis der optimalen Komplexitätsabschätzung ist die Bedingung $\kappa R < 1$ wesentlich. D.h., im Fall des *W*-Zyklus mit $R = 2$ bedeutet dies $\kappa := \max N_{l-1}/N_l < \frac{1}{2}$. Bei der Verwendung von lokal verfeinerten Gittern ist diese Bedingung oft nicht erfüllt. Um dennoch optimale Mehrgitter-Komplexität zu erhalten muss der Algorithmus modifiziert werden.

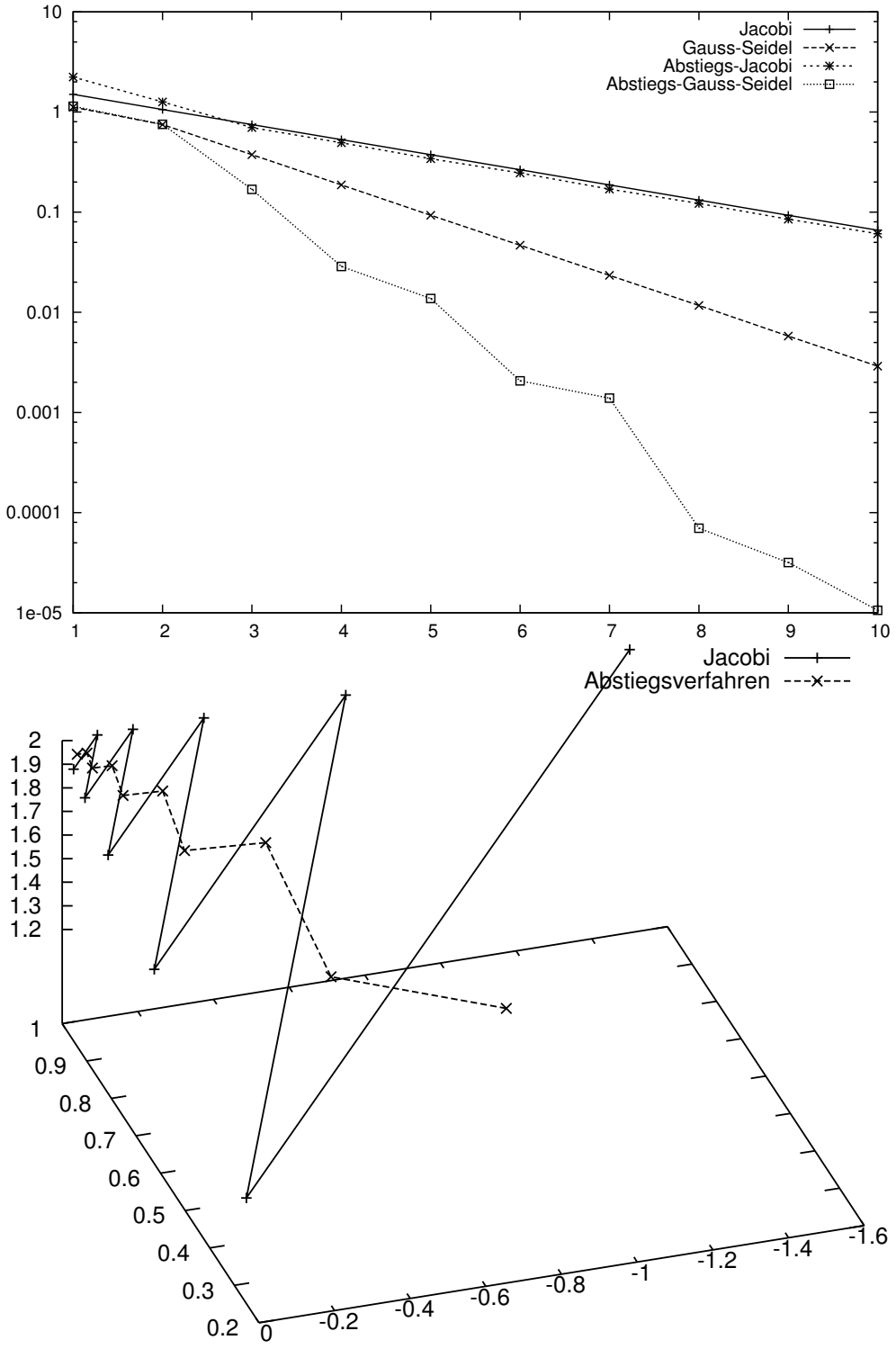


Figure 2.2: Oben: Konvergenz von Jacobi-, Gauß-Seidel- sowie den entsprechenden Abstiegsverfahren. Unten: Vergleich der Annäherung bei Jacobi- und Jacobi-Abstiegsverfahren an die Lösung $\mathbf{x} = (1, 0, 2)^{\mathsf{T}}$

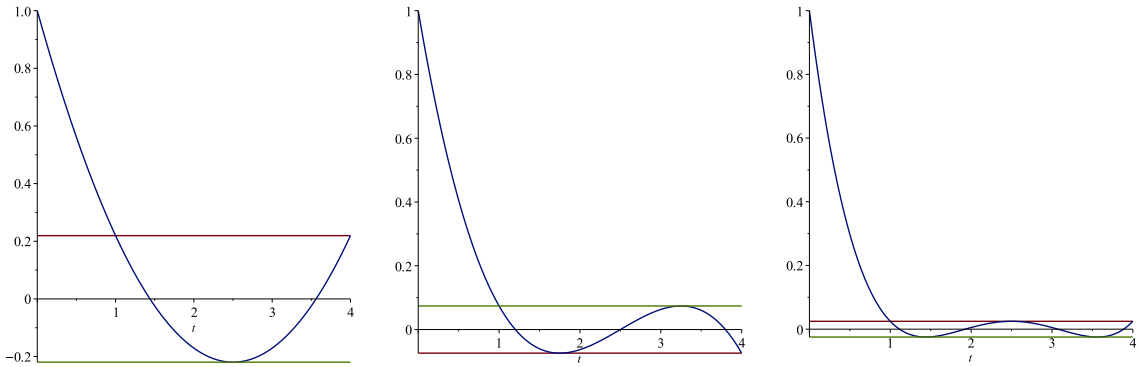


Figure 2.3: Optimale Polynome zur Beschränkung der CG-Konvergenz. Es gilt $p(0) = 1$ sowie eine Minimierung auf dem Intervall $[1, 4]$. Links $n = 2$, Mitte $n = 3$ und rechts $n = 4$.

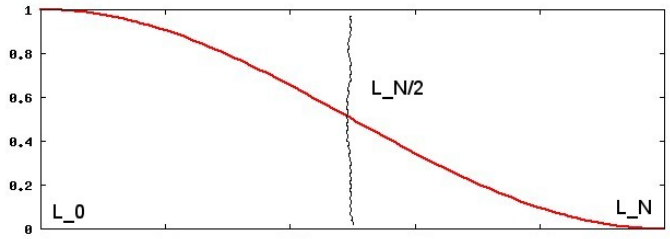


Figure 2.4: Fehlerreduktion des gedämpften Richardson-Verfahrens für die eindimensionale Poisson-Gleichung in Abhängigkeit der Fehlerfrequenz.

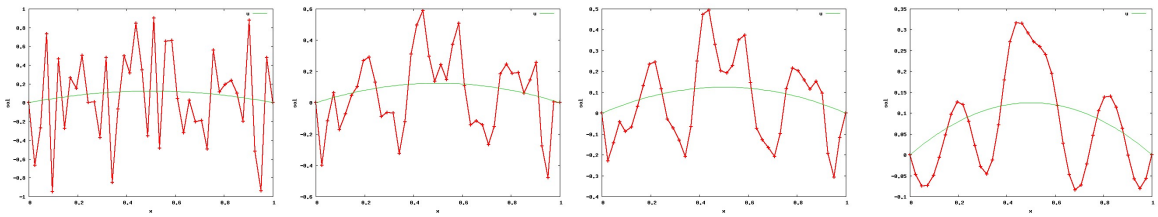


Figure 2.5: Fehlerverlauf (rot) und approximierte Lösung (grün) der gedämpften Richardson-Iteration. Von links nach rechts: Startfehler, nach einem Schritt, zwei Schritten und nach 9 Schritten.

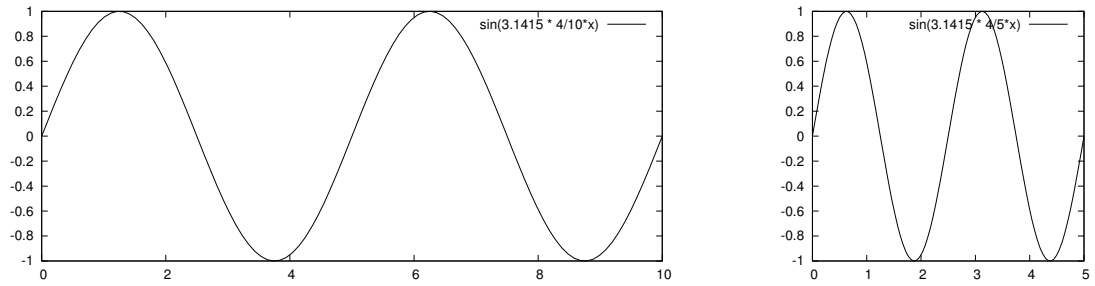


Figure 2.6: Die Fehlerfrequenz λ_4 auf einem Gitter mit 10 Elementen und auf eine Gitter mit 5 Elementen.

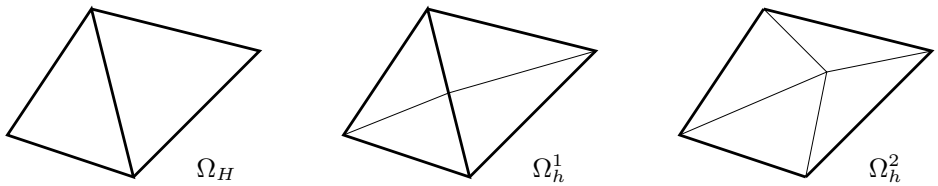


Figure 2.7: Grobgitter Ω_H und zwei feinere Gitter. Das Gitter Ω_h^1 ist durch Verfeinerung entstanden und es gilt $\Omega_H \Subset \Omega_h^1$. Das Gitter Ω_h^2 stammt nicht von Ω_H ab.

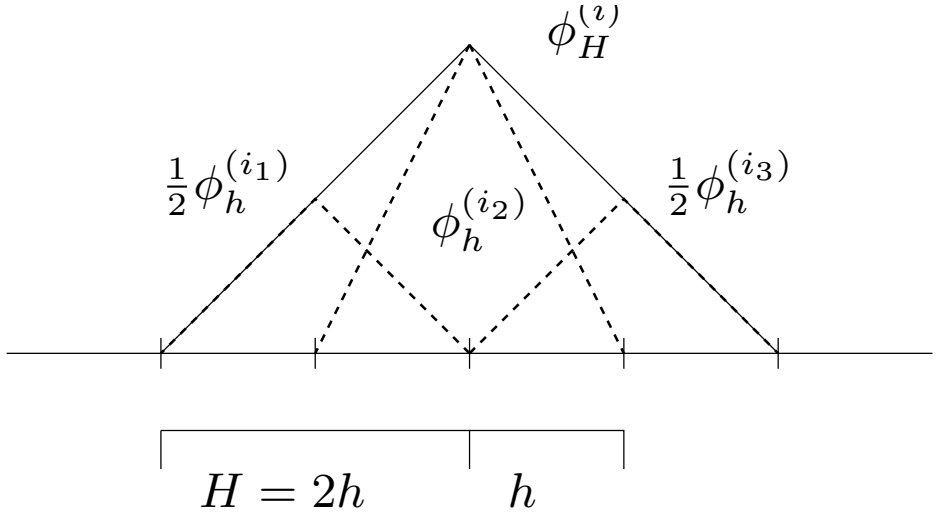


Figure 2.8: Darstellung einer groben Basisfunktion $\phi_H^{(i)}$ (durchgezogene Linie) durch drei Basisfunktionen des feinen Gitters.

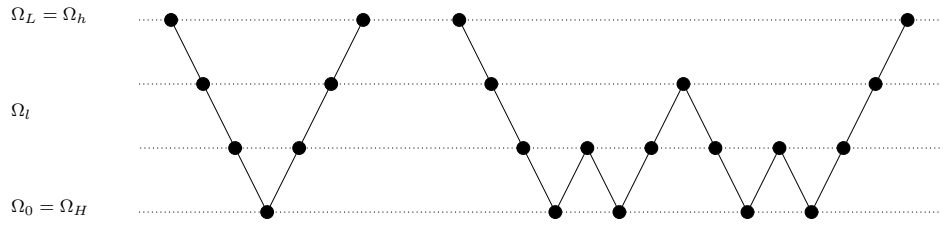


Figure 2.9: Schematische Darstellung des *V*- und *W*-Zyklus.

3 Transportstabilisierung

3.1 Elliptische Probleme

In diesem Abschnitt vertiefen wir die theoretische Analyse von elliptischen Differentialgleichungen. Auf einem Lipschitzgebiet $\Omega \subset \mathbb{R}^d$ mit $d \geqslant 2$ betrachten wir allgemein Probleme vom Typ

$$Lu = f, \quad Lu := -\sum_{i,j=1}^{d} \partial_i(a_{ij}(x)\partial_j u(x)) + \sum_{j=1}^{d} b_j(x)\partial_j u(x) + c(x)u(x), \tag{3.1}$$

mit reellen Koeffizientenfunktionen a_{ij}, b_j und c. Es gilt wegen der Vertauschbarkeit der zweiten Ableitungen ohne Einschränkung $a_{ij} = a_{ji}$. Mit der Matrix $A = (a_{ij})_{i,j}$ sowie dem Vektor $b = (b_j)_j$ schreiben wir kurz:

$$Lu = -\nabla \cdot (A\nabla u) + b \cdot \nabla u + cu.$$

Bei elliptischen Problemen sind die Eigenwerte von A ungleich Null und haben das gleiche Vorzeichen. Wir nehmen ohne Einschränkung an, dass die Koeffizientenmatrix A positiv definit ist.

Elliptische partielle Differentialgleichungen sind Randwertprobleme. Wir unterscheiden drei verschiedene Arten von Randwerten:

1. Dirichlet-Problem Auf dem Rand $\partial\Omega$ geben wir den Funktionswert der Lösung vor:

$$\text{suche } u: \quad Lu = f \text{ in } \Omega, \quad u = g \text{ auf } \partial\Omega.$$

Es sei nun $\bar{g} \in H^1(\Omega)$ eine Fortsetzung von g, d.h., g ist die Spur. Dann können wir das Dirichlet-Problem stets in ein Problem mit homogenen Randdaten transformieren:

$$\text{suche } u = \bar{u} + \bar{g}: \quad L\bar{u} = f - L\bar{g}, \quad \bar{u} = 0 \text{ auf } \partial\Omega.$$

Die strenge Voraussetzung an die Regularität der Dirichlet-Daten g ist also, dass diese Funktion als Spur einer $H^1(\Omega)$-Funktion gegeben ist. Wir bezeichnen den Raum aller Spuren von $H^1(\Omega)$ Funktionen auf $\partial\Omega$ als $H^{\frac{1}{2}}(\partial\Omega)$. Es gilt $H^{\frac{1}{2}}(\partial\Omega) \subset L^2(\partial\Omega)$.

2. Neumann-Problem Auf dem Rand von Ω geben wir die Ableitung von u in *Normal-Richtung* vor:

$$\text{suche } u: \quad Lu = f \text{ in } \Omega, \quad \partial_n u = g \text{ auf } \partial\Omega.$$

Bei Neumann-Problemen ist der Funktionswert von u auf dem Rand nicht fixiert. Die Lösung muss also im vollen Raum $H^1(\Omega)$ gesucht werden. Betrachten wir z.B. das homogene Neumann-Problem

$$-\Delta u = f \text{ in } \Omega, \quad \partial_n u = 0 \text{ auf } \partial\Omega,$$

so hat dies zur Konsequenz, dass die Lösung nicht mehr eindeutig sein muss. Denn ist $u \in H^1(\Omega)$ eine Lösung, so ist auch durch $u + c$ für jedes $c \in \mathbb{R}$ eine Lösung gegeben. Um Eindeutigkeit zu erreichen muss der Lösungsraum künstlich eingeschränkt werden, etwa durch die Wahl:

$$H^1(\Omega)/\mathbb{R} := \{v \in H^1(\Omega), \int_\Omega v \, dx = 0\}.$$

We can easily show, that Poincaré's inequality also holds on this space.

Lemma 60 (Modified Poincaré's inequality)**.** *There exists a constant* $c_p = c(\Omega) > 0$ *such that*

$$\|v\|_\Omega \leqslant c_p \left(\left| \int_\Omega v \mathbf{d}x \right| + \|\nabla v\|_\Omega \right) \quad \forall v \in H^1(\Omega).$$

Proof. Using Rellich's Theorem we can show (by contradiction), that the inequality holds for

$$\|v\|_\Omega \leqslant c\|\nabla v\|_\Omega \quad \forall v \in H^1(\Omega) \setminus \mathbb{R},$$

e.g. for all functions $v \in H^1(\Omega)$ with average zero. Then, the general case is obtained by

$$\bar{v} := \frac{1}{|\Omega|} \int_\Omega v \mathbf{d}x \quad \Rightarrow \quad \|v\|_\Omega \leqslant \|\bar{v}\|_\Omega + \|v - \bar{v}\|_\Omega \leqslant \sqrt{|\Omega|} \left| \int_\Omega v \mathbf{d}x \right| + c_p \|\nabla v\|_\Omega.$$

□

By this modification, the Neumann problem can be written as minimization problem with the functional

$$u \in H^1(\Omega) \setminus \mathbb{R}: \quad E(u) \leqslant E(v) := \frac{1}{2}\|\nabla v\|^2 - (f, v) \quad \forall v \in H^1(\Omega) \setminus \mathbb{R}.$$

3. Robin-Problem Auf dem Rand $\partial\Omega$ geben wir gemischte Randdaten vor:

$$\text{suche } u: \quad Lu = f \text{ in } \Omega, \quad \partial_n u + \alpha u = g \text{ auf } \partial\Omega,$$

mit einem $\alpha \neq 0$.

Wir werden zunächst ausschließlich das (homogene) Dirichlet-Problem betrachten. Hierzu leiten wir eine variationelle Formulierung der allgemeinen elliptischen Differentialgleichung (3.1) her:

Lemma 61 (Variationsproblem). *Die Koeffizientenmatrix* $A \in [L^\infty(\Omega)]^{d\times d}$ *sei positiv definit mit* $\langle Ax, x\rangle \geqslant \gamma|x|^2$ *mit* $\gamma > 0$*, der Vektor* $b \in [W^{1,\infty}(\Omega)]^d$ *sei divergenzfrei (also* $\nabla\cdot b = \sum_j \partial_i b_i = 0$*) und es sei* $c \in L^\infty(\Omega)$ *mit* $c \geqslant 0$*. Jede Lösung* $u \in C^2(\Omega)\cap C(\bar\Omega)$ *von (3.1) ist Lösung der Variationsgleichung:*

$$a(u,\phi) = (f,\phi) \quad \forall \phi \in H_0^1(\Omega).$$

Die Bilinearform

$$a(u,\phi) = (A\nabla u, \nabla\phi) + (b\cdot\nabla u, \phi) + (cu,\phi).$$

ist stetig

$$a(u,v) \leqslant M\|\nabla u\|\,\|\nabla v\| \quad \forall u,v \in H_0^1(\Omega)$$

mit einer Konstante

$$M = \max\{d\|A\|_\infty, c_p\sqrt{d}\|b\|_\infty, \|c\|_{L^\infty(\Omega)}c_p^2\},$$

sowie positiv definit (elliptisch)

$$a(u,u) \geqslant \gamma\|\nabla u\|^2 \quad \forall u \in H_0^1(\Omega).$$

Proof. (*i*) Wir multiplizieren (3.1) für festes $t \in \bar I$ mit einer beliebigen Funktion $\phi \in C_0^\infty(\Omega)$ und integrieren über das Gebiet:

$$\begin{aligned}(f,\phi) &= -(\nabla\cdot(A\nabla u),\phi) + (b\cdot\nabla u,\phi) + (cu,\phi)\\ &= (A\nabla u,\nabla\phi) + \underbrace{\int_{\partial\Omega} n\cdot(A\nabla u)\phi\,ds}_{=0} + (b\cdot\nabla u,\phi) + (cu,\phi).\end{aligned}$$

Wegen der Dichtheit von $C_0^\infty(\Omega)$ in $H_0^1(\Omega)$ kann der Testraum erweitert werden.

(*ii*) Weiter gilt für alle $u,v \in H_0^1(\Omega)$ wegen der Beschränktheit der Koeffizienten und mit Poincaré:

$$\begin{aligned}|a(u,v)| &\leqslant \sum_{i,j=1}^d |(a_{ij}\partial_j u, \partial_i v)| + \sum_{j=1}^d |(b_j\partial_j u, v)| + |(cu,v)|\\ &\leqslant \|A\|_\infty \sum_{ij} \|\partial_j u\|_{L^2(\Omega)}\,\|\partial_i v\|_{L^2(\Omega)}\\ &\qquad + \|b\|_\infty \sum_j \|\partial_j u\|_{L^2(\Omega)}\,\|v\|_{L^2(\Omega)} + |c|_\infty\|u\|_{L^2(\Omega)}\,\|v\|_{L^2(\Omega)}\\ &\leqslant \|A\|_\infty d\|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} + \|b\|_\infty\sqrt{d}\|\nabla u\|_{L^2(\Omega)}c_p\|\nabla v\|_{L^2(\Omega)}\\ &\qquad + |c|_\infty c_p^2\|\nabla u\|_{L^2(\Omega)}\,\|\nabla v\|_{L^2(\Omega)}\\ &\leqslant M\|\nabla u\|_{L^2(\Omega)}\,\|\nabla v\|_{L^2(\Omega)}, \quad M := \max\{\|A\|_\infty d, c_p\|b\|_\infty\sqrt{d}, |c|_\infty c_p^2\},\end{aligned}$$

wobei wir die folgende Ungleichung verwendet haben:

$$\sum_{i=1}^d |a_i| \leqslant \sqrt{d}\left(\sum_{i=1}^d |a_i|^2\right)^{\frac12}.$$

(*iii*) Zum Nachweis der *Elliptizität* untersuchen wir die verschiedenen Terme der Bilinearform getrennt. Zunächst existiert wegen der positiven Definitheit von A ein γ mit:

$$(A\nabla u, \nabla u) = \int_\Omega \langle A(x)\nabla u(x), \nabla u(x)\rangle dx \geqslant \gamma \int_\Omega |\nabla u|^2 \, dx = \gamma \|\nabla u\|^2_{L^2(\Omega)}. \tag{3.2}$$

Wegen der Divergenzfreiheit von b gilt:

$$(b \cdot \nabla u, u) = \sum_{i=1}^{d} (b\partial_i u, u) = \sum_{i=1}^{d} (b_i u, \partial_i u) = -\sum_{i=1}^{d} (\partial_i(b_i u), u) + \sum_{i=1}^{d} \underbrace{\int_{\partial\Omega} n_i (b_i u) u \, ds}_{=0}$$

$$= -\sum_{i=1}^{d} (\partial_i b_i u, u) - \sum_{j=1}^{d} (b_i \partial_i u, u) = \underbrace{-((\nabla \cdot b)u, u)}_{=0} - (b \cdot \nabla u, u),$$

und also $(b \cdot \nabla u, u) = 0$. Schließlich gilt mit $c \geqslant 0$

$$(cu, u) \geqslant \min_\Omega c \cdot \|u\| \geqslant 0$$

und zusammen mit (3.2) ergibt sich die Behauptung. □

Die stetige, positiv definite Bilinearform unterscheidet sich von einem Skalarprodukt durch die fehlende Symmetrie.

Definition 62 (Schwache Lösung)**.** *Eine Funktion $u \in H^1_0(\Omega)$ heißt verallgemeinerte (schwache) Lösung von (3.1), falls*

$$a(u, \phi) = (f, \phi) \quad \forall \phi \in H^1_0(\Omega). \tag{3.3}$$

Wie bereits argumentiert gilt:

Lemma 63. *Jede schwache Lösung $u \in H^1_0(\Omega)$ von (3.3) für welche $u \in C^2(\Omega) \cap C(\bar{\Omega})$ gilt ist auch klassische Lösung von (3.1).*

Und weiter gilt:

Lemma 64 (Minimierung des Energiefunktionals)**.** *Im Fall $b = 0$ ist die Lösung der variationellen Formulierung (3.3) äquivalent zur Minimierung des Energiefunktionals. Suche $u \in H^1_0(\Omega)$:*

$$E(u) \leqslant E(v) \quad \forall v \in V, \quad E(v) := \frac{1}{2} a(v, v) - (f, v). \tag{3.4}$$

Proof. (*0*) Die stetige und elliptische Bilinearform $a(\cdot, \cdot)$ ist symmetrisch und somit ein Skalarprodukt. Dies folgt aus $b = 0$ sowie der Symmetrie von A.

(*i*) Wir gehen zunächst davon aus, dass u eine variationelle Lösung von (3.3) ist. Dann folgt für alle $v \in H_0^1(\Omega)$:

$$\begin{aligned} E(u) - E(v) &= \frac{1}{2}a(u,u) - (f,u) - \frac{1}{2}a(v,v) + (f,v) \\ &= \frac{1}{2}a(u,u) - a(u,u) - \frac{1}{2}a(v,v) + a(u,v) \\ &= -\frac{1}{2}\Big\{a(u,u) - 2a(u,v) + a(v,v)\Big\} \\ &= -\frac{1}{2}a(u-v,u-v) \leqslant -\frac{\gamma}{2}\|\nabla(u-v)\| \leqslant 0. \end{aligned}$$

Also, $E(u) \leqslant E(v)$.

(*ii*) Nun sei u Minimum des Energiefunktionals. Dann muss gelten:

$$\frac{d}{ds}E(u+sv)\Big|_{s=0} = 0 \quad \forall v \in H_0^1(\Omega).$$

Also, wegen der Symmetrie von $a(\cdot,\cdot)$:

$$\frac{d}{ds}\Big\{\frac{1}{2}a(u+sv,u+sv) - (f,u+sv)\Big\}\Big|_{s=0} = a(u,v) - (f,v) = 0 \quad \forall v \in H_0^1(\Omega).$$

□

Existence of solutions

For considering the general (non-symmetric) case, we need some additional background from linear functional analysis.

Definition 65 (Banach space, Hilbert space)**.** *Let V be a linear space. Let $\|\cdot\| : V \to \mathbb{R}$ be a norm on V. If V is complete with $\|\cdot\|$ it is called a* Banach space*. By $(\cdot,\cdot) : V \times V \to \mathbb{R}$ we introduce a* scalar product *on V. If V is complete with respect to the induced norm $\|\cdot\|_V = (\cdot,\cdot)_{V\times V}^{1/2}$ it is called* Hilbert space*.*

Examples for Hilbert spaces are the space $L^2(\Omega)$ with the scalar product and norm

$$(f,g)_{L^2(\Omega)} := \int_\Omega fg\,dx, \quad \|f\|_{L^2(\Omega)} := \int_\Omega |f|^2 dx,$$

or the space $H_0^1(\Omega)$ with scalar product and norm

$$(f,g)_{H_0^1(\Omega)} := (\nabla f, \nabla g)_{L^2(\Omega)} = \int_\Omega \nabla f \cdot \nabla g\,dx, \quad \|f\|_{H_0^1(\Omega)} := \|\nabla f\|_{L^2(\Omega)} := \int_\Omega |\nabla f|^2 dx.$$

Further, $H^1(\Omega)$ (without zero trace) is a Hilbert space with

$$(f,g)_{H^1(\Omega)} := (f,g)_{L^2(\Omega)} + (\nabla f, \nabla g)_{L^2(\Omega)}, \quad \|f\|_{H^1(\Omega)} := \Big(\|f\|_{L^2(\Omega)}^2 + \|\nabla f\|_{L^2(\Omega)}^2\Big)^{\frac{1}{2}}.$$

Finally, $H^k(\Omega) = W^{k,2}(\Omega)$ is a Hilbert space for every k together with the scalar product

$$(f, g)_{H^k(\Omega)} = \sum_{l=0}^{k} (\nabla^l f, \nabla^l g)_{L^2(\Omega)}.$$

Definition 66 (Dual space)**.** *Let V be a vector space. By V^* we denote the* dual space *of V. The space V^* consists of all* linear functionals

$$l : V \to \mathbb{R}$$

and V^* carries the induces linear structure:

$$l(v + w) = l(v) + l(w) \quad \forall v, w \in V,$$

and

$$l(\alpha v) = \alpha l(v) \quad \forall \alpha \in \mathbb{R},\ \forall v \in V.$$

If V is a normed space with norm $\|\cdot\|_V$, the dual norm $\|\cdot\|_{V^*}$ *on V^* is induced by*

$$\|l\|_{V^*} := \sup_{v \in V,\ \|v\|_V \neq 0} \frac{|l(v)|}{\|v\|_V} = \sup_{v \in V,\ \|v\|_V \leqslant 1} |l(v)|.$$

Example 67. *Let $V = L^2(\Omega)$. A linear functional $l \in V^*$ is given by*

$$l(v) := \int_\Omega v \mathbf{d}x.$$

Linearity of $l(\cdot)$ comes from the linearity of the integral. It holds

$$\|l\|_{L^2(\Omega)^*} = \sup_{\|v\| \leqslant 1} \int_\Omega v \mathbf{d}x \leqslant \underbrace{\|v\|}_{\leqslant 1} \|1\|_\Omega \leqslant \sqrt{|\Omega|}.$$

On the other hand,

$$\|l\|_{L^2(\Omega)^*} = \sup_{\|v\| \leqslant 1} \int_\Omega v \mathbf{d}x \geqslant \int_\Omega \frac{1}{\sqrt{|\Omega|}} \mathbf{d}x = \sqrt{|\Omega|}.$$

Therefore, $\|l\|_{L^2(\Omega)^2} = \sqrt{|\Omega|}$.

Next, let $V = H_0^1(\Omega)$ and $\omega \in H_0^1(\Omega)$ be arbitrary. Then,

$$l_\omega(v) := \int_\Omega \nabla v \cdot \nabla \omega \mathbf{d}x$$

is a linear functional. Once more, linearity follows by using the linearity of the integral and the scalar product. Here, it holds

$$\|l\|_{H_0^1(\Omega)^*} = \sup_{\|\nabla v\| \leqslant 1} (\nabla v, \nabla \omega) \leqslant \|\nabla \omega\|.$$

Then,

$$\|l\|_{H_0^1(\Omega)^*} = \sup_{\|\nabla v\| \leqslant 1} (\nabla v, \nabla \omega) \geqslant \left(\frac{\nabla \omega}{\|\nabla \omega\|}, \nabla \omega \right) = \|\nabla \omega\|.$$

Hence, $\|l\|_{L^2(\Omega)^2} = \|\nabla \omega\|$.

By Hölder's inequality

$$\|fg\|_{L^1(\Omega)} \leqslant \|f\|_p \|g\|_q, \quad p, q \in [1, \infty], \quad \frac{1}{p} + \frac{1}{q} = 1,$$

we can identify the dual space of $L^p(\Omega)$ for $1 < p < \infty$ with $L^p(\Omega)^* \cong L^q(\Omega)$ where $q = (1 + 1/p)^{-1}$ is the dual exponent. For every $g \in L^q(\Omega)$, the mapping

$$f \mapsto \int_\Omega fg dx$$

is s linear functional.

One of the very import theorems in linear functional analysis is

Lemma 68 (Riesz representation theorem)**.** *Let V be a Hilbert space with scalar product $(\cdot, \cdot)_{V \times V}$. For every linear functional $l \in V^*$ there exists a unique element $w \in V$, such that*

$$l(v) = (v, w)_{V \times V} \quad \forall v \in V, \quad \|l\|_{V^*} = \|w\|_V.$$

Vice versa, every $w \in V$ defines a linear functional $l \in V^*$ by

$$l(v) := (v, w)_{V \times V}.$$

The proof is found in the literature [1].

A consequence of Riesz representation theorem is the unique existence of the Laplace solution

Lemma 69 (Existence of Laplace)**.** *Let $f \in L^2(\Omega)$. The variational solution $u \in H_0^1(\Omega)$ of the Laplace equation*

$$(\nabla v, \nabla \phi) = (f, \phi) \quad \forall \phi \in V$$

is unique.

Proof. By

$$l(\phi) := (f, \phi) \quad \forall \phi \in H_0^1(\Omega)$$

a linear functional in $H_0^1(\Omega)^*$ is given, as it is bounded

$$l(\phi) \leqslant (f, \phi) \leqslant \|f\| \, \|\phi\| \leqslant c_p \|f\| \, \|\nabla \phi\|,$$

and obviously linear. The variational formulation is the H_0^1-scalar product. The problem is therefore equivalent to the formulation

$$(u, \phi)_{H_0^1(\Omega)} = l(\phi).$$

According to Riesz, such a $u \in H_0^1(\Omega)$ exists uniquely. Furthermore it holds

$$\|u\|_{H_0^1(\Omega)} = \|\nabla u\| = \|l\|_{H^1(\Omega)^*} \leqslant c_p \|f\|.$$

□

Auf das allgemeine elliptische Problem kann der Riesz'sche Darstellungssatz nicht angewendet werden. Das Problem ist nicht durch ein Skalarprodukt gegeben. Wir benötigen den allgemeineren:

Lemma 70 (Lax-Milgram)**.** *Sei V ein Hilbertraum, $l \in V^*$ ein stetiges lineares Funktional und $a : V \times V \to \mathbb{R}$ eine stetige und elliptische Bilinearform:*

$$a(u,v) \leqslant M\|u\|_V \|v\|_V \quad \forall u,v \in V, \quad a(u,u) \geqslant \gamma\|u\|_V^2 \quad \forall u \in V.$$

Dann existiert eine eindeutige Lösung $u \in V$ der Variationsgleichung:

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

welche stetig von den Daten abhängt:

$$\|u\|_V \leqslant \frac{1}{\gamma}\|l\|_{V^*}.$$

Proof. (i) Für jedes feste $u \in V$ ist wegen der Stetigkeit der Bilinearform ein stetiges lineares Funktional $A_u \in V^*$ definiert:

$$A_u(v) := a(u,v) \leqslant M_u\|v\|_V \quad \forall v \in V.$$

Da V ein Hilbertraum ist gilt der Darstellungssatz von Riesz und es existiert ein Element $Au \in V$, so dass gilt:

$$(Au,v)_V = A_u(v) = a(u,v) \quad \forall v \in V.$$

Es ist also durch den Riesz'schen Darstellungssatz ein Operator $A : V \to V$ definiert, welcher jedem Element $u \in V$ ein Element $Au \in V$ zuordnet, so dass sich die Bilinearform durch das Skalarprodukt ausdrücken lässt. Es gilt:

$$\|Au\|_V = \|A_u\|_{V^*} = \sup_{v\in V} \frac{a(u,v)}{\|v\|_V} \leqslant \sup_{v\in V} \frac{M\|u\|_V\|v\|_V}{\|v\|_V} \leqslant M\|u\|_V. \tag{3.5}$$

(ii) Für ein festes $r \in \mathbb{R}_+$ definieren wir einen weiteren Operator $T_r : V \to V$ durch

$$(T_r u, v)_V := (u,v)_V + r\big(l(v) - (Au,v)_V\big) \quad \forall v \in V.$$

Es gilt für zwei Elemente $u, w \in V$:

$$(T_r u - T_r w, v)_V = (u - w - rA(u-w), v)_V.$$

Wir wählen die Testfunktion $v := T_r u - T_r w$

$$\begin{aligned} \|T_r u - T_r w\|_V^2 &= (u - w - rA(u-w), u - w - rA(u-w))_V \\ &= \|u-w\|_V^2 - 2r(A(u-w), u-w)_V + r^2(A(u-w), A(u-w))_V. \end{aligned} \tag{3.6}$$

Es gilt wegen der Elliptizität

$$(A(u-w), u-w)_V = a(u-w, u-w) \geqslant \gamma\|u-w\|_V^2,$$

und mit (3.5)

$$(A(u-w), A(u-w))_V = \|A(u-w)\|_V^2 \leqslant M^2\|u-w\|_V^2.$$

Zusammen folgt aus (3.6)

$$\|T_r u - T_r w\|_V^2 \leqslant |1 - 2r\gamma + M^2 r^2| \|u-w\|_V^2.$$

Und für

$$|1 - 2r\gamma + M^2 r^2| < 1 \quad \Leftrightarrow \quad r \in \left(0, \frac{2\gamma}{M^2}\right),$$

ist T_r eine Kontraktion und mit dem Banach'schen Fixpunktsatz existiert ein eindeutig bestimmtes $u \in V$ mit $T_r u = u$. Und dann:

$$T_r u = u \quad \Rightarrow \quad (Au, v) = l(v) \quad \forall v \in V.$$

(*iii*) Es existiert also eine Lösung. Angenommen es würden zwei Lösungen $u_1, u_2 \in V$ mit $w := u_1 - u_2$ existieren. Dann gilt:

$$a(w, v) = 0 \quad \forall v \in V.$$

Und wegen der Elliptizität:

$$0 = a(w,w) \geqslant \gamma \|w\|_V^2 \quad \Rightarrow \quad w = 0.$$

(*iv*) Abschließend gilt für die Lösung die *a priori* Schranke:

$$\gamma \|u\|_V^2 \leqslant a(u,u) = l(u) \leqslant \|l\|_{V^*} \|u\|_V.$$

□

Dieser Satz kann nun unmittelbar auf das allgemeine elliptische Problem angewendet werden:

Corollary 71. *Sei* L *ein elliptischer Operator mit den Eigenschaften aus Satz 61. Weiter sei* $f \in L^2(\Omega)$*. Dann hat die Gleichung:*

$$Lu = f \text{ in } \Omega, \quad u = 0 \text{ auf } \partial\Omega,$$

eine eindeutig bestimmte schwache Lösung $u \in H_0^1(\Omega)$ *und es gilt:*

$$\|\nabla u\|_{L^2(\Omega)} \leqslant \frac{c_p}{\gamma} \|f\|_{L^2(\Omega)},$$

mit der Poincaré-Konstante c_p*.*

Proof. Durch $l(v) = (f, v)$ ist ein lineares, stetiges Funktional bestimmt:

$$l(v) = (f,v) \leqslant \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leqslant c_p \|f\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega).$$

Mit dem Satz von Lax-Milgram 70 folgt die Aussage. □

This result is not optimal in terms of the right hand side. Every $f \in L^2(\Omega)$ defines a functional $l \in H_0^1(\Omega)^*$ by

$$l(\phi) = (f, \phi)_\Omega \quad \Rightarrow \quad |l(\phi)| \leqslant c_p \|f\| \, \|\nabla \phi\|.$$

It is however not necessary to choose $f \in L^2(\Omega)$. Lax-Milgram requires $f \in H_0^1(\Omega)^*$ to guarantee a solution $u \in H_0^1(\Omega)$. We can identify $L^2(\Omega)$ as a subspace of the dual space $H_0^1(\Omega)^*$. We define

Definition 72 (Dual space of $H_0^1(\Omega)$)**.** *By*

$$\|l\|_{H^{-1}(\Omega)} := \sup_{\phi \in H_0^1(\Omega)} \frac{l(\phi)}{\|\nabla \phi\|}$$

we define the H^{-1}*-dual norm. By*

$$H^{-1}(\Omega) := \{l : H_0^1(\Omega) \to \mathbb{R}, \quad \|l\|_{H^{-1}(\Omega)} < \infty\}$$

we denote the dual space of $H_0^1(\Omega)$*.*

The space $H^{-1}(\Omega)$ is the dual space of $H_0^1(\Omega)$ with respect to the scalar product $(\nabla \cdot, \nabla \cdot)_\Omega$ that only by Poincaré's inequality is a homogenous.

Further,

Definition 73 (Duality product)**.** *Let* V *be a vector space with dual space* V^**. We define the duality product*

$$\langle l, v \rangle_{V^* \times V} := l(v) \quad \forall v \in V, \ \forall l \in V^*.$$

Using this more general concept, the Laplace problem has a weak solution $u \in H_0^1(\Omega)$ for all $f \in H^{-1}(\Omega)$ and it holds

$$(\nabla u, \nabla \phi) = \langle f, \phi \rangle \quad \forall \phi \in H_0^1(\Omega) \quad \Rightarrow \quad \|\nabla u\| = \|f\|_{H^{-1}(\Omega)}.$$

One examples for a H^{-1}-functional $l \in H^{-1}(\Omega)$, that cannot be represented as an L^2-function via

$$l(\phi) = (f, \phi)_\Omega \quad \forall \phi \in H^1(\Omega),$$

is the evaluation of a line integral, i.e.

$$l(\phi) = \int_{\Gamma_{in}} \phi(x) \mathbf{d}x,$$

where $\Gamma_{in} \subset \Omega$ is a line segment within the domain. By the trace inequality we know, that such a trace exists, i.e. we know the estimate

$$|l(\phi)| = \Big| \int_{\Gamma_{in}} \phi(x) \mathbf{d}x \Big| \leqslant \|\phi\|_{\Gamma_{in}} \sqrt{|\Gamma_{in}|} \leqslant \sqrt{|\Gamma_{in}|} c_{\Gamma_{in}} \|\phi\|_{H^1(\Omega)}.$$

Hence, $l \in H^{-1}(\Omega)$. There exists however no such $f \in L^2(\Omega)$ that represents this functional. Considering the problem

$$(\nabla u, \nabla \phi) = \int_{\Gamma_{in}} \phi(x) \mathbf{d}x \quad \forall \phi \in H_0^1(\Omega),$$

we must use this generalized concept of the right hand side.

Regularität der Lösung Von entscheidender Bedeutung für das weitere Vorgehen ist die Regularität der Lösung $u \in H_0^1(\Omega)$. Falls wir z.B. $u \in H^2(\Omega)$ zeigen können, so folgt aus der kompakten Einbettung **??** $H^2(\Omega) \hookrightarrow C(\Omega)$ die Stetigkeit der Lösung. Um zu garantieren, dass u eine klassische Lösung ist, also $u \in C^2(\Omega)$ benötigen wir höhere Regularität. Der Einbettungssatz fordert in d räumlichen Dimensionen für die kompakte Einbettung $H^m(\Omega) \hookrightarrow C^2(\Omega)$:

$$m - \frac{d}{2} > 2 \quad \Leftrightarrow \quad m > 2 + \frac{d}{2},$$

also benötigen wir $m = 4$ und $u \in H^4(\Omega)$.

Zunächst gilt einfach:

Lemma 74 (Schwacher Laplace)**.** *Für die verallgemeinerte Lösung $u \in H_0^1(\Omega)$ der Laplace-Gleichung ist Δu im schwachen Sinne definiert und liegt in $L^2(\Omega)$.*

Proof. Sei $u \in H_0^1(\Omega)$ die schwache Lösung von $(\nabla u, \nabla v) = (f, v)$ für alle v. Dann gilt:

$$(f, v) = (\nabla u, \nabla v) = -(u, \Delta v) \quad \forall v \in C_0^\infty(\Omega).$$

Also ist $f = -\Delta u$ im Sinne der schwachen Ableitung mit

$$f = -\Delta u \in L^2(\Omega).$$

□

Es zeigt sich nun, dass sich die Regularität der Daten auf die Regularität der Lösung überträgt. Im Gegensatz zu gewöhnlichen Differentialgleichungen spielt jedoch die Regularität des Gebiets eine entscheidende Bedeutung. Ohne Beweis:

Lemma 75 (Regularität der Lösung)**.** *Sei Ω polygonal und konvex, oder besitze einen Rand der Klasse C^2 (lokal parametrisierbar durch eine zweimal stetig differenzierbare Funktion). Im Falle $f \in L^2(\Omega)$ gilt die* a priori *Abschätzung:*

$$\|u\|_{H^2(\Omega)} \leqslant c_s \|f\|_{L^2(\Omega)},$$

mit einer von f unabhängigen Stabilitätskonstante c_s.

Figure 3.1: Gebiet mit *einspringender Ecke*.

Dieser Satz sichert uns die Stetigkeit der Lösung $u \in H_0^1(\Omega) \cap C(\bar{\Omega})$. Die Konvexität des Gebiets ist entscheidend. Ausgeschlossen sind *einspringende Ecken*, also Kanten mit Innenwinkel ω größer $180°$. Wir betrachten auf dem "Tortenstück" aus Abbildung 3.1 die Poisson-Gleichung:

$$-\Delta u = 0, \quad u(r,\theta) = 0 \quad \theta \in [0, \frac{\pi}{\omega}], \quad u(1,\theta) = \sin(\frac{\pi}{\omega}).$$

In Polarkoordinaten ist die Lösung gegeben durch:

$$u(r,\theta) = r^{\frac{\pi}{\omega}} \sin\left(\frac{\pi}{\omega}\theta\right).$$

Für $\omega \in (\pi, 2\pi)$ gilt $u \in H^1(\Omega)$, aber $u \notin H^2(\Omega)$. Der Fall $\omega = 2\pi$ wird *Schlitz-Gebiet* genannt.

Allgemein gilt:

Lemma 76 (Höhere Regularität der Lösung)**.** *Für ein $m \geqslant 0$ sei $f \in H^{m-2}(\Omega)$. Weiter sei Ω ein Gebiet der Klasse C^{m+2}. Dann gilt für die Lösung des Poisson-Problems*

$$\|u\|_{H^m(\Omega)} \leqslant c\|f\|_{H^{m-2}(\Omega)},$$

mit einer von f unabhängigen Konstante c.

Abschließend beweisen wir als wichtige und grundlegende Aussage für elliptische Gleichungen

Lemma 77 (Elliptisches Maximumprinzip)**.** *Für den elliptischen Operator*

$$Lu := -\Delta u + au,$$

mit $a \geqslant 0$ gilt auf dem Gebiet $\Omega \subset \mathbb{R}^d$ das Maximumprinzip*. D.h., eine Funktion $u \in C^2(\Omega) \cap C(\bar{\Omega})$ mit der Eigenschaft $Lu \leqslant 0$ hat kein positives Maximum im Innern. Es gilt also*

$$u(x) \leqslant 0 \text{ für alle } x \in \Omega \text{ oder } \max_{x\in\Omega} u(x) = \max_{x\in\partial\Omega} u(x)$$

Proof. Wir beweisen den einfachen Fall $a > 0$. Sei also $Lu \leqslant 0$ und $z \in \Omega$ ein inneres Maximum mit $u(z) > 0$. Dann gilt notwendig

$$\nabla u(z) = 0, \quad \partial_{xx} u(z) \leqslant 0, \quad \partial_{zz} u(z) \leqslant 0.$$

Also folgt:

$$-\Delta u(z) \geqslant 0$$

aus $au(z) > 0$ folgt der Widerspruch. □

In einer Dimension besagt das Maximumprinzip gerade die Konvexität einer Funktion mit $\partial_{xx} u(x) > 0$.

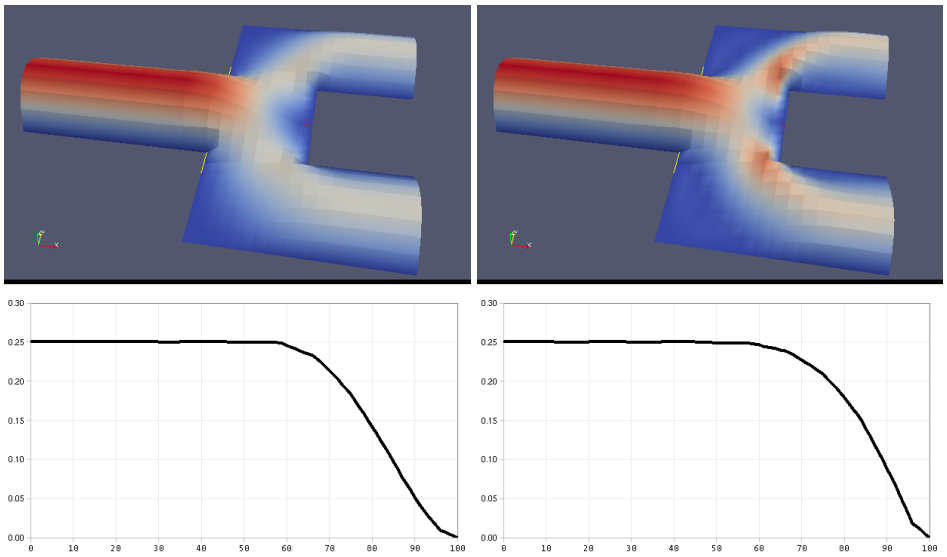


Figure 3.2: Galerkin-Approximation der Navier-Stokes Gleichungen bei Reynoldszahlen $Re = 20$ und $Re = 100$. Oben: 3d-Darstellung der Geschwindigkeit. Unten: Geschwindigkeit aufgetragen über die linie $(y{=}0)$ durch die Mitte des Gebietes.

3.2 Transportdominante Probleme

Eine wichtige Gleichung in der Anwendung sind die Navier-Stokes Gleichungen, ein System nichtlinearer partieller Differentialgleichungen für die Geschwindigkeit $\mathbf{v}$ und den Druck p, gegeben als

$$\begin{aligned}(\mathbf{v}\cdot\nabla)\mathbf{v} - \nu\Delta\mathbf{v} + \nabla p &= f,\\ \operatorname{div}\mathbf{v} &= 0.\end{aligned}$$

Die Konstante $\nu > 0$ gibt die kinematische Viskosität der Flüssigkeit an. Die Gleichung ist nichtlinear, der *Konvektionsterm* $(\mathbf{v}\cdot\nabla)\mathbf{v}$ ist die Richtungsableitung der Geschwindigkeit in Richtung des Geschwindigkeitsvektors selbst

$$(\mathbf{v}\cdot\nabla)\mathbf{v} = \sum_{i=1}^{d} \mathbf{v}_i \frac{\partial \mathbf{v}}{\partial \mathbf{x}_i}.$$

Die Navier-Stokes Gleichungen haben eine enorme Anwendung in der Strömungsmechanik und geben gute Näherungen für das Verhalten von unterschiedlichen Gasen und Flüssigkeiten an, von Blutströmungen in Gefäßen über die Umströmung von Schiffen bis hin zur Umströmung von Flugzeugen. Mathematische Schwierigkeiten ergeben sich einerseits durch den nichtlinearen Konvektionsterm, insbesondere aber durch die Nebenbedingung $\operatorname{div}\mathbf{v} = 0$, der *Inkompressibilität*. Die Navier-Stokes Gleichungen beschreiben inkompressible Materialien, also Flüssigkeiten (oder Gase), die sich durch Krafteinwirkung nicht, oder nur unwesentlich im Volumen ändern. D.h.: die Dichte des Materials bleibt konstant.

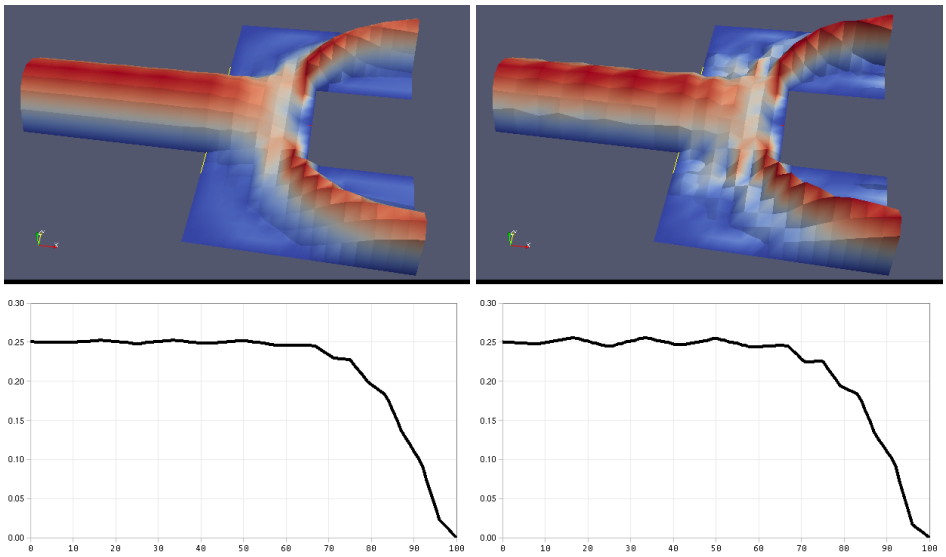


Figure 3.3: Galerkin-Approximation der Navier-Stokes Gleichungen bei Reynoldszahlen $Re = 500$ und $Re = 2500$.

Eine linear Vereinfachung der Gleichung ist die *Oseen-Linearisierung*, bei der der Konvektionsterm durch einen *Transportterm* mit fester Transportrichtung ersetzt wird

$$\begin{aligned}(\beta \cdot \nabla)\mathbf{v} - \nu\Delta\mathbf{v} + \nabla p &= f,\\ \operatorname{div} \mathbf{v} &= 0,\end{aligned}$$

wobei $\beta : \Omega \to \mathbb{R}^d$ ein Transportfeld ist, welches im üblichen wieder divergenzfrei ist, d.h.

$$\operatorname{div} \beta = 0.$$

Die Charakteristik von Strömungen ist durch die sogenannte Reynoldszahl bestimmt

$$Re = \frac{|v|L}{\nu},$$

welche das Verhältnis zwischen Geschwindigkeit $|\mathbf{v}|$, einer charakteristischen Länge L (z.B. die Größe eines umströmungen Flugzeuges) und der Viskosität angibt. Für große Reynoldszahlen heißt die Strömung *transportdominant*.

Für große Reynoldszahlen ist die Galerkin-Approximation der Navier-Stokes-Gleichungen numerisch nicht stabil. In Abbildungen 3.2 sowie 3.3 zeigen wir die Lösung der Navier-Stokes-Gleichungen in einem Kanal mit Verzweigung für die Reynoldszahlen $Re = 20, 100, 500, 2500$. Neben der x-Komponente der Strömungsgeschwindigkeit ist unten jeweils die Norm der Geschwindigkeit über eine horizontale Linie durch die Gebietsmitte aufgezeichnet.

Die Gleichung wurde jeweils auf einem relativ groben Ortsgitter gerechnet. Bei größer werdenden Reynoldszahlen ist die Lösung nicht mehr stabil. Es treten unphysikalische Oszillationen auf. Der Theorie folgend sollte mit wachsender Reynoldszahl (also kleiner werdendem Diffusionsparameter) die Breite der Grenzschicht abnehmen. Trotz Steigerung der

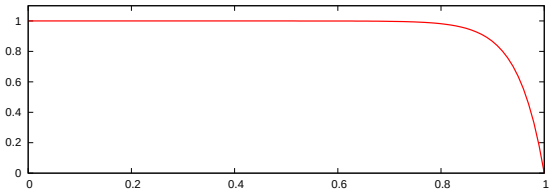


Figure 3.4: Lösung des Modellproblems.

Reynoldszahl um den Faktor 5 in jeder Rechnung bleibt die Breite der numerischen Grenzschicht in Abbildung 3.2 etwa gleich.

Eine weitere Vereinfachung der Oseen-Gleichung erhalten wir bei Vernachlässigung der Divergenzfreiheit. D.h., wir gehen zur Gleichung

$$(\beta \cdot \nabla)\mathbf{v} - \nu\Delta\mathbf{v} = f$$

über und gehen auch davon aus, dass sich die Gleichung im $\mathbb{R}^1$ abspielt. D.h., $\mathbf{v} : \Omega \to \mathbb{R}$ und wir erhalten das einfache Transport-Diffusionsproblem

$$(\beta \cdot \nabla)v - \nu\Delta v = f,$$

eine Gleichung, die formal zu den oben beschriebenen elliptischen Differentialgleichungen gehört.

3.2.1 Analyse eines Modellproblems

Um die auftretende Instabilität näher zu untersuchen betrachten wir ein eindimensionale Modellproblem auf $I = [0, 1]$:

$$-\epsilon u'' + u' = 0, \quad u(0) = 1,\ u(1) = 0. \tag{3.7}$$

Die Lösung kann für jedes $\epsilon > 0$ explizit angegeben werden, verhält sich im Wesentlichen wie $u \approx 1$ und fällt in einer Grenzschicht der Breite $O(\epsilon)$ bei $x = 1$ auf 0 ab, siehe Abbildung 3.4:

$$u(x) = \frac{\exp(1/\epsilon) - \exp(x/\epsilon)}{\exp(1/\epsilon) - 1}.$$

Zur Diskretisierung dieser Modellgleichung verwenden wir stückweise lineare Finite Elemente auf einem Gitter mit Gitterweite h: $\Omega_h = \{x_i := ih, i = 0, \ldots, N\}$. Es sei $\mathbf{u} \in \mathbb{R}^{N+1}$ der Koeffizientenvektor der Lösung, also $\mathbf{u}_i = u(x_i)$. Die Finite-Elemente Diskretisierung entspricht hier einer zentralen Differenzenapproximation:

$$-\epsilon\Delta u(x_i) \approx \epsilon\frac{2\mathbf{u}_i - \mathbf{u}_{i+1} - \mathbf{u}_{i-1}}{h^2}, \quad u'(x_i) \approx \frac{\mathbf{u}_{i+1} - \mathbf{u}_{i-1}}{2h}.$$

Die Lösung ist gegeben durch:

$$A\mathbf{u} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

mit der Matrix:

$$A = \frac{1}{h^2} \begin{pmatrix} \ddots & \ddots & 0 & & \\ -\epsilon - \frac{h}{2} & 2\epsilon & -\epsilon + \frac{h}{2} & 0 & \\ 0 & -\epsilon - \frac{h}{2} & 2\epsilon & -\epsilon + \frac{h}{2} & 0 \\ & 0 & -\epsilon - \frac{h}{2} & 2\epsilon & -\epsilon + \frac{h}{2} \\ & & 0 & \ddots & \ddots \end{pmatrix} \tag{3.8}$$

Diese Matrix ist nur für $\epsilon > h/2$ diagonaldominant. Wir haben aber gesehen, dass gerade für sehr kleine Werte von ϵ (dies entspricht der Viskosität bei Navier-Stokes) die Lösung der Gleichungen interessant sind. Die Annahme $\epsilon > h/2$ ist also zu restriktiv. Die numerische Lösung kann explizit angegeben werden. Um später auch andere Diskretisierungen der Modellgleichung untersuchen zu können betrachten wir das allgemeine Approximationsschema für die Lösung $\mathbf{u} \in \mathbb{R}^{N+1}$:

$$\mathbf{u}_0 = 1,\ \mathbf{u}_N = 0, \quad \alpha_0 \mathbf{u}_{i-1} + \alpha_1 \mathbf{u}_i + \alpha_2 \mathbf{u}_{i+1} = 0,$$

mit Koeffizienten $\alpha_0, \alpha_1, \alpha_2 \in \mathbb{R}$. (Dies sind die Koeffizienten der Matrix, also im Beispiel (3.8) $\alpha_0 = -\epsilon - h/2, \dots$). Wir bestimmen die Lösung mit dem Ansatz:

$$\mathbf{u}_i = \lambda^i,$$

Innerhalb des Gebiets ergibt sich die Gleichung:

$$\alpha_0 \lambda^{i-1} + \alpha_1 \lambda^i + \alpha_2 \lambda^{i+1} = \lambda^{i-1}(\alpha_0 + \alpha_1 \lambda + \alpha_2 \lambda^2) = 0, \ \Rightarrow\ \lambda_\pm = -\frac{\alpha_1}{2\alpha_2} \pm \sqrt{\left(\frac{\alpha_1}{2\alpha_2}\right)^2 - \frac{\alpha_0}{\alpha_2}}.$$

Mit den beiden Nullstellen λ_+ und λ_- kombinieren wir die Lösung zu:

$$\mathbf{u}_i = c_+ \lambda_+^i + c_- \lambda_-^i.$$

Die Randbedingungen $\mathbf{u}_0 = 1$ und $\mathbf{u}_N = 0$ ergeben:

$$c_- = \frac{\lambda_+^N}{\lambda_+^N - \lambda_-^N}, \quad c_+ = -\frac{\lambda_-^N}{\lambda_+^N - \lambda_-^N}.$$

Für den Fall der Galerkin-Diskretisierung gilt nun:

$$\lambda_- = 1, \quad \lambda_+ = \frac{\epsilon + \frac{h}{2}}{\epsilon - \frac{h}{2}}.$$

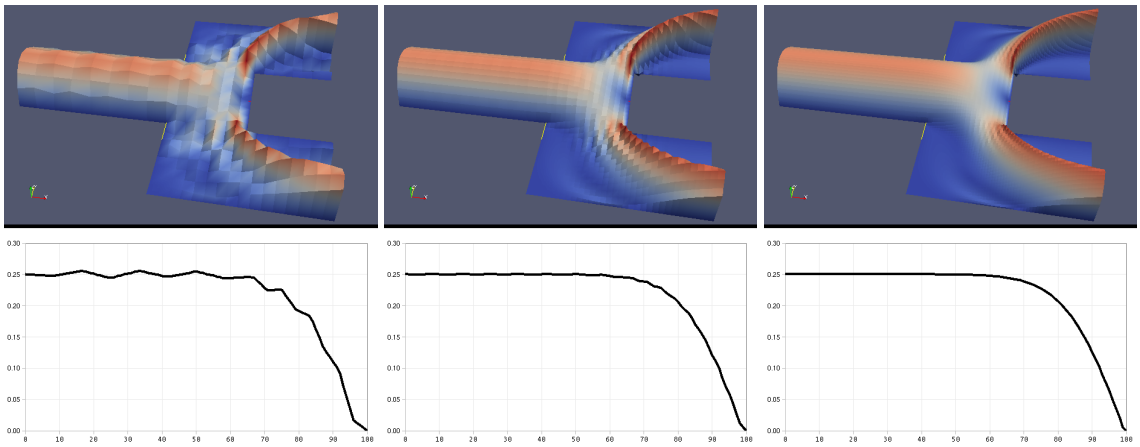


Figure 3.5: Galerkin-Approximation der Navier-Stokes Gleichungen zu $Re = 2500$ auf einer Sequenz von global verfeinerten Gittern mit Gitterweiten $h = 2^{-3}$, $h = 2^{-4}$ sowie $h = 2^{-5}$.

und die Lösung ist gegeben durch ($c_\pm$ ausrechnen!)

$$u_i = \frac{\lambda_+^N - \lambda_+^i}{\lambda_+^N - 1}.$$

Für $\epsilon < h/2$ ist $\lambda_+ < 0$, es wird also eine (nicht physikalische) oszillierende Lösung erzeugt.

In Abbildung 3.5 zeigen wir die Lösung zu $Re = 2500$ auf einer Sequenz von global verfeinerten Gittern. Ist die Gitterweite h klein genug, so wird eine Lösung ohne Oszillationen erzeugt. Bei sehr großen Reynoldszahlen und insbesondere bei dreidimensionalen Problemen ist es schon aus Speichergründen nicht möglich hinreichend kleine Gitter zu verwenden. In den folgenden Abschnitten werden wir für dieses Modellproblem Stabilisierungsmethoden vorstellen und analysieren.

Zur Untersuchung betrachten wir die Galerkin-Formulierung eines etwas verallgemeinerten Modellproblems.

$$u \in H_0^1(\Omega),\ \epsilon(\nabla u, \nabla \phi) + (\beta \cdot \nabla u, \phi) + \alpha(u, \phi) = (f, \phi) \quad \forall \phi \in H_0^1(\Omega), \tag{3.9}$$

mit einer festen Transportrichtung $\beta \in \mathbb{R}^d$ und $\alpha > 0$. Dabei sei $\epsilon > 0$, wir analysieren aber konkret den Fall $\epsilon \ll 1$, werden also stets Fehlerabschätzungen herleiten, die auch für $\epsilon \to 0$ gelten sollen. Diesem Problem ist durch diagonales Testen eine natürliche Energienorm zugeordnet:

$$\epsilon(\nabla u, \nabla u) + (\beta \cdot \nabla u, u) + \alpha(u, u) = \epsilon \|\nabla u\|^2 + \alpha \|u\|^2. \tag{3.10}$$

Der Transport-Term verschwindet bei diagonalem Testen, falls (wie hier) $\beta \in \mathbb{R}^d$ konstant ist, oder falls $\operatorname{div} \beta = 0$ vorliegt.

Lemma 78 (Galerkin-Diskretisierung des Modellproblems)**.** *Für die Finite Elemente Approximation* $u_h \in V_h^{(r)} \subset H_0^1(\Omega)$ *von (3.9) mit Polynomen von Grad* $r \geqslant 1$ *gilt*

$$\epsilon^{\frac{1}{2}}\|\nabla e_h\| + \alpha^{\frac{1}{2}}\|e_h\| \leqslant ch^r(\epsilon^{\frac{1}{2}} + h^{\frac{1}{2}})\|\nabla^{r+1}u\|.$$

Proof. (i) Es gilt mit (3.10) und mit Galerkin-Orthogonalität für $e_h = u - u_h$

$$\begin{aligned}\epsilon\|\nabla e_h\|^2 + \alpha\|e_h\|^2 &= \epsilon(\nabla e_h, \nabla(u-\phi_h)) + (\beta\cdot\nabla e_h, u-\phi_h) + \alpha(e_h, u - i_h u)\\ &\leqslant \frac{\epsilon}{2}\|\nabla e_h\|^2 + \frac{\epsilon}{2}\|\nabla(u-\phi_h)\|^2 + \frac{\alpha}{4}\|e_h\|^2 + \alpha\|u-\phi_h\|^2 + (\beta\cdot\nabla e_h, u-\phi_h).\end{aligned}$$

Der verbleibende Term wird partiell integriert. Dann folgt

$$(\beta\cdot\nabla e_h, u-\phi_h) = -(e_h, \beta\cdot\nabla(u-\phi_h)) \leqslant \|e_h\|\,|\beta|\|\nabla(u-\phi_h)\| \leqslant \frac{\alpha}{4}\|e_h\|^2 + \frac{|\beta|^2}{\alpha}\|\nabla(u-\phi_h)\|^2,$$

da sich die inneren Sprungterme aufgrund der Stetigkeit aufheben. Zusammen folgt

$$\frac{\epsilon}{2}\|\nabla e_h\|^2 + \frac{\alpha}{4}\|e_h\|^2 \leqslant c(\alpha,\beta)\Big((1+\epsilon)\|\nabla(u-\phi_h)\|^2 + \|u-\phi_h\|^2\Big)$$

Einsetzen einer Interpolation $\phi_h = i_h u$ ergibt die gewünschte Absch″zung. □

Für $h \to 0$ stellt sich die gewünschte Konvergenz ein. Im Fall $\epsilon \to 0$ geht allerdings die Kontrolle über die Ableitung $\|\nabla e_h\|$ verloren. Der Beweis suggeriert wegen $c(\alpha,\beta) \sim \alpha^{-1}$ Probleme für $\alpha \to 0$. Dieser Fall kann jedoch durch eine aufwändigere Beweisführung gerettet werden.

a) Upwind-Diskretisierung Die Stabilitätsprobleme zeigen sich in der fehlenden Diagonaldominanz der Matrix (3.8). Bei Verwendung von Differenzenapproximationen könnte zur Approximation des Terms erster Ordnung anstelle des zentralen ein einseitiger Differenzenquotient verwendet werden:

$$u'(x_i) \approx \frac{\mathbf{u}_i - \mathbf{u}_{i-1}}{h}. \tag{3.11}$$

Die Koeffizienten der Matrix sind gegeben durch

$$\alpha_0 = -\epsilon - h, \quad \alpha_1 = 2\epsilon + h, \quad \alpha_2 = -\epsilon,$$

und die Matrix ist diagonaldominant. Wir rechnen wieder die charakteristischen Nullstellen $\lambda_\pm$ aus und erhalten:

$$\lambda_- = 1, \quad \lambda_+ = \frac{\epsilon + h}{\epsilon} = 1 + \frac{h}{\epsilon}.$$

Die Lösung ist wieder gegeben durch

$$\mathbf{u}_i = \frac{\lambda_+^N - \lambda_+^i}{\lambda_+^N - 1}.$$

Jetzt ist die Wurzel λ_+ stets positiv, und die Lösung ist monoton fallend und zeigt keine Oszillationen. Es ist von großer Bedeutung, dass der einseitige Differenzenquotient (3.11) nach links, also *Rückwärts, gegen die Strömung* genommen wurde. Dies trägt der physikalischen Informationsübertragung Rechnung. Man spricht von eine *Upwind-Diskretisierung*. Diese Diskretisierung ist inhärent von 1-ter Ordnung, die Approximation kann nur $O(h)$ betragen.

Die Idee der Upwind-Diskretisierung kann auch auf die Finite-Elemente Approximation übertragen werden, ist jedoch technisch aufwendig und aufgrund der geringen Approximationsordnung nicht vielversprechend.

b) Künstliche Diffusion Diagonaldominanz kann erreicht werden, indem der Diffusionsparameter ϵ in Abhängigkeit der Gitterweite h vergrößert wird. Wir verwenden im Modellproblem (3.7) anstelle von ϵ den Parameter

$$\epsilon_h := \epsilon + \frac{h}{2}.$$

Mit diesem Parameter können wieder die Nullstellen $\lambda_\pm$ und die numerische Lösung $\mathbf{u}$ berechnet werden:

$$\lambda_- = 1,\ \lambda_+ = \frac{\epsilon + h}{\epsilon}, \quad \mathbf{u}_i = \frac{\lambda_+^N - \lambda_+^n}{\lambda_+^N - 1}.$$

In Abbildung 3.6 zeigen wir die Lösung der Navier-Stokes-Gleichungen mit künstlicher Diffusion. Das Lösungsverhalten ist glatt, zeigt also keine Oszillationen, die Lösung ist allerdings sehr *ausgeschmiert*. Obwohl die Reynoldszahl größer wird, bleibt die Breite der Grenzschicht nahezu konstant. Anstelle der physikalischen Diffusion wirkt nur die *künstliche Diffusion*. Bei $h \to 0$ konvergiert die Lösung jedoch gegen die richtige Lösung. Wieder liegt nur Konvergenz erster Ordnung $O(h)$ vor.

Übertragen auf eine Galerkin-Diskretisierung des verallgemeinerten Modellproblems führt die künstliche Diffusion zum Ansatz

$$\Big(\epsilon + h\Big)(\nabla u_h, \nabla \phi_h) + (\beta \cdot \nabla u_h, \phi_h) + \alpha(u_h, \phi_h) = (f, \phi_h) \quad \forall \phi_h \in V_h. \tag{3.12}$$

Zur Fehleranalyse muss eine gestörte Galerkin-Orthogonalität betrachtet werden:

$$(\epsilon + h)(\nabla e_h, \nabla \phi_h) + (\beta \cdot \nabla e_h, \phi_h) + \alpha(e_h, \phi_h) + h(\nabla u, \nabla \phi_h) = 0.$$

Es gilt:

Lemma 79 (Künstliche Diffusion)**.** *Für die Finite Elemente Approximation $u_h \in V_h^{(r)} \subset V$ mit Polynomgrad $r \geqslant 1$ von (3.12) gilt die a priori Abschätzung*

$$\big(\epsilon^{\frac{1}{2}} + h^{\frac{1}{2}}\big)\|\nabla e_h\| + \alpha^{\frac{1}{2}}\|e_h\| \leqslant c(\alpha, \beta) h \|\nabla^2 u\|.$$

Unabhängig vom Polynomgrad ist die Approximationsordnung durch $O(h)$ beschränkt.

Proof. (*i*) Es gilt mit gestörter Galerkin-Orthogonalität:

$$\begin{aligned}&(\epsilon + h)\|\nabla e_h\|^2 + \alpha\|e_h\|^2 \\ &= (\epsilon + h)(\nabla e_h, \nabla(u - \phi_h)) + (\beta \cdot \nabla e_h, u - \phi_h) + \alpha(e_h, u - \phi_h) + h(\nabla u, \nabla(u_h - \phi_h)).\end{aligned}$$

Mit partieller Integration im Transportterm und Abschätzen folgt

$$\begin{aligned}&\frac{1}{2}(\epsilon + h)\|\nabla e_h\|^2 + \frac{\alpha}{4}\|e_h\|^2 \\ &\qquad \leqslant (\epsilon + h)\|\nabla(u - \phi_h)\|^2 + \left(\frac{\alpha}{2} + \frac{|\beta|^2}{\alpha}\right)\|u - \phi_h\|^2 + h(\nabla u, \nabla(u_h - \phi_h)). \quad (3.13)\end{aligned}$$

(*ii*) Wir schätzen nun zunächst den Stabilisierungsterm weiter ab. Es gilt mit partieller Integration

$$h(\nabla u, \nabla(i_h u - \phi_h)) = \sum_{K \in \Omega_h} -h(\Delta u, i_h u - \phi_h)_K + \int_{\partial K} \partial_n u(i_h u - \phi_h)\,do. \quad (3.14)$$

Den ersten Teil schätzen wir ab zu

$$h(\Delta u, i_h u - \phi_h) \leqslant c(\alpha)h^2\|\Delta u\|^2 + \frac{\alpha}{8}\|u - u_h\|^2 + \frac{\alpha}{8}\|u - \phi_h\|^2. \quad (3.15)$$

Der zweite Term in (3.14) verschwindet, da $\partial_n u$ bei hinreichender Regularität ($u \in H^2(\Omega)$) stetig ist. Zusammen folgt bei $\|\Delta u\| \leqslant c\|\nabla^2 u\|$

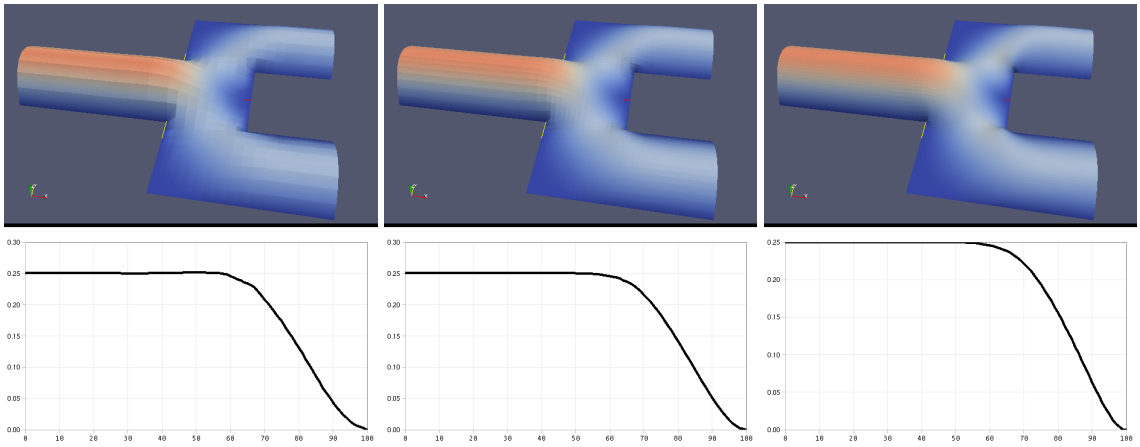
$$(\epsilon + h)\|\nabla e_h\|^2 + \alpha\|e_h\|^2 \leqslant c(\epsilon + h)\|\nabla(u - \phi_h)\|^2 + c(\alpha, \beta)\|u - \phi_h\|^2 + c(\alpha)h^2\|\nabla^2 u\|^2.$$

Mit der Wahl $\phi_h = i_h u$ folgt die Fehlerabschätzung. □

Die künstliche Diffusion liefert unabhängig vom Polynomgrad eine Fehlerabschätzung erster Ordnung. Diese erhält jedoch auch im Fall $\epsilon \to 0$ bei festen $h > 0$ Kontrolle über die Ableitung. Für beliebige Werte $\epsilon > 0$ und $h > 0$ ist die Lösung stabil. Lineare Ordnung ist jedoch im Allgemeinen nicht ausreichend, um eine hinreichende gut Approximation der Navier-Stokes Gleichungen im Fall großer Reynoldszahlen zu garantieren.

Abbildung 3.6 zeigt die Lösung der Navier-Stokes Gleichungen mit künstlicher Diffusion. Dabei ist die Reynolds-Zahl $Re = 2\,500$ und die Rechnung wird auf einer Sequenz verschiedener Gitter durchgeführt. Auch für sehr grobe Gitter ist die Lösung stabil und zeigt keine Oszillationen. Die Lösung wird jedoch am Rand zu stark geglättet, vergleiche hierzu auch Abbildung 3.5.

Figure 3.6: Approximation der Navier-Stokes Gleichungen zu $Re = 2500$ mit künstlicher Diffusion auf einer Sequenz von global verfeinerten Gittern mit Gitterweiten $h = 2^{-3}$, $h = 2^{-4}$ sowie $h = 2^{-5}$.



d) Stromliniendiffusion Das Verfahren der künstlichen Diffusion ist ähnlich zur einfachen Druckstabilisierung der Stokes-Gleichung, siehe Satz **??**. Hier konnte eine Verbesserung durch Hinzufügen eines konsistenten Stabilisierungsterms im Sinne eines Petrov-Galerkin-Verfahrens erreicht werden. Das resultierende Schema nannten wir PSPG-Verfahren. Entsprechend gehen wir nun bei der Stromliniendiffusion vor.

Wir betrachten wieder das Modellproblem

$$-\epsilon u'' + u' = 1,$$

und Testen nun zum Erreichen der Stabilität mit

$$\tilde{\phi} := \phi + \frac{h}{2}\phi'.$$

Wir erhalten die variationelle Formulierung:

$$\epsilon(u', \phi') + (u', \phi) + \frac{h}{2}(u', \phi') - \frac{\epsilon h}{2}(u'', \phi') = (1, \phi). \tag{3.16}$$

Eine Modifikation der rechten Seite entfällt wegen $(1, \phi') = 0$.

Wir wenden diesen Zugang auf das modifizierte Modellproblem $-\epsilon\Delta u + \beta \cdot \nabla u = \alpha u = f$ an und testen mit

$$\tilde{\phi} = \phi + \delta\beta \cdot \nabla\phi.$$

Der Stabilisierungsparameter δ wird später von h abhängen. Im Falle der Stromliniendiffusion kann dieser Parameter jedoch nicht immer als $\delta = h$ gewählt werden. Eine genaue Analyse folgt im Laufe der Beweise. Wir erhalten

$$A(u, \phi) + S(u, \phi) = (f, \phi) + S_f(\phi),$$

mit

$$A(u,\phi) = \epsilon(\nabla u, \nabla\phi) + (\beta\cdot\nabla u, \phi) + \alpha(u,\phi),$$

und dem Stabilisierungsterm

$$\begin{aligned} S(u,\phi) &= (-\epsilon\Delta u + \beta\cdot\nabla u + \alpha u, \delta\beta\cdot\nabla\phi) \\ &= \delta(\beta\cdot\nabla u, \beta\cdot\nabla\phi) + \delta(-\epsilon\Delta u + \alpha u, \beta\cdot\nabla\phi) \\ S_f(\phi) &= \delta(f, \beta\cdot\nabla\phi). \end{aligned} \tag{3.17}$$

Dabei dient alleine der erste Term zur Stabilisierung. Alle weiteren Terme garantieren einzig die Konsitenz des Verfahrens. Denn für $u \in C^2(\Omega)$ gilt $S(u,\phi) = 0$. Der Stabilisierungsterm führt Diffusion nur in Richtung des Transports ein. Es gilt mit $\partial_\beta := (\beta\cdot\nabla)$

$$\delta(\beta\cdot\nabla u, \beta\cdot\nabla\phi) = \delta(\partial_\beta u, \partial_\beta\phi) = -\delta(\partial_{\beta\beta} u, \phi).$$

Zur Analyse der Stromliniendiffusion führen wir zunächst eine geeignete Norm ein. Unter Beachtung von $(u, \beta\cdot\nabla u) = -(\beta\cdot\nabla u, u)$ folgt $(u, \beta\cdot\nabla u) = 0$ und es gilt für diskrete Funktionen $u_h \in V_h \subset V$:

$$A(u_h,u_h) + S(u_h,u_h) = \epsilon\|\nabla u_h\|^2 + \alpha\|u_h\|^2 + \delta\|\beta\cdot\nabla u_h\|^2 + \delta\epsilon(\Delta_h u_h, \beta\cdot\nabla u_h),$$

wobei $\Delta_h u_h$ zellweise zu verstehen ist. Für diesen letzten Term gilt mit Hilfe der inversen Abschätzung auf jeder Zelle

$$\delta\epsilon(\Delta u_h, \beta\cdot\nabla u_h)_K \leqslant c\delta h^{-1}\epsilon\|\nabla u_h\|_K\,\|\beta\cdot\nabla u\|_K \leqslant \frac{\epsilon}{2}\|\nabla u_h\|_K^2 + \frac{c^2\delta^2 h^{-2}\epsilon}{2}\|\beta\cdot\nabla u_h\|_K^2.$$

Zusammen ergibt sich

$$A(u_h,u_h) + S(u_h,u_h) \geqslant \frac{\epsilon}{2}\|\nabla u_h\|^2 + \alpha\|u_h\|^2 + \delta\left(1 - \frac{\delta\epsilon}{2h^2}\right)\|\beta\cdot\nabla u_h\|^2. \tag{3.18}$$

Lemma 80 (Elliptizität der Stromliniendiffusion)**.** *Es sei $\epsilon > 0$, $\beta \in \mathbb{R}^2$ und $\alpha > 0$ beliebig. Es sei $\epsilon < h$. Im Fall*

$$\delta \leqslant h,$$

ist die Methode der Stromliniendiffusion elliptisch mit

$$A(u_h,u_h) + S(u_h,u_h) \geqslant c|||u_h|||_\delta^2,$$

mit der Norm

$$|||u_h|||_\delta^2 := \epsilon\|\nabla u_h\|^2 + \alpha\|u_h\|^2 + \delta\|\beta\cdot\nabla u_h\|^2.$$

Proof. Wir betrachten den Fall $\epsilon < h$. Dann ist

$$\frac{\delta\epsilon}{2h^2} < \frac{\delta}{2h}.$$

Im Fall $\delta \leqslant h$ folgt

$$\left(1 - \frac{\delta\epsilon}{2h^2}\right) > 1 - \frac{\delta}{2h} \geqslant \frac{1}{2},$$

und Elliptizität folgt mit Hilfe von (3.18). □

Remark 81. *Satz 80 zeigt Elliptizität im Fall $\epsilon < h$. Der Fall $\epsilon > h$ wird hier explizit nicht behandelt. In diesem (einfachen) Fall wird jedoch überhaupt keine Stabilisierung benötigt, da die einfache Galerkin-Formulierung bereits eine optimale Approximation liefert.*

Mit diesen Vorbereitungen können wir nun einen Konvergenzsatz für die Stromliniendiffusion formulieren:

Lemma 82 (Konvergenz der Stromliniendiffusion)**.** *Es sei $\epsilon > 0$, $\alpha > 0$ sowie $\beta \in \mathbb{R}^d$. Im Fall $\delta \leqslant h$ und $\epsilon < h$ gilt für die Stromliniendiffusion bei Verwendung von Finiten Elementen vom Grad $r \geqslant 1$ die a priori Fehlerabschätzung*

$$\epsilon^{\frac{1}{2}}\|\nabla e_h\| + \alpha^{\frac{1}{2}}\|e_h\| + \delta^{\frac{1}{2}}\|\beta\cdot\nabla e_h\| \leqslant ch^{r+\frac{1}{2}}\|\nabla^{r+1}u\|.$$

Proof. (*i*) Für den Fehler in der Dreifachnorm gilt

$$|||e_h|||_\delta \leqslant |||u - i_h u|||_\delta + |||i_h u - u_h|||_\delta. \tag{3.19}$$

Der Interpolationsfehler kann unmittelbar abgeschätzt werden. Für den rein diskreten Anteil gilt mit der Elliptizität weiter

$$\begin{aligned} c|||i_h u - u_h|||_\delta^2 &\leqslant (A+S)(i_h u - u_h, i_h u - u_h) \\ &= \underbrace{(A+S)(u - u_h, i_h u - u_h)}_{=0} - (A+S)(u - i_h u, i_h u - u_h). \end{aligned} \tag{3.20}$$

Der erste Term entfällt aufgrund der Galerkin-Orthogonalität, da die Stromliniendiffusion eine konstistente Methode ist.

(*ii*) Wir schätzen jetzt den Galerkin-Aneteil $A(\cdot,\cdot)$ ab. Es gilt mit $\eta := u - i_h u$ und $\psi_h := i_h u - u_h$ und einer Konstante $c_1 > 0$

$$A(\eta,\psi_h) \leqslant \frac{\epsilon}{4c_1}\|\nabla\eta\|^2 + c_1\epsilon\|\nabla\psi_h\|^2 + \frac{\alpha}{4c_1}\|\eta\|^2 + c_1\alpha\|\psi_h\|^2 + (\beta\cdot\nabla\eta,\psi_h).$$

Partielle Integration im letzten Term liefert

$$(\beta\cdot\nabla\eta,\psi_h) = -(\eta,\beta\cdot\nabla\psi_h) \leqslant \frac{1}{4c_1\delta}\|\eta\|^2 + \delta c_1\|\beta\cdot\nabla\psi_h\|^2$$

Mit $\|\psi_h\| \leqslant \|u - i_h u\| + \|u - u_h\|$ folgt schließlich

$$A(\eta,\psi_h) \leqslant c(c_1)\Big(|||u - i_h u|||_\delta^2 + \delta^{-1}\|u - i_h u\|^2\Big) + c_1|||u - u_h|||_\delta^2. \tag{3.21}$$

Die Konstante $c_1 > 0$ kann dabei beliebig (klein) gewählt werden, um die entsprechenden Terme später auf der linken Seite absorbieren zu können.

(*iii*) Schließlich muss der Stabilisierungsterm abgeschätzt werden. Es gilt mit $\psi_h := i_h u - u_h$:

$$\begin{aligned} S(u - i_h u,\psi_h) &= \delta(\beta\cdot\nabla(u - i_h u),\beta\cdot\nabla\psi_h) - \epsilon\delta(\Delta(u - i_h u),\beta\cdot\nabla\psi_h) + \alpha\delta(u - i_h u,\beta\cdot\nabla\psi_h) \\ &\leqslant c_1\delta\|\beta\cdot\nabla\psi_h\|^2 + c(c_1)\Big(\delta\|\beta\cdot\nabla(u - i_h u)\|^2 + \epsilon^2\delta\|\Delta(u - i_h u)\|^2 + \alpha^2\delta\|u - i_h u\|^2\Big). \end{aligned}$$

Die Konstante $c_1 > 0$ kann wieder beliebig gewählt werden. Nun gilt

$$\delta\|\beta\cdot\nabla\psi_h\|^2 \leqslant \delta\|\beta\cdot\nabla e_h\|^2 + \delta\|\beta\cdot\nabla(u-i_h u)\|^2 \leqslant |||e_h|||_\delta^2 + |||u-i_h u|||_\delta^2.$$

Zusammengefasst gilt somit

$$S(u-i_h u, i_h u - u_h) \leqslant c_1|||e_h|||_\delta^2 + c_1(c_1)|||u-i_h u|||_\delta^2 + c(c_1)\epsilon^2\delta\|\nabla^2(u-i_h u)\|^2. \quad (3.22)$$

(*iv*) Wir kombinieren (3.19), (3.20), (3.21) sowie (3.22) und erhalten bei Wahl von $c_1 > 0$ klein genug

$$|||e_h|||_\delta^2 \leqslant c(\alpha)|||u-i_h u|||_\delta^2 + c\epsilon^2\delta\|\nabla^2(u-i_h u)\|^2 + c\delta^{-1}\|u-i_h u\|^2.$$

Für die Dreifachnorm gilt bei der Interpolation mit Finiten Elementen von Polynomgrad $r \geqslant 1$:

$$\begin{aligned}|||u-i_h u|||_\delta^2 &\leqslant \epsilon h^{2r}\|\nabla^{r+1}u\|^2 + \alpha h^{2r+2}\|\nabla^{r+1}u\|^2 + \delta|\beta|h^{2r}\|\nabla^{r+1}u\|^2\\ &= h^{2r}\Big(\epsilon + \alpha h^2 + \delta|\beta|\Big)\|\nabla^{r+1}u\|^2.\end{aligned}$$

Hinzu kommen die Terme

$$\epsilon^2\delta\|\nabla^2(u-i_h u)\|^2 \leqslant \epsilon^2\delta h^{2r-2}\|\nabla^{r+1}u\|^2,$$

sowie

$$\delta^{-1}\|u-i_h u\| \leqslant c\delta^{-1}h^{2r+2}\|\nabla^{r+1}u\|^2.$$

Im Fall $\epsilon < h$ folgt für $\delta \leqslant h$ die gewünschte Aussage. □

Die Stromliniendiffusion liefert für transportdominante Probleme, also im schwierigen Fall $\epsilon < h$ eine Fehlerabschätzung der Ordnung $O(h^{r+\frac{1}{2}})$. Dabei bleibt in Transportrichtung β stets Kontrolle über die Ableitung erhalten. Die Stromliniendiffusion kontrolliert nicht die Ableitung senkrecht zum Transport $\beta^\perp$, "cross-wind" genannt.

Die Stromliniendiffusion ist eine der weit verbreitesten Methoden zur Stabilisierung der Navier-Stokes Gleichungen im Fall großer Reynolds-Zahlen. Wir konkretisieren daher die variationelle Formulierung, direkt in Verbindung mit einer entsprechenden Druckstabilisierung mit Hilfe der PSPG-Methode:

$$\begin{aligned}(\mathbf{v}\cdot\nabla\mathbf{v},\phi) + \nu(\nabla\mathbf{v},\nabla\phi) - (p,\nabla\cdot\phi) + \delta(\mathbf{v}\cdot\nabla\mathbf{v},\beta\cdot\nabla\phi)&\\ +\delta(-\nu\Delta\mathbf{v}+\nabla p-\mathbf{f},\beta\cdot\nabla\phi) &= (\mathbf{f},\phi)\\ (\nabla\cdot\mathbf{v},\xi) + \alpha(\nabla p,\nabla\xi) + \alpha(-\nu\Delta\mathbf{v}+\mathbf{v}\cdot\nabla\mathbf{v}-\mathbf{f},\nabla\xi) &= 0.\end{aligned}$$

Wesentliche Anteile der Stabilisierung dienen wieder nur dem Herstellen einer konsistenten Formulierung. Im Falle der Navier-Stokes Gleichungen werden die Stabilisierungsparameter α, δ meist folgendermaßen gewählt:

$$\alpha \sim \delta \sim \min\left(\frac{h^2}{\nu}, \frac{h}{|\mathbf{v}|_\infty}\right).$$

Für die sehr aufwändige Analyse verweisen wir auf die Literatur [6].

e) Lokale Projektionen Die numerische Oszillationen zeigen sich immer auf er feinsten Gitternskala. Man vergleiche hierzu die explizit hergeleitete Lösung des eindimensionalen Modellproblems. Die Methode der lokalen Projektionen setzt nun gerade auf dieser feinsten Skala an.

Dafür sei $\tilde{V}_h$ ein *gröberer* Raum. Wir betrachten hier nur den Raum V_{2h}, also den Raum mit gleichem Polynomgrad wie V_h, jedoch auf dem Gitter Ω_{2h} mit doppelter Zellweite und die Projektion $\pi_h : V_h \to V_{2h}$. Diffusion in die Strömungsrichtung wird nun nicht voll konsistent, sondern nur für den Projektionsfehler eingeführt:

$$S(u,\phi) = \sum_{K\in\Omega_h} \delta\big(\beta\cdot\nabla(u_h - \pi_h u_h), \beta\cdot\nabla(\phi_h - \pi_h\phi_h)\big)_P.$$

Durch diagonales Testen folgt sofort die Abschätzung:

$$A(u_h,u_h) + S(u_h,u_h) = |\!|\!| u_h |\!|\!|_{lps}^2 = \epsilon\|\nabla u_h\|^2 + \alpha\|u_h\|^2 + \sum_{K\in\Omega_h}\delta\|\beta\cdot\nabla(u_h-\pi_h u_h)\|^2.$$

Die Ableitung in β-Richtung ist nun in einem schwächeren Sinne unter Kontrolle: Im Fall $u_h \in V_{2h}$ gilt $u_h - \pi_h u_h = 0$. Oszillationen werden somit nur auf der Ebene des feinsten Gitters verhindert. Für die aufwändige Analyse der LPS-Methode verweisen wir auf die Literatur [5, 6].

Wir skizzieren hier das Vorgehen. Für den Fehler gilt mit obiger Fehleridentitiät

$$|\!|\!| u-u_h |\!|\!|_{lps}^2 = A(u-u_h,u-u_h) + S(u-u_h,u-u_h).$$

Wir fügen eine Interpolation ein

$$|\!|\!| u-u_h |\!|\!|_{lps}^2 = \underbrace{A(u-u_h,u-I_h u) + S(u-u_h,u-I_h u)}_{=(I)} + \underbrace{A(u-u_h,I_h u-u_h) + S(u-u_h,I_h u-u_h)}_{=(II)}$$

Die Methode ist nicht im strengen Sinne konsistent. Es gilt die gestörte Galerkin-Orthogonalität

$$A(u-u_h,\psi_h) + S(u-u_h,\psi_h) = \underbrace{F(\psi_h) - F(\psi_h)}_{=0} + S(u,\psi_h)$$

Hiermit folgt aus dem Term (II)

$$(II) = S(u, I_h u - u_h) \leqslant \sum_K \delta\|\beta\cdot\nabla(u-\pi_h u)\|_K \|\beta\cdot\nabla\big((I_h u - u_h) - \pi_h(I_h u - u_h)\big)\|_K$$

Wir gehen davon aus, dass $\pi_h = I_{2h}$ die Interpolation in das Grobgitter ist. Dann gilt für den ersten Term

$$\|\beta\cdot\nabla(u - I_{2h}u)\| \leqslant c(\beta)(2h)^r\|\nabla^{r+1}u\| \leqslant c' h^r \|f\|_{H^{r-1}(\Omega)}.$$

Den zweiten Term schätzen wir mit $\pm u$ ab als

$$\begin{aligned}
&\|\beta\cdot\nabla\big((I_h u - u_h) - \pi_h(I_h u - u_h)\big)\|_K \\
&\qquad\leqslant \|\beta\cdot\nabla\big((u - u_h) - \pi_h(u - u_h)\big)\|_K + \|\beta\cdot\nabla\big((u - I_h u) - \pi_h(u - I_h u)\big)\|_K \\
&\quad\leqslant \|\beta\cdot\nabla\big((u - u_h) - \pi_h(u - u_h)\big)\|_K + \|\beta\cdot\nabla(u - I_h u)\|_K + \|\beta\cdot\nabla\pi_h(u - I_h u)\|_K \\
&\qquad\leqslant \|\beta\cdot\nabla\big((u - u_h) - \pi_h(u - u_h)\big)\|_K + c\|\beta\cdot\nabla(u - I_h u)\|_K,
\end{aligned}$$

wobei wir im letzten Schritt die Stabilität der Projektion π_h ausgenutzt haben. Dies erfordert unter Umständen die Verwendung eines speziellen Operators π_h, z.b. der Clement-Interpolation. Sie Kapitel 4.1. Hier gilt wieder

$$c\|\beta\cdot\nabla(u - I_h u)\|_K \leqslant ch^r\|f\|_{H^{r-1}(\mathbf{K})}$$

Zusammen gilt mit Young'scher Ungleichung

$$(II) \leqslant \sum_K \frac{\delta}{2}\|\beta\cdot\nabla\big((u - u_h) - \pi_h(u - u_h)\big)\|_K^2 + \sum_K c\delta h^{2r}\|f\|^2_{H^{r-1}(K)}$$

Der erste Term wird links in der Norm $|||u-u_h|||^2_{lps}$ versteckt. Der andere Term hat die richtige Ordnung (nach Ziehen der Wurzel bleibt $\mathcal{O}(h)$).

Der verbleibende Term (I) wird wir im Standardfall mit Hilfe der Stetigkeit von $A(\cdot,\cdot)+S(\cdot,\cdot)$ und den Interpolationsabschätzungen abgeschätzt. Bleibende Terme werden gegebenenfalls links absorbiert.

4 A posteriori Fehlerschätzung und adaptive Finite Elemente

In diesem Abschnitt befassen wir uns mit der *a posteriori Fehlerschätzung*. Hier geht es zum einen um die Frage, eine berechenbare Schranke $\eta_h \in \mathbb{R}$ für den Fehler der Finite Elemente Approximation angeben zu können:

$$|J(u - u_h)| \leqslant \eta_h(\Omega_h, u_h, f),$$

also eine Größe η_h, welche bei Kenntnis des Gitters, der Lösung und der Problemdaten berechenbar ist. J soll hier ein beliebiges Fehlerfunktional sein. Im Gegensatz zu *a priori* Fehlerabschätzungen muss auf unbekannte Konstanten soweit wie möglich verzichtet werden, d.h., zur Berechnung des Schätzers η_h dürfen nur das Gitter Ω_h, die Daten f sowie die berechnete diskrete Lösung u_h eingehen, nicht aber etwa die Lösung $u \in H_0^1(\Omega)$ oder Konstanten welche nicht berechnet werden können (wie die Konstante des Spurlemmas, der Interpolation, oder die Poincaré Konstante).

Der zweite Aspekt in diesem Kapitel ist die Berechnung von *Fehlerindikatoren* $\{\eta_T\}_{T\in\Omega_h}$. Das sind verteilte Größen, welche den lokalen Fehleranteil angeben. Lokal kann hier bedeuten, dass etwa η_T den Fehlerbeitrag der Gitterzelle $T \in \Omega_h$ angibt, oder η_i den Fehlerbeitrag des Einzugbereichs einer Finite Elemente Basisfunktion. Solche lokalen Fehlerindikatoren sind Grundlage von *adaptiven Verfahren* nach dem folgenden Muster:

1. Berechne diskrete Lösung $u_h \in V_h$
2. Schätze den Fehler η_h. Falls $\eta_h < \mathrm{TOL}$ einer vorgegebenen Fehlertoleranz, Abbruch.
3. Erstelle lokale Fehlerindikatoren $\{\eta_T\}_{T\in\Omega_h}$ und *verfeinere das Gitter* $\Omega_h \xrightarrow{\{\eta_T\}} \Omega_h'$. Weiter bei 1 mit V_h' auf Ω_h'.

Wir betrachten in diesem Abschnitt exemplarisch die Poisson-Gleichung:

$$u \in V := H_0^1(\Omega) \qquad (\nabla u, \nabla\phi) = (f, \phi) \qquad \forall \phi \in V, \tag{4.1}$$

$$u_h \in V_h \subset V \qquad (\nabla u_h, \nabla\phi_h) = (f, \phi_h) \quad \forall \phi_h \in V_h. \tag{4.2}$$

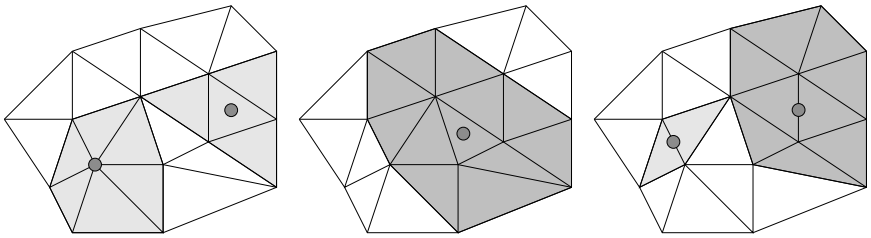


Figure 4.1: Definition der Patche. Links: Knotenpatch und "kleine Zellpatch" P_T. Mitte: "großer Zellpatch" $\tilde{P}_T$. Rechts: "kleiner Kantenpatch" P_E (hell) und "großer Kantenpatch" $\tilde{P}_E$ (dunkel).

4.1 Die Clement-Interpolation

Der natürliche Raum zur Analyse von elliptischen Differentialgleichungen ist der Sobolew-Raum $H_0^1(\Omega)$. Der Makel der Knoteninterpolation ist die Verwendung von Punktwerten zur Definition des Interpolationsoperators mittels $I_h v(a_i) = v(a_i)$. Diese Knotenfunktionale sind auf $H_0^1(\Omega)$ nicht stetig definiert. Oft benötigen wir jedoch Interpolationen von Funktionen mit dieser minimalen Regularität. Die Clement-Interpolation ist ein Interpolationsoperator welcher anstelle von Funktionsauswertungen lokale Mittelwerte verwendet. Diese Mittelwerte sind als Funktionale auf dem H^1 beschränkt. Wir definieren:

Definition 83 (Patch)**.** *Sei Ω_h ein strukturreguläres Gitter. Für jeden Knoten $a \in \Omega_h$, jedes Element $T \in \Omega_h$ und jede Kante $E \in \Omega_h$ definieren wir den* Knotenpatch $P_a \in \Omega_h$*, die* Zellpatche $P_T \in \Omega_h$ *und* $\tilde{P}_T \in \Omega_h$ *sowie die* Kantenpatche $P_E \in \Omega_h$ *und* $\tilde{P}_E \in \Omega_h$

$$
\begin{aligned}
P_a &:= \bigcup_{T \in \Omega_h,\, x_i \in \bar{T}} T \\
P_T &:= \bigcup_{T' \in \Omega_h,\, \exists E \in \Omega_h,\, E = \bar{T} \cap \bar{T}'} T', \\
\tilde{P}_T &:= \bigcup_{P_a \in \Omega_h,\, a \in \bar{T}} P_a,
\end{aligned}
$$

$$
\begin{aligned}
P_E &:= \bigcup_{T' \in \Omega_h,\, E \subset T'} T', \\
\tilde{P}_E &:= \bigcup_{P_a \in \Omega_h,\, a \in \bar{E}} P_a
\end{aligned}
$$

In Abbildung (4.1) zeigen wir ein Beispiel solcher Patche. Zur Definition des Clement-Interpolationsoperators $C_h : V \to V_h$ werden nun anstelle von Punktwerten $v(a)$ Mittelwerte über die Knotenpatche P_a verwendet. Diese Funktionale sind auf dem $V = H_0^1(\Omega)$ beschränkt.

Lemma 84 (Clement-Interpolation)**.** *Es sei Ω_h ein form- größen- und strukturreguläres Gitter. Dann gibt es einen stetigen linearen Operator $C_h : V \to V_h$ in den Raum der linearen Finiten Elemente mit den folgenden Eigenschaften:*

$$\|v - C_h v\|_{L^2(T)} \leqslant c h_T \|\nabla v\|_{L^2(\tilde{P}_T)}, \quad \|v - C_h v\|_{L^2(E)} \leqslant c h_E^{\frac{1}{2}} \|\nabla v\|_{L^2(\tilde{P}_E)} \quad \forall v \in V = H_0^1(\Omega).$$

Proof. Wir definieren zunächst die Knotenfunktionale der Clement-Interpolation. Für jeden Knoten $x_i \in \Omega_h$ sei:

$$\chi_i : L^2(P_{x_i}) \to \mathbb{R}, \quad \chi_i(v) := \begin{cases} \frac{1}{|P_{x_i}|} \int_{P_{x_i}} v(x)\,dx & x_i \notin \partial\Omega \\ 0 & x_i \in \partial\Omega \end{cases}$$

Diese Knotenfunktionale sind linear, auf $L^2(\Omega)$ und also auch auf $V = H_0^1(\Omega)$ beschränkt. Nun sei $\hat{P}_x$ ein Referenzpatch. Hier gilt:

$$|\hat{\chi}_i(\hat{v})| = \frac{1}{|\hat{P}_x|} \int_{\hat{P}_x} \hat{v}\,d\hat{x} \leqslant c\|\hat{v}\|_{\hat{P}_x}.$$

Weiter sei $T_{x_i} : \hat{P}_x \to P_{x_i}$ mit $\det(\nabla T_x) = |P_{x_i}| = O(h^2)$ und $\|\nabla T_{x_i}\|_\infty = O(h_i)$. Dann gilt:

$$\|v - \chi_i(v)\|^2_{L^2(P_{x_i})} \leqslant h^2 \|\hat{v} - \hat{\chi}_i(\hat{v})\|^2_{L^2(\hat{P}_x)}.$$

Weiter mit dem Bramble-Hilbert-Lemma und Rücktransformation:

$$\|v - \chi_i(v)\|^2_{L^2(P_{x_i})} \leqslant h^2 c_{bhl} \|\hat{\nabla}\hat{v}\|^2_{\hat{P}_x} \leqslant c_{bhl} h_i^2 \|\nabla v\|^2_{L^2(P_{x_i})}.$$

(*ii*) Wir definieren die Interpolierende als:

$$C_h v(x) := \sum_{x_i \in \Omega_h} \chi_i(v) \phi_h^{(i)}(x) \quad \forall v \in V.$$

Für die Knotenbasis gilt

$$\sum_{x_i \in \bar{T}} \phi_h^{(i)}(x)\Big|_T \equiv 1,$$

da die Interpolation auch in den Randknoten definiert ist. Weiter folgt mit $\|\phi_h^{(i)}\|_{L^\infty(T)} \leqslant 1$:

$$\begin{aligned} \|v - C_h v\|_T &= \left\| v \left(\sum_{x_i \in \bar{T}} \phi_h^{(i)} \right) - \sum_{x_i \in \bar{T}} \chi_i(v) \phi_h^{(i)} \right\|_T \\ &\leqslant \sum_{x_i \in \bar{T}} \|(v - \chi_i(v))\phi_h^{(i)}\|_T \leqslant \sum_{x_i \in \bar{T}} \|v - \chi_i(v)\|_{P_{x_i}} \leqslant \sum_{x_i \in \bar{T}} c_{bhl} h_i \|\nabla v\|_{P_{x_i}} \\ &\leqslant c_{bh} \sqrt{c(T)} \tilde{h}_T \|\nabla v\|_{\tilde{P}_T}, \end{aligned}$$

mit $\tilde{h}_T = \operatorname{diam}(\tilde{P}_T)$ und $c(T)$ der Anzahl der Zellen in einem Patch. Aus der Formregularität des Gitters folgt $c(T) \leqslant c_T$ gleichmäßig in $h \to 0$. Die Abschätzung des Interpolationsfehlers auf der Kante folgt entsprechend durch geeignete Transformation auf die Referenzkante. □

Die Clement-Interpolation wird H^1-stabil genannt. Es gilt:

Lemma 85 (H^1-Stabilität der Clement-Interpolation)**.** *Auf einem form-, größen und strukturregulärem Gitter Ω_h folgt für die Clement-Interpolation:*

$$\|\nabla C_h v\|_{L^2(T)} \leqslant c\|\nabla v\|_{\tilde{P}_T}.$$

Proof. Die Knotenfunktionale sind H^1-stabil. Deswegen kann auf gleichem Wege die H^1-Fehlerabschätzung hergeleitet werden:

$$\|\nabla(v - C_h v)\|_{L^2(T)} \leqslant c_I\|\nabla v\|_{\tilde{P}_T}.$$

Dann gilt:

$$\|\nabla C_h v\|_T \leqslant \|\nabla(v - C_h v)\|_T + \|\nabla v\|_T.$$

Hieraus folgt die Behauptung. □

4.2 Residuenbasierte Fehlerschätzer

Ziel von a posteriori Fehlerschätzern ist es, berechenbare Größen herzuleiten, die eine Abschätzung des Diskretisierungsfehlers liefern. Dies bedeutet insbesondere, dass die unbekannte Lösung $u \in V = H_0^1(\Omega)$ nicht eingeht. Ein zentraler Begriff ist das *Residuum*, ähnlich dem *Defekt* $d(x) = b - Ax$ bei linearen Gleichungssystemen.

Definition 86 (Residuum)**.** *Das Residuum der Gleichung (4.1) an der Stelle $u_h \in V_h$ ist ein Funktional $R_h : V \to \mathbb{R}$:*

$$R_h(u_h)(\phi) = (f, \phi) - (\nabla u_h, \nabla\phi) \quad \forall\phi \in V.$$

Das Residuum steht im engen Zusammenhang zum Fehler $e_h := u - u_h$ der Finite Elemente Approximation:

Lemma 87 (Residuum)**.** *Das Residuum $R_h(\cdot)$ ist ein stetiges lineares Funktional auf V. Ist $u_h \in V_h$ Lösung von (4.2) und $u \in V$ Lösung von (4.1) so gilt:*

$$R_h(u_h)(\phi_h) = 0 \quad \forall\phi_h \in V_h \tag{4.3}$$

$$\|\nabla(u - u_h)\|_{L^2(\Omega)} = \|R_h(u_h)\|_{-1}, \tag{4.4}$$

mit der Dualnorm

$$\|R_h(u_h)\|_{-1} := \sup_{\phi\in V} \frac{R_h(u_h)(\phi)}{\|\nabla\phi\|}.$$

Proof. (*i*) Wir zeigen zunächst, dass das Residuum ein stetiges lineares Funktional ist. Es gilt:

$$|R_h(u_h)(\phi)| = |(f,\phi) - (\nabla u_h, \nabla\phi)| \leqslant \|f\|\,\|\phi\| + \|\nabla u_h\|\,\|\nabla\phi\| \leqslant (c_p\|f\| + \|\nabla u_h\|)\|\nabla\phi\|,$$

mit der Poincare-Konstante c_p. Für $u_h \in V_h$ fest gilt also $R_h(u_h) \in V^*$.

(*ii*) Für die diskrete Lösung $u_h \in V_h$ gilt:

$$R_h(u_h)(\phi_h) = (f,\phi_h) - (\nabla u_h, \nabla\phi_h) = 0 \quad \forall \phi_h \in V_h.$$

Gleichung (4.3) folgt also unmittelbar aus der Definition des Residuums und der diskreten Lösung u_h.

(*iii*) Für die Lösungen $u \in V$ und $u_h \in V_h \subset V$ gilt:

$$R_h(u_h)(\phi) = (f,\phi) - (\nabla u_h, \nabla\phi) = (\nabla u - \nabla u_h, \phi) = (\nabla e_h, \nabla\phi). \tag{4.5}$$

Das heißt, für die Dualnorm des Residuums folgt:

$$\|R_h(u_h)\|_{-1} = \sup_{\phi\in V} \frac{(\nabla e_h, \nabla\phi)}{\|\nabla\phi\|} \leqslant \sup_{\phi\in V} \frac{\|\nabla e_h\|\,\|\nabla\phi\|}{\|\nabla\phi\|} = \|\nabla e_h\|.$$

Umgekehrt gilt mit (4.5):

$$\|\nabla e_h\|^2 = R_h(u_h)(e_h) = \frac{R_h(u_h)(e_h)}{\|\nabla e_h\|}\|\nabla e_h\| \leqslant \|\nabla e_h\|\,\|R_h(u_h)\|_{-1}.$$

Die letzten beiden Ungleichungen ergeben (4.4). □

Die Dualnorm des Residuums ist also eng mit der Energienorm des Fehlers verwandt. Ist $u_h \in V_h$ bekannt, so kann für jede gegebene Größe $\phi \in V$ auch das Residuum berechnet werden. Es ist im Allgemeinen jedoch nicht möglich die Dualnorm $\|R_h(u_h)\|_{-1}$ zu berechnen. Stattdessen leiten wir Abschätzungen für diese Dualnorm her. Zunächst definieren wir als Hilfsgrößen:

Definition 88 (Kantensprung)**.** *Sei* $u_h \in V_h$ *und* $E \in \Omega_h$ *die Kante einer Zelle* $T \in \Omega_h$*. Wir definieren den Kantensprung über die Normalableitung:*

$$[n_E \cdot \nabla u_h] := \begin{cases} n_{T_1} \cdot \nabla u_h\big|_{T_1} + n_{T_2} \cdot \nabla u_h\big|_{T_2} & E \subset \bar{T}_1 \cap \bar{T}_2, \quad T_1 \neq T_2 \\ 0 & E \subset \partial\Omega, \end{cases}$$

wobei n_{T_i} *die bzgl.* T_i *nach außen gerichteten Normalvektoren sind. Da* $n_{T_1} = -n_{T_2}$ *gilt folgt:*

$$|[n_E \cdot \nabla u_h]| = |n_E \cdot (\nabla u_h\big|_{T_1} - \nabla u_h\big|_{T_2})|,$$

bei beliebiger Wahl von $n_E = n_{T_i}$.

Lemma 89 (Residuenbasierter a posteriori Fehlerschätzer für den Energiefehler)**.** *Es sei* $u \in V$ *Lösung von* (4.1) *sowie* $u_h \in V_h$ *die lineare Finite Elemente Approximation gemäß* (4.2)*. Dann gilt für den Fehler* $e_h := u - u_h$

$$\|\nabla e_h\| \leqslant c\eta_h, \quad \eta_h := \left(\sum_{T \in \Omega_h} (\rho_T^2 + \sum_{E \in \partial T} \rho_E^2) \right)^{\frac{1}{2}},$$

mit den Zellresiduen ρ_T *und den Kantenresiduen* ρ_E*:*

$$\rho_T := h_T \|f\|_{L^2(T)}, \quad \rho_E := \frac{1}{2} h_E^{\frac{1}{2}} \|[n_E \cdot \nabla u_h]\|_{L^2(E)}.$$

Proof. Es gilt mit Satz 87 für die diskrete Lösung u_h:

$$\|\nabla e_h\|_{L^2(\Omega)} = \|R_h(u_h)\|_{-1}.$$

Nun seien $\phi \in V$ sowie $\phi_h \in V_h$ beliebig. Dann gilt mit (4.3)

$$\begin{aligned}
R_h(u_h)(\phi) &= R_h(u_h)(\phi - \phi_h) = (f, \phi - \phi_h) - (\nabla u_h, \nabla(\phi - \phi_h)) \\
&= \sum_{T \in \Omega_h} \left(\int_T f(\phi - \phi_h)\, dx - \int_T \nabla u_h \cdot \nabla(\phi - \phi_h)\, dx \right) \\
&= \sum_{T \in \Omega_h} \left(\int_T (f + \Delta u_h)(\phi - \phi_h)\, dx - \int_{\partial T} (n_T \cdot \nabla u_h)(\phi - \phi_h)\, ds \right) \\
&= \sum_{T \in \Omega_h} \left(\int_T f\,(\phi - \phi_h)\, dx \sum_{E \in \partial T} \frac{1}{2} \int_{\partial T} [n_E \cdot \nabla u_h](\phi - \phi_h)\, ds \right).
\end{aligned}$$

Es gilt $\Delta u_h\big|_T = 0$ wegen der Linearität der Lösung u_h.

An dieser Stelle schätzen wir mit Cauchy-Schwarz weiter ab:

$$|R_h(u_h)(\phi)| \leqslant \sum_{T \in \Omega_h} \left(\|f\|_{L^2(T)} \|\phi - \phi_h\|_{L^2(T)} + \sum_{E \in \partial T} \frac{1}{2} \|[n_E \cdot \nabla u_h]\|_{L^2(E)} \|\phi - \phi_h\|_{L^2(E)} \right)$$

Wir wollen aus den Termen $\phi - \phi_h$ positive Potenzen in der Gitterweite h gewinnen. Die Funktion ϕ nimmt Werte aus $V := H_0^1(\Omega)$ an, verfügt im Allgemeinen jedoch nicht über höhere Regularität. Wir dürfen daher nicht mit dem Ansatz $\phi_h := I_h \phi$, also der Wahl von ϕ_h als der Knoteninterpolation von ϕ weiter rechnen. Denn diese Knoteninterpolation ist im $H_0^1(\Omega)$ nicht definiert. Stattdessen wählen wir mit $\phi_h := C_h \phi$ die H^1-stabile Clement-

Interpolation aus Satz 84 und erhalten:

$$\begin{aligned}
|R_h(u_h)(\phi)| &\leqslant \sum_{T\in\Omega_h}\left(c_I h_T \|f\|_T \, \|\nabla\phi\|_{\tilde{P}_T} + \sum_{E\in\partial T}\frac{1}{2}c_I h_T^{\frac{1}{2}}\|[n_E\cdot\nabla u_h]\|\|\nabla\phi\|_{\tilde{P}_E}\right)\\
&\leqslant c_I\left(\sum_{T\in\Omega_h}\left(\rho_T^2+\sum_{E\in\partial T}\rho_E^2\right)\right)^{\frac{1}{2}}\left(\sum_{T\in\Omega_h}\|\nabla\phi\|^2_{\tilde{P}_T}+\sum_{E\in\Omega_h}\|\nabla\phi\|^2_{\tilde{P}_E}\right)^{\frac{1}{2}}\\
&\leqslant c_T c_I\left(\sum_{T\in\Omega_h}\left(\rho_T^2+\sum_{E\in\partial T}\rho_E^2\right)\right)^{\frac{1}{2}}\|\nabla\phi\|_\Omega.
\end{aligned}$$

Die Konstante c_T beschreibt den Überlappungsgrad der Patche $\tilde{P}_T$ sowie $\tilde{P}_E$. Aus der Formregularität des Gitters folgt, dass c_T unabhängig von h eine kleine Konstante ist. Aus Satz 87 und der Definition der Dualnorm folgt die Behauptung. □

Um den vorgestellten Energiefehlerschätzer auswerten zu können muss die rechte Seite f vorliegen. Darüber hinaus müssen die Kantensprünge berechnet werden. Die Zellresiduen $\rho_T := \|f + \Delta u_h\|_T$ (mit $\Delta u_h = 0$ auf jedem T) messen das Residuum der *klassischen Formulierung* der Poisson-Gleichung $-\Delta u = f$. Die Kantensprünge messen die *Glattheit* der diskreten Lösung. Für $u \in C^1(\Omega)$, also für stetige Differenzierbarkeit über die Elementkanten hinaus gilt $\rho_E(u) = 0$.

Als Unbekannte gehen in den Fehlerschätzer die Konstante der Clement-Interpolation c_I sowie die Konstante c_T ein. Die Konstante c_T kann für ein gegebenes Gitter berechnet werden. Sie misst lediglich den Überlappungsgrad der Patche P_E. Die Konstante der Clement-Interpolation kann nur in Spezialfällen bestimmt werden. Üblicherweise muss eine Schätzung $c_I \approx 0.1 - 1$ vorgenommen werden.

Der Fehlerschätzer eignet sich nun für eine Schätzung des Energiefehlers, es gilt:

$$\|\nabla e_h\| \leqslant c_I \eta_h(\Omega_h, u_h, f).$$

Weiter können wir einfach zellweise Fehlerindikatoren definieren

$$\eta_T := (\rho_T^2 + \sum_{E\in\partial T}\rho_E^2)^{\frac{1}{2}} = \left(h_T^2\|f_T\|_T^2 + \frac{1}{2}\sum_{E\in\partial T} h_E\|[n_E\cdot\nabla u_h]\|_E^2\right)^{\frac{1}{2}}, \tag{4.6}$$

und diese zur Verfeinerung des Gitters verwenden. Algorithmen zur Verfeinerung des Gitters werden in einem folgenden Abschnitt vorgestellt. Idee ist, solche Elemente T in kleinere Elemente aufzuteilen, welche einen großen Fehlerbeitrag η_T haben.

Wir haben bisher eine Abschätzung $\|\nabla e_h\| \leqslant c\eta_h$, also eine obere Schranke für den Fehler bewiesen. Diese Abschätzung ist wichtig, um die Genauigkeit der Lösung zu garantieren. Soll der Fehlerschätzer jedoch zur Verfeinerung des Gitters verwendet werden, so muss er

in gewissem Sinne "scharf" sein. D.h., wir benötigen ferner eine umgekehrte Abschätzung der Art:

$$c_1\eta_h \leqslant \|\nabla e_h\| \leqslant c_2\eta_h.$$

Ein Fehlerschätzer mit dieser Eigenschaft wird *effizient* genannt. Wenn diese Abschätzung nicht gilt, so ist es möglich, dass die lokale Gitterverfeinerung ineffizient verfeinert, dass also Bereiche verfeinert werden, welche keinen wesentlichen Fehlerbeitrag haben.

Wir benötigen als Hilfsatz eine spezielle Spurabschätzung:

Lemma 90. *Auf jedem Element $T \in \Omega_h$ gilt*

$$\|\partial_n v\|_{L^2(\partial T)} \leqslant c\left(h^{\frac{1}{2}}\|\Delta v\|_{L^2(T)} + h^{-\frac{1}{2}}\|\nabla v\|_{L^2(\Omega)}\right) \quad \forall v \in H^2(T).$$

Proof. (*i*) Zunächst sei $\hat{T}$ ein Referenzelement. Hier gilt mit der Spurabschätzung:

$$\|\partial_n \hat{v}\|_{\partial\hat{T}} \leqslant c\|\hat{v}\|_{H^2(\hat{T})}.$$

Mit der elliptischen Regularität folgt dann

$$\|\partial_n \hat{v}\|_{\partial\hat{T}} \leqslant c\|\hat{v}\|_{H^2(\hat{T})} \leqslant cc_s\left(\|\hat{\Delta}\hat{v}\|_{\hat{T}} + \|\hat{v}\|_{\hat{T}}\right)$$

(*ii*) Es sei $\bar{v} \in \mathbb{R}$ der Mittelwert von $\hat{v}$ auf $\hat{T}$. Dann gilt mit der Poincaré Ungleichung:

$$\begin{aligned}\|\partial_n \hat{v}\|_{\partial\hat{T}} &= \|\partial_n(\hat{v}-\bar{v})\|_{\partial\hat{T}} \\ &\leqslant c\|\hat{v}-\bar{v}\|_{H^2(\hat{T})} \leqslant c\left(\|\hat{\Delta}(\hat{v}-\bar{v})\|_{\hat{T}} + \|\hat{v}-\bar{v}\|_{\hat{T}}\right) \leqslant c\left(\|\hat{\Delta}\hat{v}\|_{\hat{T}} + c_p\|\hat{\nabla}\hat{v}\|_{\hat{T}}\right).\end{aligned}$$

(*iii*) Jetzt sei $T_T(\hat{x}) = B_T\hat{x}+b_T$ die affin lineare Referenztransformation. Es gilt mit $\det(\nabla T_T) = O(h^2)$, $\det(\nabla T_T\big|_{\partial T}) = O(h)$ sowie $\|B_T\|_\infty = O(h)$:

$$\begin{aligned}\|\partial_n v\|^2_{L^2(\partial T)} &= ch\,h^{-2}\|\hat{\partial}_n\hat{v}\|^2_{L^2(\partial\hat{T})} \leqslant ch^{-1}(\|\hat{\Delta}\hat{v}\|^2_{\hat{T}} + \|\hat{\nabla}\hat{v}\|^2_{\hat{T}}) \\ &= ch^{-1}(h^{-2}\,h^4\|\Delta v\|^2_T + h^{-2}\,h^2\|\nabla v\|_T) \\ &= c(h\|\Delta v\|^2_T + h^{-1}\|v\|^2_T).\end{aligned}$$

□

Jetzt beweisen wir:

Lemma 91 (Effizienz des Energiefehlerschätzers)**.** *Seien $u \in V$ und $u_h \in V_h$ Lösungen der Poisson-Gleichung. Auf einer Folge von formregulären Triangulierungen Ω_h ist der Energiefehlerschätzer asymptotisch exakt:*

$$\eta_h \leqslant c\|\nabla e_h\| + ch\|f\|.$$

Proof. Es ist:

$$\eta_h^2 = \sum_{T\in\Omega_h} \left(h_T^2\|f\|_T^2 + h_T\|[\partial_n u_h]\|_{\partial T}^2\right).$$

Für die Lösung $u \in H^2(\Omega)$ gilt auf jeder Kante $[\partial_n u]_E = 0$. Also mit Hilfsatz 90

$$\|[\partial_n u_h]\|_{\partial T}^2 = \|[\partial_n e_h]\|_{\partial T}^2 \leqslant 2\|[\partial_n e_h]\|_{\partial T}^2 \leqslant c(h\|\Delta e_h\|_T^2 + h^{-1}\|\nabla e_h\|_T^2)$$

Es ist $\|\Delta e_h\|_T = \|f + \Delta u_h\|_T = \|f\|_T$. Also:

$$\eta_h^2 \leqslant c\sum_{T\in\Omega_h} \left(h_T^2\|f\|_T^2 + h_T^2\|f\|_T^2 + \|\nabla e_h\|_T^2\right) = c\|\nabla e_h\|^2 + ch^2\|f\|_{L^2(\Omega)}.$$

□

Weiter kann bewiesen werden, dass bei der Verwendung von linearen Finiten Elementen die Sprungterme in den Fehlerindikatoren überwiegen, dass also gilt:

Lemma 92 (Dominanz der Sprünge)**.** *Für den Fehler der linearen Finite Elemente Approximation gilt für rechte Seiten* $f \in H^1(\Omega)$

$$\|\nabla e_h\| \leqslant c\left(\sum_{E\in\Omega_h} h_E\|[n_E\cdot\nabla u_h]\|_{L^2(E)}^2\right)^{\frac{1}{2}} + \left(\sum_{x_i\in\Omega_h} h_i^4\|\nabla f\|_{L^2(P_i)}^2\right)^{\frac{1}{2}}.$$

Der zweite Term konvergiert mit zweiter Ordnung in Bezug auf die Zellweite h, ist also asymptotisch zu vernachlässigen. Erstaunlicherweise dreht sich die Dominanz der lokalen Fehlerbeiträge stets um. Bei quadratischen Finiten Elementen überwiegen die Zellbeiträge.

Es stellt sich im folgenden wieder die Frage nach der Schätzung des Fehlers in anderen Fehlerfunktionalen, etwa in der L^2-Norm. Dies erfordert wieder den Aubin-Nitsche-Trick:

Lemma 93 (L^2-Norm)**.** *Für den Fehler der linearen Finite Elemente Approximation gilt auf formregulären Gittern die a posteriori Abschätzung:*

$$\|u - u_h\|_{L^2(\Omega)} \leqslant c\left(\sum_{T\in\Omega_h}\left(h_T^4\|f\|_{L^2(T)}^2 + \frac{1}{2}\sum_{E\in\bar{T}} h_E^3\|[n_E\cdot\nabla u_h]\|_{L^2(E)}^2\right)\right)^{\frac{1}{2}}.$$

Proof. (*i*) Wir betrachten das duale Problem

$$z \in V: \quad (\nabla\phi, \nabla z) = (e_h, \phi)\|e_h\|^{-1}, \quad -\Delta z = e_h\,\|e_h\|^{-1},$$

mit einer dualen Lösung $z \in H^2(\Omega)$ und mit $\|z\|_{H^2(\Omega)} \leqslant c_s\|\Delta z\| = c_s$.

(*ii*) Es gilt:

$$R_h(u_h)(\phi) = (\nabla e_h, \nabla\phi) \quad \forall\phi \in V.$$

Also für $z = \phi$

$$R_h(u_h)(z) = (\nabla e_h, \nabla z) = \|e_h\|_{L^2(\Omega)}.$$

Weiter gilt mit der Galerkin-Orthogonalität und Einschub der Knoteninterpolation:

$$\|e_h\| = (\nabla e_h, \nabla(z - I_h z)) = (f, z - I_h z) - (\nabla u_h, \nabla(z - I_h z)).$$

Durch partielle Integration auf jeder Zelle $T \in \Omega_h$ entstehen wieder Kantenterme

$$\|e_h\| = \sum_{T\in\Omega_h} \left(\int_T f\,(z - I_h z)\,dx - \int_{\partial T} n_E \cdot \nabla u_h \cdot (z - I_h z)\,ds \right),$$

welche wir zu Sprüngen zusammenfassen können:

$$\|e_h\| = \sum_{T\in\Omega_h} \left(\int_T (f + \Delta u_h)\,(z - I_h z)\,dx - \sum_{E\in\partial T} \frac{1}{2} \int_{\partial T} [n_E \cdot \nabla u_h] \cdot (z - I_h z)\,ds \right)$$

Da $z \in H^2(\Omega)$ gelten die Interpolationsabschätzungen auf jeder Zelle und auf jeder Kante und zusammen mit der Stabilität der dualen Lösung ergibt sich

$$\|z - I_h z\|_T \leqslant c_I h_T^2 \|\nabla^2 z\| \leqslant c_I c_s h_T^2, \quad \|z - I_h z\|_E \leqslant c_I h_E^{\frac{3}{2}} \|\nabla^2 z\|_{P_E} \leqslant c_I c_s h_T^{\frac{3}{2}}.$$

Weiter mit Cauchy-Schwarz:

$$\begin{aligned}
\|e_h\| &\leqslant c_I \sum_{T\in\Omega_h} \left(h_T^2 \|f\|_T \|\nabla^2 z\|_T + \sum_{E\in\partial T} \frac{1}{2} h_E^{\frac{3}{2}} \|[n_E \cdot \nabla u_h]\| \|\nabla^2 z\|_{P_E} \right) \\
&\leqslant c_I c_s c_T \left(\sum_{T\in\Omega_h} \left(h_T^4 \|f\|_T^2 + \sum_{E\in\partial T} \frac{1}{2} h_E^3 \|[n_E \cdot \nabla u_h]\|_E^2 \right) \right)^{\frac{1}{2}} \|\nabla^2 z\|_{L^2(\Omega)}.
\end{aligned}$$

Die Aussage folgt unter Verwendung der Stabilitätsabschätzung für die duale Lösung. □

4.3 Der dual gewichtete Fehlerschätzer

Bei technischen Simulationen stellt sich oft die Frage nach einer guten Approximation von speziellen Funktionalwerten. Das kann in der Strukturmechanik etwa die Spannung in einem Punkt sein, in der Strömungsmechanik die Kraft, die auf ein umströmtes Objekt wirkt. Für solche Funktionale ist die Schätzung des Fehlers in globalen Normen nur von geringem Interesse. Die Aubin-Nitsche Trick erlaubt, den Fehler in beliebigen linearen Funktionalen mit Hilfe einer dualen Lösung darzustellen. Mit der Lösung $z \in V$ des dualen Problems

$$(\nabla \phi, \nabla z) = J(\phi),$$

zu einem gegebenen Fehlerfunktional gilt wie im Beweis zu Satz 93

$$J(e_h) = R_h(u_h)(z). \tag{4.7}$$

Zur Herleitung von a priori Fehlerschätzern und auch bei der a posteriori-Schätzung des L^2-Fehlers haben wir die duale Lösung z stets mit Hilfe einer Stabilitätsabschätzung gegen die entsprechende Rechte Seite (welche a priori bekannt ist) abgeschätzt. Für allgemeine Funktionale ist dieser Zugang oft nicht möglich.

A posteriori Fehlerschätzer nutzen zur Schätzung des Fehlers die numerische Approximationen $u_h \in V_h$ der Lösung u. Für allgemeine Fehlerfunktionale wollen wir nun auch eine diskrete duale Lösung $z_h \in V_h$ verwenden. Das Duale Problem ist also nicht mehr nur ein mathematisches Hilfskonstrukt, es wird numerische approximiert und die Lösung z_h geht in die Fehlerschätzung ein:

Lemma 94 (Dual gewichteter Fehlerschätzer)**.** *Sei $J \in V^*$ ein beschränktes lineares Fehlerfunktional. Für die Finite Elemente Approximation der Poisson-Gleichung gilt die Fehleridentität*

$$J(u-u_h) = \sum_{T\in\Omega_h} \Big\{ \int_T (f+\Delta u_h)\,(z-I_h z)dx - \sum_{E\in\partial T} \frac{1}{2}\int_E [n_E\cdot\nabla u_h]\cdot(z-I_h z)\,ds\Big\}, \quad (4.8)$$

mit der Lösung $z \in V$ des dualen Problems

$$(\nabla\phi,\nabla z) = J(\phi) \quad \forall \phi \in V,$$

sowie die Fehlerabschätzung

$$|J(u-u_h)| \leqslant \sum_{T\in\Omega_h} \eta_T, \quad \eta_T := \rho_T\omega_T + \rho_{\partial T}\omega_{\partial T} \quad (4.9)$$

mit den Zell- *und* Kantenresiduen ρ_T *bzw.* $\rho_{\partial T}$ *sowie den* Zell- *und* Kantengewichten ω_T *und* $\omega_{\partial T}$*:*

$$\rho_T := \|f+\Delta u_h\|_T, \quad \rho_{\partial T} := \frac{1}{2}h_T^{-\frac{1}{2}}\|[n\cdot\nabla u_h]\|_{\partial T}, \quad \omega_T := \|z-I_h z\|_T, \quad \omega_{\partial T} := h_T^{\frac{1}{2}}\|z-I_h z\|_{\partial T}.$$

Proof. Die Fehleridentität folgt mit der Galerkin-Orthogonalität sofort aus (4.7)

$$J(e_h) = (\nabla e_h, \nabla(z-I_h z)) = \sum_{T\in\Omega_h} \Big\{ \int_T (f+\Delta u_h)\,(z-I_h z)\,dx - \int_{\partial T} n_E\cdot\nabla u_h\cdot(z-I_h z)\,ds\Big\},$$

und Übergang zu den Sprungtermen. Abschätzen mit Cauchy-Schwarz liefert

$$|J(u-u_h)| \leqslant \sum_{T\in\Omega_h} \big(\|f+\Delta u_h\|_T\,\|z-I_h z\|_T + \frac{1}{2}\|[n\cdot\nabla u_h]\|_{\partial T}\,\|z-I_h z\|_{\partial T}\big)$$

die Fehlerabschätzung mit den lokalen Fehlerindikatoren. □

Diese Fehleridentität ist noch kein a posteriori Fehlerschätzer in dem Sinne, dass er ohne unbekannte Größen auswertbar ist. Die duale Lösung $z \in V$ ist im Allgemeinen nicht verfügbar. Um den Fehlerschätzer auswerten zu können muss der Interpolationsfehler $z - I_h z$ approximiert werden.

Numerische Approximation Zunächst wäre es naheliegend, die duale Lösung $z_h \in V_h$ durch einen Finite Elemente Ansatz zu diskretisieren

$$(\nabla\phi_h, \nabla z_h) = J(\phi_h) \quad \forall \phi_h \in V_h,$$

und als Approximation $z \approx z_h$ aufzufassen. Wegen $z_h \in V_h$ und Satz 87 gilt jedoch

$$R_h(u_h)(z_h) = 0,$$

und auf diese Weise lässt sich keine Fehlerapproximation erstellen. Alternativ kann das duale Problem in einem Raum höherer Ordnung $V_h^{(*)}$ berechnet werden

$$(\nabla\phi_h^*, \nabla z_h^*) = J(\phi_h^*) \quad \forall \phi_h \in V_h^*,$$

welcher echt größer ist als der diskrete Raum V_h. Dann kann $z \approx z_h^*$ approximiert werden und der Fehlerschätzer ist auswertbar. Es gilt:

Lemma 95 (Fehlerapproximation höherer Ordnung)**.** *Sei $u_h \in V_h^{(1)}$ die lineare Finite Elemente Approximation der Poisson-Gleichung. Sei $J \in V^*$ ein Fehlerfunktional und durch $z_h^* \in V_h^{(2)}$ die quadratische Finite Elemente Approximation der dualen Lösung. Im Fall $z \in H^3(\Omega)$ ist der a posteriori Fehlerschätzer*

$$\eta_h^* := \sum_{T\in\Omega_h} \left\{ \int_T f(z_h^* - I_h z_h^*)\, dx - \sum_{E\in\partial T} \frac{1}{2} \int_E [n_E \cdot \nabla u_h] \cdot (z_h^* - I_h z_h^*)\, ds \right\}.$$

auf einer Folge von größenregulären Gittern asymptotisch effizient:

$$\frac{|\eta_h^*|}{|J(e_h)|} = 1 + O(h).$$

Proof. (*i*) Wir wollen zeigen, dass der Fehlerschätzer schneller gegen den Fehler konvergiert $|J(e_h) - \eta_h^*| \to 0$ als der Fehler $J(e_h) \to 0$ selbst. Bei der linearen Finite Elemente Approximation ist die optimale Konvergenzordnung eines linearen beschränkten Funktionals $O(h^2)$.

(*ii*) Es gilt mit der Fehlerabschätzung aus Satz 94:

$$|J(e_h) - \eta_h^*| \leqslant \sum_{T\in\Omega_h} \left\{ \|f\|_T \, \|z - z_h^*\|_T + \sum_{E\in\partial T} \frac{1}{2} \|[n_E \cdot \nabla u_h]\|_E \, \|z - z_h^*\|_E \right\}$$

Auf jeder Kante E nutzen wir das lokale Spur-Lemma:

$$\|v\|_E \leqslant c\big(h^{-\frac{1}{2}} \|v\|_{P_E} + h^{\frac{1}{2}} \|\nabla v\|_{P_E}\big).$$

Dann gilt:

$$\begin{aligned}|J(e_h) - \eta_h^*| \leqslant \sum_{T\in\Omega_h} \{&\|f\|_T \, \|z - z_h\|_T + \\ &ch^{-\frac{1}{2}} \sum_{E\in\partial T} \frac{1}{2}\|[n_E \cdot \nabla u_h]\|_E \left(\|z - z_h^*\|_{P_E} + h\|\nabla(z - z_h^*)\|_{P_E}\right)\Bigg\} \\ \leqslant \|f\|_{L^2(\Omega)} &\, \|z - z_h^*\|_{L^2(\Omega)} \\ &+ c_T c \left(\sum_{E\in\partial T} h^{-1}\|[n_E \cdot \nabla u_h]\|_E^2\right)^{\frac{1}{2}} \left(\|z - z_h^*\|_{L^2(\Omega)} + h\|\nabla(z - z_h^*)\|_{L^2(\Omega)}\right).\end{aligned}$$

Für die quadratische Finite Elemente Approximation der dualen Lösung gelten die a priori Abschätzungen:

$$\|z - z_h^*\| \leqslant ch^3\|\nabla^3 z\|, \quad \|\nabla(z - z_h^*)\| \leqslant ch^2\|\nabla^3 z\|.$$

Mit diesen folgt:

$$|J(e_h) - \eta_h^*| \leqslant c_I h^3 \left\{\|f\|_{L^2(\Omega)} + c_T c h^{-\frac{1}{2}} \left(\sum_{E\in\partial T} \|[n_E \cdot \nabla u_h]\|_E^2\right)^{\frac{1}{2}}\right\} \|\nabla^3 z\|_{L^2(\Omega)}.$$

(*iii*) Es bleibt, die Beschränktheit der Sprünge nachzuweisen. Hierzu betrachten wir eine Kante $E \in \Omega_h$ zwischen $T_1, T_2 \in \Omega_h$. Mit der Schreibweise $\partial_n := n \cdot \nabla$ gilt:

$$[n_E \cdot \nabla u_h] = \partial_{n_E} u_h\big|_{T_1} - \partial_{n_E} u_h\big|_{T_2} = h\frac{\partial_{n_E} u_h\big|_{T_1} - \partial_{n_E} u_h\big|_{T_2}}{h} \approx\approx h\frac{\partial_{n_E} u\big|_{T_1} - \partial_{n_E} u\big|_{T_2}}{h} \approx h\partial_{n_E}^2 u\big|_E.$$

Der Kantensprung über die Normalableitung verhält sich also wie die zweiten Ableitungen der Lösungen. Diese Abschätzung setzt voraus, dass die Ableitungen der diskreten Lösung u_h lokal eine gute Approximation der Ableitung von u sind. Auf regulären Gittern lässt sich eine solche Abschätzung beweisen (Super-Approximation). Zusammen gilt:

$$|J(e_h) - \eta_h^*| \leqslant c_I h^3 \left\{\|f\|_{L^2(\Omega)} + c_T c h^{\frac{1}{2}} \|u\|_{H^2(\Omega)}\right\} \|\nabla^3 z\|.$$

Der Abstand zwischen Fehler und Schätzwert konvergiert mindestens eine Ordnung besser als der Fehler selbst. □

Durch die numerische Approximation des dualen Problems mit einer erhöhten Genauigkeit kann der Fehler asymptotisch effizient geschätzt werden. Dieses Vorgehen ist jedoch im Allgemeinen nicht zu rechtfertigen, bedeutet es doch, dass zum Schätzen des Fehlers ein höherer Aufwand betrieben werden muss als zur Lösung des eigentlichen Problem selbst.

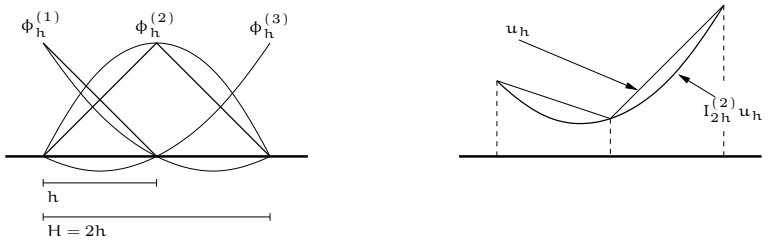


Figure 4.2: Lineare und quadratische Basisfunktionen, sowie diskrete Interpolation in den Raum höherer Ordnung.

Numerische Interpolation Eine weitere Idee zur Konstruktion von auswertbaren Fehlerschätzern ist eine nachträgliche Rekonstruktion einer dualen Lösung höherer Ordnung. Zunächst wird $z_h \in V_h$ aus dem gleichen Finite Elemente Raum der Lösung $u_h \in V_h$ berechnet. In einem zweiten Schritt wird die Lösung z_h in einen Raum von höherer Ordnung interpoliert:

$$I_h^* : V_h \to V_h^{(*)}.$$

Hier bietet sich zum Beispiel der Raum $V_{2h}^{(2)}$ an, der Raum der Finiten Elemente vom doppelten Grad auf dem doppelt so groben Gitter. Dieser Raum teil sich die gleichen Knoten wie der Raum V_h und es gilt für die Knotenbasis-Funktionen

$$\phi_h^{(i)}(x_j) = \phi_{2h}^{(2),i}(x_j) = \delta_{ij}.$$

Somit hat die Interpolierende die einfache Darstellung:

$$I_h^* u_h = \sum_{i=1}^{N} \phi_{2h}^{(2),i} \mathbf{u}_i.$$

In Abbildung 4.2 zeigen wir auf einem (eindimensionalen) Gitter einige stückweise lineare Testfunktionen $\phi_h^{(i)}$, sowie in den gleichen Gitterknoten die stückweise quadratischen Testfunktionen auf dem Gitter mit Gitterweite $H = 2h$. Rechts in der Abbildung wird die Interpolation einer diskreten Funktion in diesen Raum höherer Ordnung gezeigt.

Der Fehlerschätzer ist auswertbar als

$$\eta_h^* := \sum_{T \in \Omega_h} \left\{ \int_T f(I_h^* z_h - z_h)\,dx - \sum_{E \in \partial T} \frac{1}{2} \int_E [n_E \cdot \nabla u_h] \cdot (I_h^* z_h - z_h)\,ds \right\}. \tag{4.10}$$

Die Effizienz dieses Fehlerschätzers hängt nun an der Frage, in wie weit $I_h^* z_h$ eine bessere Approximation zu z ist als z_h selbst. Wir benötigen eine Abschätzung der Art:

$$\|z - I_h^* z\| \leqslant ch\|z - z_h\|.$$

Eine Rechtfertigung für eine nachträgliche Verbesserung der Lösung kann wieder durch das Konzept der *Superapproximation* geschehen. Bei gewisser Gitterregularität kann gezeigt

werden, dass die Finite-Elemente Lösung in den Gitterpunkten mit höherer Ordnung konvergiert. Bei linearen Finiten Elementen gilt zum Beispiel falls $f \in C^1(\Omega)$ auf gleichmäßigen Tensorproduktgittern eine Abschätzung der Art

$$|u(x_i) - u_h(x_i)| = o(h^2) \quad \text{für Gitterpunkte } x_i \in \Omega_h.$$

Diese höhere Ordnung kann dann genutzt werden, um über eine Interpolation global bessere Genauigkeit zu erreichen. In der praktischen Anwendung ist der Fehlerschätzer (4.10) höchst erfolgreich. Die Berechnung des dualen Problems ist "billig" und zum Auswerten muss lediglich die diskrete Lösung z_h mit anderen Basisfunktionen dargestellt werden.

Abschätzung der Interpolationsfehler Falls nicht die Fehleridentität (4.8), sondern nur die Abschätzung in Form (4.9) numerisch ausgewertet werden muss, so gilt es, die Residuen

$$\rho_T = \|f + \Delta u_h\|_T, \quad \rho_{\partial t} = \frac{1}{2} h^{-\frac{1}{2}} \|[n \cdot \nabla u_h]\|_{\partial T},$$

und die Gewichte

$$\omega_T = \|z - I_h z\|_T, \quad \omega_{\partial T} = h^{\frac{1}{2}} \|z - I_h z\|_T$$

zu berechnen. Die Residuen können mit der vorhandenen diskreten Lösung $u_h \in V_h$ unmittelbar ausgewertet werden. Bei den Gewichten wird zunächst die Interpolationsabschätzung im Raum $V_h^{(m-1)}$ vom Grad $m-1$ genutzt:

$$\|z - I_h z\|_T + h^{\frac{1}{2}} \|z - I_h z\|_{\partial T} \leqslant c_I h^m \|\nabla^m z\|_{P_T}.$$

Die m-ten Ableitungen von z können durch Differenzenquotienten approximiert werden:

$$\omega_T + \omega_{\partial_T} \leqslant c_I h^m \|\nabla^m z\|_T \approx c_I h^m |T|^{\frac{1}{2}} |\nabla_h^m z_h|_{T,\infty}.$$

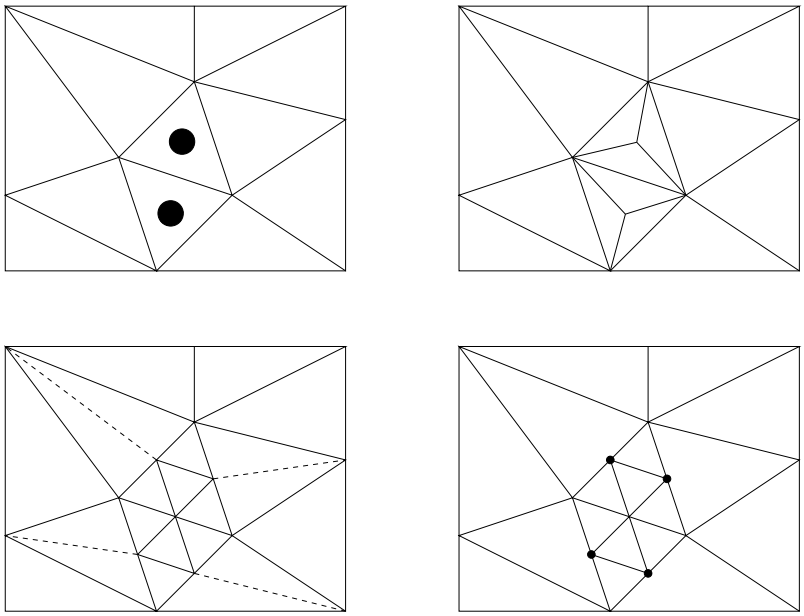


Figure 4.3: Drei verschiedene Methoden zur Verfeinerung des Gitters. Rechts oben: Verfeinerung mit neuen Inneren Knoten. Links unten: Verfeinerung mit Hilfe von *Anschlusselementen*. Rechts unten: Verfeinerung mit Hilfe von *hängenden Knoten*. Hier entstehen vier hängende Knoten.

4.4 Adaptive Gitterverfeinerung

Bei einem adaptiven Finite Elemente Verfahren werden die Fehlerindikatoren genutzt, um mit Informationen über die lokale Verteilung der Fehler die Triangulierung anzupassen und um dort die Diskretisierungsgenauigkeit zu erhöhen, wo der Fehler entsteht. Hierzu gibt es zwei alternative Optionen. Beim *Remeshing* wird mit Hilfe der Fehlerindikatoren ein komplett neues Gitter erzeugt. Hierzu wird zunächst eine *Dichtefunktion* $H(x)$ erzeugt, welche angibt, welche Gitterweite in welchem Bereich des Gebiets realisiert werden soll. Im Anschluss wird ein neues Gitter mit einem *Gittergenerator* erzeugt.

Wir betrachten hier ausschließlich die *Gitterverfeinerung*. Bei dieser Methode wird mit Hilfe der Fehlerindikatoren $\{\eta_T\}_{T\in\Omega_h}$ das Gitter Ω_h zu einem neuen Gitter Ω_h' verändert. Dabei können entweder Elemente von Ω_h zusammengefasst werden, wenn ihr Indikator-Beitrag zu dem Gesamtfehler zu vernachlässigen ist, oder aber Gitter-Elemente $T \in \Omega_h$ werden verfeinert, also in kleinere Elemente aufgeteilt.

In Abbildung 4.3 zeigen wir einige Methoden zur Verfeinerung von Dreiecks-Gittern. Bei der Gitterverfeinerung ist darauf zu achten, dass die Sequenz von Gittern Ω_h für $h \to 0$ immer noch gewissen Regularitätsbedingungen genügt:

Größenregularität Die Größenregularität, also die Forderung

$$\max_{T\in\Omega_h} h_T \leqslant c \min_{T\in\Omega_h} h_T,$$

kann sicher nicht mehr beibehalten werden. Denn das Ziel der lokalen Gitterverfeinerung ist es gerade, lokal angepasste Diskretisierungen und somit Elemente zu Verwenden. Die üblichen a priori Abschätzungen für den Fehler und die Interpolation lauten z.B.:

$$\|\nabla(u - u_h)\| \leqslant c h_{max}\|f\|,$$

mit der maximalen Gitterweite h_{max}. Abschätzungen dieser Art verlieren bei adaptiven Finiten Elementen ihre Aussagekraft, da der Gesamtfehler nicht von der maximalen Gitterweite abhängt, sondern von der optimalen Verteilung.

Formregularität Die Formregularität ist wesentlich für die Herleitung von lokalen Interpolationsabschätzungen und darf nicht durch lokale Verfeinerung verletzt werden. Bei der Aufteilung eines Dreiecks in kleinere Dreiecke muss darauf geachtet werden, dass die Innenwinkel weiterhin nicht gegen 0 oder 180 Grad gehen.

Strukturregularität In Abbildung 4.3 ist oben rechts eine Methode der Verfeinerung dargestellt, welche die Struktur-Regularität erhält. Bei dieser Verfeinerung degenerieren allerdings die Dreiecke und verletzen die Formregularität. Bei den beiden Methoden in der unteren Zeile wird die Formregularität gewahrt, die vier neuen Dreiecke haben die gleichen Winkel wie das große. Durch das Einführen von zusätzlichen Knoten auf den Eckpunkten wird jedoch die Strukturregularität verletzt.

Unten links wird die Verwendung von sogenannten *Anschlusselementen* gezeigt. Die angrenzenden Elemente werden verfeinert, diese Verfeinerung wird jedoch zurückgenommen, falls diese Elemente selbst verfeinert werden sollen. In Abbildung 4.4 werden zwei Verfeinerungsschritte bei der Verwendung von Anschlusselementen gezeigt. Im zweiten Schritt werden Anschlusselemente aufgelöst, bzw. modifiziert.

Bei der Methode unten rechts bleibt die Strukturregularität verletzt. Die neuen Knoten auf den Ecken sind sogenannte *hängende Knoten*. Diese Knoten sind nicht echte Freiheitsgrade der Triangulierung sondern werden durch den Mittelwert der beiden angrenzenden Eckknoten ersetzt.

Ziel der Gittersteuerung ist es mit Hilfe von auswertbaren Fehlerindikatoren das Gitter, also die Diskretisierung, zu verfeinern, so dass der Gesamtfehler möglichst effizient reduziert wird. Wir wählen als Beispiel die Darstellung (4.9)

$$\eta_h := \sum_{T\in\Omega_h} \eta_T, \quad \eta_T := \rho_T \omega_T,$$

welche bereits in lokalisierter Form vorliegt. Verfahren zur Gittersteuerung müssen nun solche Elemente $T \in \Omega_h$ des Gitters wählen, welche einen großen Beitrag zum Gesamtfehler haben.

Folgende Leitideen liegen jeder Gitteradaption zu Grunde:

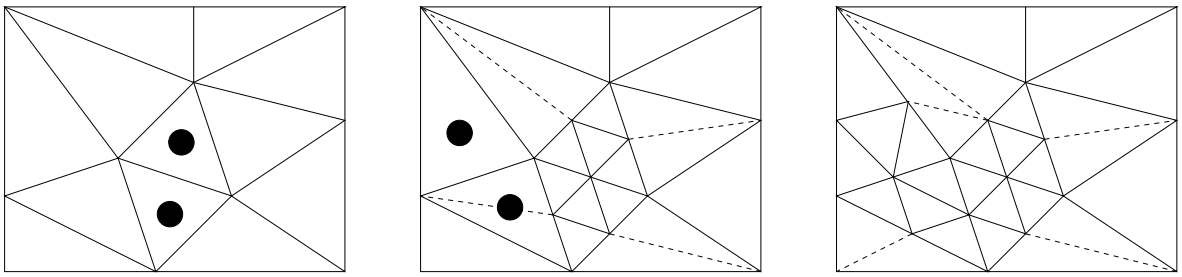


Figure 4.4: Verfeinerung bei Verwendung von Anschlusselementen.

1. Wenn ein Element $T \in \Omega_h$ verfeinert wird, so werden auch alle Elemente mit $T' \in \Omega_h$ mit einem größeren Indikatorwert $\eta_{T'} > \eta_T$ verfeinert.
2. Es wird versucht, ein Gitter mit *ausbalancierten Indikatoren* zu erreichen:

$$\eta_T \approx \eta_{T'} \quad \forall T, T', \in \Omega_h.$$

3. Wenn die Fehlerindikatoren balanciert sind, so wird global, also das ganze Gitter verfeinert.

Im Folgenden stellen wir einige Methoden zur Gitterverfeinerung vor. Dazu seien η_i für $i = 1, \ldots, N$ die Fehlerindikatoren absteigend sortiert, also mit $\eta_i \geqslant \eta_{i+1}$:

Fixed Number Es werden die $p\%$ der Elemente mit höchsten Fehlerindikatoren verfeinert.

Fixed Fraction Es werden diejenigen Elemente mit höchsten Fehlerindikatoren verfeinert, die zusammen $p\%$ des Gesamtfehlers ausmachen.

Balancierung Es werden alle Elemente Verfeinert, deren Fehlerindikator über dem Mittelwert liegt.

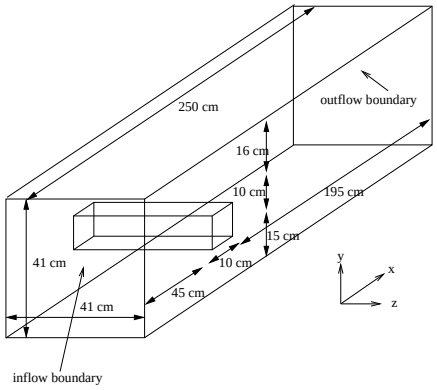


Figure 4.5: Umströmung eines Hindernis.

Ein numerisches Beispiel In einem Kanal $\Omega \subset \mathbb{R}^3$ soll die Strömung um ein Objekt mit Rand Γ_o berechnet werden. Die Konfiguration ist in Abbildung 4.5 dargestellt. Ziel der Berechnung ist es, den Strömungswiderstand des Objektes zu Berechnen. Dieser ist durch ein lineares stetiges Funktional gegeben:

$$J(\nu, p) = \int_{\Gamma_o} \nu \vec{n} \cdot \nabla \vec{v} \cdot \vec{e}_1 - np \cdot \vec{e}_1 \, ds,$$

wobei $\vec{v}$ die Geschwindigkeit der Strömung und p der Druck ist. Auf dem Rand Γ_o des Hindernis ist $\vec{n}$ der Normalvektor und $\vec{e}_1 = (1, 0, 0)$ ist der Einheitsvektor in Strömungsrichtung. Der exakte Wert des Fehlerfunktionals, also $J(u)$ ist durch Vergleichsrechnungen bekannt $J(u) \approx 7.767$.

Zunächst wird diese numerische Simulation mit stückweise quadratischen Finiten Elementen unter Verwendung von gleichmäßigen Gittern durchgeführt. Die Ergebnisse sind in Tabelle 4.1 zusammengefasst. Um eine Fehlertoleranz von 1% einzuhalten muss bereits ein Problem mit über $1\,000\,000$ Freiheitsgraden gelöst werden. D.h., wir müssen lineare Gleichungssysteme der Dimension $A \in \mathbb{R}^{1\,000\,000 \times 1\,000\,000}$ lösen!

Im Anschluss wird die gleiche Berechnung mit Hilfe von lokal verfeinerten Gittern wiederholt. Als Fehlerschätzer kommt der dual gewichtete Fehlerschätzer zum Einsatz. Das duale Problem wird im gleichen Ansatzraum approximiert wie das eigentliche Problem, also $u_h \in V_h$ und $z_h \in V_h$. Zur Approximation der Fehleridentität wird die oben beschriebene Interpolation von höherer Ordnung verwendet. In Tabelle 4.2 fassen wir die Ergebnisse zusammen.

Bei der Verwendung von adaptiven Finiten Elementen auf lokal verfeinerten Gittern wird ein relativer Fehler von 1% bereits mit $85\,000$ Freiheitsgraden erreicht. Auf global verfeinerten Gittern ist mit $1\,300\,000$ die 15 fache Anzahl von Freiheitsgraden notwendig. Wird eine Fehlertoleranz von unter 0.1% angestrebt, so ist die Ersparnis sogar ein Faktor 150. In Abbildung 4.6 zeigen wir einige Gitter und Ausschnitte von Gittern aus den Berechnungen mit lokal verfeinerten Elementen.

Gitterzellen	Freiheitsgrade	Funktionalwert	Fehler	relativer Fehler
78	3 696	13.3149	5.5479	71.4%
624	24 544	8.0450	0.2780	3.58%
4 992	177 600	7.9759	0.2089	2.69%
39 936	1 348 480	7.7878	0.0208	0.27%
319 488	10 787 840	7.7579	0.0091	0.12%
2 255 904	86 302 720	7.7612	0.0058	0.07%

Table 4.1: Umströmung eines Hindernis und Widerstandsberechnung mit quadratischen Finiten Elementen unter Verwendung von uniform verfeinerten Gittern.

Gitterzellen	Freiheitsgrade	Funktionalwert	Fehler	relativer Fehler
78	3 696	13.3149	5.5479	71.4%
624	24 544	8.0450	0.2780	3.58%
2 427	84 945	7.7942	0.0272	0.35%
7 120	242 080	7.7881	0.0211	0.27%
16 808	571 472	7.7595	0.0075	0.10%
54 880	1 811 040	7.7620	0.0050	0.06%

Table 4.2: Widerstandsberechnung mit quadratischen Finiten Elementen. Lokal verfeinerte Gitter mit dem dual gewichteten Fehlerschätzer.

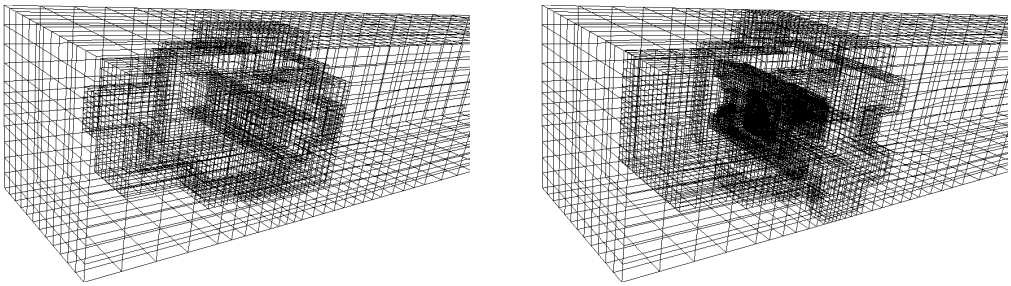


Figure 4.6: Gitterausschnitte aus der dreidimensionalen Umströmung eines Hindernis.

Bibliography

[1] H.W. Alt. *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung*. Springer Verlag, Berlin, 5. auflage edition, 2008.

[2] T. Apel. *Anisotropic finite elements: Local estimates and applications*. Advances in Numerical Mathematics. Teubner, Stuttgart, 1999.

[3] T. Apel and S. Nicaise. The finite element method with anisotropic mesh grading for elliptic problems in domains with corners and edges. *Mathematical Methods in the Applied Sciences*, 21:519–549, 1998.

[4] Y. Bazilevs, V.M. Calo, T.J.R Hughes, and Y. Zhang. Isogeometric fluid-structure interaction: theory, algorithms, and computations. *Comput Mech*, 43:3–37, 2008.

[5] M. Braack and E. Burman. Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method. *SIAM J. Numer. Anal.*, accepted, 2005.

[6] M. Braack, E. Burman, V. John, and G. Lube:. Stabilized finite element methods for the generalized oseen problem. *Comput. Methods Appl. Mech. Engrg.*, 196:853–866, 2007.

[7] D. Braess. *Finite Elemente*. Springer, 1997.

[8] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*. Applied Mathematical Sciences, 159, Springer, 2004.

[9] W. Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme*. Teubner, Stuttgart, 1991.

[10] T.J.R. Hughes, J.A. Cottrell, and Y. Bazilevs. Isogeometric analysis: Cad, finite elements, nurbs, exact geometry and mesh refinement. *Comput. Methods Appl. Mech. Engrg.*, 194(39–41):4135 – 4195, 2005.

[11] M. Lenoir. Optimal isoparametric finite elements and error estimates for domains involving curved boundaries. *SIAM Journal on Numerical Analysis*, 23(3):562–580, 1986.

[12] R. Rannacher. *Special Topics in Numerics I* (*FEM for Nonlinear Problems*). Universität Heidelberg, http://numerik.iwr.uni-heidelberg.de/~lehre/notes/, 2016. Vorlesungsskriptum.

[13] R. Rannacher and L.R. Scott. Some optimal error estimates for piecewise linear finite element approximations. *Math. Comp.*, 38:437–445, 1982.

[14] Y. Saad. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, 1996.

[15] A.H. Schatz and L.B. Wahlbin. On the quasi-optimality in L^∞ of the H^1-projection into finite element spaces. *Math. Comp.*, 38:1–22, 1982.

[16] L.R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comp.*, 54(190):483–493, 1990.