Adaptive Finite Element Methods

Lecture Notes Winter Term 2016/17

R. Verfürth

Fakultät für Mathematik, Ruhr-Universität Bochum

Contents

CHAPTER I

Introduction

I.1. Motivation

In the numerical solution of practical problems of physics or engineering such as, e.g., computational fluid dynamics, elasticity, or semiconductor device simulation one often encounters the difficulty that the overall accuracy of the numerical approximation is deteriorated by local singularities arising, e.g., from re-entrant corners, interior or boundary layers, or sharp shock-like fronts. An obvious remedy is to refine the discretization near the critical regions, i.e., to place more grid-points where the solution is less regular. The question then is how to identify those regions and how to obtain a good balance between the refined and un-refined regions such that the overall accuracy is optimal.

Another closely related problem is to obtain reliable estimates of the accuracy of the computed numerical solution. A priori error estimates, as provided, e.g., by the standard error analysis for finite element or finite difference methods, are often insufficient since they only yield information on the asymptotic error behaviour and require regularity conditions of the solution which are not satisfied in the presence of singularities as described above.

These considerations clearly show the need for an error estimator which can a posteriori be extracted from the computed numerical solution and the given data of the problem. Of course, the calculation of the a posteriori error estimate should be far less expensive than the computation of the numerical solution. Moreover, the error estimator should be local and should yield reliable upper and lower bounds for the true error in a user-specified norm. In this context one should note, that global upper bounds are sufficient to obtain a numerical solution with an accuracy below a prescribed tolerance. Local lower bounds, however, are necessary to ensure that the grid is correctly refined so that one obtains a numerical solution with a prescribed tolerance using a (nearly) minimal number of grid-points.

Disposing of an a posteriori error estimator, an adaptive mesh-refinement process has the following general structure:

ALGORITHM I.1.1. (General adaptive algorithm)

(0) *Given: The data of a partial differential equation and a tolerance* ε.

Sought: A numerical solution with an error less than ε.

(1) *Construct an initial coarse mesh $\mathcal{T}_0$ representing sufficiently well the geometry and data of the problem. Set $k = 0$.*
(2) *Solve the discrete problem on $\mathcal{T}_k$.*
(3) *For each element K in $\mathcal{T}_k$ compute an a posteriori error estimate.*
(4) *If the estimated global error is less than ε then stop. Otherwise decide which elements have to be refined and construct the next mesh $\mathcal{T}_{k+1}$. Replace k by $k+1$ and return to step (2).*

The above algorithm is best suited for stationary problems. For transient calculations, some changes have to be made:

- The accuracy of the computed numerical solution has to be estimated every few time-steps.
- The refinement process in space should be coupled with a time-step control.
- A partial coarsening of the mesh might be necessary.
- Occasionally, a complete re-meshing could be desirable.

In both stationary and transient problems, the refinement and unrefinement process may also be coupled with or replaced by a moving-point technique, which keeps the number of grid-points constant but changes there relative location.

In order to make Algorithm I.1.1 operative we must specify

- a discretization method,
- a solver for the discrete problems,
- an error estimator which furnishes the a posteriori error estimate,
- a refinement strategy which determines which elements have to be refined or coarsened and how this has to be done.

The first point is a standard one and is not the objective of these lecture notes. The second point will be addressed in Chapter IV (p. 111). The third point is the objective of Chapter II (p. 25). The last point will be addressed in Chapter III (p. 89).

In order to get a first impression of the capabilities of such an adaptive refinement strategy, we consider a simple, but typical example. We are looking for a function u which is harmonic, i.e. satisfies

$$-\Delta u = 0,$$

in the interior Ω of a circular segment centered at the origin with radius 1 and angle $\frac{3}{2}\pi$, which vanishes on the straight parts Γ_D of the boundary $\partial\Omega$, and which has normal derivative $\frac{2}{3}\sin(\frac{2}{3}\varphi)$ on the curved part Γ_N of $\partial\Omega$. Using polar co-ordinates, one easily checks that

$$u = r^{2/3}\sin\left(\frac{2}{3}\varphi\right).$$

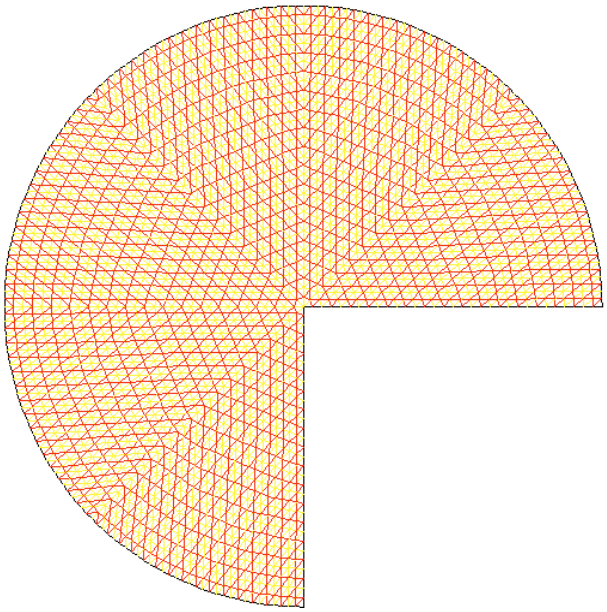

FIGURE I.1.1. Triangulation obtained by uniform refinement

We compute the Ritz projections $u_{\mathcal{T}}$ of u onto the spaces of continuous piecewise linear finite elements corresponding to the two triangulations shown in Figures I.1.1 and I.1.2, i.e., solve the problem:

Find a continuous piecewise linear function $u_{\mathcal{T}}$ such that

$$\int_\Omega \nabla u_{\mathcal{T}} \cdot \nabla v_{\mathcal{T}} = \int_{\Gamma_N} \frac{2}{3} \sin\left(\frac{2}{3}\varphi\right) v_{\mathcal{T}}$$

holds for all continuous piecewise linear functions $v_{\mathcal{T}}$.

The triangulation of Figure I.1.1 is obtained by five uniform refinements of an initial triangulation $\mathcal{T}_0$ which consists of three right-angled isosceles triangles with short sides of unit length. In each refinement step every triangle is cut into four new ones by connecting the midpoints of its edges. Moreover, the midpoint of an edge having its two endpoints on $\partial\Omega$ is projected onto $\partial\Omega$. The triangulation in Figure I.1.2 is obtained from $\mathcal{T}_0$ by applying six steps of the adaptive refinement strategy described above using the error estimator $\eta_{R,K}$ of Section II.1.9 (p. 34). A triangle $K \in \mathcal{T}_k$ is divided into four new ones if

$$\eta_{R,K} \geq 0.5 \max_{K' \in \mathcal{T}_k} \eta_{R,K'}$$

(cf. Algorithm III.1.1 (p. 89)). Midpoints of edges having their two endpoints on $\partial\Omega$ are again projected onto $\partial\Omega$. For both meshes we list in Table I.1.1 the number `NT` of triangles, the number `NN` of unknowns, and the relative error

$$\varepsilon = \frac{\|\nabla(u - u_{\mathcal{T}})\|}{\|\nabla u\|}$$

with $\|\cdot\|$ denoting the $L^2(\Omega)$-norm. It clearly shows the advantages of the adaptive refinement strategy.

TABLE I.1.1. Number of triangles NT and of unknowns NN and relative error ε for uniform and adaptive refinement

refinement	NT	NN	ε
uniform	6144	2945	0.8%
adaptive	3296	1597	0.9%

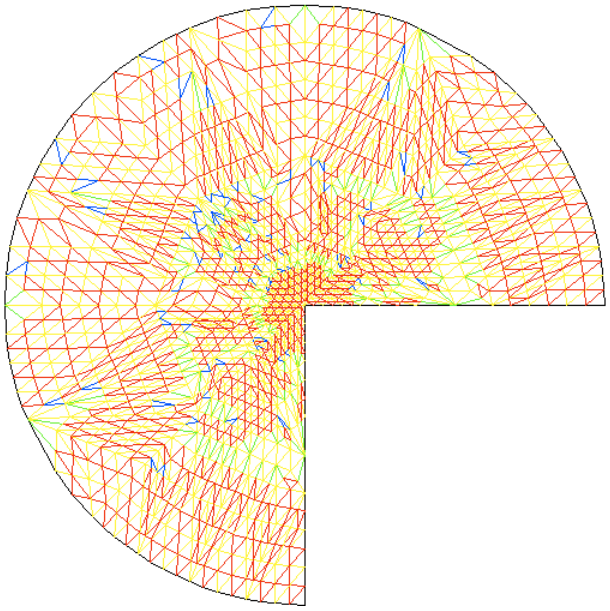

FIGURE I.1.2. Triangulation obtained by adaptive refinement

Some of the methods which are presented in these lecture notes are demonstrated in the Java applet ALF (Adaptive Linear Finite elements) and the Scilab function library AFEM (Adaptive Finite Element Methods). Both are available at the address

http://www.rub.de/num1/softwareE.html

together with short user guides in pdf-form.
ALF in particular offers the following options:

- various domains including those with curved boundaries,
- various coarsest meshes,
- various differential equations, in particular
  - the Poisson equation with smooth and singular solutions,
  - reaction-diffusion equations in particular those having solutions with interior layers,

  - convection-diffusion equations in particular those with solutions having interior and boundary layers,
- various options for building the stiffness matrix and the right-hand side,
- various solvers in particular
  - CG- and PCG-algorithms,
  - several variants of multigrid algorithms with various cycles and smoothers,
- the option to choose among uniform and and adaptive refinement based on various a posteriori error estimators.

`AFEM` has a similar functionality as `ALF`, but differs in the following details:

- It exclusively uses linear triangular elements.
- The mesh-refinement is based on the marked edge bisection of Section III.1.4 (p. 93).
- The discrete problems are solved exactly using `Scilab`'s built-in sparse matrix solver.

I.2. Sobolev and finite element spaces

I.2.1. Domains and functions. The following notations concerning domains and functions will frequently be used:

Ω open, bounded, connected set in $\mathbb{R}^d$, $d \in \{2, 3\}$;
Γ boundary of Ω, supposed to be Lipschitz-continuous;
Γ_D Dirichlet part of Ω, supposed to be non-empty;
Γ_N Neumann part of Ω, may be empty;
$\mathbf{n}$ exterior unit normal to Ω;
$p, q, r, \ldots$ scalar functions with values in $\mathbb{R}$;
$\mathbf{u}, \mathbf{v}, \mathbf{w}, \ldots$ vector-fields with values in $\mathbb{R}^d$;
$\underline{\mathbf{S}}, \underline{\mathbf{T}}, \ldots$ tensor-fields with values in $\mathbb{R}^{d\times d}$;
$\underline{\mathbf{I}}$ unit tensor;
∇ gradient;
div divergence;

$$\operatorname{div} \mathbf{u} = \sum_{i=1}^{d} \frac{\partial u_i}{\partial x_i};$$

$$\operatorname{div} \underline{\mathbf{T}} = \left(\sum_{i=1}^{d} \frac{\partial T_{ij}}{\partial x_i} \right)_{1\le j\le d};$$

$\Delta = \operatorname{div} \nabla$ Laplace operator;

$\underline{\mathbf{D}}(\mathbf{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)_{1\le i,j\le d}$ deformation tensor;

$\mathbf{u}\cdot\mathbf{v}$ inner product;

$\underline{\mathbf{S}} : \underline{\mathbf{T}}$ dyadic product (inner product of tensors).

I.2.2. Differentiation of products. The product formula for differentiation yields the following formulae for the differentiation of products of scalar functions, vector-fields and tensor-fields:

$$\begin{aligned} \operatorname{div}(p\mathbf{u}) &= \nabla p\cdot\mathbf{u} + p\operatorname{div}\mathbf{u},\\ \operatorname{div}(\underline{\mathbf{T}}\cdot\mathbf{u}) &= (\operatorname{div}\underline{\mathbf{T}})\cdot\mathbf{u} + \underline{\mathbf{T}} : \underline{\mathbf{D}}(\mathbf{u}). \end{aligned}$$

I.2.3. Integration by parts formulae. The above product formulae and the Gauss theorem for integrals give rise to the following integration by parts formulae:

$$\begin{aligned} \int_\Gamma p\mathbf{u}\cdot\mathbf{n}dS &= \int_\Omega \nabla p\cdot\mathbf{u}dx + \int_\Omega p\operatorname{div}\mathbf{u}dx,\\ \int_\Gamma \mathbf{n}\cdot\underline{\mathbf{T}}\cdot\mathbf{u}dS &= \int_\Omega (\operatorname{div}\underline{\mathbf{T}})\cdot\mathbf{u}dx + \int_\Omega \underline{\mathbf{T}} : \underline{\mathbf{D}}(\mathbf{u})dx. \end{aligned}$$

I.2.4. Weak derivatives. Recall that $\overline{A}$ denotes the closure of a set $A\subset\mathbb{R}^d$.

Example I.2.1. For the sets

$A = \{x\in\mathbb{R}^3 : x_1^2+x_2^2+x_3^2 < 1\}$ open unit ball

$B = \{x\in\mathbb{R}^3 : 0 < x_1^2+x_2^2+x_3^2 < 1\}$ punctuated open unit ball

$C = \{x\in\mathbb{R}^3 : 1 < x_1^2+x_2^2+x_3^2 < 2\}$ open annulus

we have

$\overline{A} = \{x\in\mathbb{R}^3 : x_1^2+x_2^2+x_3^2 \le 1\}$ closed unit ball

$\overline{B} = \{x\in\mathbb{R}^3 : x_1^2+x_2^2+x_3^2 \le 1\}$ closed unit ball

$\overline{C} = \{x\in\mathbb{R}^3 : 1 \le x_1^2+x_2^2+x_3^2 \le 2\}$ closed annulus.

Given a continuous function $\varphi:\mathbb{R}^d\to\mathbb{R}$, we denote its *support* by

$$\operatorname{supp}\varphi = \overline{\{x\in\mathbb{R}^d : \varphi(x)\neq 0\}}.$$

The set of all functions that are infinitely differentiable and have their support contained in Ω is denoted by $C_0^\infty(\Omega)$:

$$C_0^\infty(\Omega) = \{\varphi \in C^\infty(\Omega) : \operatorname{supp}\varphi \subset \Omega\}.$$

REMARK I.2.2. The condition "$\operatorname{supp}\varphi \subset \Omega$" is a non trivial one, since $\operatorname{supp}\varphi$ is closed and Ω is open. Functions satisfying this condition vanish at the boundary of Ω together with all their derivatives.

Given a sufficiently smooth function φ and a multi-index $\alpha \in \mathbb{N}^d$, we denote its partial derivatives by

$$D^\alpha\varphi = \frac{\partial^{\alpha_1+\ldots+\alpha_d}\varphi}{\partial x_1^{\alpha_1}\ldots\partial x_d^{\alpha_d}}.$$

Given two functions $\varphi, \psi \in C_0^\infty(\Omega)$, the Gauss theorem for integrals yields for every multi-index $\alpha \in \mathbb{N}^n$ the identity

$$\int_\Omega D^\alpha\varphi\psi dx = (-1)^{\alpha_1+\ldots+\alpha_d}\int_\Omega \varphi D^\alpha\psi.$$

This identity motivates the definition of the weak derivatives:

Given two integrable functions $\varphi, \psi \in L^1(\Omega)$ and a multi-index $\alpha \in \mathbb{N}^d$, ψ is called the α-th *weak derivative* of φ if and only if the identity

$$\int_\Omega \psi\rho dx = (-1)^{\alpha_1+\ldots+\alpha_d}\int_\Omega \varphi D^\alpha\rho$$

holds for all functions $\rho \in C_0^\infty(\Omega)$. In this case we write

$$\psi = D^\alpha\varphi.$$

REMARK I.2.3. For smooth functions, the notions of classical and weak derivatives coincide. However, there are functions which are not differentiable in the classical sense but which have a weak derivative (cf. Example I.2.4 below).

EXAMPLE I.2.4. The function $|x|$ is not differentiable in $(-1, 1)$, but it is differentiable in the weak sense. Its weak derivative is the piecewise constant function which equals -1 on $(-1, 0)$ and 1 on $(0, 1)$.

I.2.5. Sobolev spaces and norms. We will frequently use the following *Sobolev spaces* and norms:

$$H^k(\Omega) = \{\varphi \in L^2(\Omega) : D^\alpha\varphi \in L^2(\Omega) \text{ for all } \alpha \in \mathbb{N}^d \text{ with } \alpha_1 + \ldots + \alpha_d \le k\},$$

$$|\varphi|_k = \left\{ \sum_{\substack{\alpha\in\mathbb{N}^d \\ \alpha_1+\ldots+\alpha_d=k}} \|D^\alpha\varphi\|^2_{L^2(\Omega)} \right\}^{\frac{1}{2}},$$

$$\|\varphi\|_k = \left\{ \sum_{\ell=0}^{k} |\varphi|^2_\ell \right\}^{\frac{1}{2}} = \left\{ \sum_{\substack{\alpha\in\mathbb{N}^d \\ \alpha_1+\ldots+\alpha_d\le k}} \|D^\alpha\varphi\|^2_{L^2(\Omega)} \right\}^{\frac{1}{2}},$$

$$H^1_0(\Omega) = \{\varphi \in H^1(\Omega) : \varphi = 0 \ \text{ on } \ \Gamma\},$$

$$H^1_D(\Omega) = \{\varphi \in H^1(\Omega) : \varphi = 0 \ \text{ on } \ \Gamma_D\},$$

$$H^{\frac{1}{2}}(\Gamma) = \{\psi \in L^2(\Gamma) : \psi = \varphi\big|_\Gamma \ \text{ for some } \varphi \in H^1(\Omega)\},$$

$$\|\psi\|_{\frac{1}{2},\Gamma} = \inf\{\|\varphi\|_1 : \varphi \in H^1(\Omega), \varphi\big|_\Gamma = \psi\}.$$

Note that all derivatives are to be understood in the weak sense.

REMARK I.2.5. The space $H^{\frac{1}{2}}(\Gamma)$ is called *trace space* of $H^1(\Omega)$, its elements are called *traces* of functions in $H^1(\Omega)$.

REMARK I.2.6. Except in one dimension, $d = 1$, H^1 functions are in general not continuous and do not admit point values (cf. Example I.2.7 below). A function, however, which is piecewise differentiable is in $H^1(\Omega)$ if and only if it is globally continuous. This is crucial for finite element functions.

EXAMPLE I.2.7. The function $|x|$ is not differentiable, but it is in $H^1((-1,1))$. In two dimensions, the function $\ln\left(\ln\left(\sqrt{x_1^2+x_2^2}\right)\right)$ is an example of an H^1-function that is not continuous and which does not admit a point value in the origin. In three dimensions, a similar example is given by $\ln(\sqrt{x_1^2+x_2^2+x_3^2})$.

EXAMPLE I.2.8. Consider the open unit ball

$$\Omega = \{x \in \mathbb{R}^d : x_1^2 + \ldots + x_d^2 < 1\}$$

in $\mathbb{R}^d$ and the functions

$$\varphi_\alpha(x) = \{x_1^2 + \ldots + x_d^2\}^{\frac{\alpha}{2}}, \quad \alpha \in \mathbb{R}.$$

Then we have

$$\varphi_\alpha \in H^1(\Omega) \iff \begin{cases} \alpha \ge 0 & \text{if } d = 2, \\ \alpha > 1 - \frac{d}{2} & \text{if } d > 2. \end{cases}$$

I.2.6. Friedrichs and Poincaré inequalities. The following inequalities are fundamental:

$$\|\varphi\|_0 \le c_\Omega |\varphi|_1 \quad \text{for all } \varphi \in H^1_D(\Omega), \quad \textit{Friedrichs inequality}$$

$$\|\varphi\|_0 \le c'_\Omega |\varphi|_1 \quad \text{for all } \varphi \in H^1(\Omega) \text{ with } \int_\Omega \varphi = 0 \quad \textit{Poincaré inequality.}$$

The constants c_Ω and c'_Ω depend on the domain Ω and are proportional to its diameter.

I.2.7. Finite element partitions. The finite element discretizations are based on partitions of the domain Ω into non-overlapping simple subdomains. The collection of these subdomains is called a *partition* and is labeled $\mathcal{T}$. The members of $\mathcal{T}$, i.e. the subdomains, are called *elements* and are labeled K.

Any partition $\mathcal{T}$ has to satisfy the following conditions:

- $\Omega \cup \Gamma$ is the union of all elements in $\mathcal{T}$.
- *(Affine equivalence)* Each $K \in \mathcal{T}$ is either a triangle or a parallelogram, if $d = 2$, or a tetrahedron or a parallelepiped, if $d = 3$.
- *(Admissibility)* Any two elements in $\mathcal{T}$ are either disjoint or share a vertex or a complete edge or – if $d = 3$ – a complete face.
- *(Shape-regularity)* For any element K, the ratio of its diameter h_K to the diameter ρ_K of the largest ball inscribed into K is bounded independently of K.

REMARK I.2.9. In two dimensions, $d = 2$, shape regularity means that the smallest angles of all elements stay bounded away from zero. In practice one usually not only considers a single partition $\mathcal{T}$, but complete families of partitions which are often obtained by successive local or global refinements. Then, the ratio h_K/ρ_K must be bounded *uniformly* with respect to *all elements and all partitions*.

With every partition $\mathcal{T}$ we associate its *shape parameter*

$$\mathrm{C}_\mathcal{T} = \max_{K \in \mathcal{T}} \frac{h_K}{\rho_K}.$$

REMARK I.2.10. In two dimensions triangles and parallelograms may be mixed (cf. Figure I.2.1). In three dimensions tetrahedrons and parallelepipeds can be mixed provided prismatic elements are also incorporated. The condition of affine equivalence may be dropped. It, however, considerably simplifies the analysis since it implies constant Jacobians for all element transformations.

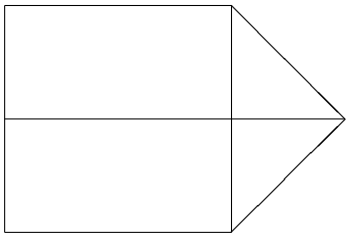

FIGURE I.2.1. Mixture of triangular and quadrilateral elements

With every partition $\mathcal{T}$ and its elements K we associate the following sets:

$\mathcal{N}_K$: the vertices of K,
$\mathcal{E}_K$: the edges or faces of K,
$\mathcal{N}$: the vertices of all elements in $\mathcal{T}$, i.e.

$$\mathcal{N} = \bigcup_{K\in\mathcal{T}} \mathcal{N}_K,$$

$\mathcal{E}$: the edges or faces of all elements in $\mathcal{T}$, i.e.

$$\mathcal{E} = \bigcup_{K\in\mathcal{T}} \mathcal{E}_K,$$

$\mathcal{N}_E$: the vertices of an edge or face $E \in \mathcal{E}$,
$\mathcal{N}_\Gamma$: the vertices on the boundary,
$\mathcal{N}_{\Gamma_D}$: the vertices on the Dirichlet boundary,
$\mathcal{N}_{\Gamma_N}$: the vertices on the Neumann boundary,
$\mathcal{N}_\Omega$: the vertices in the interior of Ω,
$\mathcal{E}_\Gamma$: the edges or faces contained in the boundary,
$\mathcal{E}_{\Gamma_D}$: the edges or faces contained in the Dirichlet boundary,
$\mathcal{E}_{\Gamma_N}$: the edges or faces contained in the Neumann boundary,
$\mathcal{E}_\Omega$: the edges or faces having at least one endpoint in the interior of Ω.

For every element, face, or edge $S \in \mathcal{T} \cup \mathcal{E}$ we denote by h_S its diameter. Note that the shape regularity of $\mathcal{T}$ implies that for all elements K and K' and all edges E and E' that share at least one

vertex the ratios $\frac{h_K}{h_{K'}}$, $\frac{h_E}{h_{E'}}$ and $\frac{h_K}{h_E}$ are bounded from below and from above by constants which only depend on the shape parameter $\mathrm{C}_{\mathcal{T}}$ of $\mathcal{T}$.

With any element K, any edge or face E, and any vertex x we associate the following sets (cf. figures I.2.2 and I.2.3)

$$\omega_K = \bigcup_{\mathcal{E}_K \cap \mathcal{E}_{K'} \neq \emptyset} K', \qquad \widetilde{\omega}_K = \bigcup_{\mathcal{N}_K \cap \mathcal{N}_{K'} \neq \emptyset} K',$$
$$\omega_E = \bigcup_{E \in \mathcal{E}_{K'}} K', \qquad \widetilde{\omega}_E = \bigcup_{\mathcal{N}_E \cap \mathcal{N}_{K'} \neq \emptyset} K',$$
$$\omega_x = \bigcup_{x \in \mathcal{N}_{K'}} K'.$$

Due to the shape-regularity of $\mathcal{T}$ the diameter of any of these sets can be bounded by a multiple of the diameter of any element or edge contained in that set. The constant only depends on the the shape parameter $\mathrm{C}_{\mathcal{T}}$ of $\mathcal{T}$.

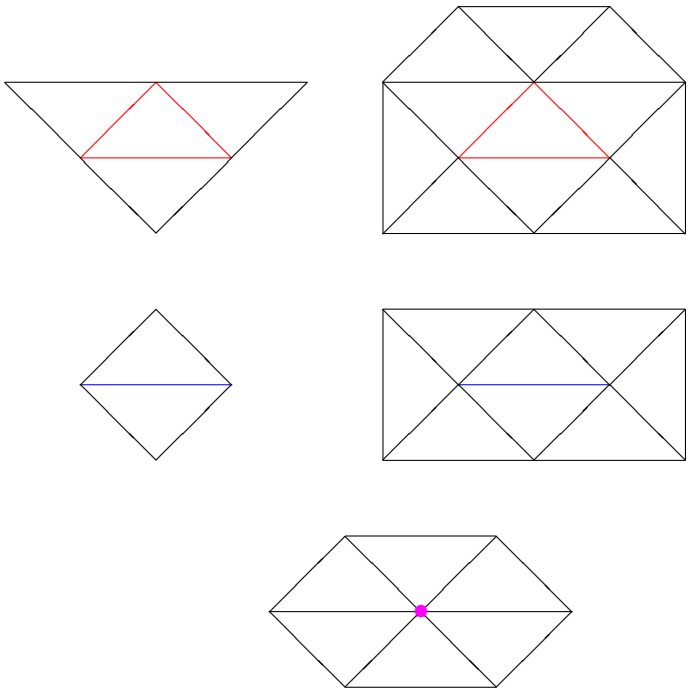

FIGURE I.2.2. Some domains ω_K, $\widetilde{\omega}_K$, ω_E, $\widetilde{\omega}_E$, and ω_x

I.2.8. Finite element spaces. For any multi-index $\alpha \in \mathbb{N}^d$ we set for abbreviation

$$|\alpha|_1 = \alpha_1 + \ldots + \alpha_d,$$

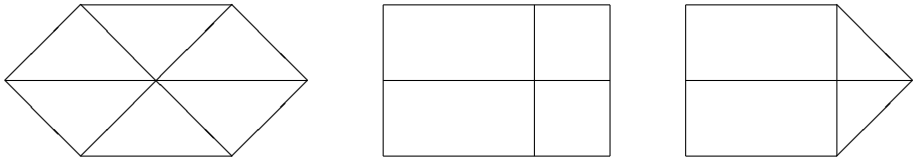

FIGURE I.2.3. Some examples of domains ω_x

$$|\alpha|_\infty = \max\{\alpha_i : 1 \le i \le d\},$$
$$x^\alpha = x_1^{\alpha_1} \cdot \ldots \cdot x_d^{\alpha_d}.$$

Denote by

$$\widehat{K} = \{\widehat{x} \in \mathbb{R}^d : x_1 + \ldots + x_d \le 1, x_i \ge 0, 1 \le i \le d\}$$

the *reference simplex* for a partition into triangles or tetrahedra and by

$$\widehat{K} = [0,1]^d$$

the *reference cube* for a partition into parallelograms or parallelepipeds. Then every element $K \in \mathcal{T}$ is the image of $\widehat{K}$ under an affine mapping F_K. For every integer number k set

$$R_k(\widehat{K}) = \begin{cases} \text{span}\{x^\alpha : |\alpha|_1 \le k\} & \text{if } K \text{ is the reference simplex,} \\ \text{span}\{x^\alpha : |\alpha|_\infty \le k\} & \text{if } K \text{ is the reference cube} \end{cases}$$

and set

$$R_k(K) = \left\{\widehat{p} \circ F_K^{-1} : \widehat{p} \in \widehat{R}_k\right\}.$$

With this notation we define finite element spaces by

$$\begin{aligned} S^{k,-1}(\mathcal{T}) &= \{\varphi : \Omega \to \mathbb{R} : \varphi\big|_K \in R_k(K) \ \text{ for all } K \in \mathcal{T}\}, \\ S^{k,0}(\mathcal{T}) &= S^{k,-1}(\mathcal{T}) \cap C(\overline{\Omega}), \\ S_0^{k,0}(\mathcal{T}) &= S^{k,0}(\mathcal{T}) \cap H_0^1(\Omega) = \{\varphi \in S^{k,0}(\mathcal{T}) : \varphi = 0 \ \text{ on } \Gamma\}. \\ S_D^{k,0}(\mathcal{T}) &= S^{k,0}(\mathcal{T}) \cap H_D^1(\Omega) = \{\varphi \in S^{k,0}(\mathcal{T}) : \varphi = 0 \ \text{ on } \Gamma_D\}. \end{aligned}$$

Note, that k may be 0 for the first space, but must be at least 1 for the other spaces.

EXAMPLE I.2.11. For the reference triangle, we have

$$R_1(\widehat{K}) = \text{span}\{1, x_1, x_2\},$$
$$R_2(\widehat{K}) = \text{span}\{1, x_1, x_2, x_1^2, x_1x_2, x_2^2\}.$$

For the reference square on the other hand, we have

$$R_1(\widehat{K}) = \text{span}\{1, x_1, x_2, x_1x_2\},$$
$$R_2(\widehat{K}) = \text{span}\{1, x_1, x_2, x_1x_2, x_1^2, x_1^2x_2, x_1^2x_2^2, x_1x_2^2, x_2^2\}.$$

I.2.9. Approximation properties. The finite element spaces defined above satisfy the following approximation properties:

$$\begin{aligned}
&\inf_{\varphi_{\mathcal{T}}\in S^{k,-1}(\mathcal{T})} \|\varphi-\varphi_{\mathcal{T}}\|_0 \le ch^{k+1}|\varphi|_{k+1} && \varphi\in H^{k+1}(\Omega),\ k\in\mathbb{N},\\
&\inf_{\varphi_{\mathcal{T}}\in S^{k,0}(\mathcal{T})} |\varphi-\varphi_{\mathcal{T}}|_j \le ch^{k+1-j}|\varphi|_{k+1} && \varphi\in H^{k+1}(\Omega),\\
& && j\in\{0,1\},\ k\in\mathbb{N}^*,\\
&\inf_{\varphi_{\mathcal{T}}\in S_0^{k,0}(\mathcal{T})} |\varphi-\varphi_{\mathcal{T}}|_j \le ch^{k+1-j}|\varphi|_{k+1} && \varphi\in H^{k+1}(\Omega)\cap H_0^1(\Omega),\\
& && j\in\{0,1\},\ k\in\mathbb{N}^*.
\end{aligned}$$

I.2.10. Nodal shape functions. Recall that $\mathcal{N}$ denotes the set of all element vertices.

For any vertex $x\in\mathcal{N}$ the associated *nodal shape function* is denoted by λ_x. It is the unique function in $S^{1,0}(\mathcal{T})$ that equals 1 at vertex x and that vanishes at all other vertices $y\in\mathcal{N}\backslash\{x\}$.

The support of a nodal shape function λ_x is the set ω_x and consists of all elements that share the vertex x (cf. Figure I.2.3).

The nodal shape functions can easily be computed element-wise from the co-ordinates of the element's vertices.

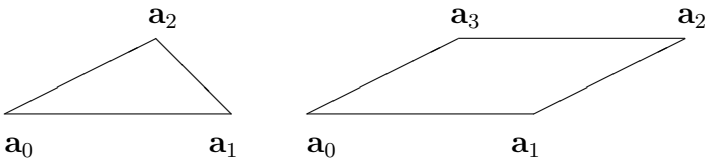

FIGURE I.2.4. Enumeration of vertices of triangles and parallelograms

EXAMPLE I.2.12. (1) Consider a triangle K with vertices $\mathbf{a}_0,\dots,\mathbf{a}_2$ numbered counterclockwise (cf. Figure I.2.4). Then the restrictions to K of the nodal shape functions $\lambda_{\mathbf{a}_0},\dots,\lambda_{\mathbf{a}_2}$ are given by

$$\lambda_{\mathbf{a}_i}(x)=\frac{\det(x-\mathbf{a}_{i+1}\,,\,\mathbf{a}_{i+2}-\mathbf{a}_{i+1})}{\det(\mathbf{a}_i-\mathbf{a}_{i+1}\,,\,\mathbf{a}_{i+2}-\mathbf{a}_{i+1})}\quad i=0,\dots,2,$$

where all indices have to be taken modulo 3.
(2) Consider a parallelogram K with vertices $\mathbf{a}_0,\dots,\mathbf{a}_3$ numbered counterclockwise (cf. Figure I.2.4). Then the restrictions to K of the nodal shape functions $\lambda_{\mathbf{a}_0},\dots,\lambda_{\mathbf{a}_3}$ are given by

$$\lambda_{\mathbf{a}_i}(x)=\frac{\det(x-\mathbf{a}_{i+2}\,,\,\mathbf{a}_{i+3}-\mathbf{a}_{i+2})}{\det(\mathbf{a}_i-\mathbf{a}_{i+2}\,,\,\mathbf{a}_{i+3}-\mathbf{a}_{i+2})}\cdot\frac{\det(x-\mathbf{a}_{i+2}\,,\,\mathbf{a}_{i+1}-\mathbf{a}_{i+2})}{\det(\mathbf{a}_i-\mathbf{a}_{i+2}\,,\,\mathbf{a}_{i+1}-\mathbf{a}_{i+2})}$$

$$i = 0, \dots, 3,$$

where all indices have to be taken modulo 4.
(3) Consider a tetrahedron K with vertices $\mathbf{a}_0, \dots, \mathbf{a}_3$ enumerated as in Figure I.2.5. Then the restrictions to K of the nodal shape functions $\lambda_{\mathbf{a}_0}, \dots, \lambda_{\mathbf{a}_3}$ are given by

$$\lambda_{\mathbf{a}_i}(x) = \frac{\det(x - \mathbf{a}_{i+1}\,,\, \mathbf{a}_{i+2} - \mathbf{a}_{i+1}\,,\, \mathbf{a}_{i+3} - \mathbf{a}_{i+1})}{\det(\mathbf{a}_i - \mathbf{a}_{i+1}\,,\, \mathbf{a}_{i+2} - \mathbf{a}_{i+1}\,,\, \mathbf{a}_{i+3} - \mathbf{a}_{i+1})} \quad i = 0, \dots, 3,$$

where all indices have to be taken modulo 4.
(4) Consider a parallelepiped K with vertices $\mathbf{a}_0, \dots, \mathbf{a}_7$ enumerated as in Figure I.2.5. Then the restrictions to K of the nodal shape functions $\lambda_{\mathbf{a}_0}, \dots, \lambda_{\mathbf{a}_7}$ are given by

$$\begin{aligned}\lambda_{\mathbf{a}_i}(x) = {} & \frac{\det(x - \mathbf{a}_{i+1}\,,\, \mathbf{a}_{i+3} - \mathbf{a}_{i+1}\,,\, \mathbf{a}_{i+5} - \mathbf{a}_{i+1})}{\det(\mathbf{a}_i - \mathbf{a}_{i+1}\,,\, \mathbf{a}_{i+3} - \mathbf{a}_{i+1}\,,\, \mathbf{a}_{i+5} - \mathbf{a}_{i+1})} \cdot \\ & \frac{\det(x - \mathbf{a}_{i+2}\,,\, \mathbf{a}_{i+3} - \mathbf{a}_{i+2}\,,\, \mathbf{a}_{i+6} - \mathbf{a}_{i+2})}{\det(\mathbf{a}_i - \mathbf{a}_{i+2}\,,\, \mathbf{a}_{i+3} - \mathbf{a}_{i+2}\,,\, \mathbf{a}_{i+6} - \mathbf{a}_{i+2})} \cdot \\ & \frac{\det(x - \mathbf{a}_{i+4}\,,\, \mathbf{a}_{i+5} - \mathbf{a}_{i+4}\,,\, \mathbf{a}_{i+6} - \mathbf{a}_{i+4})}{\det(\mathbf{a}_i - \mathbf{a}_{i+4}\,,\, \mathbf{a}_{i+5} - \mathbf{a}_{i+4}\,,\, \mathbf{a}_{i+6} - \mathbf{a}_{i+4})} \\ & i = 0, \dots, 7,\end{aligned}$$

where all indices have to be taken modulo 8.

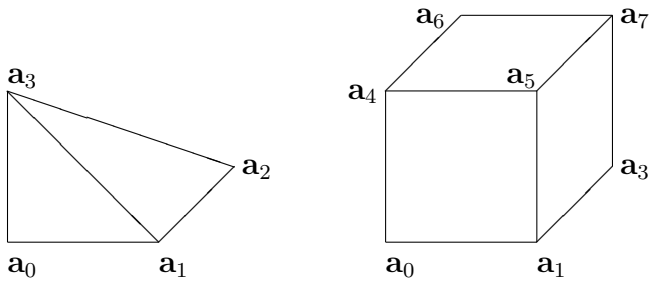

FIGURE I.2.5. Enumeration of vertices of tetrahedra and parallelepipeds (The vertex $\mathbf{a}_2$ of the parallelepiped is hidden.)

REMARK I.2.13. For every element (triangle, parallelogram, tetrahedron, or parallelepiped) the sum of all nodal shape functions corresponding to the element's vertices is identical equal to 1 on the element.

The functions λ_x, $x \in \mathcal{N}$, form a bases of $S^{1,0}(\mathcal{T})$. The bases of higher-order spaces $S^{k,0}(\mathcal{T})$, $k \geq 2$, consist of suitable products of functions λ_x corresponding to appropriate vertices x.

EXAMPLE I.2.14. (1) Consider a again a triangle K with its vertices numbered as in Example I.2.12 (1). Then the nodal basis of $S^{2,0}(\mathcal{T})\big|_K$ consists of the functions

$$\begin{aligned} \lambda_{\mathbf{a}_i}[\lambda_{\mathbf{a}_i} - \lambda_{\mathbf{a}_{i+1}} - \lambda_{\mathbf{a}_{i+2}}] \quad & i = 0, \dots, 2 \\ 4\lambda_{\mathbf{a}_i}\lambda_{\mathbf{a}_{i+1}} \quad & i = 0, \dots, 2, \end{aligned}$$

where the functions $\lambda_{\mathbf{a}_\ell}$ are as in Example I.2.12 (1) and where all indices have to be taken modulo 3. An other basis of $S^{2,0}(\mathcal{T})\big|_K$, called *hierarchical basis*, consists of the functions

$$\begin{aligned} \lambda_{\mathbf{a}_i} \quad & i = 0, \dots, 2 \\ 4\lambda_{\mathbf{a}_i}\lambda_{\mathbf{a}_{i+1}} \quad & i = 0, \dots, 2. \end{aligned}$$

(2) Consider a again a parallelogram K with its vertices numbered as in Example I.2.12 (2). Then the nodal basis of $S^{2,0}(\mathcal{T})\big|_K$ consists of the functions

$$\begin{aligned} \lambda_{\mathbf{a}_i}[\lambda_{\mathbf{a}_i} - \lambda_{\mathbf{a}_{i+1}} + \lambda_{\mathbf{a}_{i+2}} - \lambda_{\mathbf{a}_{i+3}}] \quad & i = 0, \dots, 3 \\ 4\lambda_{\mathbf{a}_i}[\lambda_{\mathbf{a}_{i+1}} - \lambda_{\mathbf{a}_{i+2}}] \quad & i = 0, \dots, 3 \\ 16\lambda_{\mathbf{a}_0}\lambda_{\mathbf{a}_2} & \end{aligned}$$

where the functions $\lambda_{\mathbf{a}_\ell}$ are as in Example I.2.12 (2) and where all indices have to be taken modulo 3. The *hierarchical basis* of $S^{2,0}(\mathcal{T})\big|_K$ consists of the functions

$$\begin{aligned} \lambda_{\mathbf{a}_i} \quad & i = 0, \dots, 3 \\ 4\lambda_{\mathbf{a}_i}[\lambda_{\mathbf{a}_{i+1}} - \lambda_{\mathbf{a}_{i+2}}] \quad & i = 0, \dots, 3 \\ 16\lambda_{\mathbf{a}_0}\lambda_{\mathbf{a}_2} \quad & . \end{aligned}$$

(3) Consider a again a tetrahedron K with its vertices numbered as in Example I.2.12 (3). Then the nodal basis of $S^{2,0}(\mathcal{T})\big|_K$ consists of the functions

$$\begin{aligned} \lambda_{\mathbf{a}_i}[\lambda_{\mathbf{a}_i} - \lambda_{\mathbf{a}_{i+1}} - \lambda_{\mathbf{a}_{i+2}} - \lambda_{\mathbf{a}_{i+3}}] \quad & i = 0, \dots, 3 \\ 4\lambda_{\mathbf{a}_i}\lambda_{\mathbf{a}_j} \quad & 0 \le i < j \le 3, \end{aligned}$$

where the functions $\lambda_{\mathbf{a}_\ell}$ are as in Example I.2.12 (3) and where all indices have to be taken modulo 4. The hierarchical basis consists of the functions

$$\begin{aligned} \lambda_{\mathbf{a}_i} \quad & i = 0, \dots, 3 \\ 4\lambda_{\mathbf{a}_i}\lambda_{\mathbf{a}_j} \quad & 0 \le i < j \le 3. \end{aligned}$$

I.2.11. A quasi-interpolation operator. We will frequently use the *quasi-interpolation operator* $I_{\mathcal{T}} : L^1(\Omega) \to S^{1,0}_D(\mathcal{T})$ which is defined by

$$I_{\mathcal{T}}\varphi = \sum_{x\in\mathcal{N}_\Omega\cup\mathcal{N}_{\Gamma_N}} \lambda_x \frac{1}{|\omega_x|}\int_{\omega_x}\varphi dx.$$

Here, $|\omega_x|$ denotes the area, if $d=2$, respectively volume, if $d=3$, of the set ω_x.
The operator $I_{\mathcal{T}}$ satisfies the following local error estimates for all $\varphi \in H^1_D(\Omega)$ and all elements $K\in\mathcal{T}$:

$$\begin{aligned}\|\varphi - I_{\mathcal{T}}\varphi\|_{L^2(K)} &\le c_{A1}h_K\|\varphi\|_{H^1(\widetilde{\omega}_K)},\\ \|\varphi - I_{\mathcal{T}}\varphi\|_{L^2(\partial K)} &\le c_{A2}h_K^{\frac{1}{2}}\|\varphi\|_{H^1(\widetilde{\omega}_K)}.\end{aligned}$$

Here, $\widetilde{\omega}_K$ denotes the set of all elements that share at least a vertex with K (cf. Figure I.2.6). The constants c_{A1} and c_{A2} only depend on the shape parameter $\mathrm{C}_{\mathcal{T}}$ of $\mathcal{T}$.

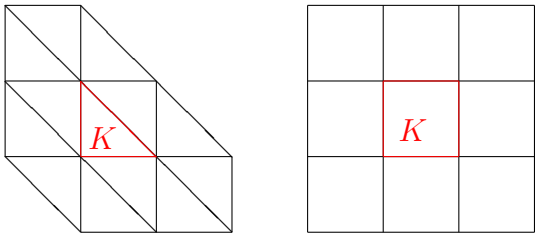

FIGURE I.2.6. Examples of domains $\widetilde{\omega}_K$

REMARK I.2.15. The operator $I_{\mathcal{T}}$ is called a quasi-interpolation operator since it *does not interpolate* a given function φ at the vertices $x\in\mathcal{N}$. In fact, point values *are not defined* for H^1-functions. For functions with more regularity which are at least in $H^2(\Omega)$, the situation is different. For those functions point values do exist and the classical nodal interpolation operator $J_{\mathcal{T}} : H^2(\Omega)\cap H^1_D(\Omega)\to S^{1,0}_D(\mathcal{T})$ can be defined by the relation $(J_{\mathcal{T}}(\varphi))(x)=\varphi(x)$ for all vertices $x\in\mathcal{N}$.

I.2.12. Bubble functions. For any element $K\in\mathcal{T}$ we define an *element bubble function* by

$$\psi_K = \alpha_K \prod_{x\in\mathcal{N}_K}\lambda_x\,,$$

$$\alpha_K = \begin{cases} 27 & \text{if } K \text{ is a triangle,}\\ 256 & \text{if } K \text{ is a tetrahedron,}\\ 16 & \text{if } K \text{ is a parallelogram,}\\ 64 & \text{if } K \text{ is a parallelepiped.}\end{cases}$$

It has the following properties:

$$\begin{aligned} 0 \le \psi_K(x) \le 1 \quad & \text{for all } x \in K, \\ \psi_K(x) = 0 \quad & \text{for all } x \notin K, \\ \max_{x\in K} \psi_K(x) = 1. & \end{aligned}$$

For every polynomial degree k there are constants $c_{I1,k}$ and $c_{I2,k}$, which only depend on the degree k and the shape parameter $\mathrm{C}_{\mathcal{T}}$ of $\mathcal{T}$, such that the following inverse estimates hold for all polynomials φ of degree k:

$$\begin{aligned} c_{I1,k}\|\varphi\|_K &\le \|\psi_K^{\frac{1}{2}}\varphi\|_K, \\ \|\nabla(\psi_K\varphi)\|_K &\le c_{I2,k}h_K^{-1}\|\varphi\|_K. \end{aligned}$$

Recall that we denote by $\mathcal{E}$ the set of all edges, if $d = 2$, and of all faces, if $d = 3$, of all elements in $\mathcal{T}$ and by $\mathcal{N}_E$ the vertices of any $E \in \mathcal{E}$. With each edge respectively face $E \in \mathcal{E}$ we associate an *edge* respectively *face bubble function* by

$$\psi_E = \beta_E \prod_{x\in\mathcal{N}_E} \lambda_x \,,$$

$$\beta_E = \begin{cases} 4 & \text{if } E \text{ is a line segment,} \\ 27 & \text{if } E \text{ is a triangle,} \\ 16 & \text{if } E \text{ is a parallelogram.} \end{cases}$$

It has the following properties:

$$\begin{aligned} 0 \le \psi_E(x) \le 1 \quad & \text{for all } x \in \omega_E, \\ \psi_E(x) = 0 \quad & \text{for all } x \notin \omega_E, \\ \max_{x\in\omega_E} \psi_E(x) = 1. & \end{aligned}$$

For every polynomial degree k there are constants $c_{I3,k}$, $c_{I4,k}$, and $c_{I5,k}$, which only depend on the degree k and the shape parameter $\mathrm{C}_{\mathcal{T}}$ of $\mathcal{T}$, such that the following inverse estimates hold for all polynomials φ of degree k:

$$\begin{aligned} c_{I3,k}\|\varphi\|_E &\le \|\psi_E^{\frac{1}{2}}\varphi\|_E, \\ \|\nabla(\psi_E\varphi)\|_{\omega_E} &\le c_{I4,k}h_E^{-\frac{1}{2}}\|\varphi\|_E, \\ \|\psi_E\varphi\|_{\omega_E} &\le c_{I5,k}h_E^{\frac{1}{2}}\|\varphi\|_E. \end{aligned}$$

Here ω_E is the union of all elements that share E (cf. Figure I.2.7). Note that ω_E consists of two elements, if E is not contained in the boundary Γ, and of exactly one element, if E is a subset of Γ.

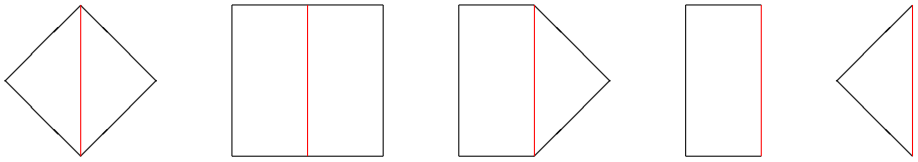

FIGURE I.2.7. Examples of domains ω_E

With each edge respectively face $E \in \mathcal{E}$ we finally associate a unit vector $\mathbf{n}_E$ orthogonal to E and denote by $\mathbb{J}_E(\cdot)$ the jump across E in direction $\mathbf{n}_E$, i.e.

$$\mathbb{J}_E(\varphi)(x) = \lim_{t \to 0+} \varphi(x + t\mathbf{n}_E) - \lim_{t \to 0+} \varphi(x - t\mathbf{n}_E).$$

If E is contained in the boundary Γ the orientation of $\mathbf{n}_E$ is fixed to be the one of the exterior normal. Otherwise it is not fixed.

REMARK I.2.16. $\mathbb{J}_E(\cdot)$ depends on the orientation of $\mathbf{n}_E$ but quantities of the form $\mathbb{J}_E(\mathbf{n}_E \cdot \varphi)$ are independent of this orientation.

CHAPTER II

A posteriori error estimates

In this chapter we will describe various possibilities for a posteriori error estimation. In order to keep the presentation as simple as possible we will consider in Sections II.1 and II.2 a simple model problem: the two-dimensional Poisson equation (cf. Equation (II.1.1) (p. 26)) discretized by continuous linear or bilinear finite elements (cf. Equation (II.1.3) (p. 26)). We will review several a posteriori error estimators and show that – in a certain sense – they are all equivalent and yield lower and upper bounds on the error of the finite element discretization. The estimators can roughly be classified as follows:

- *Residual estimates*: Estimate the error of the computed numerical solution by a suitable norm of its residual with respect to the strong form of the differential equation (Section II.1.9 (p. 34)).
- *Solution of auxiliary local problems*: On small patches of elements, solve auxiliary discrete problems similar to, but simpler than the original problem and use appropriate norms of the local solutions for error estimation (Section II.2.1 (p. 36)).
- *Hierarchical basis error estimates*: Evaluate the residual of the computed finite element solution with respect to another finite element space corresponding to higher order elements or to a refined grid (Section II.2.2 (p. 42)).
- *Averaging methods*: Use some local extrapolate or average of the gradient of the computed numerical solution for error estimation (Section II.2.3 (p. 47)).
- $H(\mathrm{div})$*-lifting*: Sweeping through the elements sharing a given vertex construct a vector field such that its divergence equals the residual (Section II.2.4 (p. 49)).

In Section II.2.5 (p. 52), we shortly address the question of asymptotic exactness, i.e., whether the ratio of the estimated and the exact error remains bounded or even approaches 1 when the mesh-size converges to 0. In Section II.2.6 (p. 54) we finally show that an adaptive method based on a suitable error estimator and a suitable mesh-refinement strategy converges to the true solution of the differential equation.

II.1. A residual error estimator for the model problem

II.1.1. The model problem. As a model problem we consider the Poisson equation with mixed Dirichlet-Neumann boundary conditions

$$\text{(II.1.1)}\qquad \boxed{\begin{aligned} -\Delta u &= f && \text{in } \Omega \\ u &= 0 && \text{on } \Gamma_D \\ \frac{\partial u}{\partial n} &= g && \text{on } \Gamma_N \end{aligned}}$$

in a connected, bounded, polygonal domain $\Omega \subset \mathbb{R}^2$ with boundary Γ consisting of two disjoint parts Γ_D and Γ_N. We assume that the Dirichlet boundary Γ_D is closed relative to Γ and has a positive length and that f and g are square integrable functions on Ω and Γ_N, respectively. The Neumann boundary Γ_N may be empty.

II.1.2. Variational formulation. The standard weak formulation of problem (II.1.1) is:

$$\text{(II.1.2)}\qquad \boxed{\begin{aligned} &\text{Find } u \in H^1_D(\Omega) \text{ such that} \\ &\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v + \int_{\Gamma_N} g v \\ &\text{for all } v \in H^1_D(\Omega). \end{aligned}}$$

It is well-known that problem (II.1.2) admits a unique solution.

II.1.3. Finite element discretization. We choose an affine equivalent, admissible and shape-regular partition $\mathcal{T}$ of Ω as in Section I.2.7 (p. 15) and consider the following finite element discretization of problem (II.1.2):

$$\text{(II.1.3)}\qquad \boxed{\begin{aligned} &\text{Find } u_{\mathcal{T}} \in S^{1,0}_D(\mathcal{T}) \text{ such that} \\ &\int_\Omega \nabla u_{\mathcal{T}} \cdot \nabla v_{\mathcal{T}} = \int_\Omega f v_{\mathcal{T}} + \int_{\Gamma_N} g v_{\mathcal{T}} \\ &\text{for all } v_{\mathcal{T}} \in S^{1,0}_D(\mathcal{T}). \end{aligned}}$$

Again it is well-known that problem (II.1.3) admits a unique solution.

II.1.4. Equivalence of error and residual. In what follows we always denote by $u \in H^1_D(\Omega)$ and $u_{\mathcal{T}} \in S^{1,0}_D(\mathcal{T})$ the exact solutions of problems (II.1.2) and (II.1.3), respectively. They satisfy the identity

$$\int_\Omega \nabla(u-u_{\mathcal{T}})\cdot\nabla v = \int_\Omega fv + \int_{\Gamma_N} gv - \int_\Omega \nabla u_{\mathcal{T}}\cdot\nabla v$$

for all $v \in H^1_D(\Omega)$. The right-hand side of this equation implicitly defines the *residual* of $u_{\mathcal{T}}$ as an element of the dual space of $H^1_D(\Omega)$.

The Friedrichs and Cauchy-Schwarz inequalities imply for all $v \in H^1_D(\Omega)$

$$\frac{1}{\sqrt{1+c_\Omega^2}}\|v\|_{H^1(\Omega)} \le \sup_{\substack{w\in H^1_D(\Omega)\\ \|w\|_{H^1(\Omega)}=1}} \int_\Omega \nabla v\cdot\nabla w \le \|v\|_{H^1(\Omega)}.$$

This corresponds to the fact that the bilinear form

$$H^1_D(\Omega)\ni v,w \mapsto \int_\Omega \nabla v\cdot\nabla w$$

defines an isomorphism of $H^1_D(\Omega)$ onto its dual space. The constants multiplying the first and last term in this inequality are related to the norm of this isomorphism and of its inverse.

The definition of the residual and the above inequality imply the estimate

$$\begin{aligned}
&\sup_{\substack{w\in H^1_D(\Omega)\\ \|w\|_{H^1(\Omega)}=1}} \left\{\int_\Omega fw + \int_{\Gamma_N} gw - \int_\Omega \nabla u_{\mathcal{T}}\cdot\nabla w\right\}\\
&\le \|u-u_{\mathcal{T}}\|_{H^1(\Omega)}\\
&\le \sqrt{1+c_\Omega^2}\sup_{\substack{w\in H^1_D(\Omega)\\ \|w\|_{H^1(\Omega)}=1}} \left\{\int_\Omega fw + \int_{\Gamma_N} gw - \int_\Omega \nabla u_{\mathcal{T}}\cdot\nabla w\right\}.
\end{aligned}$$

Since the sup-term in this inequality is equivalent to the norm of the residual in the dual space of $H^1_D(\Omega)$, we have proved:

> The norm in $H^1_D(\Omega)$ of the error is, up to multiplicative constants, bounded from above and from below by the norm of the residual in the dual space of $H^1_D(\Omega)$.

Most a posteriori error estimators try to estimate this dual norm of the residual by quantities that can more easily be computed from f, g, and $u_{\mathcal{T}}$.

II.1.5. Galerkin orthogonality. Since $S^{1,0}_D(\mathcal{T}) \subset H^1_D(\Omega)$, the error is orthogonal to $S^{1,0}_D(\mathcal{T})$:

$$\int_\Omega \nabla(u - u_\mathcal{T}) \cdot \nabla w_\mathcal{T} = 0$$

for all $w_\mathcal{T} \in S^{1,0}_D(\mathcal{T})$. Using the definition of the residual, this can be written as

$$\int_\Omega f w_\mathcal{T} + \int_{\Gamma_N} g w_\mathcal{T} - \int_\Omega \nabla u_\mathcal{T} \cdot \nabla w_\mathcal{T} = 0$$

for all $w_\mathcal{T} \in S^{1,0}_D(\mathcal{T})$. This identity reflects the fact that the discretization (II.1.3) is consistent and that no additional errors are introduced by numerical integration or by inexact solution of the discrete problem. It is often referred to as *Galerkin orthogonality*.

II.1.6. L^2-representation of the residual. Integration by parts element-wise yields for all $w \in H^1_D(\Omega)$

$$\begin{aligned}
&\int_\Omega fw + \int_{\Gamma_N} gw - \int_\Omega \nabla u_\mathcal{T} \cdot \nabla w \\
&\quad = \int_\Omega fw + \int_{\Gamma_N} gw - \sum_{K\in\mathcal{T}} \int_K \nabla u_\mathcal{T} \cdot \nabla w \\
&\quad = \int_\Omega fw + \int_{\Gamma_N} gw + \sum_{K\in\mathcal{T}} \left\{ \int_K \Delta u_\mathcal{T} w - \int_{\partial K} \mathbf{n}_K \cdot \nabla u_\mathcal{T} w \right\} \\
&\quad = \sum_{K\in\mathcal{T}} \int_K (f + \Delta u_\mathcal{T}) w + \sum_{E\in\mathcal{E}_{\Gamma_N}} \int_E (g - \mathbf{n}_E \cdot \nabla u_\mathcal{T}) w \\
&\qquad - \sum_{E\in\mathcal{E}_\Omega} \int_E \mathbb{J}_E(\mathbf{n}_E \cdot \nabla u_\mathcal{T}) w.
\end{aligned}$$

Here, $\mathbf{n}_K$ denotes the unit exterior normal to the element K. Note that $\Delta u_\mathcal{T}$ vanishes on all triangles.

For abbreviation, we define *element* and *edge residuals* by

$$R_K(u_\mathcal{T}) = f + \Delta u_\mathcal{T}$$

and

$$R_E(u_\mathcal{T}) = \begin{cases} -\mathbb{J}_E(\mathbf{n}_E \cdot \nabla u_\mathcal{T}) & \text{if } E \in \mathcal{E}_\Omega, \\ g - \mathbf{n}_E \cdot \nabla u_\mathcal{T} & \text{if } E \in \mathcal{E}_{\Gamma_N}, \\ 0 & \text{if } E \in \mathcal{E}_{\Gamma_D}. \end{cases}$$

Then we obtain the following *L^2-representation of the residual*

$$\int_\Omega fw + \int_{\Gamma_N} gw - \int_\Omega \nabla u_\mathcal{T} \cdot \nabla w = \sum_{K\in\mathcal{T}} \int_K R_K(u_\mathcal{T})w + \sum_{E\in\mathcal{E}} \int_E R_E(u_\mathcal{T})w.$$

Together with the Galerkin orthogonality this implies

$$\int_\Omega fw + \int_{\Gamma_N} gw - \int_\Omega \nabla u_\mathcal{T} \cdot \nabla w = \sum_{K\in\mathcal{T}} \int_K R_K(u_\mathcal{T})(w - w_\mathcal{T}) + \sum_{E\in\mathcal{E}} \int_E R_E(u_\mathcal{T})(w - w_\mathcal{T})$$

for all $w \in H^1_D(\Omega)$ and all $w_\mathcal{T} \in S^{1,0}_D(\mathcal{T})$.

II.1.7. Upper error bound. We fix an arbitrary function $w \in H^1_D(\Omega)$ and choose $w_\mathcal{T} = I_\mathcal{T} w$ with the quasi-interpolation operator of Section I.2.11 (p. 21). The Cauchy-Schwarz inequality for integrals and the properties of $I_\mathcal{T}$ then yield

$$\begin{aligned}
&\int_\Omega fw + \int_{\Gamma_N} gw - \int_\Omega \nabla u_\mathcal{T} \cdot \nabla w \\
&\quad = \sum_{K\in\mathcal{T}} \int_K R_K(u_\mathcal{T})(w - I_\mathcal{T} w) + \sum_{E\in\mathcal{E}} \int_E R_E(u_\mathcal{T})(w - I_\mathcal{T} w) \\
&\quad \le \sum_{K\in\mathcal{T}} \|R_K(u_\mathcal{T})\|_K \|w - I_\mathcal{T} w\|_K + \sum_{E\in\mathcal{E}} \|R_E(u_\mathcal{T})\|_E \|w - I_\mathcal{T} w\|_E \\
&\quad \le \sum_{K\in\mathcal{T}} \|R_K(u_\mathcal{T})\|_K c_{A1} h_K \|w\|_{H^1(\widetilde{\omega}_K)} \\
&\qquad + \sum_{E\in\mathcal{E}} \|R_E(u_\mathcal{T})\|_E c_{A2} h_E^{\frac{1}{2}} \|w\|_{H^1(\widetilde{\omega}_E)}.
\end{aligned}$$

Invoking the Cauchy-Schwarz inequality for sums this gives

$$\begin{aligned}
&\int_\Omega fw + \int_{\Gamma_N} gw - \int_\Omega \nabla u_\mathcal{T} \cdot \nabla w \\
&\qquad \le \max\{c_{A1}, c_{A2}\} \left\{ \sum_{K\in\mathcal{T}} h_K^2 \|R_K(u_\mathcal{T})\|_K^2 \right.
\end{aligned}$$

$$+ \sum_{E\in\mathcal{E}} h_E \|R_E(u_\mathcal{T})\|_E^2 \Bigg\}^{\frac{1}{2}} \cdot$$

$$\cdot \left\{ \sum_{K\in\mathcal{T}} \|w\|^2_{H^1(\widetilde{\omega}_K)} + \sum_{E\in\mathcal{E}} \|w\|^2_{H^1(\widetilde{\omega}_E)} \right\}^{\frac{1}{2}}.$$

In a last step we observe that the shape-regularity of $\mathcal{T}$ implies

$$\left\{ \sum_{K\in\mathcal{T}} \|w\|^2_{H^1(\widetilde{\omega}_K)} + \sum_{E\in\mathcal{E}} \|w\|^2_{H^1(\widetilde{\omega}_E)} \right\}^{\frac{1}{2}} \leq c\|w\|_{H^1(\Omega)}$$

with a constant c which only depends on the shape parameter $\mathrm{C}_\mathcal{T}$ of $\mathcal{T}$ and which takes into account that every element is counted several times on the left-hand side of this inequality.

Combining these estimates with the equivalence of error and residual, we obtain the following upper bound on the error

$$\|u - u_\mathcal{T}\|_{H^1(\Omega)} \leq c^* \left\{ \sum_{K\in\mathcal{T}} h_K^2 \|R_K(u_\mathcal{T})\|_K^2 + \sum_{E\in\mathcal{E}} h_E \|R_E(u_\mathcal{T})\|_E^2 \right\}^{\frac{1}{2}}$$

with

$$c^* = \sqrt{1 + c_\Omega^2}\, \max\{c_{A1}, c_{A2}\} c.$$

The right-hand side of this estimate can be used as an *a posteriori error estimator* since it only involves the known data f and g, the solution $u_\mathcal{T}$ of the discrete problem, and the geometrical data of the partition. The above inequality implies that the a posteriori error estimator is *reliable* in the sense that an inequality of the form "error estimator $\leq$ tolerance" implies that the true error is also less than the tolerance up to the multiplicative constant c^*. We want to show that the error estimator is also *efficient* in the sense that an inequality of the form "error estimator $\geq$ tolerance" implies that the true error is also greater than the tolerance possibly up to another multiplicative constant.

For general functions f and g the exact evaluation of the integrals occurring on the right-hand side of the above estimate may be prohibitively expensive or even impossible. The integrals then must be approximated by suitable quadrature formulae. Alternatively the functions f and g may be approximated by simpler functions, e.g., piecewise polynomial ones, and the resulting integrals be evaluated exactly. Often, both approaches are equivalent.

II.1.8. Lower error bound. In order to prove the announced efficiency, we denote for every element K by f_K the mean value of f on K

$$\boxed{f_K = \frac{1}{|K|}\int_K f \mathrm{d}x}$$

and for every edge E on the Neumann boundary by g_E the mean value of g on E

$$\boxed{g_E = \frac{1}{|E|}\int_E g \mathrm{d}S.}$$

We fix an arbitrary element K and insert the function

$$w_K = (f_K + \Delta u_\mathcal{T})\psi_K$$

in the L^2-representation of the residual. Taking into account that $\operatorname{supp} w_K \subset K$ we obtain

$$\int_K R_K(u_\mathcal{T}) w_K = \int_K \nabla(u - u_\mathcal{T}) \cdot \nabla w_K.$$

We add $\int_K (f_K - f) w_K$ on both sides of this equation and obtain

$$\begin{aligned}\int_K (f_K + \Delta u_\mathcal{T})^2 \psi_K &= \int_K (f_K + \Delta u_\mathcal{T}) w_K \\ &= \int_K \nabla(u - u_\mathcal{T}) \cdot \nabla w_K - \int_K (f - f_K) w_K.\end{aligned}$$

The results of Section I.2.12 (p. 22) imply for the left hand-side of this equation

$$\int_K (f_K + \Delta u_\mathcal{T})^2 \psi_K \geq c_{I1}^2 \|f_K + \Delta u_\mathcal{T}\|_K^2$$

and for the two terms on its right-hand side

$$\begin{aligned}\int_K \nabla(u - u_\mathcal{T}) \cdot \nabla w_K &\leq \|\nabla(u - u_\mathcal{T})\|_K \|\nabla w_K\|_K \\ &\leq \|\nabla(u - u_\mathcal{T})\|_K c_{I2} h_K^{-1} \|f_K + \Delta u_\mathcal{T}\|_K \\ \int_K (f - f_K) w_K &\leq \|f - f_K\|_K \|w_K\|_K \\ &\leq \|f - f_K\|_K \|f_K + \Delta u_\mathcal{T}\|_K.\end{aligned}$$

This proves that

$$\begin{aligned}h_K \|f_K + \Delta u_\mathcal{T}\|_K \leq\ & c_{I1}^{-2} c_{I2} \|\nabla(u - u_\mathcal{T})\|_K \\ & + c_{I1}^{-2} h_K \|f - f_K\|_K.\end{aligned} \tag{II.1.4}$$

Next, we consider an arbitrary interior edge $E \in \mathcal{E}_\Omega$ and insert the function

$$w_E = R_E(u_\mathcal{T})\psi_E$$

in the L^2-representation of the residual. This gives

$$\begin{aligned}
\int_E \mathbb{J}_E(\mathbf{n}_E \cdot \nabla u_\mathcal{T})^2 \psi_E &= \int_E R_E(u_\mathcal{T}) w_E \\
&= \int_{\omega_E} \nabla(u - u_\mathcal{T}) \cdot \nabla w_E \\
&\quad - \sum_{\substack{K \in \mathcal{T} \\ E \in \mathcal{E}_K}} \int_K R_K(u_\mathcal{T}) w_E \\
&= \int_{\omega_E} \nabla(u - u_\mathcal{T}) \cdot \nabla w_E \\
&\quad - \sum_{\substack{K \in \mathcal{T} \\ E \in \mathcal{E}_K}} \int_K (f_K + \Delta u_\mathcal{T}) w_E \\
&\quad - \sum_{\substack{K \in \mathcal{T} \\ E \in \mathcal{E}_K}} \int_K (f - f_K) w_E
\end{aligned}$$

The results of Section I.2.12 (p. 22) imply for the left-hand side of this equation

$$\int_E \mathbb{J}_E(\mathbf{n}_E \cdot \nabla u_\mathcal{T})^2 \psi_E \geq c_{I3}^2 \|\mathbb{J}_E(\mathbf{n}_E \cdot \nabla u_\mathcal{T})\|_E^2$$

and for the three terms on its right-hand side

$$\begin{aligned}
\int_{\omega_E} \nabla(u - u_\mathcal{T}) \cdot \nabla w_E &\leq \|\nabla(u - u_\mathcal{T})\|_{H^1(\omega_E)} \|\nabla w_E\|_{H^1(\omega_E)} \\
&\leq \|\nabla(u - u_\mathcal{T})\|_{H^1(\omega_E)} \\
&\qquad \cdot c_{I4} h_E^{-\frac{1}{2}} \|\mathbb{J}_E(\mathbf{n}_E \cdot \nabla u_\mathcal{T})\|_E \\
\sum_{\substack{K \in \mathcal{T} \\ E \in \mathcal{E}_K}} \int_K (f_K + \Delta u_\mathcal{T}) w_E &\leq \sum_{\substack{K \in \mathcal{T} \\ E \in \mathcal{E}_K}} \|f_K + \Delta u_\mathcal{T}\|_K \|w_E\|_K \\
&\leq \sum_{\substack{K \in \mathcal{T} \\ E \in \mathcal{E}_K}} \|f_K + \Delta u_\mathcal{T}\|_K \\
&\qquad \cdot c_{I5} h_E^{\frac{1}{2}} \|\mathbb{J}_E(\mathbf{n}_E \cdot \nabla u_\mathcal{T})\|_E
\end{aligned}$$

$$\sum_{\substack{K \in \mathcal{T} \\ E \in \mathcal{E}_K}} \int_K (f - f_K) w_E \leq \sum_{\substack{K \in \mathcal{T} \\ E \in \mathcal{E}_K}} \|f - f_K\|_K \|w_E\|_K$$

$$\leq \sum_{\substack{K\in\mathcal{T}\\ E\in\mathcal{E}_K}} \|f-f_K\|_K \cdot c_{I5} h_E^{\frac{1}{2}} \|\mathbb{J}_E(\mathbf{n}_E\cdot\nabla u_\mathcal{T})\|_E$$

and thus yields

$$\begin{aligned} c_{I3}^2\|\mathbb{J}_E(\mathbf{n}_E\cdot\nabla u_\mathcal{T})\|_E &\leq c_{I4}h_E^{-\frac{1}{2}}\|\nabla(u-u_\mathcal{T})\|_{H^1(\omega_E)} \\ &\quad + \sum_{\substack{K\in\mathcal{T}\\ E\in\mathcal{E}_K}} c_{I5}h_E^{\frac{1}{2}}\|f_K+\Delta u_\mathcal{T}\|_K \\ &\quad + \sum_{\substack{K\in\mathcal{T}\\ E\in\mathcal{E}_K}} c_{I5}h_E^{\frac{1}{2}}\|f-f_K\|_K. \end{aligned}$$

Combining this estimate with inequality (II.1.4) we obtain

$$\begin{aligned} &h_E^{\frac{1}{2}}\|\mathbb{J}_E(\mathbf{n}_E\cdot\nabla u_\mathcal{T})\|_E \\ &\quad\leq c_{I3}^{-2}c_{I5}\left[c_{I4}+c_{I1}^{-2}c_{I2}\right]\|\nabla(u-u_\mathcal{T})\|_{H^1(\omega_E)} \\ &\qquad + c_{I3}^{-2}c_{I5}\left[1+c_{I1}^{-2}\right]h_E\sum_{\substack{K\in\mathcal{T}\\ E\in\mathcal{E}_K}}\|f-f_K\|_K. \end{aligned} \tag{II.1.5}$$

Finally, we fix an edge E on the Neumann boundary, denote by K the adjacent element and insert the function

$$w_E = (g_E - \mathbf{n}_E\cdot\nabla u_\mathcal{T})\psi_E$$

in L^2-representation of the residual. This gives

$$\int_E R_E(u_\mathcal{T})w_E = \int_K \nabla(u-u_\mathcal{T})\cdot\nabla w_E - \int_K R_K(u_\mathcal{T})w_E.$$

We add $\int_E (g_E-g)w_E$ on both sides of this equation and obtain

$$\begin{aligned} \int_E (g_E-\mathbf{n}_E\cdot\nabla u_\mathcal{T})^2\psi_E &= \int_E (g_E-\mathbf{n}_E\cdot\nabla u_\mathcal{T})w_E \\ &= \int_K \nabla(u-u_\mathcal{T})\cdot\nabla w_E \\ &\quad - \int_K (f_K+\Delta u_\mathcal{T})w_E - \int_K (f-f_K)w_E \\ &\quad - \int_E (g-g_E)w_E. \end{aligned}$$

Invoking once again the results of Section I.2.12 (p. 22) and using the same arguments as above this implies that

$$
\begin{aligned}
&h_E^{\frac12}\|g_E-\mathbf{n}_E\cdot\nabla u_{\mathcal T}\|_E\\
&\quad\le c_{I3}^{-2}c_{I5}\left[c_{I4}+c_{I1}^{-2}c_{I2}\right]\|\nabla(u-u_{\mathcal T})\|_K\\
&\qquad+c_{I3}^{-2}c_{I5}\left[1+c_{I1}^{-2}\right]h_K\|f-f_K\|_K\\
&\qquad+c_{I3}^{-2}h_E^{\frac12}\|g-g_E\|_E.
\end{aligned}
\tag{II.1.6}
$$

Estimates (II.1.4), (II.1.5), and (II.1.6) prove the announced efficiency of the a posteriori error estimate:

$$
\begin{aligned}
&\Big\{h_K^2\|f_K+\Delta u_{\mathcal T}\|_K^2\\
&\qquad+\frac12\sum_{E\in\mathcal E_K\cap\mathcal E_\Omega}h_E\|\mathbb J_E(\mathbf n_E\cdot\nabla u_{\mathcal T})\|_E^2\\
&\qquad+\sum_{E\in\mathcal E_K\cap\mathcal E_{\Gamma_N}}h_E\|g_E-\mathbf n_E\cdot\nabla u_{\mathcal T}\|_E^2\Big\}^{\frac12}\\
&\quad\le c_*\Big\{\|u-u_{\mathcal T}\|_{H^1(\omega_K)}^2\\
&\qquad+\sum_{\substack{K'\in\mathcal T\\ \mathcal E_{K'}\cap\mathcal E_K\neq\emptyset}}h_{K'}^2\|f-f_{K'}\|_{H^1(K')}^2\\
&\qquad+\sum_{E\in\mathcal E_K\cap\mathcal E_{\Gamma_N}}h_E\|g-g_E\|_E^2\Big\}^{\frac12}.
\end{aligned}
$$

The constant c_* only depends on the shape parameter $\mathrm{C}_{\mathcal T}$.

II.1.9. Residual a posteriori error estimate. The results of the preceding sections can be summarized as follows:

Denote by $u\in H^1_D(\Omega)$ and $u_{\mathcal T}\in S^{1,0}_D(\mathcal T)$ the unique solutions of problems (II.1.2) (p. 26) and (II.1.3) (p. 26), respectively. For every element $K\in\mathcal T$ define the *residual a posteriori error estimator* $\eta_{R,K}$ by

$$
\begin{aligned}
\eta_{R,K}=&\Big\{h_K^2\|f_K+\Delta u_{\mathcal T}\|_K^2\\
&+\frac12\sum_{E\in\mathcal E_K\cap\mathcal E_\Omega}h_E\|\mathbb J_E(\mathbf n_E\cdot\nabla u_{\mathcal T})\|_E^2\\
&+\sum_{E\in\mathcal E_K\cap\mathcal E_{\Gamma_N}}h_E\|g_E-\mathbf n_E\cdot\nabla u_{\mathcal T}\|_E^2\Big\}^{\frac12},
\end{aligned}
$$

where f_K and g_E are the mean values of f and g on K and E, respectively. There are two constants c^* and c_*, which only depend on the shape parameter $\mathrm{C}_{\mathcal{T}}$, such that the estimates

$$\|u-u_{\mathcal{T}}\|_{H^1(\Omega)} \le c^*\Big\{\sum_{K\in\mathcal{T}}\eta_{R,K}^2 + \sum_{K\in\mathcal{T}} h_K^2\|f-f_K\|_K^2 + \sum_{E\in\mathcal{E}_{\Gamma_N}} h_E\|g-g_E\|_E^2\Big\}^{\frac{1}{2}}$$

and

$$\eta_{R,K} \le c_*\Big\{\|u-u_{\mathcal{T}}\|^2_{H^1(\omega_K)} + \sum_{\substack{K'\in\mathcal{T}\\ \mathcal{E}_{K'}\cap\mathcal{E}_K\neq\emptyset}} h_{K'}^2\|f-f_{K'}\|^2_{H^1(K')} + \sum_{E\in\mathcal{E}_K\cap\mathcal{E}_{\Gamma_N}} h_E\|g-g_E\|_E^2\Big\}^{\frac{1}{2}}$$

hold for all $K\in\mathcal{T}$.

REMARK II.1.1. The factor $\frac{1}{2}$ multiplying the second term in $\eta_{R,K}$ takes into account that each interior edge is counted twice when adding all $\eta_{R,K}^2$. Note that $\Delta u_{\mathcal{T}} = 0$ on all triangles.

REMARK II.1.2. The first term in $\eta_{R,K}$ is related to the residual of $u_{\mathcal{T}}$ with respect to the strong form of the differential equation. The second and third term in $\eta_{R,K}$ are related to that boundary operator which links the strong and weak form of the differential equation. These boundary terms are crucial when considering low order finite element discretizations as done here. Consider e.g. problem (II.1.1) (p. 26) in the unit square $(0,1)^2$ with Dirichlet boundary conditions on the left and bottom part and exact solution $u(x) = x_1x_2$. When using a triangulation consisting of right angled isosceles triangles and evaluating the line integrals by the trapezoidal rule, the solution of problem (II.1.3) (p. 26) satisfies $u_{\mathcal{T}}(x) = u(x)$ for all $x\in\mathcal{N}$ but $u_{\mathcal{T}}\neq u$. The second and third term in $\eta_{R,K}$ reflect the fact that $u_{\mathcal{T}}\notin H^2(\Omega)$ and that $u_{\mathcal{T}}$ does not exactly satisfy the Neumann boundary condition.

REMARK II.1.3. The correction terms

$$h_K\|f-f_K\|_K \quad\text{and}\quad h_E^{\frac{1}{2}}\|g-g_E\|_E$$

in the above a posteriori error estimate are in general higher order perturbations of the other terms. In special situations, however, they can be dominant. To see this, assume that $\mathcal{T}$ contains at least one triangle, choose a triangle $K_0 \in \mathcal{T}$ and a non-zero function $\varrho_0 \in C_0^\infty(\overset{\circ}{K}_0)$, and consider problem (II.1.1) (p. 26) with $f = -\Delta\varrho_0$ and $\Gamma_D = \Gamma$. Since

$$\int_{K_0} f = -\int_{K_0} \Delta\varrho_0 = 0$$

and $f = 0$ outside K_0, we have

$$f_K = 0$$

for all $K \in \mathcal{T}$. Since

$$\int_\Omega f v_\mathcal{T} = -\int_{K_0} \Delta\varrho_0 v_\mathcal{T} = -\int_{K_0} \varrho_0 \Delta v_\mathcal{T} = 0$$

for all $v_\mathcal{T} \in S_D^{1,0}(\mathcal{T})$, the exact solution of problem (II.1.3) (p. 26) is

$$u_\mathcal{T} = 0.$$

Hence, we have

$$\eta_{R,K} = 0$$

for all $K \in \mathcal{T}$, but

$$\|u - u_\mathcal{T}\|_{H^1(\Omega)} \neq 0.$$

This effect is not restricted to the particular approximation of f considered here. Since $\varrho_0 \in C_0^\infty(\overset{\circ}{K}_0)$ is completely arbitrary, we will always encounter similar difficulties as long as we do not evaluate $\|f\|_K$ exactly – which in general is impossible. Obviously, this problem is cured when further refining the mesh.

II.2. A catalogue of error estimators for the model problem

II.2.1. Solution of auxiliary local discrete problems. The results of Section II.1 show that we must reliably estimate the norm of the residual as an element of the dual space of $H_D^1(\Omega)$. This could be achieved by lifting the residual to a suitable subspace of $H_D^1(\Omega)$ by solving auxiliary problems similar to, but simpler than the original discrete problem (II.1.3) (p. 26). Practical considerations and the results of the Section II.1 suggest that the auxiliary problems should satisfy the following conditions:

- In order to get an information on the local behaviour of the error, they should involve only small subdomains of Ω.
- In order to yield an accurate information on the error, they should be based on finite element spaces which are more accurate than the original one.
- In order to keep the computational work at a minimum, they should involve as few degrees of freedom as possible.

- To each edge and, if need be, to each element there should correspond at least one degree of freedom in at least one of the auxiliary problems.
- The solution of all auxiliary problems should not cost more than the assembly of the stiffness matrix of problem (II.1.3) (p. 26).

There are many possible ways to satisfy these conditions. Here, we present three of them. To this end we denote by $\mathbb{P}_1 = \operatorname{span}\{1, x_1, x_2\}$ the space of linear polynomials in two variables.

II.2.1.1. *Dirichlet problems associated with vertices.* First, we decide to impose Dirichlet boundary conditions on the auxiliary problems. The fourth condition then implies that the corresponding subdomains must consist of more than one element. A reasonable choice is to consider all nodes $x \in \mathcal{N}_\Omega \cup \mathcal{N}_{\Gamma_N}$ and the corresponding domains ω_x (cf. Figures I.2.2 (p. 17) and I.2.3 (p. 18)). The above conditions then lead to the following definition:
Set for all $x \in \mathcal{N}_\Omega \cup \mathcal{N}_{\Gamma_N}$

$$\begin{aligned} V_x = \operatorname{span}\{\varphi\psi_K, \rho\psi_E, \sigma\psi_{E'} : {} & K \in \mathcal{T}, x \in \mathcal{N}_K, \\ & E \in \mathcal{E}_\Omega, x \in \mathcal{N}_E, \\ & E' \in \mathcal{E}_{\Gamma_N}, E' \subset \partial\omega_x, \\ & \varphi, \rho, \sigma \in \mathbb{P}_1\} \end{aligned}$$

and

$$\eta_{D,x} = \|\nabla v_x\|_{\omega_x}$$

where $v_x \in V_x$ is the unique solution of

$$\begin{aligned} \int_{\omega_x} \nabla v_x \cdot \nabla w = {} & \sum_{\substack{K \in \mathcal{T} \\ x \in \mathcal{N}_K}} \int_K f_K w + \sum_{\substack{E \in \mathcal{E}_{\Gamma_N} \\ E \subset \partial\omega_x}} \int_E g_E w \\ & - \int_{\omega_x} \nabla u_{\mathcal{T}} \cdot \nabla w \end{aligned}$$

for all $w \in V_x$.

In order to get a different interpretation of the above problem, set

$$u_x = u_{\mathcal{T}} + v_x.$$

Then

$$\eta_{D,x} = \|\nabla(u_x - u_{\mathcal{T}})\|_{\omega_x}$$

and $u_x \in u_{\mathcal{T}} + V_x$ is the unique solution of

$$\int_{\omega_x} \nabla u_x \cdot \nabla w = \sum_{\substack{K \in \mathcal{T} \\ x \in \mathcal{N}_K}} \int_K f_K w + \sum_{\substack{E \in \mathcal{E}_{\Gamma_N} \\ E \subset \partial\omega_x}} \int_E g_E w$$

for all $w \in V_x$. This is a discrete analogue of the following Dirichlet problem

$$\begin{aligned} -\Delta\varphi &= f && \text{in } \omega_x \\ \varphi &= u_{\mathcal{T}} && \text{on } \partial\omega_x \backslash \Gamma_N \\ \frac{\partial\varphi}{\partial n} &= g && \text{on } \partial\omega_x \cap \Gamma_N. \end{aligned}$$

Hence, we can interpret the error estimator $\eta_{D,x}$ in two ways:

- We solve a local analogue of the residual equation using a higher order finite element approximation and use a suitable norm of the solution as error estimator.
- We solve a local discrete analogue of the original problem using a higher order finite element space and compare the solution of this problem to the one of problem (II.1.3) (p. 26).

Thus, in a certain sense, $\eta_{D,x}$ is based on an extrapolation technique. It can be proven that it yields upper and lower bounds on the error $u - u_{\mathcal{T}}$ and that it is comparable to the estimator $\eta_{R,T}$.

Denote by $u \in H^1_D(\Omega)$ and $u_{\mathcal{T}} \in S^{1,0}_D(\Omega)$ the unique solutions of problems (II.1.2) (p. 26) and (II.1.3) (p. 26). There are constants $c_{\mathcal{N},1}, \dots, c_{\mathcal{N},4}$, which only depend on the shape parameter $\mathrm{C}_{\mathcal{T}}$, such that the estimates

$$\eta_{D,x} \le c_{\mathcal{N},1} \Big\{ \sum_{\substack{K \in \mathcal{T} \\ x \in \mathcal{N}_K}} \eta^2_{R,K} \Big\}^{\frac{1}{2}},$$

$$\eta_{R,K} \le c_{\mathcal{N},2} \Big\{ \sum_{x \in \mathcal{N}_K \backslash \mathcal{N}_{\Gamma_D}} \eta^2_{D,x} \Big\}^{\frac{1}{2}},$$

$$\begin{aligned} \eta_{D,x} \le c_{\mathcal{N},3} \Big\{ & \|u - u_{\mathcal{T}}\|^2_{H^1(\omega_x)} \\ & + \sum_{\substack{K \in \mathcal{T} \\ x \in \mathcal{N}_K}} h^2_K \|f - f_K\|^2_K \\ & + \sum_{\substack{E \in \mathcal{E}_{\Gamma_N} \\ E \subset \partial\omega_x}} h_E \|g - g_E\|^2_E \Big\}^{\frac{1}{2}}, \end{aligned}$$

$$\begin{aligned} \|u - u_{\mathcal{T}}\|_{H^1(\Omega)} \le c_{\mathcal{N},4} \Big\{ & \sum_{x \in \mathcal{N}_\Omega \cup \mathcal{N}_{\Gamma_N}} \eta^2_{D,x} \\ & + \sum_{K \in \mathcal{T}} h^2_K \|f - f_K\|^2_K \end{aligned}$$

$$+ \sum_{E\in\mathcal{E}_{\Gamma_N}} h_E\|g-g_E\|_E^2\Big\}^{\frac{1}{2}}$$

hold for all $x \in \mathcal{N}_\Omega \cup \mathcal{N}_{\Gamma_N}$ and all $K \in \mathcal{T}$. Here, f_K, g_E, and $\eta_{R,K}$ are as in Sections II.1.8 (p. 31) and II.1.9 (p. 34).

II.2.1.2. *Dirichlet problems associated with elements.* We now consider an estimator which is a slight variation of the preceding one. Instead of all $x \in \mathcal{N}_\Omega \cup \mathcal{N}_{\Gamma_N}$ and the corresponding domains ω_x we consider all $K \in \mathcal{T}$ and the corresponding sets ω_K (cf. Figure I.2.2 (p. 17)). The considerations from the beginning of this section then lead to the following definition:
Set for all $K \in \mathcal{T}$

$$\begin{aligned}\widetilde{V}_K = \operatorname{span}\{\varphi\psi_{K'}, \rho\psi_E, \sigma\psi_{E'} : {} & K' \in \mathcal{T}, \mathcal{E}_{K'} \cap \mathcal{E}_K \neq \emptyset,\\ & E \in \mathcal{E}_K \cap \mathcal{E}_\Omega,\\ & E' \in \mathcal{E}_{\Gamma_N}, E' \subset \partial\omega_K,\\ & \varphi, \rho, \sigma \in \mathbb{P}_1\}\end{aligned}$$

and

$$\eta_{D,K} = \|\nabla\widetilde{v}_K\|_{\omega_K}$$

where $\widetilde{v}_K \in \widetilde{V}_K$ is the unique solution of

$$\begin{aligned}\int_{\omega_K} \nabla\widetilde{v}_K \cdot \nabla w = {} & \sum_{\substack{K'\in\mathcal{T}\\ \mathcal{E}_{K'}\cap\mathcal{E}_K\neq\emptyset}} \int_{K'} f_{K'}w + \sum_{\substack{E'\in\mathcal{E}_{\Gamma_N}\\ E'\subset\partial\omega_K}} \int_{E'} g_{E'}w\\ & - \int_{\omega_K} \nabla u_\mathcal{T} \cdot \nabla w\end{aligned}$$

for all $w \in \widetilde{V}_K$.

As before we can interpret $u_\mathcal{T} + \widetilde{v}_K$ as an approximate solution of the following Dirichlet problem

$$\begin{aligned} -\Delta\varphi &= f && \text{in } \omega_K\\ \varphi &= u_\mathcal{T} && \text{on } \partial\omega_K\backslash\Gamma_N\\ \frac{\partial\varphi}{\partial n} &= g && \text{on } \partial\omega_K \cap \Gamma_N.\end{aligned}$$

It can be proven that $\eta_{D,K}$ also yields upper and lower bounds on the error $u - u_\mathcal{T}$ and that it is comparable to $\eta_{D,x}$ and $\eta_{R,K}$.

Denote by $u \in H^1_D(\Omega)$ and $u_{\mathcal{T}} \in S^{1,0}_D(\mathcal{T})$ the unique solutions of problem (II.1.2) (p. 26) and (II.1.3) (p. 26). There are constants $c_{\mathcal{E},1}, \dots, c_{\mathcal{E},4}$, which only depend on the shape parameter $\mathrm{C}_{\mathcal{T}}$, such that the estimates

$$\eta_{D,K} \leq c_{\mathcal{E},1}\Big\{ \sum_{\substack{K' \in \mathcal{T} \\ \mathcal{E}_{K'} \cap \mathcal{E}_K \neq \emptyset}} \eta^2_{R,K'} \Big\}^{\frac{1}{2}},$$

$$\eta_{R,K} \leq c_{\mathcal{E},2}\Big\{ \sum_{\substack{K' \in \mathcal{T} \\ \mathcal{E}_{K'} \cap \mathcal{E}_K \neq \emptyset}} \eta^2_{D,K'} \Big\}^{\frac{1}{2}},$$

$$\begin{aligned} \eta_{D,K} \leq c_{\mathcal{E},3}\Big\{ & \|u - u_{\mathcal{T}}\|^2_{H^1(\omega_K)} \\ & + \sum_{\substack{K' \in \mathcal{T} \\ \mathcal{E}_{K'} \cap \mathcal{E}_K \neq \emptyset}} h^2_{K'} \|f - f_{K'}\|^2_{K'} \\ & + \sum_{\substack{E' \in \mathcal{E}_{\Gamma_N} \\ E' \subset \partial\omega_K}} h_{E'} \|g - g_{E'}\|^2_{E'} \Big\}^{\frac{1}{2}}, \end{aligned}$$

$$\begin{aligned} \|u - u_{\mathcal{T}}\|_{H^1(\Omega)} \leq c_{\mathcal{E},4}\Big\{ & \sum_{K \in \mathcal{T}} \eta^2_{D,K} \\ & + \sum_{K \in \mathcal{T}} h^2_K \|f - f_K\|^2_K \\ & + \sum_{E \in \mathcal{E}_{\Gamma_N}} h_E \|g - g_E\|^2_E \Big\}^{\frac{1}{2}} \end{aligned}$$

hold for all $K \in \mathcal{T}$. Here, f_K, g_E, $\eta_{R,K}$ are as in Sections II.1.8 (p. 31) and II.1.9 (p. 34).

II.2.1.3. *Neumann problems.* For the third estimator we decide to impose Neumann boundary conditions on the auxiliary problems. Now it is possible to choose the elements in $\mathcal{T}$ as the corresponding subdomain. This leads to the definition:
Set for alle $K \in \mathcal{T}$

$$V_K = \operatorname{span}\{\varphi\psi_K, \rho\psi_E : E \in \mathcal{E}_K \backslash \mathcal{E}_{\Gamma_D}, \varphi, \rho \in \mathbb{P}_1\}$$

and

$$\eta_{N,K} = \|\nabla v_K\|_K$$

where v_K is the unique solution of

$$\begin{aligned}\int_K \nabla v_K \cdot \nabla w = \int_K (f_K + \Delta u_\mathcal{T}) w \\ - \frac{1}{2} \sum_{E \in \mathcal{E}_K \cap \mathcal{E}_\Omega} \int_E \mathbb{J}_E (\mathbf{n}_E \cdot \nabla u_\mathcal{T}) w \\ + \sum_{E \in \mathcal{E}_K \cap \mathcal{E}_{\Gamma_N}} \int_E (g_E - \mathbf{n}_E \cdot \nabla u_\mathcal{T}) w\end{aligned}$$

for all $w \in V_K$.

Note, that the factor $\frac{1}{2}$ multiplying the residuals on interior edges takes into account that interior edges are counted twice when summing the contributions of all elements.

The above problem can be interpreted as a discrete analogue of the following Neumann problem

$$\begin{aligned} -\Delta \varphi &= R_K(u_\mathcal{T}) && \text{in } K \\ \frac{\partial \varphi}{\partial n} &= \frac{1}{2} R_E(u_\mathcal{T}) && \text{on } \partial K \cap \Omega \\ \frac{\partial \varphi}{\partial n} &= R_E(u_\mathcal{T}) && \text{on } \partial K \cap \Gamma_N \\ \varphi &= 0 && \text{on } \partial K \cap \Gamma_D. \end{aligned}$$

Again it can be proven that $\eta_{N,K}$ also yields upper and lower bounds on the error and that it is comparable to $\eta_{R,K}$.

Denote by $u \in H^1_D(\Omega)$ and $u_\mathcal{T} \in S^{1,0}_D(\mathcal{T})$ the unique solutions of problem (II.1.2) (p. 26) and (II.1.3) (p. 26). There are constants $c_{\mathcal{E},5}, \dots, c_{\mathcal{E},8}$, which only depend on the shape parameter $\mathrm{C}_\mathcal{T}$, such that the estimates

$$\begin{aligned} \eta_{N,K} &\le c_{\mathcal{E},5} \eta_{R,K}, \\ \eta_{R,K} &\le c_{\mathcal{E},6} \Big\{ \sum_{\substack{K' \in \mathcal{T} \\ \mathcal{E}_{K'} \cap \mathcal{E}_K \neq \emptyset}} \eta^2_{N,K'} \Big\}^{\frac{1}{2}}, \\ \eta_{N,K} &\le c_{\mathcal{E},7} \Big\{ \|u - u_\mathcal{T}\|^2_{H^1(\omega_K)} \\ &\qquad + \sum_{\substack{K' \in \mathcal{T} \\ \mathcal{E}_{K'} \cap \mathcal{E}_K \neq \emptyset}} h^2_{K'} \|f - f_{K'}\|^2_{K'} \\ &\qquad + \sum_{E \in \mathcal{E}_K \cap \mathcal{E}_{\Gamma_N}} h_E \|g - g_E\|^2_E \Big\}^{\frac{1}{2}}, \\ \|u - u_\mathcal{T}\|_{H^1(\Omega)} &\le c_{\mathcal{E},8} \Big\{ \sum_{K \in \mathcal{T}} \eta^2_{N,K} \end{aligned}$$

$$+ \sum_{K\in\mathcal{T}} h_K^2 \|f - f_K\|_K^2$$
$$+ \sum_{E\in\mathcal{E}_{\Gamma_N}} h_E \|g - g_E\|_E^2 \Big\}^{\frac{1}{2}}$$

hold for all $K \in \mathcal{T}$. Here, f_K, g_E, $\eta_{R,K}$ are as in Sections II.1.8 (p. 31) and II.1.9 (p. 34).

REMARK II.2.1. When $\mathcal{T}$ exclusively consists of triangles $\Delta u_\mathcal{T}$ vanishes element-wise and the normal derivatives $\mathbf{n}_E \cdot \nabla u_\mathcal{T}$ are edge-wise constant. In this case the functions φ, ρ, and σ can be dropped in the definitions of V_x, $\widetilde{V}_K$, and V_K. This considerably reduces the dimension of the spaces V_x, $\widetilde{V}_K$, and V_K and thus of the discrete auxiliary problems. Figures I.2.2 (p. 17) and I.2.3 (p. 18) show typical examples of domains ω_x and ω_K. From this it is obvious that in general the above auxiliary discrete problems have at least the dimensions 12, 7, and 4, respectively. In any case the computation of $\eta_{D,x}$, $\eta_{D,K}$, and $\eta_{N,K}$ is more expensive than the one of $\eta_{R,K}$. This is sometimes payed off by an improved accuracy of the error estimate.

II.2.2. Hierarchical error estimates. The key-idea of the hierarchical approach is to solve problem (II.1.2) (p. 26) approximately using a more accurate finite element space and to compare this solution with the solution of problem (II.1.3) (p. 26). In order to reduce the computational cost of the new problem, the new finite element space is decomposed into the original one and a nearly orthogonal higher order complement. Then only the contribution corresponding to the complement is computed. To further reduce the computational cost, the original bilinear form is replaced by an equivalent one which leads to a diagonal stiffness matrix.

To describe this idea in detail, we consider a finite element space $Y_\mathcal{T}$ which satisfies $S_D^{1,0}(\mathcal{T}) \subset Y_\mathcal{T} \subset H_D^1(\Omega)$ and which either consists of higher order elements or corresponds to a refinement of $\mathcal{T}$. We then denote by $w_\mathcal{T} \in Y_\mathcal{T}$ the unique solution of

$$\int_\Omega \nabla w_\mathcal{T} \cdot \nabla v_\mathcal{T} = \int_\Omega f v_\mathcal{T} + \int_{\Gamma_N} g v_\mathcal{T} \tag{II.2.1}$$

for all $v_\mathcal{T} \in Y_\mathcal{T}$.

To compare the solutions $w_\mathcal{T}$ of problem (II.2.1) and $u_\mathcal{T}$ of problem (II.1.3) (p. 26) we subtract $\int_\Omega \nabla u_\mathcal{T} \cdot \nabla v_\mathcal{T}$ on both sides of equation (II.2.1) and take the Galerkin orthogonality into account. We thus

obtain

$$\int_\Omega \nabla(w_\mathcal{T} - u_\mathcal{T}) \cdot \nabla v_\mathcal{T} = \int_\Omega f v_\mathcal{T} + \int_{\Gamma_N} g v_\mathcal{T} - \int_\Omega \nabla u_\mathcal{T} \cdot \nabla v_\mathcal{T}$$
$$= \int_\Omega \nabla(u - u_\mathcal{T}) \cdot \nabla v_\mathcal{T}$$

for all $v_\mathcal{T} \in Y_\mathcal{T}$, where $u \in H^1_D(\Omega)$ is the unique solution of problem (II.1.2) (p. 26). Since $S^{1,0}_D(\mathcal{T}) \subset Y_\mathcal{T}$, we may insert $v_\mathcal{T} = w_\mathcal{T} - u_\mathcal{T}$ as a test-function in this equation. The Cauchy-Schwarz inequality for integrals then implies

$$\|\nabla(w_\mathcal{T} - u_\mathcal{T})\| \le \|\nabla(u - u_\mathcal{T})\|.$$

To prove the converse estimate, we assume that the space $Y_\mathcal{T}$ satisfies a *saturation assumption*, i.e., there is a constant β with $0 \le \beta < 1$ such that

$$\|\nabla(u - w_\mathcal{T})\| \le \beta \|\nabla(u - u_\mathcal{T})\|. \tag{II.2.2}$$

From the saturation assumption (II.2.2) and the triangle inequality we immediately conclude that

$$\|\nabla(u - u_\mathcal{T})\| \le \|\nabla(u - w_\mathcal{T})\| + \|\nabla(w_\mathcal{T} - u_\mathcal{T})\|$$
$$\le \beta \|\nabla(u - u_\mathcal{T})\| + \|\nabla(w_\mathcal{T} - u_\mathcal{T})\|$$

and therefore

$$\|\nabla(u - u_\mathcal{T})\| \le \frac{1}{1-\beta} \|\nabla(w_\mathcal{T} - u_\mathcal{T})\|.$$

Thus, we have proven the two-sided error bound

$$\|\nabla(w_\mathcal{T} - u_\mathcal{T})\| \le \|\nabla(u - u_\mathcal{T})\|$$
$$\le \frac{1}{1-\beta} \|\nabla(w_\mathcal{T} - u_\mathcal{T})\|.$$

Hence, we may use $\|\nabla(w_\mathcal{T} - u_\mathcal{T})\|$ as an a posteriori error estimator.

This device, however, is not efficient since the computation of $w_\mathcal{T}$ is at least as costly as the one of $u_\mathcal{T}$. In order to obtain a more efficient error estimation, we use a hierarchical splitting

$$Y_\mathcal{T} = S^{1,0}_D(\mathcal{T}) \oplus Z_\mathcal{T}$$

and assume that the spaces $S^{1,0}_D(\mathcal{T})$ and $Z_\mathcal{T}$ are nearly orthogonal and satisfy a *strengthened Cauchy-Schwarz inequality*, i.e., there is a constant γ with $0 \le \gamma < 1$ such that

$$\left| \int_\Omega \nabla v_\mathcal{T} \cdot \nabla z_\mathcal{T} \right| \le \gamma \|\nabla v_\mathcal{T}\| \|\nabla z_\mathcal{T}\| \tag{II.2.3}$$

holds for all $v_\mathcal{T} \in S^{1,0}_D(\mathcal{T})$, $z_\mathcal{T} \in Z_\mathcal{T}$.

Now, we write $w_{\mathcal{T}} - u_{\mathcal{T}}$ in the form $\overline{v}_{\mathcal{T}} + \overline{z}_{\mathcal{T}}$ with $\overline{v}_{\mathcal{T}} \in S^{1,0}_D(\mathcal{T})$ and $\overline{z}_{\mathcal{T}} \in Z_{\mathcal{T}}$. From the strengthened Cauchy-Schwarz inequality we then deduce that

$$\begin{aligned}(1-\gamma)\{\|\nabla\overline{v}_{\mathcal{T}}\|^2 + \|\nabla\overline{z}_{\mathcal{T}}\|^2\}&\\ \le \|\nabla(w_{\mathcal{T}} - u_{\mathcal{T}})\|^2&\\ \le (1+\gamma)\{\|\nabla\overline{v}_{\mathcal{T}}\|^2 + \|\nabla\overline{z}_{\mathcal{T}}\|^2\}&\end{aligned}$$

and in particular

$$\|\nabla\overline{z}_{\mathcal{T}}\| \le \frac{1}{\sqrt{1-\gamma}}\|\nabla(w_{\mathcal{T}} - u_{\mathcal{T}})\|. \tag{II.2.4}$$

Denote by $z_{\mathcal{T}} \in Z_{\mathcal{T}}$ the unique solution of

$$\int_\Omega \nabla z_{\mathcal{T}} \cdot \nabla\zeta_{\mathcal{T}} = \int_\Omega f\zeta_{\mathcal{T}} + \int_{\Gamma_N} g\zeta_{\mathcal{T}} - \int_\Omega \nabla u_{\mathcal{T}} \cdot \nabla\zeta_{\mathcal{T}} \tag{II.2.5}$$

for all $\zeta_{\mathcal{T}} \in Z_{\mathcal{T}}$.

From the definitions (II.1.2) (p. 26), (II.1.3) (p. 26), (II.2.1), and (II.2.5) of u, $u_{\mathcal{T}}$, $w_{\mathcal{T}}$, and $z_{\mathcal{T}}$ we infer that

$$\begin{aligned}\int_\Omega \nabla z_{\mathcal{T}} \cdot \nabla\zeta_{\mathcal{T}} &= \int_\Omega \nabla(u - u_{\mathcal{T}}) \cdot \nabla\zeta_{\mathcal{T}}\\ &= \int_\Omega \nabla(w_{\mathcal{T}} - u_{\mathcal{T}}) \cdot \nabla\zeta_{\mathcal{T}}\end{aligned} \tag{II.2.6}$$

for all $\zeta_{\mathcal{T}} \in Z_{\mathcal{T}}$ and

$$\int_\Omega \nabla(w_{\mathcal{T}} - u_{\mathcal{T}}) \cdot \nabla v_{\mathcal{T}} = 0 \tag{II.2.7}$$

for all $v_{\mathcal{T}} \in S^{1,0}_D(\mathcal{T})$. We insert $\zeta_{\mathcal{T}} = z_{\mathcal{T}}$ in equation (II.2.6). The Cauchy-Schwarz inequality for integrals then yields

$$\|\nabla z_{\mathcal{T}}\| \le \|\nabla(u - u_{\mathcal{T}})\|.$$

On the other hand, we conclude from inequality (II.2.4) and equations (II.2.6) and (II.2.7) with $\zeta_{\mathcal{T}} = \overline{z}_{\mathcal{T}}$ that

$$\begin{aligned}\|\nabla(w_{\mathcal{T}} - u_{\mathcal{T}})\|^2 &= \int_\Omega \nabla(w_{\mathcal{T}} - u_{\mathcal{T}}) \cdot \nabla(w_{\mathcal{T}} - u_{\mathcal{T}})\\ &= \int_\Omega \nabla(w_{\mathcal{T}} - u_{\mathcal{T}}) \cdot \nabla(\overline{v}_{\mathcal{T}} + \overline{z}_{\mathcal{T}})\\ &= \int_\Omega \nabla(w_{\mathcal{T}} - u_{\mathcal{T}}) \cdot \nabla\overline{z}_{\mathcal{T}}\\ &= \int_\Omega \nabla z_{\mathcal{T}} \cdot \nabla\overline{z}_{\mathcal{T}}\\ &\le \|\nabla z_{\mathcal{T}}\|\|\nabla\overline{z}_{\mathcal{T}}\|\end{aligned}$$

$$\leq \frac{1}{\sqrt{1-\gamma}}\|\nabla z_{\mathcal{T}}\|\|\nabla(w_{\mathcal{T}}-u_{\mathcal{T}})\|$$

and hence

$$\begin{aligned}\|\nabla(u-u_{\mathcal{T}})\| &\leq \frac{1}{1-\beta}\|\nabla(w_{\mathcal{T}}-u_{\mathcal{T}})\| \\ &\leq \frac{1}{(1-\beta)\sqrt{1-\gamma}}\|\nabla z_{\mathcal{T}}\|.\end{aligned}$$

Thus, we have established the two-sided error bound

$$\begin{aligned}\|\nabla z_{\mathcal{T}}\| &\leq \|\nabla(u-u_{\mathcal{T}})\| \\ &\leq \frac{1}{(1-\beta)\sqrt{1-\gamma}}\|\nabla z_{\mathcal{T}}\|.\end{aligned}$$

Therefore, $\|\nabla z_{\mathcal{T}}\|$ can be used as an error estimator.

At first sight, its computation seems to be cheaper than the one of $w_{\mathcal{T}}$ since the dimension of $Z_{\mathcal{T}}$ is smaller than that of $Y_{\mathcal{T}}$. The computation of $z_{\mathcal{T}}$, however, still requires the solution of a global system and is therefore as expensive as the calculation of $u_{\mathcal{T}}$ and $w_{\mathcal{T}}$. Yet, in most applications the functions in $Z_{\mathcal{T}}$ vanish at the vertices of $\mathcal{E}$ since $Z_{\mathcal{T}}$ is the hierarchical complement of $S^{1,0}_D(\mathcal{T})$ in $Y_{\mathcal{T}}$. This in particular implies that the stiffness matrix corresponding to $Z_{\mathcal{T}}$ is spectrally equivalent to a suitably scaled lumped mass matrix. Therefore, $z_{\mathcal{T}}$ can be replaced by a quantity $z^*_{\mathcal{T}}$ which can be computed by solving a diagonal linear system of equations.

More precisely, we assume that there is a bilinear form b on $Z_{\mathcal{T}}\times Z_{\mathcal{T}}$ which has a diagonal stiffness matrix and which defines an equivalent norm to $\|\nabla\cdot\|$ on $Z_{\mathcal{T}}$, i.e.,

$$\lambda\|\nabla\zeta_{\mathcal{T}}\|^2 \leq b(\zeta_{\mathcal{T}},\zeta_{\mathcal{T}}) \leq \Lambda\|\nabla\zeta_{\mathcal{T}}\|^2 \tag{II.2.8}$$

holds for all $\zeta_{\mathcal{T}}\in Z_{\mathcal{T}}$ with constants $0<\lambda\leq\Lambda$.

The conditions on b imply that there is a unique function $z^*_{\mathcal{T}}\in Z_{\mathcal{T}}$ which satisfies

$$b(z^*_{\mathcal{T}},\zeta_{\mathcal{T}}) = \int_\Omega f\zeta_{\mathcal{T}} + \int_{\Gamma_N} g\zeta_{\mathcal{T}} - \int_\Omega \nabla u_{\mathcal{T}}\cdot\nabla\zeta_{\mathcal{T}} \tag{II.2.9}$$

for all $\zeta_{\mathcal{T}}\in Z_{\mathcal{T}}$.

The Galerkin orthogonality and equation (II.2.5) imply

$$\begin{aligned}b(z^*_{\mathcal{T}},\zeta_{\mathcal{T}}) &= \int_\Omega \nabla(u-u_{\mathcal{T}})\cdot\nabla\zeta_{\mathcal{T}} \\ &= \int_\Omega \nabla z_{\mathcal{T}}\cdot\nabla\zeta_{\mathcal{T}}\end{aligned}$$

for all $\zeta_{\mathcal{T}} \in Z_{\mathcal{T}}$. Inserting $\zeta_{\mathcal{T}} = z_{\mathcal{T}}$ and $\zeta_{\mathcal{T}} = z^*_{\mathcal{T}}$ in this identity and using estimate (II.2.8) we infer that

$$\begin{aligned} b(z^*_{\mathcal{T}}, z^*_{\mathcal{T}}) &= \int_\Omega \nabla(u - u_{\mathcal{T}}) \cdot \nabla z^*_{\mathcal{T}} \\ &\leq \|\nabla(u - u_{\mathcal{T}})\| \|\nabla z^*_{\mathcal{T}}\| \\ &\leq \|\nabla(u - u_{\mathcal{T}})\| \frac{1}{\sqrt{\lambda}} b(z^*_{\mathcal{T}}, z^*_{\mathcal{T}})^{\frac{1}{2}} \end{aligned}$$

and

$$\begin{aligned} \|\nabla z_{\mathcal{T}}\|^2 &= b(z^*_{\mathcal{T}}, z_{\mathcal{T}}) \\ &\leq b(z^*_{\mathcal{T}}, z^*_{\mathcal{T}})^{\frac{1}{2}} b(z_{\mathcal{T}}, z_{\mathcal{T}})^{\frac{1}{2}} \\ &\leq b(z^*_{\mathcal{T}}, z^*_{\mathcal{T}})^{\frac{1}{2}} \sqrt{\Lambda} \|\nabla z_{\mathcal{T}}\|. \end{aligned}$$

This proves the two-sided error bound

$$\begin{aligned} \sqrt{\lambda} b(z^*_{\mathcal{T}}, z^*_{\mathcal{T}})^{\frac{1}{2}} &\leq \|\nabla(u - u_{\mathcal{T}})\| \\ &\leq \frac{\sqrt{\Lambda}}{(1-\beta)\sqrt{1-\gamma}} b(z^*_{\mathcal{T}}, z^*_{\mathcal{T}})^{\frac{1}{2}}. \end{aligned}$$

We may summarize the results of this section as follows:

Denote by $u \in H^1_D(\Omega)$ and $u_{\mathcal{T}} \in S^{1,0}_D(\mathcal{T})$ the unique solutions of problems (II.1.2) (p. 26) and (II.1.3) (p. 26), respectively. Assume that the space $Y_{\mathcal{T}} = S^{1,0}_D(\mathcal{T}) \oplus Z_{\mathcal{T}}$ satisfies the saturation assumption (II.2.2) and the strengthened Cauchy-Schwarz inequality (II.2.3) and admits a bilinear form b on $Z_{\mathcal{T}} \times Z_{\mathcal{T}}$ which has a diagonal stiffness matrix and which satisfies estimate (II.2.8). Denote by $z^*_{\mathcal{T}} \in Z_{\mathcal{T}}$ the unique solution of problem (II.2.9) and define the *hierarchical a posteriori error estimator* η_H by

$$\eta_H = b(z^*_{\mathcal{T}}, z^*_{\mathcal{T}})^{\frac{1}{2}}.$$

Then the a posteriori error estimates

$$\|\nabla(u - u_{\mathcal{T}})\| \leq \frac{\sqrt{\Lambda}}{(1-\beta)\sqrt{1-\gamma}} \eta_H$$

and

$$\eta_H \leq \frac{1}{\sqrt{\lambda}} \|\nabla(u - u_{\mathcal{T}})\|$$

are valid.

REMARK II.2.2. When considering families of partitions obtained by successive refinement, the constants β and γ in the saturation assumption and the strengthened Cauchy-Schwarz inequality should be

uniformly less than 1. Similarly, the quotient $\frac{\Lambda}{\lambda}$ should be uniformly bounded.

REMARK II.2.3. The bilinear form b can often be constructed as follows. The hierarchical complement $Z_{\mathcal{T}}$ can be chosen such that its elements vanish at the element vertices $\mathcal{E}$. Standard scaling arguments then imply that on $Z_{\mathcal{T}}$ the H^1-semi-norm $\|\nabla\cdot\|$ is equivalent to a scaled L^2-norm. Similarly, one can then prove that the mass-matrix corresponding to this norm is spectrally equivalent to a lumped mass-matrix. The lumping process in turn corresponds to a suitable numerical quadrature. The bilinear form b then is given by the inner-product corresponding to the weighted L^2-norm evaluated with the quadrature rule.

REMARK II.2.4. The strengthened Cauchy-Schwarz inequality, e.g., holds if $Y_{\mathcal{T}}$ consists of continuous piecewise quadratic or biquadratic functions. Often it can be established by transforming to the reference element and solving a small eigenvalue-problem there.

REMARK II.2.5. The saturation assumption (II.2.2) is used to establish the reliability of the error estimator η_H. One can prove that the reliability of η_H in turn implies the saturation assumption (II.2.2). If the space $Y_{\mathcal{T}}$ contains the functions w_K and w_E of Section II.1.8 (p. 31) one may repeat the proofs of estimates (II.1.4) (p. 31), (II.1.5) (p. 33), and (II.1.6) (p. 34) and obtains that – up to perturbation terms of the form $h_K\|f - f_K\|_K$ and $h_E^{\frac{1}{2}}\|g - g_E\|_E$ – the quantity $\|\nabla z_{\mathcal{T}}^*\|_{\omega_K}$ is bounded from below by $\eta_{R,K}$ for every element K. Together with the results of Section II.1.9 (p. 34) and inequality (II.2.9) this proves – up to the perturbation terms – the reliability of η_H without resorting to the saturation assumption. In fact, this result may be used to prove that the saturation assumption holds if the right-hand sides f and g of problem (II.1.1) (p. 26) are piecewise constant on $\mathcal{T}$ and $\mathcal{E}_{\Gamma_N}$, respectively.

II.2.3. Averaging techniques. To avoid unnecessary technical difficulties and to simplify the presentation, we consider in this section problem (II.1.1) (p. 26) with pure Dirichlet boundary conditions, i.e. $\Gamma_N = \emptyset$, and assume that the partition $\mathcal{T}$ exclusively consists of triangles.

The error estimator of this chapter is based on the following ideas. Denote by u and $u_{\mathcal{T}}$ the unique solutions of problems (II.1.2) (p. 26) and (II.1.3) (p. 26). Suppose that we dispose of an easily computable approximation $Gu_{\mathcal{T}}$ of $\nabla u_{\mathcal{T}}$ such that

$$\|\nabla u - Gu_{\mathcal{T}}\| \leq \beta\|\nabla u - \nabla u_{\mathcal{T}}\| \tag{II.2.10}$$

holds with a constant $0 \le \beta < 1$. We then have

$$\frac{1}{1+\beta}\|Gu_{\mathcal{T}} - \nabla u_{\mathcal{T}}\| \le \|\nabla u - \nabla u_{\mathcal{T}}\| \le \frac{1}{1-\beta}\|Gu_{\mathcal{T}} - \nabla u_{\mathcal{T}}\|$$

and may therefore choose $\|Gu_{\mathcal{T}} - \nabla u_{\mathcal{T}}\|$ as an error estimator. Since $\nabla u_{\mathcal{T}}$ is a piecewise constant vector-field we may hope that its L^2-projection onto the continuous, piecewise linear vector-fields satisfies inequality (II.2.10). The computation of this projection, however, is as expensive as the solution of problem (II.1.3) (p. 26). We therefore replace the L^2-scalar product by an approximation which leads to a more tractable auxiliary problem.

In order to make these ideas more precise, we denote by $W_{\mathcal{T}}$ the space of all piecewise linear vector-fields and set $V_{\mathcal{T}} = W_{\mathcal{T}} \cap C(\Omega, \mathbb{R}^2)$. Note that $\nabla X_{\mathcal{T}} \subset W_{\mathcal{T}}$. We define a mesh-dependent scalar product $(\cdot,\cdot)_{\mathcal{T}}$ on $W_{\mathcal{T}}$ by

$$(\mathbf{v}, \mathbf{w})_{\mathcal{T}} = \sum_{K \in \mathcal{T}} \frac{|K|}{3} \Big\{ \sum_{x \in \mathcal{N}_K} \mathbf{v}|_K(x) \cdot \mathbf{w}|_K(x) \Big\}.$$

Here, $|K|$ denotes the area of K and

$$\boldsymbol{\varphi}|_K(x) = \lim_{\substack{y \to x \\ y \in K}} \boldsymbol{\varphi}(y)$$

for all $\boldsymbol{\varphi} \in W_{\mathcal{T}}$, $K \in \mathcal{T}$, $x \in \mathcal{N}_K$.

Since the quadrature formula

$$\int_K \varphi \approx \frac{|K|}{3} \sum_{x \in \mathcal{N}_K} \varphi(x)$$

is exact for all linear functions, we have

$$(\mathbf{v}, \mathbf{w})_{\mathcal{T}} = \int_\Omega \mathbf{v} \cdot \mathbf{w} \tag{II.2.11}$$

if both arguments are elements of $W_{\mathcal{T}}$ and at least one of them is piecewise constant. Moreover, one easily checks that

$$\frac{1}{4}\|\mathbf{v}\|^2 \le (\mathbf{v}, \mathbf{v})_{\mathcal{T}} \le \|\mathbf{v}\|^2$$

for all $\mathbf{v} \in W_{\mathcal{T}}$ and

$$(\mathbf{v}, \mathbf{w})_{\mathcal{T}} = \frac{1}{3} \sum_{x \in \mathcal{E}} |\omega_x| \mathbf{v}(x) \cdot \mathbf{w}(x) \tag{II.2.12}$$

for all $\mathbf{v}, \mathbf{w} \in V_{\mathcal{T}}$.

Denote by $Gu_{\mathcal{T}} \in V_{\mathcal{T}}$ the $(\cdot,\cdot)_{\mathcal{T}}$-projection of $\nabla u_{\mathcal{T}}$ onto $V_{\mathcal{T}}$, i.e.,

$$(Gu_{\mathcal{T}}, \mathbf{v}_{\mathcal{T}})_{\mathcal{T}} = (\nabla u_{\mathcal{T}}, \mathbf{v}_{\mathcal{T}})_{\mathcal{T}}$$

for all $\mathbf{v}_\mathcal{T} \in V_\mathcal{T}$. Equations (II.2.11) and (II.2.12) imply that

$$Gu_\mathcal{T}(x) = \sum_{\substack{K\in\mathcal{T}\\ x\in\mathcal{N}_K}} \frac{|K|}{|\omega_x|}\nabla u_\mathcal{T}|_K$$

for all $x \in \mathcal{E}$. Thus, $Gu_\mathcal{T}$ may be computed by a local averaging of $\nabla u_\mathcal{T}$.

We finally set

$$\eta_{Z,K} = \|Gu_\mathcal{T} - \nabla u_\mathcal{T}\|_K$$

and

$$\eta_Z = \left\{\sum_{K\in\mathcal{T}} \eta_{Z,K}^2\right\}^{\frac{1}{2}}.$$

One can prove that η_Z yields upper and lower bounds for the error and that it is comparable to the residual error estimator $\eta_{R,K}$ of Section II.1.9 (p. 34).

II.2.4. $H(\mathrm{div})$-lifting. The basic idea is to construct a piece-wise linear vector field $\rho_\mathcal{T}$ such that

$$\begin{aligned} -\operatorname{div}\rho_\mathcal{T} &= f && \text{on every } K\in\mathcal{T}\\ \mathbb{J}_E(\mathbf{n}_E\cdot\rho_\mathcal{T}) &= -\mathbb{J}_E(\mathbf{n}_E\cdot\nabla u_\mathcal{T}) && \text{on every } E\in\mathcal{E}_\Omega\\ \mathbf{n}\cdot\rho_\mathcal{T} &= g - \mathbf{n}\cdot\nabla u_\mathcal{T} && \text{on every } E\in\mathcal{E}_{\Gamma_N}. \end{aligned} \tag{II.2.13}$$

Then the vector field $\rho = \rho_\mathcal{T} + \nabla u_\mathcal{T}$ is contained in $H(\mathrm{div};\Omega)$ and satisfies

$$\begin{aligned} -\operatorname{div}\rho &= f \quad \text{in } \Omega\\ \rho\cdot\mathbf{n} &= g \quad \text{on } \Gamma_N. \end{aligned} \tag{II.2.14}$$

since $\Delta u_\mathcal{T}$ vanishes element-wise.

To simplify the presentation we assume for the rest of this section that

- $\mathcal{T}$ exclusively consists of triangles,
- f is piece-wise constant,
- g is piece-wise constant.

Parallelograms could be treated by changing the definition (II.2.15) of the vector fields $\gamma_{K,E}$. General functions f and g introduce additional data errors.

For every triangle K and every edge E thereof we denote by $a_{K,E}$ the vertex of K which is not contained in E and set

$$\gamma_{K,E}(x) = \frac{\mu_1(E)}{2\mu_2(K)}(x - a_{K,E}), \tag{II.2.15}$$

where $\mu_1(E)$ is the length of E and $\mu_2(K)$ the area of K. The vector fields $\gamma_{K,E}$ are the shape functions of the lowest order Raviart-Thomas space and have the following properties

$$\begin{aligned} \operatorname{div} \gamma_{K,E} &= \frac{\mu_1(E)}{\mu_2(K)} && \text{on } K, \\ \mathbf{n}_K \cdot \gamma_{K,E} &= 0 && \text{on } \partial K \setminus E, \\ \mathbf{n}_K \cdot \gamma_{K,E} &= 1 && \text{on } E, \\ \|\gamma_{K,E}\|_K &\leq c h_K, \end{aligned} \tag{II.2.16}$$

where $\mathbf{n}_K$ denotes the unit exterior normal of K and where the constant c only depends on the shape parameter of $\mathcal{T}$.

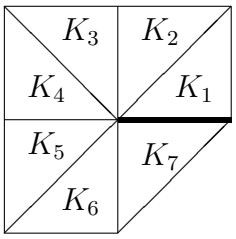

FIGURE II.2.1. Enumeration of elements in ω_z

Now, we consider an arbitrary interior vertex $z \in \mathcal{N}_\Omega$. We enumerate the triangles in ω_z from 1 to n and the edges emanating from z from 0 to n such that (cf. Figure II.2.1)

- $E_0 = E_n$,
- E_{i-1} and E_i are edges of K_i for every i.

We define

$$\alpha_0 = 0$$

and recursively for $i = 1, \dots, n$

$$\alpha_i = -\frac{\mu_2(K_i)}{3\mu_1(E_i)} f + \frac{\mu_1(E_{i-1})}{2\mu_1(E_i)} \mathbb{J}_{E_{i-1}}(\mathbf{n}_{E_{i-1}} \cdot \nabla u_\mathcal{T}) + \frac{\mu_1(E_{i-1})}{\mu_1(E_i)} \alpha_{i-1}.$$

By induction we obtain

$$\mu_1(E_n)\alpha_n = -\sum_{i=1}^{n} \frac{\mu_2(K_i)}{3} f + \sum_{j=0}^{n-1} \frac{\mu_1(E_j)}{2} \mathbb{J}_{E_j}(\mathbf{n}_{E_j} \cdot \nabla u_\mathcal{T}).$$

Since

$$\int_{K_i} \lambda_z = \frac{\mu_2(K_i)}{3}$$

for every $i \in \{1, \dots, n\}$ and since

$$\int_{E_j} \lambda_z = \frac{\mu_1(E_j)}{2}$$

for every $j \in \{0, \dots, n-1\}$, we conclude – using the assumption that f and g are piece-wise constant – that

$$-\sum_{i=1}^{n} \frac{\mu_2(K_i)}{3} f + \sum_{j=0}^{n-1} \frac{\mu_1(E_j)}{2} \mathbb{J}_{E_j}(\mathbf{n}_{E_j} \cdot \nabla u_{\mathcal{T}}) = -\int_\Omega r\lambda_z - \int_\Sigma j\lambda_z = 0.$$

Hence we have $\alpha_n = 0$. Therefore we can define a vector field ρ_z by setting for every $i \in \{1, \dots, n\}$

$$\rho_z|_{K_i} = \alpha_i \gamma_{K_i,E_i} - \Big(\mathbb{J}_{E_{i-1}}(\mathbf{n}_{E_{i-1}} \cdot \nabla u_{\mathcal{T}}) + \alpha_{i-1}\Big)\gamma_{K_i,E_{i-1}}. \tag{II.2.17}$$

Equations (II.2.16) and the definition of the α_i imply that

$$\begin{aligned} -\operatorname{div} \rho_z &= \frac{1}{3} f && \text{on } K_i \\ \mathbb{J}_{E_i}(\rho_z \cdot \mathbf{n}_{E_i}) &= -\frac{1}{2} \mathbb{J}_E(\mathbf{n}_{E_i} \cdot \nabla u_{\mathcal{T}}) && \text{on } E_i \end{aligned} \tag{II.2.18}$$

holds for every $i \in \{1, \dots, n\}$.

For a vertex on the boundary Γ, the construction of ρ_z must be modified as follows:

- For every edge on the Neumann boundary Γ_N we must replace $-\mathbb{J}_E(\mathbf{n}_E \cdot \nabla u_{\mathcal{T}})$ by $g - \mathbf{n} \cdot \nabla u_{\mathcal{T}}$.
- If z is a vertex on the Dirichlet boundary, there is at least one edge emanating from z which is contained in Γ_D. We must choose the enumeration of the edges such that E_n is one of these edges.

With these modifications, equations (II.2.17) and (II.2.18) carry over although in general $\alpha_n \neq 0$ for vertices on the boundary Γ.

In a final step, we extend the vector fields ρ_z by zero outside ω_z and set

$$\rho_{\mathcal{T}} = \sum_{z \in \mathcal{N}} \rho_z. \tag{II.2.19}$$

Since every triangle has three vertices and every edge has two vertices, we conclude from equations (II.2.18) that $\rho_{\mathcal{T}}$ has the desired properties (II.2.13).

The last inequality in (II.2.16), the definition of the α_i, and the observation that $\Delta u_{\mathcal{T}}$ vanishes element-wise imply that

$$\begin{aligned} \|\rho_z\|_{\omega_z} \leq c\Big\{ &\sum_{K \subset \omega_z} h_K^2 \|f + \Delta u_{\mathcal{T}}\|_K^2 \\ &+ \sum_{E \subset \sigma_z \cap \Omega} h_E \|\mathbb{J}_E(\mathbf{n}_E \cdot \nabla u_{\mathcal{T}})\|_E^2 \\ &+ \sum_{E \subset \sigma_z \cap \Gamma_N} h_E \|g - \mathbf{n}_E \cdot \nabla u_{\mathcal{T}}\|_E^2 \Big\}^{\frac{1}{2}} \end{aligned}$$

holds for every vertex $z \in \mathcal{N}$ with a constant which only depends on the shape parameter of $\mathcal{T}$.

Combining these results, we arrive at the following a posteriori error estimates:

$$\|\nabla(u - u_{\mathcal{T}})\| \leq \|\rho_{\mathcal{T}}\|$$
$$\|\rho_{\mathcal{T}}\| \leq c_* \|\nabla(u - u_{\mathcal{T}})\|$$

II.2.5. Asymptotic exactness. The quality of an a posteriori error estimator is often measured by its *efficiency index*, i.e., the ratio of the estimated error and of the true error. An error estimator is called *efficient* if its efficiency index together with its inverse remain bounded for all mesh-sizes. It is called *asymptotically exact* if its efficiency index tends to one when the mesh-size converges to zero.

In the generic case we have

$$\Big\{ \sum_{K \in \mathcal{T}} h_K^2 \|f - f_K\|_K^2 \Big\}^{\frac{1}{2}} = o(h)$$

and

$$\Big\{ \sum_{E \in \mathcal{E}_{\Gamma_N}} h_E \|g - g_E\|_E^2 \Big\}^{\frac{1}{2}} = o(h),$$

where

$$h = \max_{K \in \mathcal{T}} h_K$$

denotes the maximal mesh-size. On the other hand, the solutions of problems (II.1.2) (p. 26) and (II.1.3) (p. 26) satisfy

$$\|u - u_{\mathcal{T}}\|_{H^1(\Omega)} \geq ch$$

always but in trivial cases. Hence, the results of Sections II.1.9 (p. 34), II.2.1.1 (p. 37), II.2.1.2 (p. 39), II.2.1.3 (p. 40), II.2.3 (p. 47), and II.2.3 (p. 47) imply that the corresponding error estimators are efficient. Their efficiency indices can in principle be estimated explicitly since the constants in the above sections only depend on the constants in the quasi-interpolation error estimate of Section I.2.11 (p. 21) and the inverse inequalities of Section I.2.12 (p. 22) for which sharp bounds can be derived.

Using super-convergence results one can also prove that on special meshes the error estimators of Sections II.1.9 (p. 34), II.2.1.1 (p. 37), II.2.1.2 (p. 39), II.2.1.3 (p. 40), II.2.3 (p. 47), and II.2.3 (p. 47) are asymptotically exact.

The following example shows that asymptotic exactness may not hold on general meshes even if they are strongly structured.

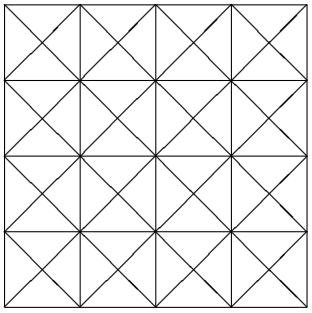

FIGURE II.2.2. Triangulation of Example II.2.6 corresponding to $n = 4$

EXAMPLE II.2.6. Consider problem (II.1.1) (p. 26) on the unit square

$$\Omega = (0,1)^2$$

with

$$\begin{aligned} \Gamma_N &= (0,1) \times \{0\} \cup (0,1) \times \{1\}, \\ g &= 0, \end{aligned}$$

and

$$f = 1.$$

The exact solution is

$$u(x,y) = \frac{1}{2} x(1-x).$$

The triangulation $\mathcal{T}$ is obtained as follows (cf. Figure II.2.2): Ω is divided into n^2 squares with sides of length $h = \frac{1}{n}$, $n \in \mathbb{N}^*$; each square is cut into four triangles by drawing the two diagonals. This triangulation is often called a *criss-cross grid*. Since the solution u of problem (II.1.1) (p. 26) is quadratic and the Neumann boundary conditions are homogeneous, one easily checks that the solution $u_{\mathcal{T}}$ of problem (II.1.3) (p. 26) is given by

$$u_{\mathcal{T}}(x) = \begin{cases} u(x) & \text{if } x \text{ is a vertex of a square,} \\ u(x) - \frac{h^2}{24} & \text{if } x \text{ is a midpoint of a square.} \end{cases}$$

Using this expression for $u_{\mathcal{T}}$ one can explicitly calculate the error and the error estimator. After some computations one obtains for any square Q, which is disjoint from Γ_N,

$$\Big\{ \sum_{\substack{K \in \mathcal{T} \\ K \subset Q}} \eta_{N,K}^2 \Big\}^{\frac{1}{2}} \Big/ \|\nabla e\|_Q = \sqrt{\frac{17}{6}} \approx 1.68.$$

Hence, the error estimator cannot be asymptotically exact.

II.2.6. Convergence. Assume that we dispose of an error estimator η_K which yields global upper and local lower bounds for the error of the solution of problem (II.1.2) (p. 26) and its finite element discretization (II.1.3) (p. 26) and that we apply the general adaptive algorithm I.1.1 (p. 7) with one of the refinement strategies of Algorithms III.1.1 (p. 89) and III.1.2 (p. 90). Then one can prove that the error decreases linearly. More precisely: If u denotes the solution of problem (II.1.2) and if u_i denotes the solution of the discrete problem (II.1.3) corresponding to the i-th partition $\mathcal{T}_i$, then there is a constant $0 < \beta < 1$, which only depends on the constants c^* and c_* in the error bounds, such that

$$\boxed{\|\nabla(u - u_i)\| \leq \beta^i \|\nabla(u - u_0)\|.}$$

II.3. Elliptic problems

II.3.1. Scalar linear elliptic equations. In this section we consider scalar linear elliptic partial differential equations in their general form

$$\boxed{\begin{aligned} -\operatorname{div}(A\nabla u) + \mathbf{a}\cdot\nabla u + \alpha u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma_D \\ \mathbf{n}\cdot A\nabla u &= g \quad \text{on } \Gamma_N \end{aligned}}$$

where the diffusion $A(x)$ is for every $x \in \Omega$ a symmetric positive definite matrix. We assume that the data satisfy the following conditions:

- The diffusion A is continuously differentiable and uniformly elliptic and uniformly isotropic, i.e.,
$$\varepsilon = \inf_{x\in\Omega} \min_{z\in\mathbb{R}^d\setminus\{0\}} \frac{z^t A(x) z}{z^t z} > 0$$
and
$$\kappa = \varepsilon^{-1} \sup_{x\in\Omega} \max_{z\in\mathbb{R}^d\setminus\{0\}} \frac{z^t A(x) z}{z^t z}$$
is of moderate size.
- The convection $\mathbf{a}$ is a continuously differentiable vector field and scaled such that
$$\sup_{x\in\Omega} |\mathbf{a}(x)| \leq 1.$$
- The reaction α is a continuous non-negative scalar function.
- There is a constant $\beta \geq 0$ such that
$$\alpha - \frac{1}{2}\operatorname{div}\mathbf{a} \geq \beta$$

for all $x \in \Omega$. Moreover there is a constant $c_b \geq 0$ of moderate size such that

$$\sup_{x\in\Omega} \alpha(x) \leq c_b \beta.$$

- The Dirichlet boundary Γ_D has positive $(d-1)$-dimensional measure and includes the inflow boundary $\{x \in \Gamma : \mathbf{a}(x) \cdot \mathbf{n}(x) < 0\}$.

With these assumptions we can distinguish different regimes:

- *dominant diffusion*: $\sup_{x\in\Omega}|\mathbf{a}(x)| \leq c_c\varepsilon$ and $\beta \leq c_b'\varepsilon$ with constants of moderate size;
- *dominant reaction*: $\sup_{x\in\Omega}|\mathbf{a}(x)| \leq c_c\varepsilon$ and $\beta \gg \varepsilon$ with a constant c_c of moderate size;
- *dominant convection*: $\beta \gg \varepsilon$.

II.3.1.1. *Variational formulation.* The variational formulation of the above differential equation is given by:

Find $u \in H^1_D(\Omega)$ such that

$$\int_\Omega \{\nabla u \cdot A\nabla v + \mathbf{a}\cdot\nabla u v + \alpha u v\} = \int_\Omega fv + \int_{\Gamma_N} gv$$

holds for all $v \in H^1_D(\Omega)$.

The above assumptions on the differential equation imply that this variational problem admits a unique solution and that the corresponding natural energy norm is given by

$$|||v||| = \Big\{\varepsilon\|\nabla v\|^2 + \beta\|v\|^2\Big\}^{\frac{1}{2}}.$$

The corresponding dual norm is denoted by $|||\cdot|||_*$ and is given by

$$|||w|||_* = \sup_{v\in H^1_D(\Omega)\setminus\{0\}} \frac{1}{|||v|||}\int_\Omega \{\varepsilon\nabla v\cdot\nabla w + \beta v w\}.$$

II.3.1.2. *Finite element discretization.* The finite element discretization of the above differential equation is given by:

Find $u_{\mathcal{T}} \in S^{k,0}_D(\mathcal{T})$ such that

$$\begin{aligned}\int_\Omega &\{\nabla u_{\mathcal{T}} \cdot A\nabla v_{\mathcal{T}} + \mathbf{a}\cdot\nabla u_{\mathcal{T}} v_{\mathcal{T}} + \alpha u_{\mathcal{T}} v_{\mathcal{T}}\} \\ &\quad + \sum_{K\in\mathcal{T}} \delta_K \int_K \{-\operatorname{div}(A\nabla u_{\mathcal{T}}) + \mathbf{a}\cdot\nabla u_{\mathcal{T}} + \alpha u_{\mathcal{T}}\}\mathbf{a}\cdot\nabla v_{\mathcal{T}} \\ &= \int_\Omega f v_{\mathcal{T}} + \int_{\Gamma_N} g v_{\mathcal{T}}\end{aligned}$$

$$+ \sum_{K \in \mathcal{T}} \delta_K \int_K f \mathbf{a} \cdot \nabla v_{\mathcal{T}}$$

holds for all $v_{\mathcal{T}} \in S_D^{k,0}(\mathcal{T})$.

The δ_K are non-negative stabilization parameters. The case $\delta_K = 0$ for all elements K corresponds to the standard finite element scheme. This choice is appropriate for the diffusion dominated and reaction dominated regimes. In the case of a dominant convection, however, the δ_K should be chosen strictly positive in order to stabilize the discretization. In this case the discretization is often referred to as *streamline upwind Petrov-Galerkin discretization* or in short *SUPG discretization*.

With an appropriate choice of the stabilization parameters δ_K one can prove that the above discrete problem admits a unique solution for all regimes described above.

II.3.1.3. *Residual error estimates.* Denote by u and $u_{\mathcal{T}}$ the solutions of the variational problem and of its discretization. As in Section II.1.6 (p. 28) we define element and edge or face residuals by

$$R_K(u_{\mathcal{T}}) = f_K + \operatorname{div}(A \nabla u_{\mathcal{T}}) - \mathbf{a} \cdot \nabla u_{\mathcal{T}} - \alpha u_{\mathcal{T}}$$

$$R_E(u_{\mathcal{T}}) = \begin{cases} -\mathbb{J}_E(\mathbf{n}_E \cdot A \nabla u_{\mathcal{T}}) & \text{if } E \in \mathcal{E}_\Omega, \\ g - \mathbf{n}_E \cdot A \nabla u_{\mathcal{T}} & \text{if } E \in \mathcal{E}_{\Gamma_N}, \\ 0 & \text{if } E \in \mathcal{E}_{\Gamma_D}. \end{cases}$$

Here, as in Section I.2.8 (p. 17), f_K and g_E denote the average of f on K and the average of g on E, respectively. The residual error estimator is then given by

$$\eta_{R,K} = \left\{ \alpha_K^2 \|R_K(u_{\mathcal{T}})\|_K^2 + \sum_{E \in \mathcal{E}_K} \varepsilon^{-\frac{1}{2}} \alpha_E \|R_E(u_{\mathcal{T}})\|_E^2 \right\}^{\frac{1}{2}}$$

with

$$\alpha_S = \min\{\varepsilon^{-\frac{1}{2}} h_S, \beta^{-\frac{1}{2}}\} \quad \text{for } S \in \mathcal{T} \cup \mathcal{E}_{\mathcal{T}}.$$

One can prove that $\eta_{R,K}$ yields global upper and lower bounds for the error measured in the norm $|||e||| + |||\mathbf{a} \cdot \nabla e|||_*$. In the case of dominant diffusion or dominant reaction, the dual norm $|||\cdot|||_*$ can be dropped. In this case, also lower error bounds can be established.

II.3.1.4. *Other error estimators.* The error estimators of Sections II.2.1.1 (p. 37), II.2.1.2 (p. 39) and II.2.1.3 (p. 40) which are based

on the solution of auxiliary local discrete problems can easily be extended to the present situation. One only has to replace the differential operator $u \mapsto -\Delta u$ by the the actual differential operator $u \mapsto -\operatorname{div}(A\nabla u) + \mathbf{a} \cdot \nabla u + \alpha u$ and to use the above definition of the element and edge or face residuals.

In the cases of dominant diffusion or of dominant reaction, the same remark applies to the hierarchical estimator of Section II.2.2 (p. 42). It has difficulties in the case of dominant convection, due to the lacking symmetry of the bilinear form associated with the variational problem.

In the case of dominant diffusion, the averaging technique of Section II.2.3 (p. 47) can easily be extended to the present situation. One only has to replace the gradient ∇u by the oblique derivative $A\nabla u$. In the cases of dominant reaction or of dominant convection, however, the averaging technique is not appropriate since it is based on the diffusive part of the differential operator which is no longer dominant.

II.3.2. Mixed formulation of the Poisson equation. In this section we once again consider the model problem. But now we impose pure homogeneous Dirichlet boundary conditions and – most important – write the problem as a first order system by introducing ∇u as an additional unknown:

$$\boxed{\begin{aligned} \operatorname{div}\sigma &= -f \quad \text{in } \Omega \\ \sigma &= \nabla u \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma. \end{aligned}}$$

Our interest in this problem is twofold:

- Its finite element discretization introduced below allows the direct approximation of ∇u without resorting to a differentiation of the finite element approximation $u_{\mathcal{T}}$ considered so far.
- Its analysis prepares the a posteriori error analysis of the equations of linear elasticity considered in the next two sections where mixed methods are mandatory to avoid locking phenomena.

For the variational formulation, we introduce the space

$$H(\operatorname{div};\Omega) = \left\{\sigma \in L^2(\Omega)^d : \operatorname{div}\sigma \in L^2(\Omega)\right\}$$

and its norm

$$\|\sigma\|_H = \left\{\|\sigma\|^2 + \|\operatorname{div}\sigma\|^2\right\}^{\frac{1}{2}}.$$

Next, we multiply the first equation of the above differential equation by a function $v \in L^2(\Omega)$ and the second equation by a vector field

$\tau \in H(\mathrm{div};\Omega)$, integrate both expressions over Ω and use integration by parts for the integral involving ∇u. We thus arrive at the problem:

Find $\sigma \in H(\mathrm{div};\Omega)$ and $u \in L^2(\Omega)$ such that

$$\int_\Omega \sigma \cdot \tau + \int_\Omega u \,\mathrm{div}\,\tau = 0$$
$$\int_\Omega \mathrm{div}\,\sigma v = -\int_\Omega f v$$

holds for all $\tau \in H(\mathrm{div};\Omega)$ and $v \in L^2(\Omega)$.

The differential equation and its variational formulation are equivalent in the usual weak sense: Every classical solution of the differential equation is a solution of the variational problem and every solution of the variational problem which is sufficiently regular is a classical solution of the differential equation.

To keep the exposition as simple as possible, we only consider the simplest discretization which is given by the lowest order *Raviart-Thomas spaces*. For every element $K \in \mathcal{T}$ we set

$$\mathrm{RT}_0(K) = R_0(K)^d + R_0(K) \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

and

$$\mathrm{RT}_0(\mathcal{T}) = \Big\{ \sigma_\mathcal{T} : \sigma_{\mathcal{T}\big|_K} \in \mathrm{RT}_0(K) \text{ for all } K \in \mathcal{T},$$
$$\int_E \mathbb{J}_E(\mathbf{n}_E \cdot \sigma_\mathcal{T}) = 0 \text{ for all } E \in \mathcal{E}_\Omega \Big\}.$$

The degrees of freedom associated with $\mathrm{RT}_0(\mathcal{T})$ are the values of the normal components of the $\sigma_\mathcal{T}$ evaluated at the midpoints of edges, if $d = 2$, or the barycentres of faces, if $d = 3$. Since the normal components $\mathbf{n}_E \cdot \sigma_\mathcal{T}$ are constant on the edges respective faces, the condition

$$\int_E \mathbb{J}_E(\mathbf{n}_E \cdot \sigma_\mathcal{T}) = 0 \quad \text{for all } E \in \mathcal{E}_\Omega$$

ensures that the space $\mathrm{RT}_0(\mathcal{T})$ is contained in $H(\mathrm{div};\Omega)$. The *mixed finite element approximation* of the model problem is then given by:

Find $\sigma_\mathcal{T} \in \mathrm{RT}_0(\mathcal{T})$ and $u_\mathcal{T} \in S^{0,-1}(\mathcal{T})$ such that

$$\int_\Omega \sigma_\mathcal{T} \cdot \tau_\mathcal{T} + \int_\Omega u_\mathcal{T} \,\mathrm{div}\,\tau_\mathcal{T} = 0$$
$$\int_\Omega \mathrm{div}\,\sigma_\mathcal{T} v_\mathcal{T} = -\int_\Omega f v_\mathcal{T}$$

holds for all $\tau_{\mathcal{T}} \in \mathrm{RT}_0(\mathcal{T})$ and $v_{\mathcal{T}} \in S^{0,-1}(\mathcal{T})$.

The a posteriori error analysis of the mixed formulation of the model problem relies on the *Helmholtz decomposition* of vector fields. For its description we define the so-called *curl operator* curl by

$$\begin{aligned}
\operatorname{curl}\tau &= \nabla \times \tau \\
&= \begin{pmatrix} \dfrac{\partial \tau_3}{\partial x_2} - \dfrac{\partial \tau_2}{\partial x_3} \\ \dfrac{\partial \tau_1}{\partial x_3} - \dfrac{\partial \tau_3}{\partial x_1} \\ \dfrac{\partial \tau_2}{\partial x_1} - \dfrac{\partial \tau_1}{\partial x_2} \end{pmatrix} && \text{if } \tau : \Omega \to \mathbb{R}^3, \\
\operatorname{curl}\tau &= \frac{\partial \tau_2}{\partial x_1} - \frac{\partial \tau_1}{\partial x_2} && \text{if } \tau : \Omega \to \mathbb{R}^2, \\
\operatorname{curl} v &= \begin{pmatrix} \dfrac{\partial v}{\partial x_2} \\ -\dfrac{\partial v}{\partial x_1} \end{pmatrix} && \text{if } v : \Omega \to \mathbb{R} \text{ and } d = 2.
\end{aligned}$$

The Helmholtz decomposition then states that every vector field can be split into a gradient and a rotational component. More precisely, there are two continuous linear operators R and G such that every vector field τ can be split in the form

$$\tau = \nabla(G\tau) + \operatorname{curl}(R\tau).$$

Using the Helmholtz decomposition one can prove the following *a posteriori error estimate*:

$$\begin{aligned}
&\Big\{\|\sigma - \sigma_{\mathcal{T}}\|_H^2 + \|u - u_{\mathcal{T}}\|^2\Big\}^{\frac{1}{2}} \\
&\qquad \le c^* \Big\{\sum_{K \in \mathcal{T}} \eta_{R,K}^2\Big\}^{\frac{1}{2}} \\
&\eta_{R,K} \le c_* \Big\{\|\sigma - \sigma_{\mathcal{T}}\|_{H(\operatorname{div};\omega_K)}^2 + \|u - u_{\mathcal{T}}\|_{\omega_K}^2\Big\}^{\frac{1}{2}}
\end{aligned}$$

with

$$\begin{aligned}
\eta_{R,K} = \Big\{ & h_K^2 \|\operatorname{curl}\sigma_{\mathcal{T}}\|_K^2 + h_K^2 \|\sigma_{\mathcal{T}}\|_K^2 \\
& + \|f + \operatorname{div}\sigma_{\mathcal{T}}\|_K^2
\end{aligned}$$

$$+\frac{1}{2}\sum_{E\in\mathcal{E}_K\cap\mathcal{E}_\Omega} h_E\|\mathbb{J}_E(\sigma_\mathcal{T}-(\sigma_\mathcal{T}\cdot\mathbf{n}_E)\mathbf{n}_E)\|_E^2$$
$$+\sum_{E\in\mathcal{E}_K\cap\mathcal{E}_\Gamma} h_E\|\sigma_\mathcal{T}-(\sigma_\mathcal{T}\cdot\mathbf{n}_E)\mathbf{n}_E\|_E^2\Big\}^{\frac{1}{2}}.$$

REMARK II.3.1. The terms

$$\begin{aligned}\|\operatorname{curl}\sigma_\mathcal{T}\|_K &\quad\text{with } K\in\mathcal{T},\\ \|\mathbb{J}_E(\sigma_\mathcal{T}-(\sigma_\mathcal{T}\cdot\mathbf{n}_E)\mathbf{n}_E)\|_E &\quad\text{with } E\in\mathcal{E}_\Omega,\\ \|\sigma_\mathcal{T}-(\sigma_\mathcal{T}\cdot\mathbf{n}_E)\mathbf{n}_E\|_E &\quad\text{with } E\in\mathcal{E}_\Gamma\end{aligned}$$

in $\eta_{R,K}$ are the residuals of $\sigma_\mathcal{T}$ corresponding to the equation $\operatorname{curl}\sigma=0$. Due to the condition $\sigma=\nabla u$, this equation is a redundant one for the analytical problem. For the discrete problem, however, it is an extra condition which is not incorporated.

II.3.3. Displacement form of the equations of linearized elasticity. The equations of linearized elasticity are given by the boundary value problem

$$\begin{aligned}\varepsilon &= D\mathbf{u} && \text{in } \Omega\\ \varepsilon &= C^{-1}\sigma && \text{in } \Omega\\ -\operatorname{div}\sigma &= \mathbf{f} && \text{in } \Omega\\ \operatorname{as}\sigma &= 0 && \text{in } \Omega\\ \mathbf{u} &= 0 && \text{on } \Gamma_D\\ \sigma\cdot\mathbf{n} &= 0 && \text{on } \Gamma_N\end{aligned}$$

where the various quantities are

- $\mathbf{u}:\Omega\to\mathbb{R}^d$ the *displacement*,
- $D\mathbf{u}=\frac{1}{2}(\nabla\mathbf{u}+\nabla\mathbf{u}^t)=\frac{1}{2}(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i})_{1\le i,j\le d}$ the *deformation tensor* or *symmetric gradient*,
- $\varepsilon:\Omega\to\mathbb{R}^{d\times d}$ the *strain tensor*,
- $\sigma:\Omega\to\mathbb{R}^{d\times d}$ the *stress tensor*,
- C the *elasticity tensor*,
- $\mathbf{f}:\Omega\to\mathbb{R}^d$ the given *body load*, and
- $\operatorname{as}\tau=\tau-\tau^t$ the *skew symmetric part* of a given tensor.

The most important example of an elasticity tensor is given by

$$C\varepsilon=\lambda\operatorname{tr}(\varepsilon)I+2\mu\varepsilon$$

where $I\in\mathbb{R}^{d\times d}$ is the unit tensor, $\operatorname{tr}(\varepsilon)$ denotes the trace of ε, and λ, $\mu>0$ are the *Lamé parameters*. To simplify the presentation we

assume throughout this section that C takes the above form. We are mainly interested in estimates which are uniform with respect to the Lamé parameters.

II.3.3.1. *Displacement formulation.* The simplest discretization of the equations of linearized elasticity is based on its *displacement formulation*:

$$\begin{aligned} -\operatorname{div}(CD\mathbf{u}) &= \mathbf{f} \quad \text{in } \Omega \\ \mathbf{u} &= 0 \quad \text{on } \Gamma_D \\ \mathbf{n} \cdot CD\mathbf{u} &= 0 \quad \text{on } \Gamma_N. \end{aligned}$$

The corresponding variational problem is given by:

Find $\mathbf{u} \in H^1_D(\Omega)^d$ such that

$$\int_\Omega D\mathbf{u} : CD\mathbf{v} = \int_\Omega \mathbf{f} \cdot \mathbf{v}$$

holds for all $\mathbf{v} \in H^1_D(\Omega)^d$.

Here, $\sigma : \tau$ denotes the inner product of two tensors, i.e.,

$$\sigma : \tau = \sum_{1 \le i,j \le d} \sigma_{ij} \tau_{ij}.$$

The variational problem is the Euler-Lagrange equation corresponding to the problem of minimizing the *total energy*

$$J(\mathbf{u}) = \frac{1}{2} \int_\Omega D\mathbf{u} : CD\mathbf{u} - \int_\Omega \mathbf{f} \cdot \mathbf{u}.$$

II.3.3.2. *Finite element discretization.* The finite element discretization of the displacement formulation of the equations of linearized elasticity is given by:

Find $\mathbf{u}_\mathcal{T} \in S^{k,0}_D(\mathcal{T})^d$ such that

$$\int_\Omega D\mathbf{u}_\mathcal{T} : CD\mathbf{v}_\mathcal{T} = \int_\Omega \mathbf{f} \cdot \mathbf{v}_\mathcal{T}$$

holds for all $\mathbf{v}_\mathcal{T} \in S^{k,0}_D(\mathcal{T})^d$.

It is well known that this problem admits a unique solution.

II.3.3.3. *Residual error estimates.* The methods of Sections II.1 (p. 26) and II.3.1 (p. 54) directly carry over to this situation. They yield the residual a posteriori error estimator

$$\eta_{R,K} = \Big\{ h_K^2 \|f_{\mathcal{T}} + \operatorname{div}(CD\mathbf{u}_{\mathcal{T}})\|_K^2 + \frac{1}{2} \sum_{E \in \mathcal{E}_K \cap \mathcal{E}_\Omega} h_E \|\mathbb{J}_E(\mathbf{n}_E \cdot CD\mathbf{u}_{\mathcal{T}})\|_E^2 + \sum_{E \in \mathcal{E}_K \cap \mathcal{E}_{\Gamma_N}} h_E \|\mathbf{n}_E \cdot CD\mathbf{u}_{\mathcal{T}}\|_E^2 \Big\}^{\frac{1}{2}}$$

which gives global upper and local lower bounds on the H^1-norm of the error in the displacements.

II.3.3.4. *Other error estimators.* Similarly the methods of Sections II.2.1.1 (p. 37), II.2.1.2 (p. 39), II.2.1.3 (p. 40), II.2.2 (p. 42), and II.2.3 (p. 47) can be extended to the displacement formulation of the equations of linearized elasticity. Now, of course, the auxiliary problems are elasticity problems in displacement form and terms of the form $\nabla u_{\mathcal{T}}$ must be replaced by the stress tensor $\sigma(\mathbf{u}_{\mathcal{T}})$.

II.3.4. Mixed formulation of the equations of linearized elasticity. Though appealing by its simplicity the displacement formulation of the equations of linearized elasticity and the corresponding finite element discretizations suffer from serious drawbacks:

- The displacement formulation and its discretization break down for nearly incompressible materials which is reflected by the so-called locking phenomenon.
- The quality of the a posteriori error estimates deteriorates in the incompressible limit. More precisely, the constants in the upper and lower error bounds depend on the Lamé parameter λ and tend to infinity for large values of λ.
- Often, the displacement field is not of primary interest but the stress tensor is of physical interest. This quantity, however, is not directly discretized in displacement methods and must a posteriori be extracted from the displacement field which often leads to unsatisfactory results.

These drawbacks can be overcome by suitable mixed formulations of the equations of linearized elasticity and appropriate mixed finite element discretizations. Correspondingly these are in the focus of the following subsections. We are primarily interested in discretizations and a posteriori error estimates which are robust in the sense that their quality does not deteriorate in the incompressible limit. This is reflected by the need for estimates which are uniform with respect to the Lamé parameter λ.

II.3.4.1. *The Hellinger-Reissner principle.* To simplify the notation we introduce the spaces

$$\begin{aligned} H &= H(\operatorname{div};\Omega)^d \\ &= \{\sigma \in L^2(\Omega)^{d\times d} \mid \operatorname{div}\sigma \in L^2(\Omega)^d\}, \\ V &= L^2(\Omega)^d, \\ W &= \{\gamma \in L^2(\Omega)^{d\times d} \mid \gamma + \gamma^t = 0\} \end{aligned}$$

and equip them with their natural norms

$$\begin{aligned} \|\sigma\|_H &= \left\{\|\sigma\|^2 + \|\operatorname{div}\sigma\|^2\right\}^{\frac{1}{2}}, \\ \|\mathbf{u}\|_V &= \|\mathbf{u}\|, \\ \|\gamma\|_W &= \|\gamma\|. \end{aligned}$$

Here, the divergence of a tensor σ is taken row by row, i.e.,

$$(\operatorname{div}\sigma)_i = \sum_{1\le j\le n} \frac{\partial \sigma_{ij}}{\partial x_j}.$$

The *Hellinger-Reissner principle* is a mixed variational formulation of the equations of linearized elasticity, in which the strain ε is eliminated. It is given by:

Find $\sigma \in H$, $\mathbf{u} \in V$, $\gamma \in W$ such that

$$\begin{aligned} \int_\Omega C^{-1}\sigma : \tau + \int_\Omega \operatorname{div}\tau \cdot \mathbf{u} + \int_\Omega \tau : \gamma &= 0 \\ \int_\Omega \operatorname{div}\sigma \cdot \mathbf{v} &= -\int_\Omega \mathbf{f}\cdot\mathbf{v} \\ \int_\Omega \sigma : \eta &= 0 \end{aligned}$$

holds for all $\tau \in H$, $\mathbf{v} \in V$, $\eta \in W$.

It can be proven that the bilinear form corresponding to the left hand-sides of the above problem is uniformly continuous and coercive with respect to the Lamé parameter λ. Thanks to this stability result, all forthcoming constants are independent of λ. Hence, the corresponding estimates are robust for nearly incompressible materials.

II.3.4.2. *PEERS and BDMS elements.* We consider two types of mixed finite element discretizations of the Hellinger-Reissner principle:

- the *PEERS element* and
- the *BDMS elements*.

Both families have proven to be particularly well suited to avoid locking phenomena. They are based on the curl operators of Section II.3.2 (p. 57) and the Raviart-Thomas space

$$\mathrm{RT}_0(K) = R_0(K)^d + R_0(K)\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}.$$

For any integer $k \in \mathbb{N}$ and any element $K \in \mathcal{T}$ we then set

$$\begin{aligned} B_k(K) &= \{\sigma \in \mathbb{R}^{d\times d} : (\sigma_{i1}, \dots, \sigma_{id}) = \mathrm{curl}(\psi_K w_i), \\ &\qquad w_i \in R_k(K)^{2d-3}, 1 \le i \le d\}, \\ \mathrm{BDM}_k(K) &= R_k(K)^d. \end{aligned}$$

Both types of discretizations are obtained by replacing the spaces H, V and W in the variational problem corresponding to the Hellinger-Reissner principle by discrete counterparts $H_\mathcal{T}$, $V_\mathcal{T}$ and $W_\mathcal{T}$, respectively. They differ by choice of these spaces $H_\mathcal{T}$, $V_\mathcal{T}$ and $W_\mathcal{T}$ which are given by

$$\begin{aligned} H_\mathcal{T} &= \{\sigma_\mathcal{T} \in H : \sigma_\mathcal{T}|_K \in \mathrm{RT}_0(K)^d \oplus B_0(K), K \in \mathcal{T}, \\ &\qquad \sigma_\mathcal{T} \cdot n = 0 \text{ on } \Gamma_N\}, \\ V_\mathcal{T} &= \{\mathbf{v}_\mathcal{T} \in V : \mathbf{v}_\mathcal{T}|_K \in R_0(K)^d, K \in CT\}, \\ W_\mathcal{T} &= \{\eta_\mathcal{T} \in W \cap C(\Omega)^{d\times d} : \eta_\mathcal{T}|_K \in R_1(K)^{d\times d}, K \in \mathcal{T}\}, \end{aligned}$$

for the *PEERS element* and

$$\begin{aligned} H_\mathcal{T} &= \{\sigma_\mathcal{T} \in H : \sigma_\mathcal{T}|_K \in \mathrm{BDM}_k(K)^d \oplus B_{k-1}(K), K \in \mathcal{T}, \\ &\qquad \sigma_\mathcal{T} \cdot n = 0 \text{ on } \Gamma_N\}, \\ V_\mathcal{T} &= \{\mathbf{v}_\mathcal{T} \in V : \mathbf{v}_\mathcal{T}|_K \in R_{k-1}(K)^d, K \in CT\}, \\ W_\mathcal{T} &= \{\eta_\mathcal{T} \in W \cap C(\Omega)^{d\times d} : \eta_\mathcal{T}|_K \in R_k(K)^{d\times d}, K \in \mathcal{T}\} \end{aligned}$$

for the *BDMS elements*.

For both discretizations it can be proven that they admit a unique solution.

II.3.4.3. *Residual a posteriori error estimates.* In what follows we always denote by $(\sigma, \mathbf{u}, \gamma) \in H \times V \times W$ the unique solution of the variational problem corresponding to the Hellinger-Reissner principle and by $(\sigma_\mathcal{T}, \mathbf{u}_\mathcal{T}, \gamma_\mathcal{T}) \in H_\mathcal{T} \times V_\mathcal{T} \times W_\mathcal{T}$ its finite element approximation using the PEERS or BDMS elements.

For every edge of face E and every tensor field $\tau : \Omega \to \mathbb{R}^{d\times d}$ we denote by

$$\gamma_E(\tau) = \tau - (\mathbf{n} \cdot \tau \cdot \mathbf{n})\mathbf{n} \otimes \mathbf{n}$$

the *tangential component* of τ.

With these notations we define for every element $K \in \mathcal{T}$ the *residual a posteriori error estimator* $\eta_{R,K}$ by

$$\begin{aligned}\eta_{R,K} = \Big\{ & h_K^2 \|C^{-1}\sigma_{\mathcal{T}} + \gamma_{\mathcal{T}} - \nabla \mathbf{u}_{\mathcal{T}}\|_K^2 \\ & + \frac{1}{\mu^2}\|f + \operatorname{div}\sigma_{\mathcal{T}}\|_K^2 + \frac{1}{\mu^2}\|\operatorname{as}(\sigma_{\mathcal{T}})\|_K^2 \\ & + h_K^2 \|\operatorname{curl}(C^{-1}\sigma_{\mathcal{T}} + \gamma_{\mathcal{T}})\|_K^2 \\ & + \sum_{E \in \mathcal{E}_K \cap \mathcal{E}_\Omega} h_E \|\mathbb{J}_E(\gamma_E(C^{-1}\sigma_{\mathcal{T}} + \gamma_{\mathcal{T}}))\|_E^2 \\ & + \sum_{E \in \mathcal{E}_K \cap \mathcal{E}_\Gamma} h_E \|\gamma_E(C^{-1}\sigma_{\mathcal{T}} + \gamma_{\mathcal{T}})\|_E^2 \Big\}^{\frac{1}{2}}.\end{aligned}$$

It can be proven that this estimator yields upper and lower bounds on the error

$$\Big\{ \frac{1}{\mu^2}\|\sigma - \sigma_{\mathcal{T}}\|_H^2 + \|\mathbf{u} - \mathbf{u}_{\mathcal{T}}\|^2 + \|\gamma - \gamma_{\mathcal{T}}\|^2 \Big\}^{\frac{1}{2}}$$

up to multiplicative constants which are independent of the Lamé parameters λ and μ.

II.3.4.4. *Local Neumann problems.* We want to treat local auxiliary problems, which are based on a single element $K \in \mathcal{T}$. Furthermore we want to impose pure Neumann boundary conditions. Since the displacement of a linear elasticity problem with pure Neumann boundary conditions is unique only up to *rigid body motions*, we must factor out the rigid body motions R_K of the element K. These are given by

$$R_K = \begin{cases} \{\mathbf{v} = (a,b) + c(-x_2, x_1) : a, b, c \in \mathbb{R}\} & \text{if } d = 2, \\ \{\mathbf{v} = a + b \times x : a, b \in \mathbb{R}^3\} & \text{if } d = 3. \end{cases}$$

We set

$$\begin{aligned} H_K &= \mathrm{BDM}_m(K)^d \oplus B_{m-1}(K), \\ V_K &= R_{m-1}(K)^d / R_K \\ W_K &= \{\eta_K \in R_m(K)^{d\times d} : \eta_K + \eta_K^t = 0\} \end{aligned}$$

and

$$X_K = H_K \times V_K \times W_K,$$

where $m \geq k + 2d$. With this definition of spaces it can be proven that the following auxiliary local discrete problem admits a unique solution:

Find $(\sigma_K, \mathbf{u}_K, \gamma_K) \in X_K$ such that

$$\begin{aligned}\int_K C^{-1}\sigma_K : \tau_K + \int_K \operatorname{div}\tau_K \cdot \mathbf{u}_K + \int_K \tau_K : \gamma_K \\ = -\int_K C^{-1}\sigma_{\mathcal{T}} : \tau_K - \int_K \operatorname{div}\tau_K \cdot \mathbf{u}_{\mathcal{T}} \\ - \int_K \tau_K : \gamma_{\mathcal{T}}\end{aligned}$$

$$\int_K \operatorname{div}\sigma_K \cdot \mathbf{v}_K = -\int_K \mathbf{f}\cdot\mathbf{v}_K - \int_K \operatorname{div}\sigma_{\mathcal{T}}\cdot\mathbf{v}_K$$

$$\int_K \sigma_K : \eta_K = -\int_K \sigma_{\mathcal{T}} : \eta_K$$

holds for all $(\tau_K, \mathbf{v}_K, \eta_K) \in X_K$.

With the solution of this problem we define the error estimator $\eta_{N,K}$ by

$$\eta_{N,K} = \Big\{ \frac{1}{\mu^2}\|\sigma_K\|^2_{H(\operatorname{div};K)} + \|\mathbf{v}_K\|^2_K + \|\gamma_K\|^2_K \Big\}^{\frac{1}{2}}.$$

It yields upper and lower bounds on the error up to multiplicative constants which are independent of the Lamé parameters λ and μ.

REMARK II.3.2. The above auxiliary problem is a discrete linear elasticity problem with pure Neumann boundary conditions on the single element K. In order to implement the error estimator $\eta_{N,K}$ one has to construct a basis for the space V_K. This can be done by taking the standard basis of $R_m(K)^d$ and dropping those degrees of freedom that belong to the rigid body motions. Afterwards one has to compute the stiffness matrix for each element K and solve the associated local auxiliary problem.

II.3.4.5. *Local Dirichlet problems.* Now we want to construct an error estimator which is similar to the estimator $\eta_{D,K}$ of Section II.2.1.2 (p. 39) and which is based on the solution of discrete linear elasticity problems with Dirichlet boundary conditions on the patches ω_K. To this end we associate with every element K the spaces

$$\begin{aligned}\widetilde{H}_K = \{\sigma_K \in H(\operatorname{div};\omega_K)^d : \sigma_{\mathcal{T}}|_{K'} \in \mathrm{BDM}_m(K')^d \oplus B_{m-1}(K'), \\ K' \in \mathcal{T}, \mathcal{E}_{K'} \cap \mathcal{E}_K \neq \emptyset\},\end{aligned}$$

$$\begin{aligned}\widetilde{V}_K = \{\mathbf{v}_{\mathcal{T}} \in L^2(\omega_K)^d : \mathbf{v}_{\mathcal{T}}|_{K'} \in R_{m-1}(K')^d, \\ K' \in \mathcal{T}, \mathcal{E}_{K'} \cap \mathcal{E}_K \neq \emptyset\},\end{aligned}$$

$$\widetilde{W}_K = \{\eta_{\mathcal{T}} \in L^2(\omega_K)^{d\times d} \cap C(\omega_K)^{d\times d} : \eta_{\mathcal{T}} + \eta_{\mathcal{T}}^t = 0,$$

$$\eta_{\mathcal{T}}|_{K'} \in R_m(K')^{d\times d},$$
$$K' \in \mathcal{T}, \mathcal{E}_{K'} \cap \mathcal{E}_K \neq \emptyset\}$$

and

$$\widetilde{X}_K = \widetilde{H}_K \times \widetilde{V}_K \times \widetilde{W}_K$$

and consider the following auxiliary local problem

Find $(\widetilde{\sigma}_K, \widetilde{\mathbf{u}}_K, \widetilde{\gamma}_K) \in X_K$ such that

$$\begin{aligned}
\int_{\omega_K} C^{-1}\widetilde{\sigma}_K : \tau_K + \int_{\omega_K} \operatorname{div} \tau_K \cdot \widetilde{\mathbf{u}}_K & \\
+ \int_{\omega_K} \tau_K : \widetilde{\gamma}_K &= -\int_{\omega_K} C^{-1}\sigma_{\mathcal{T}} : \tau_K \\
&\quad - \int_{\omega_K} \operatorname{div} \tau_K \cdot \mathbf{u}_{\mathcal{T}} \\
&\quad - \int_{\omega_K} \tau_K : \gamma_{\mathcal{T}} \\
\int_{\omega_K} \operatorname{div} \widetilde{\sigma}_K \cdot \mathbf{v}_K &= -\int_{\omega_K} \mathbf{f} \cdot \mathbf{v}_K \\
&\quad - \int_{\omega_K} \operatorname{div} \widetilde{\sigma}_{\mathcal{T}} \cdot \mathbf{v}_K \\
\int_{\omega_K} \widetilde{\sigma}_K : \eta_K &= -\int_{\omega_K} \sigma_{\mathcal{T}} : \eta_K
\end{aligned}$$

holds for all $(\tau_K, \mathbf{v}_K, \eta_K) \in X_K$.

Again it can be proven that this problem admits a unique solution. With it we define the error estimator $\eta_{D,K}$ by

$$\eta_{D,K} = \left\{ \frac{1}{\mu^2} \|\widetilde{\sigma}_K\|^2_{H(\operatorname{div};K)} + \|\widetilde{\mathbf{v}}_K\|^2_K + \|\widetilde{\gamma}_K\|^2_K \right\}^{\frac{1}{2}}.$$

It yields upper and lower bounds on the error up to multiplicative constants which are independent of the Lamé parameters λ and μ.

II.3.5. Non-linear problems. For non-linear elliptic problems, residual a posteriori error estimators are constructed in the same way as for linear problems. The estimators consist of two ingredients:

- element residuals which consist of the element-wise residual of the actual discrete solution with respect to the strong form of the differential equation,
- edge or face residuals which consist of the inter-element jump of that trace operator which links the strong and weak form

of the differential equation where all differential operators are evaluated at the current discrete solution.

EXAMPLE II.3.3. If the differential equation takes the form

$$\begin{aligned} -\operatorname{div}\mathbf{a}(x,u,\nabla u)+b(x,u,\nabla u) &= 0 && \text{in } \Omega \\ u &= 0 && \text{on } \Gamma_D \\ \mathbf{n}\cdot\mathbf{a}(x,u,\nabla u) &= g && \text{on } \Gamma_N \end{aligned}$$

with a suitable differentiable vector-field $\mathbf{a}:\Omega\times\mathbb{R}\times\mathbb{R}^d\to\mathbb{R}^d$ and a suitable continuous function $b:\Omega\times\mathbb{R}\times\mathbb{R}^d\to\mathbb{R}$, the element and edge or face residuals are given by

$$R_K(u_{\mathcal{T}}) = -\operatorname{div}\mathbf{a}(x,u_{\mathcal{T}},\nabla u_{\mathcal{T}})+b(x,u_{\mathcal{T}},\nabla u_{\mathcal{T}})$$

and

$$R_E(u_{\mathcal{T}}) = \begin{cases} \mathbb{J}_E(\mathbf{n}_E\cdot\mathbf{a}(x,u_{\mathcal{T}},\nabla u_{\mathcal{T}})) & \text{if } E\in\mathcal{E}_\Omega \\ g_E-\mathbf{n}_E\cdot\mathbf{a}(x,u_{\mathcal{T}},\nabla u_{\mathcal{T}}) & \text{if } E\in\mathcal{E}_{\Gamma_N} \\ 0 & \text{if } E\in\mathcal{E}_{\Gamma_D} \end{cases}$$

respectively, where g_E denotes the mean value of g on E.

Some peculiarities arise from the non-linearity and must be taken into account:

- The error estimation only makes sense if the discrete problem is based on a well-posed variational formulation of the differential equation. Here, well-posedness means that the non-linear mapping associated with the variational problem must at least be continuous. In order to fulfil this requirement one often has to leave the Hilbert setting and to replace $H^1(\Omega)$ by general Sobolev spaces $W^{1,p}(\Omega)$ with a Lebesgue exponent $p\neq 2$, typically $p>d$. The choice of the Lebesgue exponent p is not at the disposal of the user, it is dictated by the nature of the non-linearity such as, e.g., its growth. When leaving the Hilbert setting, the L^2-norms used in the error estimators must be replaced by corresponding L^p-norms and the weighting factors must be adapted too. Thus, in a general L^p-setting, a typical residual error estimator takes the form

$$\eta_{R,K} = \Big\{ h_K^p \|R_K(u_{\mathcal{T}})\|_{p;K}^p + \frac{1}{2}\sum_{E\in\mathcal{E}_K} h_E \|R_E(u_{\mathcal{T}})\|_{p;E}^p \Big\}^{\frac{1}{p}}.$$

- Non-linear problems in general have multiple solutions. Therefore any error estimator can at best control the error of the actual discrete solution with respect to a near-by solution of the variational problem. Moreover, an error control often is possible only if the actual grid is fine enough. Unfortunately,

the notions "near-by" and "fine enough" can in general not be quantified.
- Non-linear problems often inhibit bifurcation or turning points. Of course, one would like to keep track of these phenomena with the help of the error estimator. Unfortunately, this is often possible only on a heuristic bases. Rigourous arguments in general require additional a priori information on the structure of the solution manifold which often is not available.

For non-linear problems, error estimators based on the solution of auxiliary discrete problems can be devised as in Section II.2.1 (p. 36) for the model problem. Their solution is considerably simplified by the following observations:

- The non-linearity only enters on the right-hand side of the auxiliary problems via the element and edge or face residuals described above.
- The left-hand sides of the auxiliary problems correspond to *linear* differential operators which are obtained by linearizing the non-linear problem at the current discrete solution.
- Variable coefficients can be frozen at the current discrete solution and suitable points of the local patch such as, e.g., the barycentres of the elements and edges or faces.

The error estimators of Sections II.2.2 (p. 42) and II.2.3 (p. 47) can in general be applied to non-linear problems only on a heuristic bases; rigourous results are at present only available for some of these estimators applied to particular model problems.

II.4. Parabolic problems

II.4.1. Scalar linear parabolic equations. In this section we extend the results of Section II.3 (p. 54) to general linear parabolic equations of second order:

$$\begin{aligned} \partial_t u - \operatorname{div}(A\nabla u) + \mathbf{a}\cdot\nabla u + \alpha u &= f && \text{in } \Omega\times(0,T] \\ u &= 0 && \text{on } \Gamma_D\times(0,T] \\ \mathbf{n}\cdot A\nabla u &= g && \text{on } \Gamma_N\times(0,T] \\ u &= u_0 && \text{in } \Omega. \end{aligned}$$

Here, $\Omega\subset\mathbb{R}^d$, $d\geq 2$, is a bounded polygonal cross-section with a Lipschitz boundary Γ consisting of two disjoint parts Γ_D and Γ_N. The final time T is arbitrary, but kept fixed in what follows.

We assume that the data satisfy the following conditions:

- The diffusion A is a continuously differentiable matrix-valued function and symmetric, uniformly positive definite and uniformly isotropic, i.e.,
$$\varepsilon = \inf_{0<t\le T, x\in\Omega} \min_{z\in\mathbb{R}^d\setminus\{0\}} \frac{z^T A(x,t) z}{z^T z} > 0$$
and
$$\kappa = \varepsilon^{-1} \sup_{0<t\le T, x\in\Omega} \max_{z\in\mathbb{R}^d\setminus\{0\}} \frac{z^T A(x,t) z}{z^T z}$$
is of moderate size.
- The convection $\mathbf{a}$ is a continuously differentiable vector-field and scaled such that
$$\sup_{0<t\le T, x\in\Omega} |\mathbf{a}(x,t)| \le 1.$$
- The reaction α is a continuos non-negative scalar function.
- There is a constant $\beta \ge 0$ such that
$$\alpha - \frac{1}{2}\operatorname{div}\mathbf{a} \ge \beta$$
for almost all $x \in \Omega$ and $0 < t \le T$. Moreover there is a constant $c_b \ge 0$ of moderate size such that
$$\sup_{0<t\le T, x\in\Omega} |\alpha(x,t)| \le c_b \beta.$$
- The Dirichlet boundary Γ_D has positive $(d-1)$-dimensional measure and includes the inflow boundary
$$\bigcup_{0<t\le T} \{x \in \Gamma : \mathbf{a}(x,t)\cdot n(x) < 0\}.$$

With these assumptions we can distinguish different regimes:

- *dominant diffusion*: $\sup_{0<t\le T, x\in\Omega}|\mathbf{a}(x,t)| \le c_c\varepsilon$ and $\beta \le c_b'\varepsilon$ with constants of moderate size;
- *dominant reaction*: $\sup_{0<t\le T, x\in\Omega}|\mathbf{a}(x,t)| \le c_c\varepsilon$ and $\beta \gg \varepsilon$ with a constant c_c of moderate size;
- *dominant convection*: $\beta \gg \varepsilon$.

II.4.2. Variational formulation. The variational formulation of the above parabolic differential equation is given by:

Find $u : (0,T) \to H^1_D(\Omega)$ such that
$$\int_0^T \|\nabla u(x,t)\|^2 dt < \infty,$$
$$\int_0^T \Big\{ \sup_{\substack{v\in H^1_d(\Omega)\setminus\{0\}\\ \|\nabla v\|=1}} \int_\Omega \partial_t u(x,t) v(x) dx \Big\}^2 dt < \infty,$$

$$u(\cdot,0) = u_0$$

and for almost every $t \in (0,T)$ and all $v \in H^1_D(\Omega)$

$$\int_\Omega \partial_t uv + \int_\Omega \nabla u \cdot A \nabla v + \int_\Omega \mathbf{a}\cdot\nabla uv + \int_\Omega \alpha uv = \int_\Omega fv + \int_{\Gamma_N} gv.$$

The assumptions of Section II.4.1 imply that this problem admits a unique solution.

The error estimation of the following sections is based on the energy norm associated with this variational problem

$$|||v||| = \left\{\varepsilon\|\nabla v\|^2 + \beta\|v\|^2\right\}^{\frac{1}{2}}$$

and the corresponding dual norm

$$|||\varphi|||_* = \sup_{v\in H^1_D(\Omega)\setminus\{0\}} \frac{1}{|||v|||}\int_\Omega \nabla\varphi\cdot\nabla v.$$

II.4.3. An overview of discretization methods for parabolic equations. Within the finite element framework, there are three main approaches to discretize parabolic equations:

- *Method of lines*: One chooses a fixed spatial mesh and applies a standard finite element scheme to the spatial part of the differential equation. This gives rise to a large system of ordinary differential equations. The size of this system is given by the number of degrees of freedom of the finite element space, the unknowns are the (now time-dependent) coefficients of the finite element functions. The system of ordinary differential equations is then solved by a standard ODE-solver such as, e.g., the Crank-Nicolson scheme or some Runge-Kutta method.
- *Rothe's method*: In this approach the order of temporal and spatial discretization is interchanged. The parabolic partial differential equation is interpreted as an ordinary differential equation with respect to time with functions having their temporal values in suitable infinite dimensional function spaces such as, e.g., $H^1(\Omega)$. One applies a standard ODE-solver to this system of ordinary differential equations. At each time-step this gives rise to a stationary elliptic partial differential equation. These elliptic equations are then discretized by a standard finite element scheme.

- *Space-time finite elements*: In this approach space and time are discretized simultaneously.

All three approaches often lead to the same discrete problem. Yet, they considerably differ in their analysis and – most important in our context – their potential for adaptivity. With respect to the latter, the space-time elements are clearly superior.

II.4.4. Space-time finite elements. In what follows we consider partitions

$$\mathcal{I} = \{[t_{n-1}, t_n] : 1 \le n \le N_{\mathcal{I}}\}$$

of the time-interval $[0, T]$ into subintervals satisfying

$$0 = t_0 < \ldots < t_{N_{\mathcal{I}}} = T.$$

For every n with $1 \le n \le N_{\mathcal{I}}$ we denote by

$$I_n = [t_{n-1}, t_n]$$

the n-th subinterval and by

$$\tau_n = t_n - t_{n-1}$$

its length.

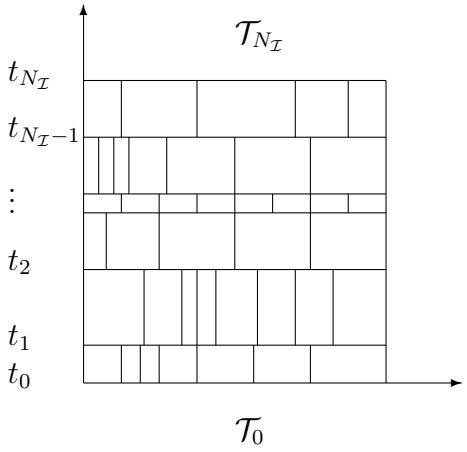

FIGURE II.4.1. Space-time partition

With every intermediate time t_n, $0 \le n \le N_{\mathcal{I}}$, we associate an admissible, affine equivalent, shape regular partition $\mathcal{T}_n$ of Ω (cf. Figure II.4.1) and a corresponding finite element space X_n. In addition to the conditions of Sections I.2.7 (p. 15) and I.2.8 (p. 17) the partitions $\mathcal{I}$ and $\mathcal{T}_n$ and the spaces X_n must satisfy the following assumptions:

- *Non-degeneracy*: Every time-interval has a positive length, i.e., $\tau_n > 0$ for all $1 \le n \le N_{\mathcal{I}}$ and all $\mathcal{I}$.

- *Transition condition*: For every n with $1 \le n \le N_{\mathcal{I}}$ there is an affine equivalent, admissible, and shape-regular partition $\widetilde{\mathcal{T}}_n$ such that it is a refinement of both $\mathcal{T}_n$ and $\mathcal{T}_{n-1}$ and such that
$$\sup_{1\le n\le N_{\mathcal{I}}} \sup_{K\in\widetilde{\mathcal{T}}_n} \sup_{\substack{K'\in\mathcal{T}_n\\ K\subset K'}} \frac{h_{K'}}{h_K} < \infty$$
uniformly with respect to all partitions $\mathcal{I}$ which are obtained by adaptive or uniform refinement of any initial partition of $[0,T]$.
- *Degree condition*: Each X_n consists of continuous functions which are piecewise polynomials, the degrees being at least one and being bounded uniformly with respect to all partitions $\mathcal{T}_n$ and $\mathcal{I}$.

The non-degeneracy is an obvious requirement to exclude pathological situations.
The transition condition is due to the simultaneous presence of finite element functions defined on different grids. Usually the partition $\mathcal{T}_n$ is obtained from $\mathcal{T}_{n-1}$ by a combination of refinement and of coarsening. In this case the transition condition only restricts the coarsening: it must not be too abrupt nor too strong.
The lower bound on the polynomial degrees is needed for the construction of suitable quasi-interpolation operators. The upper bound ensures that the constants in inverse estimates similar to those of Section I.2.12 (p. 22) are uniformly bounded.

For every n with $0 \le n \le N_{\mathcal{I}}$ we finally denote by π_n the L^2-projection of $L^2(\Omega)$ onto X_n.

II.4.5. Finite element discretization. For the finite element discretization we choose a partition $\mathcal{I}$ of $[0,T]$, corresponding partitions $\mathcal{T}_n$ of Ω and associated finite element spaces X_n as above and a parameter $\theta \in [\frac{1}{2}, 1]$. With the abbreviation
$$\begin{aligned} A^n &= A(\cdot, t_n),\\ \mathbf{a}^n &= \mathbf{a}(\cdot, t_n),\\ \alpha^n &= \alpha(\cdot, t_n),\\ f^n &= f(\cdot, t_n),\\ g^n &= g(\cdot, t_n), \end{aligned}$$
the finite element discretization is then given by:

Find $u^n_{\mathcal{T}_n} \in X_n$, $0 \le n \le N_{\mathcal{I}}$, such that
$$u^0_{\mathcal{T}_0} = \pi_0 u_0$$

and, for $n = 1, \dots, N_{\mathcal{I}}$,

$$\begin{aligned}\int_\Omega \frac{1}{\tau_n}\big(u^n_{\mathcal{T}_n} - u^{n-1}_{\mathcal{T}_{n-1}}\big)v_{\mathcal{T}_n} &+ \int_\Omega \big(\theta\nabla u^n_{\mathcal{T}_n} + (1-\theta)\nabla u^{n-1}_{\mathcal{T}_{n-1}}\big)\cdot A^n\nabla v_{\mathcal{T}_n}\\ &+ \int_\Omega \mathbf{a}^n\cdot\nabla\big(\theta u^n_{\mathcal{T}_n} + (1-\theta)u^{n-1}_{\mathcal{T}_{n-1}}\big)v_{\mathcal{T}_n}\\ &+ \int_\Omega \alpha^n\big(\theta u^n_{\mathcal{T}_n} + (1-\theta)u^{n-1}_{\mathcal{T}_{n-1}}\big)v_{\mathcal{T}_n}\\ &= \int_\Omega \big(\theta f^n + (1-\theta)f^{n-1}\big)v_{\mathcal{T}_n}\\ &\quad + \int_{\Gamma_N}\big(\theta g^n + (1-\theta)g^{n-1}\big)v_{\mathcal{T}_n}\end{aligned}$$

for all $v_{\mathcal{T}_n} \in X_n$.

This is the popular A-stable *θ-scheme* which in particular yields the *Crank-Nicolson scheme* if $\theta = \frac{1}{2}$ and the *implicit Euler scheme* if $\theta = 1$.

The assumptions of Section II.4.1 imply that the discrete problem admits a unique solution $(u^n_{\mathcal{T}_n})_{0\le n\le N_{\mathcal{I}}}$. With this sequence we associate the function $u_{\mathcal{I}}$ which is *piecewise affine* on the time-intervals $[t_{n-1}, t_n]$, $1 \le n \le N_{\mathcal{I}}$, and which equals $u^n_{\mathcal{T}_n}$ at time t_n, $0 \le n \le N_{\mathcal{I}}$, i.e.,

$$u_{\mathcal{I}}(\cdot,t) = \frac{1}{\tau_n}\big((t_n - t)u^{n-1}_{\mathcal{T}_{n-1}} + (t - t_{n-1})u^n_{\mathcal{T}_n}\big) \quad \text{on } [t_{n-1}, t_n].$$

Note that

$$\partial_t u_{\mathcal{I}} = \frac{1}{\tau_n}(u^n_{\mathcal{T}_n} - u^{n-1}_{\mathcal{T}_{n-1}}) \quad \text{on } [t_{n-1}, t_n].$$

Similarly we denote by $f_{\mathcal{I}}$ and $g_{\mathcal{I}}$ the functions which are *piecewise constant* on the time-intervals and which, on each interval $(t_{n-1}, t_n]$, are equal to the L^2-projection of $\theta f^n + (1-\theta)f^{n-1}$ and $\theta g^n + (1-\theta)g^{n-1}$, respectively onto the finite element space X_n, i.e.,

$$\begin{aligned} f_{\mathcal{I}}(\cdot,t) &= \pi_n\big(\theta f(\cdot,t_n) + (1-\theta)f(\cdot,t_{n-1})\big)\\ g_{\mathcal{I}}(\cdot,t) &= \pi_n\big(\theta g(\cdot,t_n) + (1-\theta)g(\cdot,t_{n-1})\big)\end{aligned}$$

on $[t_{n-1}, t_n]$.

II.4.6. A preliminary residual error estimator. Similarly to elliptic problems we define element residuals by

$$\begin{aligned} R_K = f_{\mathcal{I}} &- \frac{1}{\tau_n}(u^n_{\mathcal{T}_n} - u^{n-1}_{\mathcal{T}_{n-1}}) + \operatorname{div}(A^n(\theta u^n_{\mathcal{T}_n} + (1-\theta)u^{n-1}_{\mathcal{T}_{n-1}}))\\ &- \mathbf{a}^n\cdot\nabla(\theta u^n_{\mathcal{T}_n} + (1-\theta)u^{n-1}_{\mathcal{T}_{n-1}}) - \alpha^n(\theta u^n_{\mathcal{T}_n} + (1-\theta)u^{n-1}_{\mathcal{T}_{n-1}}),\end{aligned}$$

and edge or face residuals by

$$R_E = \begin{cases} -\mathbb{J}_E(\mathbf{n}_E \cdot A^n \nabla(\theta u^n_{\mathcal{T}_n} + (1-\theta) u^{n-1}_{\mathcal{T}_{n-1}}) & \text{if } E \in \widetilde{\mathcal{E}}_{n,\Omega}, \\ g_{\mathcal{I}} - \mathbf{n}_E \cdot A^n \nabla(\theta u^n_{\mathcal{T}_n} + (1-\theta) u^{n-1}_{\mathcal{T}_{n-1}}) & \text{if } E \in \widetilde{\mathcal{E}}_{n,\Gamma_N}, \\ 0 & \text{if } E \in \widetilde{\mathcal{E}}_{n,\Gamma_D} \end{cases}$$

with $\widetilde{\mathcal{E}}_n$ denoting the collection of all edges or faces of $\widetilde{\mathcal{T}}_n$ and weighting factors by

$$\alpha_S = \min\{h_S \varepsilon^{-\frac{1}{2}}, \beta^{-\frac{1}{2}}\}$$

for all elements, edges or faces $S \in \mathcal{T} \cup \mathcal{E}$. Here we use the convention that $\beta^{-\frac{1}{2}} = \infty$ if $\beta = 0$.

With these notations a preliminary residual space-time error estimator for the parabolic equation is given by

$$\widehat{\eta}_{\mathcal{I}} = \Bigg\{ \|u_0 - \pi_0 u_0\|^2 + \sum_{n=1}^{N_{\mathcal{I}}} \tau_n \Big[\Big(\eta^n_{\mathcal{T}_n}\Big)^2 + |||u^n_{\mathcal{T}_n} - u^{n-1}_{\mathcal{T}_{n-1}}|||^2 + |||\mathbf{a}^n \cdot \nabla(u^n_{\mathcal{T}_n} - u^{n-1}_{\mathcal{T}_{n-1}})|||^2_* \Big] \Bigg\}^{\frac{1}{2}}$$

with

$$\eta^n_{\mathcal{T}_n} = \Big\{ \sum_{K \in \widetilde{\mathcal{T}}_n} \alpha_K^2 \|R_K\|_K^2 + \sum_{E \in \widetilde{\mathcal{E}}_n} \varepsilon^{-\frac{1}{2}} \alpha_E \|R_E\|_E^2 \Big\}^{\frac{1}{2}}.$$

One can prove that $\widehat{\eta}_{\mathcal{I}}$ yields upper and lower bounds for the error measured in the norm

$$\Big\{ \sup_{0 \le t \le T} \|u - u_{\mathcal{I}}\|^2 + \int_0^T |||u - u_{\mathcal{I}}|||^2 + \int_0^T |||\frac{\partial}{\partial t}(u - u_{\mathcal{I}}) + \mathbf{a} \cdot \nabla(u - u_{\mathcal{I}})|||^2_* \Big\}^{\frac{1}{2}}.$$

We call $\widehat{\eta}_{\mathcal{I}}$ a preliminary error estimator since it is not suited for practical computations due to the presence of the dual norm $|||\cdot|||_*$ which is not computable. To obtain the final computable error estimator we

must replace this dual norm be a computable quantity. For achieving this goal we must distinguish two cases:

- *small convection*: $\sup_{0\le t\le T}\|\mathbf{a}(\cdot,t)\| \lessapprox \varepsilon^{\frac{1}{2}}\max\{\varepsilon,\beta\}^{\frac{1}{2}}$;
- *large convection*: $\sup_{0\le t\le T}\|\mathbf{a}(\cdot,t)\| \gg \varepsilon^{\frac{1}{2}}\max\{\varepsilon,\beta\}^{\frac{1}{2}}$.

II.4.7. A residual error estimator for the case of small convection. In this case, we use an inverse estimate to bound the critical term

$$|||\mathbf{a}^n\cdot\nabla(u^n_{\mathcal{T}_n}-u^{n-1}_{\mathcal{T}_{n-1}})|||_*$$

by

$$|||u^n_{\mathcal{T}_n}-u^{n-1}_{\mathcal{T}_{n-1}}|||$$

times a constant of moderate size. We thus obtain the residual error estimator

$$\eta_{\mathcal{I}}=\Bigg\{\|u_0-\pi_0u_0\|^2 + \sum_{n=1}^{N_{\mathcal{I}}}\tau_n\Big[\Big(\eta^n_{\mathcal{T}_n}\Big)^2+|||u^n_{\mathcal{T}_n}-u^{n-1}_{\mathcal{T}_{n-1}}|||^2\Big]\Bigg\}^{\frac{1}{2}}$$

with

$$\eta^n_{\mathcal{T}_n}=\Big\{\sum_{K\in\widetilde{\mathcal{T}}_n}\alpha_K^2\|R_K\|_K^2+\sum_{E\in\widetilde{\mathcal{E}}_n}\varepsilon^{-\frac{1}{2}}\alpha_E\|R_E\|_E^2\Big\}^{\frac{1}{2}}.$$

It is easy to compute and yields upper and lower bounds for the error measured in the norm of Section II.4.6.

II.4.8. A residual error estimator for the case of large convection. In this case we cannot bound the dual norm by an inverse estimate. If we would do so, we would lose a factor $\varepsilon^{-\frac{1}{2}}$ in the error estimates. To avoid this undesirable phenomenon we must invest some additional work. The basic idea is as follows:

> Due to the definition of the dual norm, its contribution equals the energy norm of the weak solution of a suitable stationary reaction-diffusion equation. This solution is approximated by a suitable finite element function. The error of this approximation is estimated by an error estimator for stationary reaction-diffusion equations.

To make these ideas precise, we denote for every integer n between 1 and $N_{\mathcal{I}}$ by

$$\widetilde{X}_n=S^{1,0}_D(\widetilde{\mathcal{T}}_n)$$

the space of continuous piecewise linear functions corresponding to $\widetilde{\mathcal{T}}_n$ and vanishing on Γ_D and by $\widetilde{u}^n_{\mathcal{T}_n} \in \widetilde{X}_n$ the unique solution of the discrete reaction-diffusion problem

$$\varepsilon \int_\Omega \nabla \widetilde{u}^n_{\mathcal{T}_n} \cdot \nabla v_{\mathcal{T}_n} + \beta \int_\Omega \widetilde{u}^n_{\mathcal{T}_n} v_{\mathcal{T}_n} = \int_\Omega \mathbf{a}^n \cdot \nabla (u^n_{\mathcal{T}_n} - u^{n-1}_{\mathcal{T}_{n-1}}) v_{\mathcal{T}_n}$$

for all $v_{\mathcal{T}_n} \in \widetilde{X}_n$. Further we define an error estimator $\widetilde{\eta}^n_{\mathcal{T}_n}$ by

$$\widetilde{\eta}^n_{\mathcal{T}_n} = \Big\{ \sum_{K \in \widetilde{\mathcal{T}}_n} \alpha_K^2 \|\mathbf{a}^n \cdot \nabla (u^n_{\mathcal{T}_n} - u^{n-1}_{\mathcal{T}_{n-1}}) + \varepsilon \Delta \widetilde{u}^n_{\mathcal{T}_n} - \beta \widetilde{u}^n_{\mathcal{T}_n}\|^2_K + \sum_{E \in \widetilde{\mathcal{E}}_{n,\Omega} \cup \widetilde{\mathcal{E}}_{n,\Gamma_N}} \varepsilon^{-\frac{1}{2}} \alpha_E \|\mathbb{J}_E (n_E \cdot \nabla \widetilde{u}^n_{\mathcal{T}_n})\|^2_E \Big\}^{\frac{1}{2}}.$$

With these notations the error estimator for the parabolic equation is the given by

$$\eta_{\mathcal{I}} = \Bigg\{ \|u_0 - \pi_0 u_0\|^2 + \sum_{n=1}^{N_{\mathcal{I}}} \tau_n \Big[\Big(\eta^n_{\mathcal{T}_n}\Big)^2 + |||u^n_{\mathcal{T}_n} - u^{n-1}_{\mathcal{T}_{n-1}}|||^2 + \Big(\widetilde{\eta}^n_{\widetilde{\mathcal{T}}_n}\Big)^2 + |||\widetilde{u}^n_{\mathcal{T}_n}|||^2 \Big] \Bigg\}^{\frac{1}{2}}$$

with

$$\eta^n_{\mathcal{T}_n} = \Big\{ \sum_{K \in \widetilde{\mathcal{T}}_n} \alpha_K^2 \|R_K\|^2_K + \sum_{E \in \mathcal{E}_{\widetilde{\mathcal{T}}_n}} \varepsilon^{-\frac{1}{2}} \alpha_E \|R_E\|^2_E \Big\}^{\frac{1}{2}}.$$

Compared to the case of small convection we must solve on each time-level an additional discrete problem to compute $\widetilde{u}^n_{\mathcal{T}_n}$. The computational work associated with these additional problems corresponds to doubling the number of time-steps for the discrete parabolic problem.

II.4.9. Space-time adaptivity. When considering the error estimators $\eta_{\mathcal{I}}$ of the preceding two sections, the term

$$\tau_n^{\frac{1}{2}} \eta^n_{\mathcal{T}_n}$$

can be interpreted as a spatial error indicator, whereas the other terms can be interpreted as temporal error indicators. These different contributions can be used to control the adaptive process in space and time.

To make things precise and to simplify the notation we set

$$\eta_h^n = \eta_{\mathcal{T}_n}^n$$

and

$$\eta_\tau^n = |||u_{\mathcal{T}_n}^n - u_{\mathcal{T}_{n-1}}^{n-1}|||$$

in the case of small convection and

$$\eta_\tau^n = \left\{ |||u_{\mathcal{T}_n}^n - u_{\mathcal{T}_{n-1}}^{n-1}|||^2 + \left(\widetilde{\eta}_{\widetilde{\mathcal{T}}_n}^n\right)^2 + |||\widetilde{u}_{\mathcal{T}_n}^n|||^2 \right\}^{\frac{1}{2}}$$

in the case of large convection. Thus, η_h^n is our measure for the spatial error and η_τ^n does the corresponding job for the temporal error.

II.4.9.1. *Time adaptivity.* Assume that we have solved the discrete problem up to time-level $n-1$ and that we have computed the error estimators η_h^{n-1} and η_τ^{n-1}. Then we set

$$t_n = \begin{cases} \min\{T, t_{n-1} + \tau_{n-1}\} & \text{if } \eta_\tau^{n-1} \approx \eta_h^{n-1}, \\ \min\{T, t_{n-1} + 2\tau_{n-1}\} & \text{if } \eta_\tau^{n-1} \leq \frac{1}{2}\eta_h^{n-1}. \end{cases}$$

In the first case we retain the previous time-step; in the second case we try a larger time step.

Next, we solve the discrete problem on time-level n with the current value of t_n and compute the error estimators η_h^n and η_τ^n.

If $\eta_\tau^n \approx \eta_h^n$, we accept the current time-step and continue with the space adaptivity, which is described in the next sub-section.

If $\eta_\tau^n \geq 2\eta_h^n$, we reject the current time-step. We replace t_n by $\frac{1}{2}(t_{n-1} + t_n)$ and repeat the solution of the discrete problem on time-level n and the computation of the error estimators.

The described strategy obviously aims at balancing the two contributions η_h^n and η_τ^n of the error estimator.

II.4.9.2. *Space adaptivity.* For time-dependent problems the spatial adaptivity must also allow for a local mesh coarsening. Hence, the marking strategies of Section III.1.1 (p. 89) must be modified accordingly, cf. Algorithm III.1.3 (p. 94).

Assume that we have solved the discrete problem on time-level n with an actual time-step τ_n and an actual partition $\mathcal{T}_n$ of the spatial domain Ω and that we have computed the estimators η_h^n and η_τ^n. Moreover, suppose that we have accepted the current time-step and want to optimize the partition $\mathcal{T}_n$.

We may assume that $\mathcal{T}_n$ currently is the finest partition in a hierarchy $\mathcal{T}_n^0, \ldots, \mathcal{T}_n^\ell$ of nested, successively refined partitions, i.e. $\mathcal{T}_n = \mathcal{T}_n^\ell$ and $\mathcal{T}_n^j$ is a (local) refinement of $\mathcal{T}_n^{j-1}$, $1 \leq j \leq \ell$.

Now, we go back m generations in the grid-hierarchy to the partition $\mathcal{T}_n^{\ell-m}$. Due to the nestedness of the partitions, each element $K \in \mathcal{T}_n^{\ell-m}$ is the union of several elements $K' \in \mathcal{T}_n$. Each K' gives a contribution to η_h^n. We add these contributions and thus obtain for every $K \in \mathcal{T}_n^{\ell-m}$ an error estimator η_K. With these estimators we then perform M steps of one of the marking strategies of Section III.1.1 (p. 89). This yields a new partition $\mathcal{T}_n^{\ell-m+M}$ which usually is different from $\mathcal{T}_n$. We replace $\mathcal{T}_n$ by this partition, solve the corresponding discrete problem on time-level n and compute the new error estimators η_h^n and η_τ^n.

If the newly calculated error estimators satisfy $\eta_h^n \approx \eta_\tau^n$, we accept the current partition $\mathcal{T}_n$ and proceed with the next time-level.

If $\eta_h^n \geq 2\eta_\tau^n$, we successively refine the partition $\mathcal{T}_n$ as described in Section III.1 (p. 89) with the η_h^n as error estimators until we arrive at a partition which satisfies $\eta_h^n \approx \eta_\tau^n$. When this goal is achieved we accept the spatial discretization and proceed with the next time-level.

Typical values for the parameters m and M are $1 \leq m \leq 3$ and $m \leq M \leq m+2$.

II.4.10. The method of characteristics. The *method of characteristics* can be interpreted as a modification of space-time finite elements designed for problems with a large convection term. The main idea is to split the discretization of the *material derivative* consisting of the time derivative and the convective derivative from the remaining terms.

To simplify the description of the method of characteristics we assume that we use *linear finite elements*, have pure Dirichlet boundary conditions, i.e. $\Gamma_D = \Gamma$, and that the convection satisfies the slightly more restrictive condition

$$\begin{aligned} \operatorname{div} \mathbf{a} &= 0 \quad \text{in } \Omega \times (0,T], \\ \mathbf{a} &= 0 \quad \text{on } \Gamma \times (0,T]. \end{aligned}$$

Since the function $\mathbf{a}$ is Lipschitz continuous with respect to the spatial variable and vanishes on the boundary Γ, for every $(x^*,t^*) \in \Omega \times (0,T]$, standard global existence results for the flows of ordinary differential equations imply that the *characteristic equation*

$$\boxed{\begin{aligned} \frac{d}{dt} x(t;x^*,t^*) &= \mathbf{a}(x(t;x^*,t^*),t), \quad t \in (0,t^*), \\ x(t^*;x^*,t^*) &= x^* \end{aligned}}$$

has a unique solution $x(\cdot;x^*,t^*)$ which exists for all $t \in [0,t^*]$ and stays within $\Omega \cup \Gamma$. Hence, we may set $U(x^*,t) = u(x(t;x^*,t^*),t)$. The total derivative $d_t U$ satisfies

$$d_t U = \partial_t u + \mathbf{a} \cdot \nabla u.$$

Therefore, the parabolic equation can equivalently be written as

$$d_t U - \operatorname{div}(D\nabla u) + bu = f \quad \text{in } \Omega \times (0,T).$$

The discretization by the method of characteristics relies on a separate treatment of these two equations.

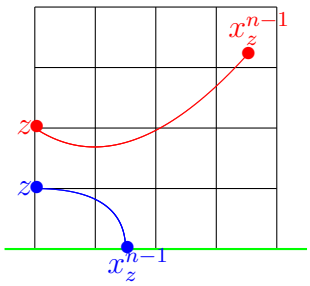

FIGURE II.4.2. Computation of x_z^{n-1} in the method of characteristics

For every intermediate time t_n, $1 \le n \le N_{\mathcal{I}}$, and every node $z \in \mathcal{N}_{n,\Omega}$ we compute an approximation x_z^{n-1} to $x(t_{n-1}; z, t_n)$ (cf. Figure II.4.2) by applying an arbitrary but fixed ODE-solver such as e.g. the explicit Euler scheme to the characteristic equation with $(x^*, t^*) = (z, t_n)$. We assume that the time-step τ_n and the ODE-solver are chosen such that x_z^{n-1} lies within $\Omega \cup \Gamma$ for every $n \in \{1, \dots, N_{\mathcal{I}}\}$ and every $z \in \mathcal{N}_{n,\Omega}$. The assumptions on the convection $\mathbf{a}$ in particular imply that this condition is satisfied for a single explicit Euler step if $\tau_n < 1/\|\mathbf{a}(\cdot, t_n)\|_{L^{1,\infty}(\Omega)}$. Denote by $\pi_n : L^2(\Omega) \to X_n$ a suitable quasi-interpolation operator, e.g. the L^2-projection. Then the method of characteristics takes the form:

Set

$$u_{\mathcal{T}_0}^0 = \pi_0 u_0.$$

For $n = 1, \dots, N_{\mathcal{I}}$ successively compute $\widetilde{u}_{\mathcal{T}_n}^{n-1} \in X_n$ such that

$$\widetilde{u}_{\mathcal{T}_n}^{n-1}(z) = \begin{cases} u_{\mathcal{T}_{n-1}}^{n-1}(x_z^{n-1}) & \text{if } z \in \mathcal{N}_{n,\Omega}, \\ 0 & \text{if } z \in \mathcal{N}_{n,\Gamma}, \end{cases}$$

and find $u_{\mathcal{T}_n}^n \in X_n$ such that

$$\int_\Omega \frac{1}{\tau_n}\big(u_{\mathcal{T}_n}^n - \widetilde{u}_{\mathcal{T}_n}^{n-1}\big) v_{\mathcal{T}_n} + \int_\Omega \nabla u_{\mathcal{T}_n}^n \cdot A^n \cdot \nabla v_{\mathcal{T}_n} + \int_\Omega \alpha^n u_{\mathcal{T}_n}^n v_{\mathcal{T}_n} = \int_\Omega f^n v_{\mathcal{T}_n}$$

holds for all $v_{\mathcal{T}_n} \in X_n$.

II.4.11. Finite volume methods. Finite volume methods are a different popular approach for solving parabolic problems in particular those with large convection. For this type of discretizations, the theory of a posteriori error estimation and adaptivity is much less developed than for finite element methods. Yet, there is an important particular case where finite volume methods can easily profit from finite element techniques. This is the case of so-called *dual finite volume meshes*.

II.4.11.1. *Systems in divergence form.* Finite volume methods are tailored for *systems in divergence form* where we are looking for a vector field $\mathbf{U}$ defined on a subset Ω of $\mathbb{R}^d$ having values in $\mathbb{R}^m$ which satisfies the differential equation

$$\begin{aligned} \frac{\partial \mathbf{M}(\mathbf{U})}{\partial t} + \operatorname{div} \underline{\mathbf{F}}(\mathbf{U}) &= \mathbf{g}(\mathbf{U}, x, t) \quad \text{in } \Omega \times (0, \infty) \\ \mathbf{U}(\cdot, 0) &= \mathbf{U}_0 \qquad\quad \text{in } \Omega. \end{aligned}$$

Here, $\mathbf{g}$, the *source*, is a vector field on $\mathbb{R}^m \times \Omega \times (0, \infty)$ with values in $\mathbb{R}^m$, $\mathbf{M}$, the *mass*, is a vector field on $\mathbb{R}^m$ with values in $\mathbb{R}^m$, $\underline{\mathbf{F}}$ the *flux* is a matrix valued function on $\mathbb{R}^m$ with values in $\mathbb{R}^{m \times d}$ and $\mathbf{U}_0$, the *initial value*, is a vector field on Ω with values in $\mathbb{R}^m$. The differential equation of course has to be completed with suitable boundary conditions. These, however, will be ignored in what follows.

Notice that the divergence has to be taken row-wise

$$\operatorname{div} \underline{\mathbf{F}}(\mathbf{U}) = \Big(\sum_{j=1}^{d} \frac{\partial \underline{\mathbf{F}}(\mathbf{U})_{i,j}}{\partial x_j} \Big)_{1 \le i \le m}.$$

The flux $\underline{\mathbf{F}}$ can be slit into two contributions

$$\underline{\mathbf{F}} = \underline{\mathbf{F}}_{\text{adv}} + \underline{\mathbf{F}}_{\text{visc}}.$$

$\underline{\mathbf{F}}_{\text{adv}}$is called *advective flux* and does not contain any derivatives. $\underline{\mathbf{F}}_{\text{visc}}$ is called *viscous flux* and contains spatial derivatives. The advective flux models transport or convection phenomena while the viscous flux is responsible for diffusion phenomena.

Example II.4.1. A linear parabolic equation of 2nd order

$$\frac{\partial u}{\partial t} - \operatorname{div}(A \nabla u) + \mathbf{a} \cdot \nabla u + \alpha u = f,$$

is a system in divergence form with

$$\begin{gathered} m = 1, \qquad \mathbf{U} = u, \qquad \mathbf{M}(\mathbf{U}) = u, \\ \underline{\mathbf{F}}_{\text{adv}}(\mathbf{U}) = \mathbf{a}u, \quad \underline{\mathbf{F}}_{\text{visc}}(\mathbf{U}) = -A \nabla u, \quad \mathbf{g}(\mathbf{U}) = f - \alpha u + (\operatorname{div} \mathbf{a}) u. \end{gathered}$$

Example II.4.2. *Burger's equation*

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0$$

is a system in divergence form with

$$m = d = 1, \qquad \mathbf{u} = u, \qquad \mathbf{M}(\mathbf{U}) = u,$$
$$\underline{\mathbf{F}}_{\text{adv}}(u) = \frac{1}{2}u^2, \qquad \underline{\mathbf{F}}_{\text{visc}}(\mathbf{U}) = 0, \qquad \mathbf{g}(\mathbf{U}) = 0.$$

Other important examples of systems in divergence form are the *Euler equations* and *Navier-Stokes equations* for non-viscous respective viscous fluids. Here we have $d = 2$ or $d = 3$ and $m = d+2$. The vector $\mathbf{U}$ consists of the density, velocity and the internal energy of the fluid.

II.4.11.2. *Basic idea of the finite volume method.* Choose a time step $\tau > 0$ and a partition $\mathcal{T}$ of Ω consisting of arbitrary non-overlapping polyhedra. Here, the elements may have more complicated shapes than in the finite element method (cf. Figures II.4.3 (p. 84) and II.4.4 (p. 84)). Moreover, hanging nodes are allowed.

Now we choose an integer $n \geq 1$ and an element $K \in \mathcal{T}$ and keep both fixed in what follows. First we integrate the differential equation on $K \times [(n-1)\tau, n\tau]$

$$\int_{(n-1)\tau}^{n\tau}\int_K \frac{\partial \mathbf{M}(\mathbf{U})}{\partial t}dxdt + \int_{(n-1)\tau}^{n\tau}\int_K \operatorname{div}\underline{\mathbf{F}}(\mathbf{U})dxdt = \int_{(n-1)\tau}^{n\tau}\int_K \mathbf{g}(\mathbf{U},x,t)dxdt.$$

Next we use integration by parts for the terms on the left-hand side

$$\int_{(n-1)\tau}^{n\tau}\int_K \frac{\partial \mathbf{M}(\mathbf{U})}{\partial t}dxdt = \int_K \mathbf{M}(\mathbf{U}(x,n\tau))dx - \int_K \mathbf{M}(\mathbf{U}(x,(n-1)\tau))dx,$$
$$\int_{(n-1)\tau}^{n\tau}\int_K \operatorname{div}\underline{\mathbf{F}}(\mathbf{U})dxdt = \int_{(n-1)\tau}^{n\tau}\int_{\partial K} \underline{\mathbf{F}}(\mathbf{U})\cdot\mathbf{n}_K dSdt.$$

For the following steps we assume that $\mathbf{U}$ is piecewise constant with respect to space and time. We denote by $\mathbf{U}_K^n$ and $\mathbf{U}_K^{n-1}$ the value of $\mathbf{U}$ on K at times $n\tau$ und $(n-1)\tau$, respectively. Then we have

$$\int_K \mathbf{M}(\mathbf{U}(x,n\tau))dx \approx |K|\mathbf{M}(\mathbf{U}_K^n)$$
$$\int_K \mathbf{M}(\mathbf{U}(x,(n-1)\tau))dx \approx |K|\mathbf{M}(\mathbf{U}_K^{n-1})$$
$$\int_{(n-1)\tau}^{n\tau}\int_{\partial K} \underline{\mathbf{F}}(\mathbf{U})\cdot\mathbf{n}_K dSdt \approx \tau\int_{\partial K}\underline{\mathbf{F}}(\mathbf{U}_K^{n-1})\cdot\mathbf{n}_K dS$$
$$\int_{(n-1)\tau}^{n\tau}\int_K \mathbf{g}(\mathbf{U},x,t)dxdt \approx \tau|K|\mathbf{g}(\mathbf{U}_K^{n-1},x_K,(n-1)\tau).$$

Here, $|K|$ denotes the area of K, if $d = 2$, or the volume of K, if $d = 3$, respectively.

In a last step we approximate the boundary integral for the flux by a *numerical flux*

$$\begin{aligned}\tau \int_{\partial K} \underline{\mathbf{F}}(\mathbf{U}_K^{n-1}) \cdot \mathbf{n}_K d \\ \approx \tau \sum_{\substack{K' \in \mathcal{T} \\ \partial K \cap \partial K' \in \mathcal{E}}} |\partial K \cap \partial K'| \mathbf{F}_{\mathcal{T}}(\mathbf{U}_K^{n-1}, \mathbf{U}_{K'}^{n-1}).\end{aligned}$$

All together we obtain the following *finite volume method*

For every element $K \in \mathcal{T}$ compute

$$\mathbf{U}_K^0 = \frac{1}{|K|} \int_K \mathbf{U}_0(x).$$

For $n = 1, 2, \ldots$ successively compute for every element $K \in \mathcal{T}$

$$\begin{aligned}\mathbf{M}(\mathbf{U}_K^n) = \mathbf{M}(\mathbf{U}_K^{n-1}) \\ - \tau \sum_{\substack{K' \in \mathcal{T} \\ \partial K \cap \partial K' \in \mathcal{E}}} \frac{|\partial K \cap \partial K'|}{|K|} \mathbf{F}_{\mathcal{T}}(\mathbf{U}_K^{n-1}, \mathbf{U}_{K'}^{n-1}) \\ + \tau \mathbf{g}(\mathbf{U}_K^{n-1}, x_K, (n-1)\tau).\end{aligned}$$

Here, $|\partial K \cap \partial K'|$ denotes the length respective area of the common boundary of $K \cap K'$.

This method may easily be modified as follows:

- The time step may be variable.
- The partition of Ω may change from one time step to the other.
- The approximation $\mathbf{U}_K^n$ must not be piecewise constant.

In order to obtain an operating discretization, we still have to make precise the following points:

- construction of $\mathcal{T}$,
- choice of $\underline{\mathbf{F}}_{\mathcal{T}}$.

Moreover we have to take into account boundary conditions. This item, however, will not be addressed in what follows.

II.4.11.3. *Construction of dual finite volume meshes.* For constructing the finite volume mesh $\mathcal{T}$, we start from a standard finite element partition $\widetilde{\mathcal{T}}$ which satisfies the conditions of Section I.2.7 (p. 15). Then we subdivide each element $\widetilde{K} \in \widetilde{\mathcal{T}}$ into smaller elements by either

- drawing the perpendicular bisectors at the midpoints of edges of $\widetilde{K}$ (cf. Figure II.4.3) or by
- connecting the barycentre of $\widetilde{K}$ with its midpoints of edges (cf. Figure II.4.4).

Then the elements in $\mathcal{T}$ consist of the unions of all small elements that share a common vertex in the partition $\widetilde{\mathcal{T}}$.

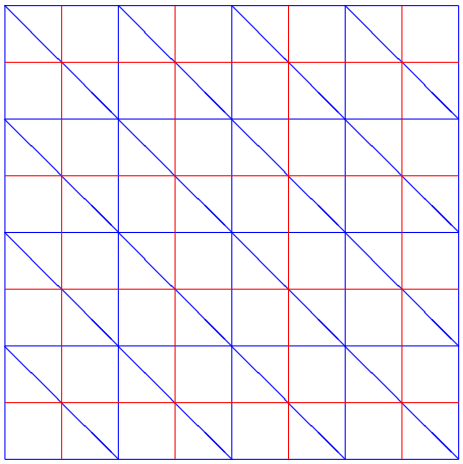

FIGURE II.4.3. Dual mesh (red) via perpendicular bisectors of primal mesh (blue)

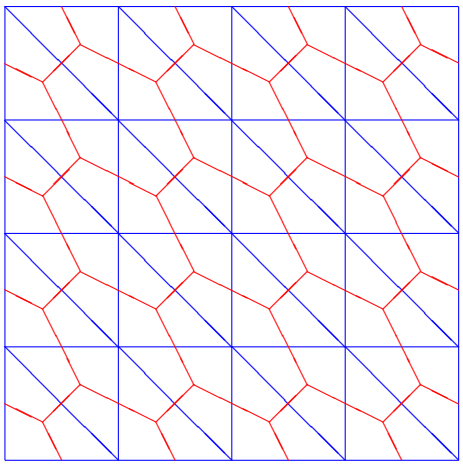

FIGURE II.4.4. Dual mesh (red) via barycentres of primal mesh (blue)

Thus the elements in $\mathcal{T}$ can be associated with the vertices in $\widetilde{\mathcal{N}}$. Moreover, we may associate with each edge or face in $\mathcal{E}$ exactly two vertices in $\widetilde{\mathcal{N}}$ such that the line connecting these vertices intersects the given edge or face, respectively.

The first construction has the advantage that this intersection is orthogonal. But this construction also has some disadvantages which are not present with the second construction:

- The perpendicular bisectors of a triangle may intersect in a point outside the triangle. The intersection point is within the triangle only if its largest angle is at most a right one.

- The perpendicular bisectors of a quadrilateral may not intersect at all. They intersect in a common point inside the quadrilateral only if it is a rectangle.
- The first construction has no three dimensional analogue.

II.4.11.4. *Construction of numerical fluxes.* For the construction of numerical fluxes we assume that $\mathcal{T}$ is a dual mesh corresponding to a primal finite element partition $\widetilde{\mathcal{T}}$. With every edge or face E of $\mathcal{T}$ we denote by K_1 and K_2 the adjacent volumes, by $\mathbf{U}_1$ and $\mathbf{U}_2$ the values $\mathbf{U}_{K_1}^{n-1}$ and $\mathbf{U}_{K_2}^{n-1}$, respectively and by x_1, x_2 vertices of $\widetilde{\mathcal{T}}$such that the segment $\overline{x_1\, x_2}$ intersects E.

As in the analytical case, we split the numerical flux $\underline{\mathbf{F}}_{\mathcal{T}}(\mathbf{U}_1, \mathbf{U}_2)$ into a *viscous numerical flux* $\underline{\mathbf{F}}_{\mathcal{T},\mathrm{visc}}(\mathbf{U}_1, \mathbf{U}_2)$ and an *advective numerical flux* $\underline{\mathbf{F}}_{\mathcal{T},\mathrm{adv}}(\mathbf{U}_1, \mathbf{U}_2)$ which are constructed separately.

We first construct the *numerical viscous fluxes*. To this end we introduce a local co-ordinate system $\eta_1, \dots, \eta_d$ such that η_1 is parallel to $\overline{x_1\, x_2}$ and such that the remaining co-ordinates are tangential to E (cf. Figure II.4.5). Next we express all derivatives in $\underline{\mathbf{F}}_{\mathrm{visc}}$ in terms of partial derivatives corresponding to the new co-ordinates and suppress all derivatives which do not pertain to η_1. Finally we approximate derivatives corresponding to η_1 by differences of the form $\frac{\varphi_1-\varphi_2}{|x_1-x_2|}$.

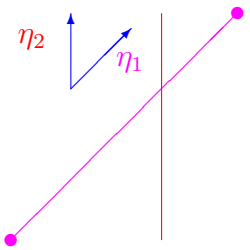

FIGURE II.4.5. Local co-ordinate system for the approximation of viscous fluxes

We now construct the *numerical advective fluxes*. To this end we denote by

$$C(\mathbf{V}) = D(\underline{\mathbf{F}}_{\mathrm{adv}}(\mathbf{V}) \cdot \mathbf{n}_{K_1}) \in \mathbb{R}^{m\times m}$$

the derivative of $\underline{\mathbf{F}}_{\mathrm{adv}}(\mathbf{V}) \cdot \mathbf{n}_{K_1}$ with respect to $\mathbf{V}$ and suppose that this matrix can be diagonalized, i.e., there is an invertible matrix $Q(\mathbf{V}) \in \mathbb{R}^{m\times m}$ and a diagonal matrix $\Delta(\mathbf{V}) \in \mathbb{R}^{m\times m}$ such that

$$Q(\mathbf{V})^{-1} C(\mathbf{V}) Q(\mathbf{V}) = \Delta(\mathbf{V}).$$

This assumption is, e.g., satisfied for the Euler and Navier-Stokes equations. With any real number z we then associate its positive and negative part

$$z^+ = \max\{z, 0\}, \quad z^- = \min\{z, 0\}$$

and set

$$\Delta(\mathbf{V})^{\pm} = \operatorname{diag}\big(\Delta(\mathbf{V})^{\pm}_{11}, \dots, \Delta(\mathbf{V})^{\pm}_{mm}\big),$$
$$C(\mathbf{V})^{\pm} = Q(\mathbf{V})\Delta(\mathbf{V})^{\pm}Q(\mathbf{V})^{-1}.$$

With these notations the *Steger-Warming scheme* for the approximation of advective fluxes is given by

$$\mathbf{F}_{\mathcal{T},\mathrm{adv}}(\mathbf{U}_1, \mathbf{U}_2) = C(\mathbf{U}_1)^{+}\mathbf{U}_1 + C(\mathbf{U}_2)^{-}\mathbf{U}_2.$$

A better approximation is the *van Leer scheme*

$$\begin{aligned}&\mathbf{F}_{\mathcal{T},\mathrm{adv}}(\mathbf{U}_1, \mathbf{U}_2)\\ &= \Big[\frac{1}{2}C(\mathbf{U}_1) + C(\frac{1}{2}(\mathbf{U}_1+\mathbf{U}_2))^{+} - C(\frac{1}{2}(\mathbf{U}_1+\mathbf{U}_2))^{-}\Big]\mathbf{U}_1\\ &\quad + \Big[\frac{1}{2}C(\mathbf{U}_2) - C(\frac{1}{2}(\mathbf{U}_1+\mathbf{U}_2))^{+} + C(\frac{1}{2}(\mathbf{U}_1+\mathbf{U}_2))^{-}\Big]\mathbf{U}_2.\end{aligned}$$

Both approaches require the computation of $D\underline{\mathbf{F}}_{\mathrm{adv}}(\mathbf{V}) \cdot \mathbf{n}_{K_1}$ together with its eigenvalues and eigenvectors for suitable values of $\mathbf{V}$. In general the van Leer scheme is more costly than the Steger-Warming scheme since it requires three evaluations of $C(\mathbf{V})$ instead of two. For the Euler and Navier-Stokes equations, however, this extra cost can be avoided profiting from the particular structure $\underline{\mathbf{F}}_{\mathrm{adv}}(\mathbf{V})\cdot\mathbf{n}_{K_1} = C(\mathbf{V})\mathbf{V}$ of these equations.

Example II.4.3. When applied to Burger's equation of Example II.4.2 (p. 81) the Steger-Warming scheme takes the form

$$\underline{\mathbf{F}}_{\mathcal{T},\mathrm{adv}}(u_1, u_2) = \begin{cases} u_1^2 & \text{if } u_1 \geq 0, u_2 \geq 0\\ u_1^2 + u_2^2 & \text{if } u_1 \geq 0, u_2 \leq 0\\ u_2^2 & \text{if } u_1 \leq 0, u_2 \leq 0\\ 0 & \text{if } u_1 \leq 0, u_2 \geq 0\end{cases}$$

while the van Leer scheme reads

$$\underline{\mathbf{F}}_{\mathcal{T},\mathrm{adv}}(u_1, u_2) = \begin{cases} u_1^2 & \text{if } u_1 \geq -u_2\\ u_2^2 & \text{if } u_1 \leq -u_2.\end{cases}$$

II.4.11.5. *Relation to finite element methods.* The fact that the elements of a dual mesh can be associated with the vertices of a finite element partition gives a link between finite volume and finite element methods:

> Consider a function φ that is piecewise constant on the dual mesh $\mathcal{T}$, i.e. $\varphi \in S^{0,-1}(\mathcal{T})$. With φ we associate a continuous piecewise linear function $\Phi \in S^{1,0}(\widetilde{\mathcal{T}})$ corresponding to the finite element partition $\widetilde{\mathcal{T}}$ such that

$\Phi(x_K) = \varphi_K$ for the vertex $x_K \in \mathcal{N}_{\widetilde{\mathcal{T}}}$ corresponding to $K \in \mathcal{T}$.

This link considerably simplifies the analysis of finite volume methods and suggests a very simple and natural approach to a posteriori error estimation and mesh adaptivity for finite volume methods:

- Given the solution φ of the finite volume scheme compute the corresponding finite element function Φ.
- Apply a standard a posteriori error estimator to Φ.
- Given the error estimator apply a standard mesh refinement strategy to the finite element mesh $\widetilde{\mathcal{T}}$ and thus construct a new, locally refined partition $\widehat{\mathcal{T}}$.
- Use $\widehat{\mathcal{T}}$ to construct a new dual mesh $\mathcal{T}'$. This is the refinement of $\mathcal{T}$.

II.4.12. Discontinuous Galerkin methods. These methods can be interpreted as a mixture of finite element and finite volume methods. The basic idea of discontinuous Galerkin methods can be described as follows:

- Approximate $\mathbf{U}$ by discontinuous functions which are polynomials with respect to space and time on small space-time cylinders of the form $K \times [(n-1)\tau, n\tau]$ with $K \in \mathcal{T}$.
- For every such cylinder multiply the differential equation by a corresponding test-polynomial and integrate the result over the cylinder.
- Use integration by parts for the flux term.
- Accumulate the contributions of all elements in $\mathcal{T}$.
- Compensate for the illegal integration by parts by adding appropriate jump-terms across the element boundaries.
- Stabilize the scheme in a Petrov-Galerkin way by adding suitable element residuals.

In their simplest form these ideas lead to the following discrete problem:

Compute $\mathbf{U}_{\mathcal{T}}^0$, the L^2-projection of $\mathbf{U}_0$ onto $S^{k,-1}(\mathcal{T})$.
For $n \geq 1$ successively find $\mathbf{U}_{\mathcal{T}}^n \in S^{k,-1}(\mathcal{T})$ such that

$$\begin{aligned}
&\sum_{K\in\mathcal{T}} \frac{1}{\tau} \int_K M(\mathbf{U}_{\mathcal{T}}^n) \cdot \mathbf{V}_{\mathcal{T}} - \sum_{K\in\mathcal{T}} \int_K \underline{\mathbf{F}}(\mathbf{U}_{\mathcal{T}}^n) : \nabla \mathbf{V}_{\mathcal{T}} \\
&+ \sum_{E\in\mathcal{E}} \delta_E h_E \int_E \mathbb{J}_E(\mathbf{n}_E \cdot \underline{\mathbf{F}}(\mathbf{U}_{\mathcal{T}}^n) \mathbf{V}_{\mathcal{T}}) \\
&+ \sum_{K\in\mathcal{T}} \delta_K h_K^2 \int_K \operatorname{div} \underline{\mathbf{F}}(\mathbf{U}_{\mathcal{T}}^n) \cdot \operatorname{div} \underline{\mathbf{F}}(\mathbf{V}_{\mathcal{T}}) \\
&= \sum_{K\in\mathcal{T}} \frac{1}{\tau} \int_K M(\mathbf{U}_{\mathcal{T}}^{n-1}) \cdot \mathbf{V}_{\mathcal{T}} + \sum_{K\in\mathcal{T}} \int_K \mathbf{g}(\cdot, n\tau) \cdot \mathbf{V}_{\mathcal{T}}
\end{aligned}$$

$$+ \sum_{K \in \mathcal{T}} \delta_K h_K^2 \int_K \mathbf{g}(\cdot, n\tau) \cdot \operatorname{div} \underline{\mathbf{F}}(\mathbf{V}_\mathcal{T})$$

holds for all $\mathbf{V}_\mathcal{T}$.

This discretization can easily be generalized as follows:

- The jump and stabilization terms can be chosen more judiciously.
- The time-step may not be constant.
- The spatial mesh may depend on time.
- The functions $\mathbf{U}_\mathcal{T}$ and $\mathbf{V}_\mathcal{T}$ may be piecewise polynomials of higher order with respect to to time. Then the term

$$\sum_{K \in \mathcal{T}} \int_{(n-1)\tau}^{n\tau} \int_K \frac{\partial M(\mathbf{U}_\mathcal{T})}{\partial t} \cdot \mathbf{V}_\mathcal{T}$$

must be added on the left-hand side and terms of the form

$$\frac{\partial M(\mathbf{U}_\mathcal{T})}{\partial t} \cdot \mathbf{V}_\mathcal{T}$$

must be added to the element residuals.

CHAPTER III

Implementation

III.1. Mesh-refinement techniques

For step (4) of the adaptive algorithm I.1.1 (p. 7) we must provide a device that constructs the next mesh $\mathcal{T}_{k+1}$ from the current mesh $\mathcal{T}_k$ disposing of an error estimate η_K for every element $K \in \mathcal{T}_k$. This requires two key-ingredients:

- a *marking strategy* that decides which elements should be refined and
- *refinement rules* which determine the actual subdivision of a single element.

Since we want to ensure the admissibility of the next partition, we have to avoid *hanging nodes* (cf. Figure III.1.1). Therefore, the refinement process will proceed in two stages:

- In the first stage we determine a subset $\widetilde{\mathcal{T}}_k$ of $\mathcal{T}_k$ consisting of all those elements that must be refined due to a too large value of η_K. The refinement of these elements usually is called *regular*.
- In the second stage additional elements are refined in order to eliminate the hanging nodes which may be created during the first stage. The refinement of these additional elements is sometimes referred to as *irregular*.

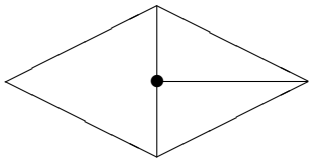

FIGURE III.1.1. Hanging node •

III.1.1. Marking strategies. There are two popular marking strategies for determining the set $\widetilde{\mathcal{T}}_k$: the *maximum strategy* and the *equilibration strategy*.

ALGORITHM III.1.1. (Maximum strategy)

(0) *Given: a partition* $\mathcal{T}$*, error estimates* η_K *for the elements* $K \in \mathcal{T}$*, and a threshold* $\theta \in (0,1)$*.*
Sought: a subset $\widetilde{\mathcal{T}}$ *of* marked *elements that should be refined.*

(1) *Compute*

$$\eta_{\mathcal{T},\max} = \max_{K\in\mathcal{T}} \eta_K.$$

(2) *If*

$$\eta_K \geq \theta \eta_{\mathcal{T},\max}$$

mark K for refinement and put it into the set $\widetilde{\mathcal{T}}$.

Algorithm III.1.2. (Equilibration strategy)

(0) *Given: a partition $\mathcal{T}$, error estimates η_K for the elements $K \in \mathcal{T}$, and a threshold $\theta \in (0,1)$.*
Sought: a subset $\widetilde{\mathcal{T}}$ of marked *elements that should be refined.*

(1) *Compute*

$$\Theta_{\mathcal{T}} = \sum_{K\in\mathcal{T}} \eta_K^2.$$

Set

$$\Sigma_{\mathcal{T}} = 0 \quad \textit{and} \quad \widetilde{\mathcal{T}} = \emptyset.$$

(2) *If*

$$\Sigma_{\mathcal{T}} \geq \theta \Theta_{\mathcal{T}}$$

return $\widetilde{\mathcal{T}}$; stop. Otherwise go to step 3.

(3) *Compute*

$$\widetilde{\eta}_{\mathcal{T},\max} = \max_{K\in\mathcal{T}\setminus\widetilde{\mathcal{T}}} \eta_K.$$

(4) *For all elements $K \in \mathcal{T}\setminus\widetilde{\mathcal{T}}$ check whether*

$$\eta_K = \widetilde{\eta}_{\mathcal{T},\max}.$$

If this is the case, put K in $\widetilde{\mathcal{T}}$ and add η_K^2 to $\Sigma_{\mathcal{T}}$. Otherwise skip K.
When all elements have been treated, return to step 2.

At the end of this algorithm the set $\widetilde{\mathcal{T}}$ satisfies

$$\sum_{K\in\widetilde{\mathcal{T}}} \eta_K^2 \geq \theta \sum_{K\in\mathcal{T}} \eta_K^2.$$

Both marking strategies yield comparable results. The maximum strategy obviously is cheaper than the equilibration strategy. In the maximum strategy, a large value of θ leads to small sets $\widetilde{\mathcal{T}}$, i.e. very few elements are marked and a small value of θ leads to large sets $\widetilde{\mathcal{T}}$, i.e. nearly all elements are marked. In the equilibration strategy on the contrary, a small value of θ leads to small sets $\widetilde{\mathcal{T}}$, i.e. very few elements are marked and a large value of θ leads to large sets $\widetilde{\mathcal{T}}$, i.e. nearly all elements are marked. A popular and well established choice for both strategies is $\theta \approx 0.5$.

In many applications, one encounters the difficulty that very few elements have an extremely large estimated error, whereas the remaining ones split into the vast majority with an extremely small estimated

error and a third group of medium size consisting of elements which have an estimated error much less than the error of the elements in the first group and much larger than the error of the elements in the second group. In this situation Algorithms III.1.1 and III.1.2 will only refine the elements of the first group. This deteriorates the performance of the adaptive algorithm. It can substantially be enhanced by a simple modification:

> Given a small percentage ε, we first mark the $\varepsilon\%$ elements with largest estimated error for refinement and then apply Algorithms III.1.1 and III.1.2 only to the remaining elements.

III.1.2. Regular refinement. Elements that are marked for refinement often are refined by connecting their midpoints of edges. The resulting elements are called *red*.

Triangles and quadrilaterals are thus subdivided into four smaller triangles and quadrilaterals that are similar to the parent element and have the same angles. Thus the shape parameter $\frac{h_K}{\rho_K}$ of the elements does not change.

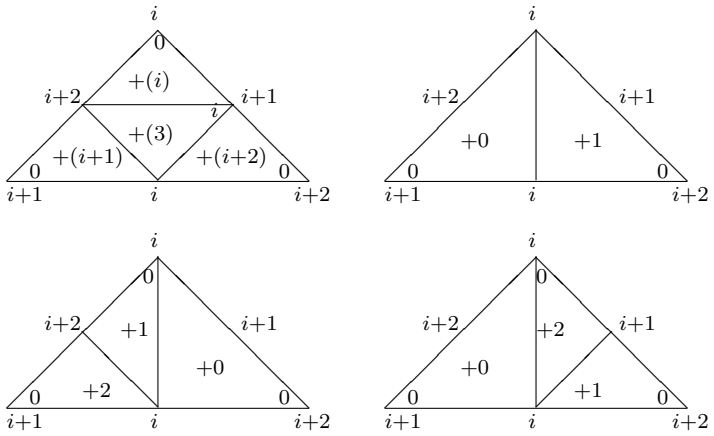

FIGURE III.1.2. Refinement of triangles

This refinement is illustrated by the top-left triangle of Figure III.1.2 and by the top square of Figure III.1.3. The numbers outside the elements indicate the local enumeration of edges and vertices of the parent element. The numbers inside the elements close to the vertices indicate the local enumeration of the vertices of the child elements. The numbers +0, +1 etc. inside the elements give the enumeration of the children.

Note that the enumeration of new elements and new vertices is chosen in such a way that triangles and quadrilaterals may be treated simultaneously without case selections.

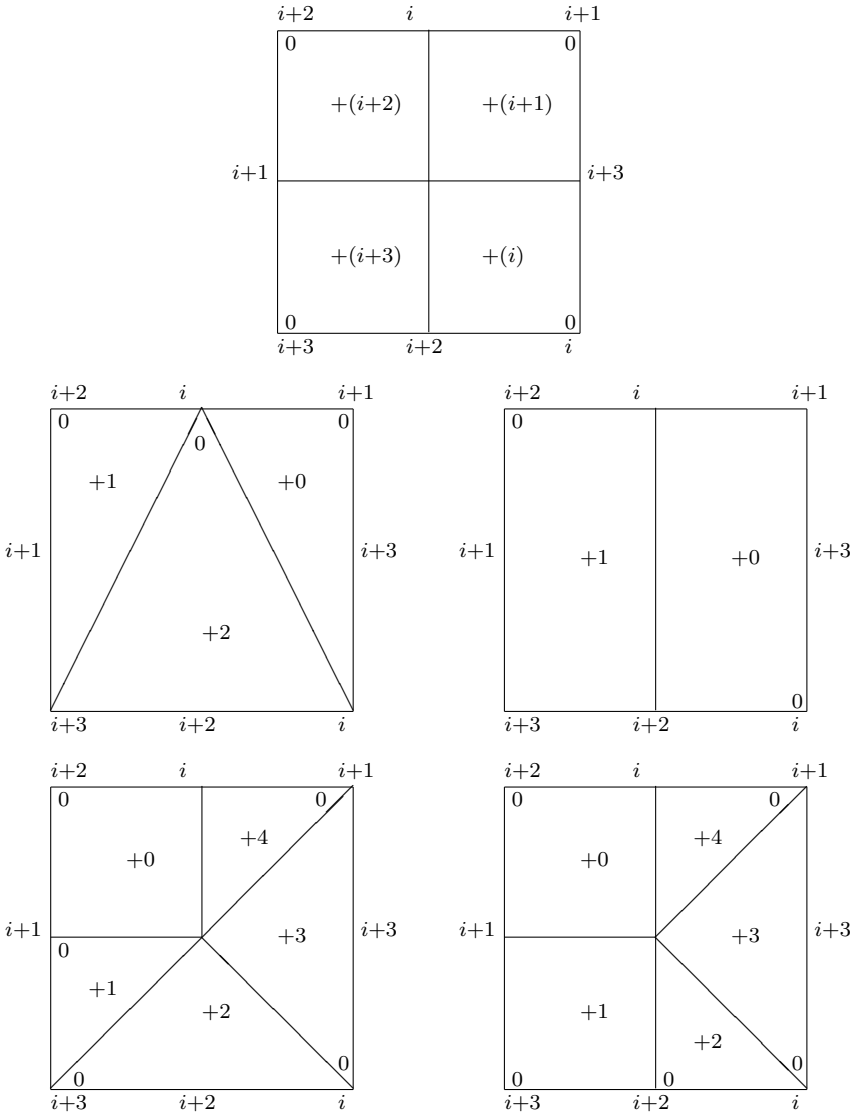

FIGURE III.1.3. Refinement of quadrilaterals

Parallelepipeds are also subdivided into eight smaller similar parallelepipeds by joining the midpoints of edges.

For tetrahedrons, the situation is more complicated. Joining the midpoints of edges introduces four smaller similar tetrahedrons at the vertices of the parent tetrahedron plus a double pyramid in its interior. The latter one is subdivided into four small tetrahedrons by cutting it along two orthogonal planes. These tetrahedrons, however, are not similar to the parent tetrahedron. Still there are rules which determine the cutting planes such that a repeated refinement according to these rules leads to at most four similarity classes of elements originating

from a parent element. Thus these rules guarantee that the shape parameter of the partition does not deteriorate during a repeated adaptive refinement procedure.

III.1.3. Additional refinement. Since not all elements are refined regularly, we need additional refinement rules in order to avoid *hanging nodes* (cf. Figure III.1.1) and to ensure the admissibility of the refined partition. These rules are illustrated in Figures III.1.2 and III.1.3.

For abbreviation we call the resulting elements *green*, *blue*, and *purple*. They are obtained as follows:

- a green element by bisecting exactly one edge,
- a blue element by bisecting exactly two edges,
- a purple quadrilateral by bisecting exactly three edges.

In order to avoid too acute or too abstuse triangles, the blue and green refinement of triangles obey to the following two rules:

- In a blue refinement of a triangle, the longest one of the refinement edges is bisected first.
- Before performing a green refinement of a triangle it is checked whether the refinement edge is part of an edge which has been bisected during the last `ng` generations. If this is the case, a blue refinement is performed instead.

The second rule is illustrated in Figure III.1.4. The cross in the left part represents a hanging node which should be eliminated by a green refinement. The right part shows the blue refinement which is performed instead. Here the cross represents the new hanging node which is created by the blue refinement. Numerical experiments indicate that the optimal value of `ng` is 1. Larger values result in an excessive blow-up of the refinement zone.

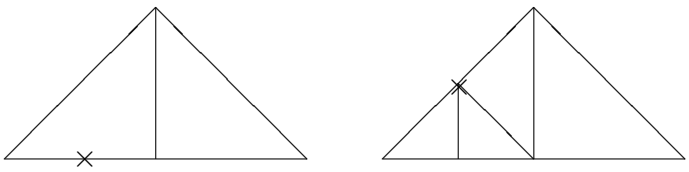

FIGURE III.1.4. Forbidden green refinement and substituting blue refinement

III.1.4. Marked edge bisection. The *marked edge bisection* is an alternative to the described regular red refinement which does not require additional refinement rules for avoiding hanging nodes. It is performed according to the following rules:

- The coarsest mesh $\mathcal{T}_0$ is constructed such that the longest edge of any element is also the longest edge of the adjacent element unless it is a boundary edge.

- The longest edges of the elements in $\mathcal{T}_0$ are marked.
- Given a partition $\mathcal{T}_k$ and an element thereof which should be refined, it is bisected by joining the mid-point of its marked edge with the vertex opposite to this edge.
- When besecting the edge of an element, its two remaining edges become the marked edges of the two resulting new triangles.

This process is illustrated in Figure III.1.5. The marked edges are labeled by •.

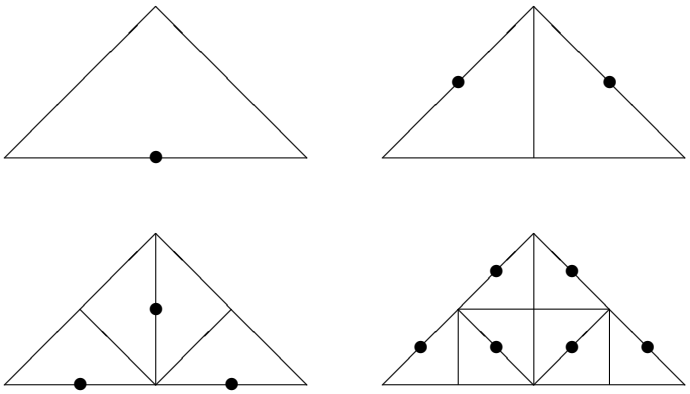

FIGURE III.1.5. Subsequent marked edge bisection, the marked edges are labeled by •

III.1.5. Mesh-coarsening. The adaptive Algorithm I.1.1 (p. 7) in combination with the marking strategies of Algorithms III.1.1 (p. 89) and III.1.2 (p. 90) produces a sequence of increasingly refined partitions. In many situations, however, some elements must be coarsened in the course of the adaptive process. For time dependent problems this is obvious: A critical region, e.g. an interior layer, may move through the spatial domain in the course of time. For stationary problems this is less obvious. Yet, for elliptic problems one can prove that a possible coarsening is mandatory to ensure the optimal complexity of the adaptive process.

The basic idea of the coarsening process is to go back in the hierarchy of partitions and to cluster elements with too small an error. The following algorithm goes m generations backwards, accumulates the error indicators, and then advances $n > m$ generations using the marking strategies of Algorithms III.1.1 (p. 89) and III.1.2 (p. 90). For stationary problems, typical values are $m = 1$ and $n = 2$. For time dependent problems one may choose $m > 1$ and $n > m + 1$ to enhance the temporal movement of the refinement zone.

ALGORITHM III.1.3. *Given: A hierarchy $\mathcal{T}_0, \ldots, \mathcal{T}_k$ of adaptively refined partitions, error indicators η_K for the elements K of $\mathcal{T}_k$, and*

parameters $1 \le m \le k$ *and* $n > m$.
Sought: A new partition $\mathcal{T}_{k-m+n}$.

(1) *For every element* $K \in \mathcal{T}_{k-m}$ *set* $\widetilde{\eta}_K = 0$.
(2) *For every element* $K \in \mathcal{T}_k$ *determine its ancestor* $K' \in \mathcal{T}_{k-m}$ *and add* η_K^2 *to* $\widetilde{\eta}_{K'}^2$.
(3) *Successively apply Algorithms III.1.1 (p. 89) or III.1.2 (p. 90)* n *times with* $\widetilde{\eta}$ *as error indicator. In this process, equally distribute* $\widetilde{\eta}_K$ *over the siblings of* K *once an element* K *is subdivided.*

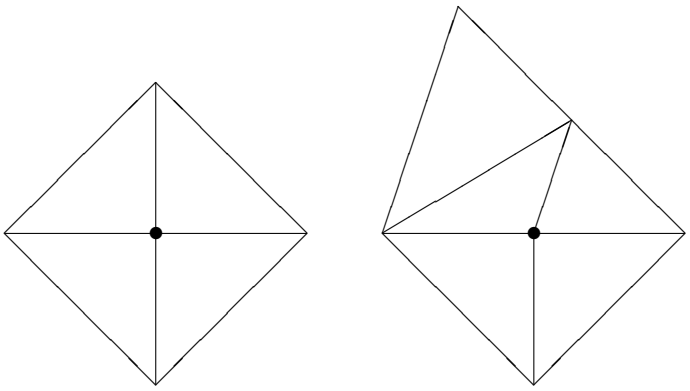

FIGURE III.1.6. The vertex marked • is resolvable in the left patch but not in the right one.

The following algorithm is particularly suited for the marked edge bisection of Section III.1.4. In the framework of Algorithm III.1.3 its parameters are $m = 1$ and $n = 2$, i.e., it constructs the partition of the next level simultaneously refining and coarsening elements of the current partition. For its description we need some notations:

- An element K of the current partition $\mathcal{T}$ has *refinement level* ℓ if it is obtained by subdividing ℓ times an element of the coarsest partition.
- Given a triangle K of the current partition $\mathcal{T}$ which is obtained by bisecting a parent triangle K', the vertex of K which is not a vertex of K' is called the *refinement vertex* of K.
- A vertex $z \in \mathcal{N}$ of the current partition $\mathcal{T}$ and the corresponding patch ω_z are called *resolvable* (cf. Figure III.1.6) if
  - z is the refinement vertex of all elements contained in ω_z and
  - all elements contained in ω_z have the same refinement level.

ALGORITHM III.1.4. *Given: A partition* $\mathcal{T}$, *error indicators* η_K *for all elements* K *of* $\mathcal{T}$, *and parameters* $0 < \theta_1 < \theta_2 < 1$.
Sought: Subsets $\mathcal{T}_c$ *and* $\mathcal{T}_r$ *of elements that should be coarsened and refined, respectively.*

(1) *Set* $\mathcal{T}_c = \emptyset$, $\mathcal{T}_r = \emptyset$ *and compute* $\eta_{\mathcal{T},\max} = \max_{K\in\mathcal{T}} \eta_K$.
(2) *For all* $K \in \mathcal{T}$ *check whether* $\eta_K \geq \theta_2 \eta_{\mathcal{T},\max}$. *If this is the case, put* K *into* $\mathcal{T}_r$.
(3) *For all vertices* $z \in \mathcal{N}$ *check whether* z *is resolvable. If this is the case and if* $\max_{K\subset\omega_z} \eta_K \leq \theta_1 \eta_{\mathcal{T},\max}$, *put all elements contained in* ω_z *into* $\mathcal{T}_c$.

Remark III.1.5. Algorithm III.1.4 obviously is a modification of the maximum strategy of Algorithm III.1.1 (p. 89). A coarsening of elements can also be incorporated in the equilibration strategy of Algorithm III.1.2 (p. 90).

III.1.6. Mesh-smoothing. In this section we describe mesh-smoothing strategies which try to improve the quality of a partition while retaining its topological structure. The vertices of the partition are moved, but the number of elements and their adjacency remain unchanged. All strategies use a process similar to the well-known Gauss-Seidel algorithm to optimize a suitable *quality function* q over the class of all partitions having the same topological structure. They differ in the choice of the quality function. The strategies of this section do not replace the mesh-refinement methods of the previous sections, they complement them. In particular an improved partition may thus be obtained when a further refinement is impossible due to an exhausted storage.

In order to simplify the presentation, we assume throughout this section that all partitions exclusively consist of triangles.

III.1.6.1. *The Optimization Process.* We first describe the optimization process. To this end we assume that we dispose of a quality function q which associates with every element a non-negative number such that a larger value of q indicates a better quality. Given a partition $\mathcal{T}$ we want to find an improved partition $\widetilde{\mathcal{T}}$ with the same number of elements and the same adjacency such that

$$\min_{\widetilde{K}\in\widetilde{\mathcal{T}}} q(\widetilde{K}) > \min_{K\in\mathcal{T}} q(K).$$

To this end we perform several iterations of the following *smoothing procedure* similar to the Gauß-Seidel iteration:

> For every vertex z in the current partition $\mathcal{T}$, fix the vertices of $\partial\omega_z$ and find a new vertex $\widetilde{z}$ inside ω_z such that
>
> $$\min_{\widetilde{K}\subset\omega_{\widetilde{z}}} q(\widetilde{K}) > \min_{K\subset\omega_z} q(K).$$

The practical solution of the local optimization problem depends on the choice of the quality function q. In what follows we will present three possible choices for q.

III.1.6.2. *A Quality Function Based on Geometrical Criteria.* The first choice is purely based on the geometry of the partitions and tries to obtain a partition which consists of equilateral triangles. To describe this approach, we enumerate the vertices and edges of a given triangle consecutively in counter-clockwise order from 0 to 2 such that edge i is opposite to vertex i (cf. Figures III.1.2 (p. 91) and III.1.3 (p. 92)). Then edge i has the vertices $i+1$ and $i+2$ as its endpoints where all expressions have to be taken modulo 3. With these notations we define the *geometric quality function* q_G by

$$q_G(K) = \frac{4\sqrt{3}\mu_2(K)}{\mu_1(E_0)^2 + \mu_1(E_1)^2 + \mu_1(E_2)^2},$$

where $\mu_2(K)$ is the area of K and $\mu_1(E)$ the length of E. The function q_G is normalized such that it attains its maximal value 1 for an equilateral triangle.

To obtain a more explicit representation of q_G and to solve the optimization problem, we denote by $x_0 = (x_{0,1}, x_{0,2})$, $x_1 = (x_{1,1}, x_{1,2})$, and $x_2 = (x_{2,1}, x_{2,2})$ the co-ordinates of the vertices. Then we have

$$\mu_2(K) = \frac{1}{2}\big\{(x_{1,1} - x_{0,1})(x_{2,2} - x_{0,2}) - (x_{2,1} - x_{0,1})(x_{1,2} - x_{0,2})\big\}$$

and

$$\mu_1(E_i)^2 = (x_{i+2,1} - x_{i+1,1})^2 + (x_{i+2,2} - x_{i+1,2})^2$$

for $i = 0, 1, 2$. There are two main possibilities to solve the optimization problem for q_G.

In the first approach, we determine a triangle K_1 in ω_z such that

$$q_G(K_1) = \min_{K \subset \omega_z} q_G(K)$$

and start the enumeration of its vertices at the vertex z. Then we determine a point z' such that the points z', x_1, and x_2 are the vertices of an equilateral triangle and that this enumeration of vertices is in counter-clockwise order. Now, we try to find a point $\widetilde{z}$ approximately solving the optimization problem by a line search on the straight line segment connecting z and z'.

In the second approach, we determine two triangles K_1 and K_2 in ω_z such that

$$q_G(K_1) = \min_{K \subset \omega_z} q_G(K) \quad \text{and} \quad q_G(K_2) = \min_{K \subset \omega_z \setminus K_1} q_G(K).$$

Then, we determine the unique point z' such that the two triangles corresponding to K_1 and K_2 with z replaced by z' have equal qualities q_G. This point can be computed explicitly from the co-ordinates of K_1 and K_2 which remain unchanged. If z' is within ω_z, it is the optimal solution $\widetilde{z}$ of the optimization problem. Otherwise we again try to find $\widetilde{z}$ by a line search on the straight line segment connecting z and z'.

III.1.6.3. *A Quality Function Based on Interpolation.* Our second candidate for a quality function is given by

$$q_I(K) = \|\nabla(u_Q - u_L)\|_K^2,$$

where u_Q and u_L denote the quadratic and linear interpolant, respectively of u. Using the functions ψ_E of Section I.2.12 (p. 22) we have

$$u_Q - u_L = \sum_{i=0}^{2} d_i \psi_{E_i}$$

with

$$d_i = u\left(\frac{1}{2}(x_{i+1} + x_{i+2})\right) - \frac{1}{2}u(x_{i+1}) - \frac{1}{2}u(x_{i+2})$$

for $i = 0, 1, 2$ where again all indices have to be taken modulo 3. Hence, we have

$$q_I(K) = v^t B v \quad \text{with} \quad v = \begin{pmatrix} d_0 \\ d_1 \\ d_2 \end{pmatrix} \quad \text{and} \quad B_{ij} = \int_K \nabla\psi_{E_i} \cdot \nabla\psi_{E_j}$$

for $i, j = 0, 1, 2$. A straightforward calculation yields

$$B_{ii} = \frac{\mu_1(E_0)^2 + \mu_1(E_1)^2 + \mu_1(E_2)^2}{3\mu_2(K)} = \frac{4}{\sqrt{3}}\frac{1}{q_G(K)}$$

for all i and

$$B_{ij} = \frac{2(x_{i+2} - x_{i+1}) \cdot (x_{j+2} - x_{j+1})}{3\mu_2(K)}$$

for $i \neq j$. Since B is spectrally equivalent to its diagonal, we approximate $q_I(K)$ by

$$\widetilde{q}_I(K) = \frac{1}{q_G(K)} \sum_{i=0}^{2} d_i^2.$$

To obtain an explicit representation of $\widetilde{q}_I$ in terms of the geometrical data of K, we assume that the second derivatives of u are constant on K. Denoting by H_K the Hessian matrix of u on K, Taylor's formula then yields

$$d_i = -\frac{1}{8}(x_{i+2} - x_{i+1})^t H_K (x_{i+2} - x_{i+1})$$

for $i = 0, 1, 2$. Hence, with this assumption, $\widetilde{q}_I$ is a rational function with quadratic polynomials in the nominator and denominator. The optimization problem can therefore be solved approximately with a few steps of a damped Newton iteration. Alternatively we may adopt our previous geometrical reasoning with q_G replaced by $\widetilde{q}_I$.

III.1.6.4. *A Quality Function Based on an Error Indicator.* The third choice of a quality function is given by

$$q_E(K) = \int_K \Big| \sum_{i=0}^{2} e_i \nabla \psi_{E_i} \Big|^2,$$

where the coefficients e_0, e_1 and e_2 are computed from an error indicator η_K. Once we dispose of these coefficients, the optimization problem for the function q_E may be solved in the same way as for q_I.

The computation of the coefficients e_0, e_1 and e_2 is particularly simple for the error indicator $\eta_{N,K}$ of Section II.2.1.3 (p. 40) which is based on the solution of local Neumann problems on the elements. Denoting by v_K the solution of the auxiliary problem, we compute the e_i by solving the least-squares problem

$$\text{minimize} \int_K \Big| \nabla v_K - \sum_{i=0}^{2} e_i \nabla \psi_{E_i} \Big|^2.$$

For the error indicator $\eta_{D,K}$ of Section II.2.1.2 (p. 39) which is based on the solution of an auxiliary discrete Dirichlet problem on the patch ω_K, we may proceed in a similar way and simply replace v_K by the restriction to K of $\widetilde{v}_K$, the solution of the auxiliary problem.

For the residual error indicator $\eta_{R,K}$ of Section II.1 (p. 26), finally, we replace the function v_K by

$$h_K \big(\overline{f}_K + \Delta u_{\mathcal{T}} \big) \psi_K - \sum_{i=0}^{2} h_{E_i}^{\frac{1}{2}} \mathbb{J}_E (n_{E_i} \cdot \nabla u_{\mathcal{T}}) \psi_{E_i}$$

with the obvious modifications for edges on the boundary Γ.

III.2. Data structures

In this section we shortly describe the required data structures for a `Java` or `C++` implementation of an adaptive finite element algorithm. For simplicity we consider only the two-dimensional case. Note that the data structures are independent of the particular differential equation and apply to all engineering problems which require the approximate solution of partial differential equations. The described data structures are realized in the `Java` applet `ALF` (Adaptive Linear Finite elements). It is available at the address

`http://www.rub.de/num1/softwareE.html`

together with a user guide in pdf-format.

III.2.1. Nodes. The class `NODE` realizes the concept of a node, i.e., of a vertex of a grid. It has three members `c`, `t`, and `d`.
The member `c` stores the co-ordinates in Euclidean 2-space. It is a double array of length 2.
The member `t` stores the type of the node. It equals 0 if it is an

interior point of the computational domain. It is k, $k > 0$, if the node belongs to the k-th component of the Dirichlet boundary part of the computational domain. It equals $-k$, $k > 0$, if the node is on the k-th component of the Neumann boundary.
The member `d` gives the address of the corresponding degree of freedom. It equals -1 if the corresponding node is not a degree of freedom since, e.g., it lies on the Dirichlet boundary. This member takes into account that not every node actually is a degree of freedom.

III.2.2. Elements. The class `ELEMENT` realizes the concept of an element. Its member `nv` determines the element type, i.e., triangle or quadrilateral. Its members `v` and `e` realize the vertex and edge informations, respectively. Both are integer arrays of length 4.
The vertices are enumerated consecutively in counter-clockwise order, `v`$[i]$ gives the global number of the i-th vertex. It is assumed that `v`$[3] = -1$ if `nv`$= 3$.
The edges are also enumerated consecutively in counter-clockwise order such that the i-th edge has the vertices $i + 1 \bmod$ `nv` and $i + 2 \bmod$ `nv` as its endpoints. Thus, in a triangle, edge i is opposite vertex i.
A value `e`$[i] = -1$ indicates that the corresponding edge is on a straight part of the boundary. Similarly `e`$[i] = -k - 2$, $k \geq 0$, indicates that the endpoints of the corresponding edge are on the k-th curved part of the boundary. A value `e`$[i] = j \geq 0$ indicates that edge i of the current element is adjacent to element number j. Thus the member `e` decribes the neighbourhood relation of elements.
The members `p`, `c`, and `t` realize the grid hierarchy and give the number of the parent, the number of the first child, and the refinement type, respectively. In particular we have

$$\texttt{t} \in \begin{cases} \{0\} & \text{if the element is not refined} \\ \{1, \ldots, 4\} & \text{if the element is refined green} \\ \{5\} & \text{if the element is refined red} \\ \{6, \ldots, 24\} & \text{if the element is refined blue} \\ \{25, \ldots, 100\} & \text{if the element is refined purple.} \end{cases}$$

At first sight it may seem strange to keep the information about nodes and elements in different classes. But this approach has several advantages:

- It minimizes the storage requirement. The co-ordinates of a node must be stored only once. If nodes and elements are represented by a common structure, these co-ordinates are stored $4 - 6$ times.
- The elements represent the topology of the grid which is independent of the particular position of the nodes. If nodes and

elements are represented by different structures it is much easier to implement mesh smoothing algorithms which affect the position of the nodes but do not change the mesh topology.

III.2.3. Grid hierarchy. When creating a hierarchy of adaptively refined grids, the nodes are completely hierarchical, i.e., a node of grid $\mathcal{T}_i$ is also a node of any grid $\mathcal{T}_j$ with $j > i$. Since in general the grids are only partly refined, the elements are not completely hierarchical. Therefore, all elements of all grids are stored.

The information about the different grids is implemented by the class `LEVEL`. Its members `nn`, `nt`, `nq`, and `ne` give the number of nodes, triangles, quadrilaterals, and edges, resp. of a given grid. The members `first` and `last` give the addresses of the first element of the current grid and of the first element of the next grid, respectively. The member `dof` yields the number of degrees of freedom of the corresponding discrete finite element problems.

III.3. Numerical examples

The examples of this section are computed with the demonstration `Java`-applet `ALF` on a MacIntosh G4 powerbook. The linear systems are solved with a multi-grid V-cycle algorithm in Examples III.3.1 – III.3.3 and a multi-grid W-cycle algorithm in Example III.3.4. In Examples III.3.1 and III.3.2 we use one Gauß-Seidel forward sweep for pre-smoothing and one backward Gauß-Seidel sweep for post-smoothing. In Example III.3.3 the smoother is one symmetric Gauß-Seidel sweep for a downwind re-enumeration for the unknowns. In Example III.3.4 we use two steps of a symmetric Gauß-Seidel algorithm for pre- and post-smoothing. Tables III.3.1 – III.3.4 give for all examples the following quantities:

- `L:` the number of refinement levels,
- `NN:` the number of unknowns,
- `NT:` the number of triangles,
- `NQ:` the number of quadrilaterals,
- ε`:` the true relative error $\frac{\|u-u_{\mathcal{T}}\|_{H^1(\Omega)}}{\|u\|_{H^1(\Omega)}}$, if the exact solution is known, and the estimated relative error $\frac{\eta}{\|u_{\mathcal{T}}\|_{H^1(\Omega)}}$, if the exact solution is unknown,
- `q:` the efficiency index $\frac{\eta}{\|u-u_{\mathcal{T}}\|_{H^1(\Omega)}}$ of the error estimator provided the exact solution is known.

Example III.3.1. We consider the Poisson equation

$$\begin{aligned} -\Delta u &= 0 \quad \text{in } \Omega \\ u &= g \quad \text{on } \Gamma \end{aligned}$$

in the L-shaped domain $\Omega = (-1,1)^2 \backslash (0,1) \times (-1,0)$. The boundary data g are chosen such that the exact solution in polar co-ordinates is

$$u = r^{\frac{2}{3}} \sin\left(\frac{3}{2}\pi\varphi\right).$$

The coarsest mesh for a partition into quadrilaterals consists of three squares with sides of length 1. The coarsest triangulation is obtained by dividing each of these squares into two triangles by joining the top-left and bottom-right corner of the square. For both coarsest meshes we first apply a uniform refinement until the storage capacity is exhausted. Then we apply an adaptive refinement strategy based on the residual error estimator $\eta_{R,K}$ of Section II.1.9 (p. 34) and the maximum strategy of Algorithm III.1.1 (p. 89). The refinement process is stopped as soon as we obtain a solution with roughly the same relative error as the solution on the finest uniform mesh. The corresponding numbers are given in Table III.3.1. Figures III.3.1 and III.3.2 show the finest meshes obtained by the adaptive process.

TABLE III.3.1. Comparison of uniform and adaptive refinement for Example III.3.1

	triangles		quadrilaterals	
	uniform	adaptive	uniform	adaptive
L	5	5	5	5
NN	2945	718	2945	405
NT	6144	1508	0	524
NQ	0	0	3072	175
$\varepsilon(\%)$	1.3	1.5	3.6	3.9
q	-	1.23	-	0.605

EXAMPLE III.3.2. Now we consider the reaction-diffusion equation

$$\begin{aligned} -\Delta u + \kappa^2 u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma \end{aligned}$$

in the square $\Omega = (-1,1)^2$. The reaction parameter κ is chosen equal to 100. The right-hand side f is such that the exact solution is

$$u = \tanh\left(\kappa(x^2 + y^2 - \frac{1}{4})\right).$$

It exhibits an interior layer along the boundary of the circle of radius $\frac{1}{2}$ centered at the origin. The coarsest mesh for a partition into squares consists of 4 squares with sides of length 1. The coarsest triangulation again is obtained by dividing each square into two triangles by joining the top-left and right-bottom corners of the square. For the comparison of adaptive and uniform refinement we proceed as in the previous

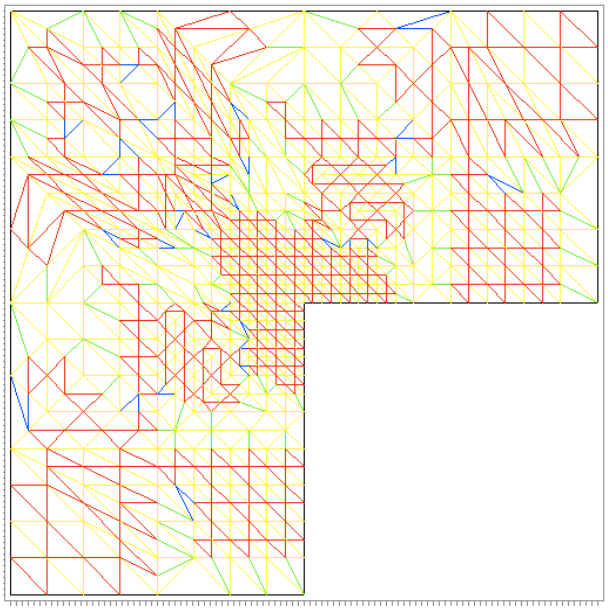

FIGURE III.3.1. Adaptively refined triangulation of level 5 for Example III.3.1

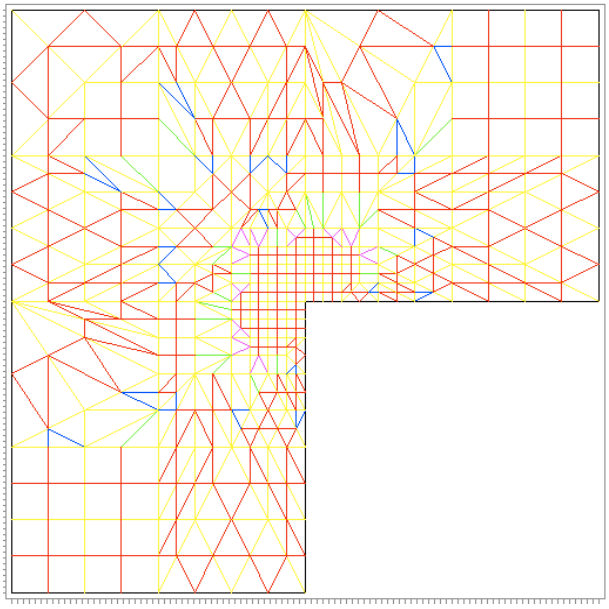

FIGURE III.3.2. Adaptively refined partition into squares of level 5 for Example III.3.1

example. In order to take account of the reaction term, the error estimator now is the modified residual estimator $\eta_{R;K}$ of Section II.3.1.3 (p. 56).

TABLE III.3.2. Comparison of uniform and adaptive refinement for Example III.3.2

	triangles		quadrilaterals	
	uniform	adaptive	uniform	adaptive
L	5	6	5	6
NN	3969	1443	3969	2650
NT	8192	2900	0	1600
NQ	0	0	4096	1857
$\varepsilon(\%)$	3.8	3.5	6.1	4.4
q	-	0.047	-	0.041

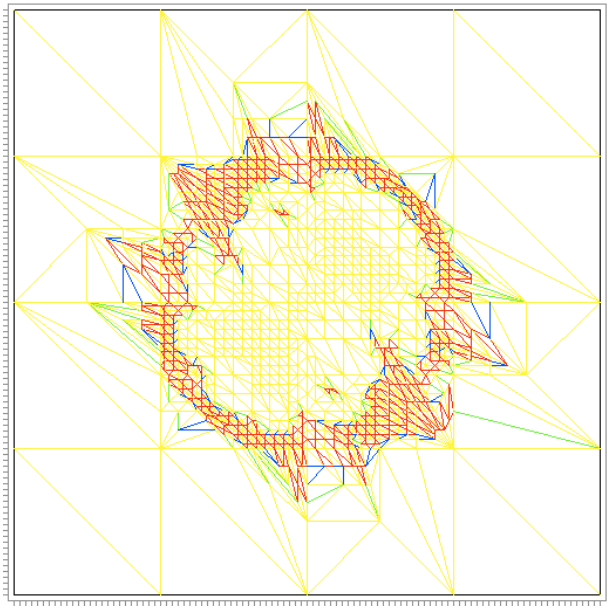

FIGURE III.3.3. Adaptively refined triangulation of level 6 for Example III.3.2

TABLE III.3.3. Comparison of uniform and adaptive refinement for Example III.3.3

	triangles		quadrilaterals	
	adaptive excess 0%	adaptive excess 20%	adaptive excess 0%	adaptive excess 20%
L	8	6	9	7
NN	5472	2945	2613	3237
NT	11102	6014	1749	3053
NQ	0	0	1960	1830
$\varepsilon(\%)$	0.4	0.4	0.6	1.2

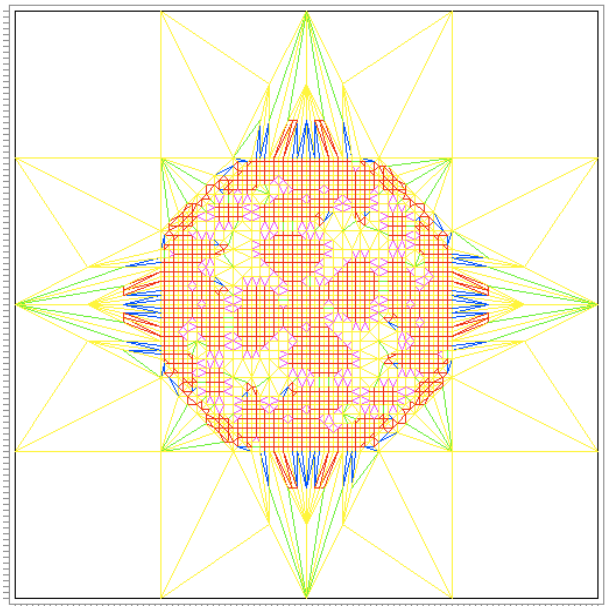

FIGURE III.3.4. Adaptively refined partition into squares of level 6 for Example III.3.2

EXAMPLE III.3.3. Next we consider the convection-diffusion equation

$$\begin{aligned} -\varepsilon\Delta u + \mathbf{a}\cdot\nabla u &= 0 \quad \text{in } \Omega \\ u &= g \quad \text{on } \Gamma \end{aligned}$$

in the square $\Omega = (-1,1)^2$. The diffusion parameter is

$$\varepsilon = \frac{1}{100},$$

the convection is

$$\underline{a} = \begin{pmatrix} 2 \\ 1 \end{pmatrix},$$

and the boundary condition is

$$g = \begin{cases} 0 & \text{on the left and top boundary,} \\ 100 & \text{on the bottom and right boundary.} \end{cases}$$

The exact solution of this problem is unknown, but it is known that it exhibits an exponential boundary layer at the boundary $x = 1$, $y > 0$ and a parabolic interior layer along the line connecting the points $(-1,-1)$ and $(1,0)$. The coarsest meshes are determined as in Example III.3.2. Since the exact solution is unknown, we cannot give the efficiency index $\mathtt{q}$ and perform only an adaptive refinement. The error estimator is the one of Section II.3.1.3 (p. 56). Since the exponential layer is far stronger than the parabolic one, the maximum strategy of

Algorithm III.1.1 (p. 89) leads to a refinement preferably close to the boundary $x = 1$, $y > 0$ and has difficulties in catching the parabolic interior layer. This is in particular demonstrated by Figure III.3.7. We therefor also apply the modified maximum strategy of Section III.3 with an excess ε of 20%, i.e., the 20% elements with largest error are first refined regularly and the maximum strategy is then applied to the remaining elements.

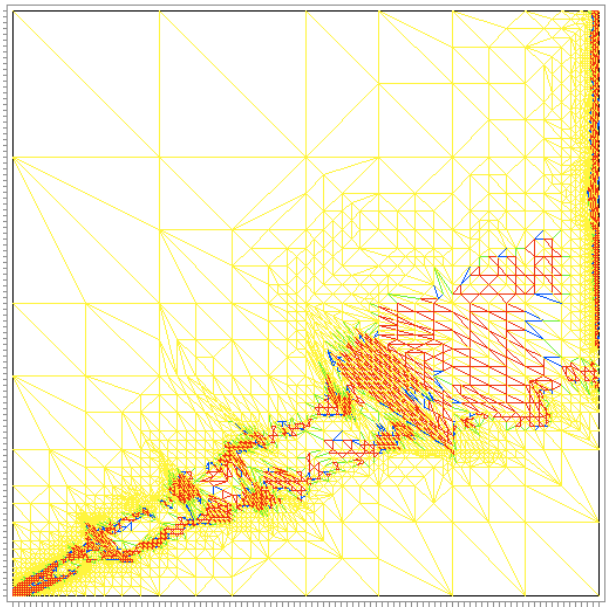

FIGURE III.3.5. Adaptively refined triangulation of Example III.3.3 with refinement based on the maximum strategy

EXAMPLE III.3.4. Finally we consider a diffusion equation

$$\begin{aligned} -\operatorname{div}(A \operatorname{grad} u) &= 1 \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma \end{aligned}$$

in the square $\Omega = (-1,1)^2$ with a discontinuous diffusion

$$A = \begin{cases} \begin{pmatrix} 10 & \frac{90}{11} \\ \frac{90}{11} & 10 \end{pmatrix} & \text{in } 4x^2 + 16y^2 < 1, \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & \text{in } 4x^2 + 16y^2 \geq 1. \end{cases}$$

The exact solution of this problem is not known. Hence we cannot give the efficiency index `q`. The coarsest meshes are as in Examples III.3.2 and III.3.3. The adaptive process is based on the error estimator of Section II.3.1.3 (p. 56) and the maximum strategy of Algorithm III.1.1 (p. 89).

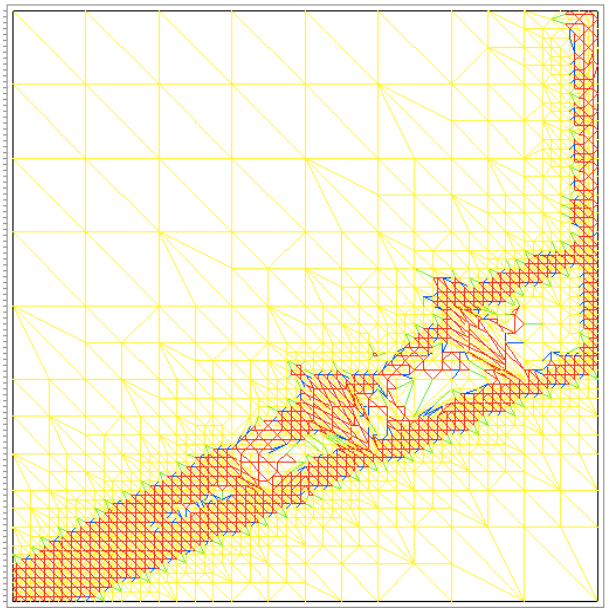

FIGURE III.3.6. Adaptively refined triangulation of Example III.3.3 with refinement based on the modified maximum strategy with excess of 20%

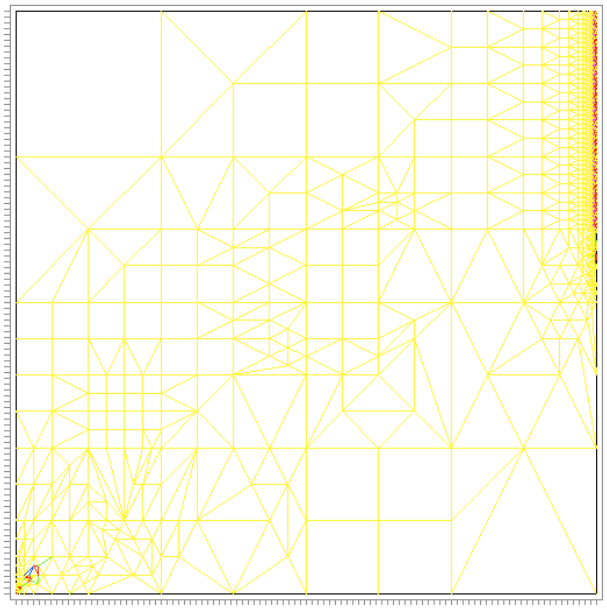

FIGURE III.3.7. Adaptively refined partition into squares of Example III.3.3 with refinement based on the maximum strategy

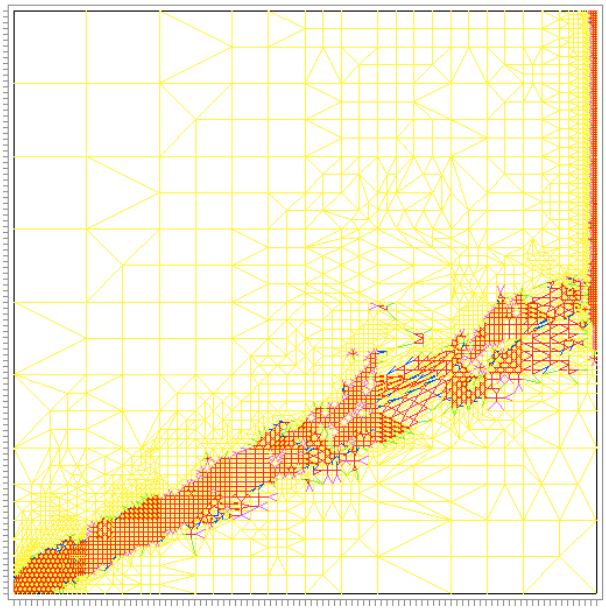

FIGURE III.3.8. Adaptively refined partition into squares of Example III.3.3 with refinement based on the modified maximum strategy with excess of 20%

TABLE III.3.4. Comparison of uniform and adaptive refinement for Example III.3.4

	triangles		quadrilaterals	
	uniform	adaptive	uniform	adaptive
L	5	6	5	6
NN	3969	5459	3969	2870
NT	8192	11128	0	1412
NQ	0	0	4096	2227
$\varepsilon(\%)$	-	2.5	-	14.6

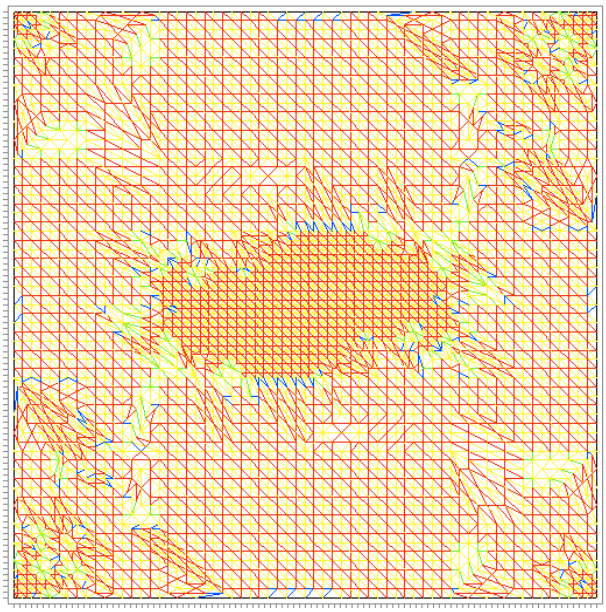

FIGURE III.3.9. Adaptively refined triangulation of Example III.3.4

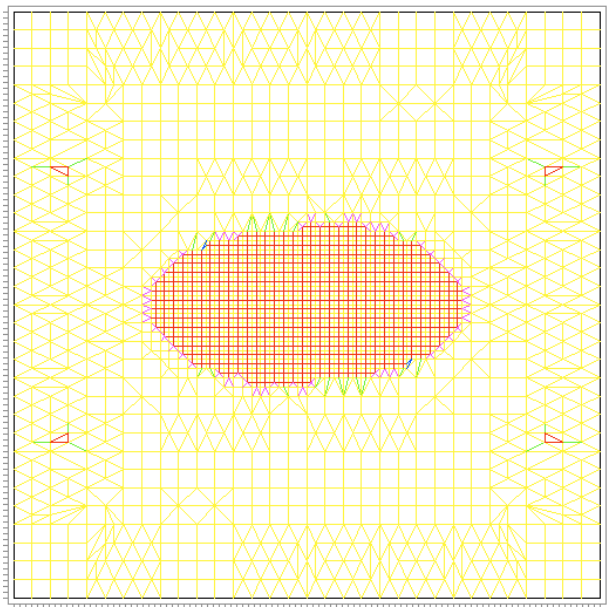

FIGURE III.3.10. Adaptively refined partition into squares of Example III.3.4

CHAPTER IV

Solution of the discrete problems

IV.1. Overview

To get an overview of the particularities of the solution of finite element problems, we consider a simple, but instructive model situation: the model problem of Section II.1.2 (p. 26) on the unit square $(0,1)^2$ ($d=2$) or the unit cube $(0,1)^3$ ($d=3$) discretized by linear elements (Section II.1.3 (p. 26) with $k=1$) on a mesh that consists of squares ($d=2$) or cubes ($d=3$) with edges of length $h=\frac{1}{n}$.

The number of unknowns is

$$N_h = \Big(\frac{1}{n-1}\Big)^d.$$

The stiffness matrix L_h is symmetric positive definite and sparse; every row contains at most 3^d non-zero elements. The total number of non-zero entries in L_h is

$$e_h = 3^d N_h.$$

The ratio of non-zero entries to the total number of entries in L_h is

$$p_h = \frac{e_h}{N_h^2} \approx 3^d N_h^{-1}.$$

The stiffness matrix is a band matrix with bandwidth

$$b_h = h^{-d+1} \approx N_h^{1-\frac{1}{d}}.$$

Therefore the Gaussian elimination, the LR-decomposition or the Cholesky decomposition require

$$s_h = b_h N_h \approx N_h^{2-\frac{1}{d}}$$

bytes for storage and

$$z_h = b_h^2 N_h \approx N_h^{3-\frac{2}{d}}$$

arithmetic operations.

These numbers are collected in Table IV.1.1. It clearly shows that direct methods are not suited for the solution of large finite element problems both with respect to the storage requirement as with respect to the computational work. Therefore one usually uses iterative methods for the solution of large finite element problems. Their efficiency is essentially determined by the following considerations:

TABLE IV.1.1. Storage requirement and arithmetic operations of the Cholesky decomposition applied to the linear finite element discretization of the model problem on $(0,1)^d$

d	h	N_h	e_h	b_h	s_h	z_h
2	$\frac{1}{16}$	225	$1.1 \cdot 10^3$	15	$3.3 \cdot 10^3$	$7.6 \cdot 10^5$
	$\frac{1}{32}$	961	$4.8 \cdot 10^3$	31	$2.9 \cdot 10^4$	$2.8 \cdot 10^7$
	$\frac{1}{64}$	$3.9 \cdot 10^3$	$2.0 \cdot 10^4$	63	$2.5 \cdot 10^5$	$9.9 \cdot 10^8$
	$\frac{1}{128}$	$1.6 \cdot 10^4$	$8.0 \cdot 10^4$	127	$2.0 \cdot 10^6$	$3.3 \cdot 10^{10}$
3	$\frac{1}{16}$	$3.3 \cdot 10^3$	$2.4 \cdot 10^4$	225	$7.6 \cdot 10^5$	$1.7 \cdot 10^8$
	$\frac{1}{32}$	$3.0 \cdot 10^4$	$2.1 \cdot 10^5$	961	$2.8 \cdot 10^7$	$2.8 \cdot 10^{10}$
	$\frac{1}{64}$	$2.5 \cdot 10^5$	$1.8 \cdot 10^6$	$3.9 \cdot 10^3$	$9.9 \cdot 10^8$	$3.9 \cdot 10^{12}$
	$\frac{1}{128}$	$2.0 \cdot 10^6$	$1.4 \cdot 10^7$	$1.6 \cdot 10^4$	$3.3 \cdot 10^{10}$	$5.3 \cdot 10^{14}$

- The exact solution of the finite element problem is an approximation of the solution of the differential equation, which is the quantity of interest, with an error $O(h^k)$ where k is the polynomial degree of the finite element space. Therefore it is sufficient to compute an approximate solution of the discrete problem which has the same accuracy.
- If the mesh $\mathcal{T}_1$ is a global or local refinement of the mesh $\mathcal{T}_0$, the interpolate of the approximate discrete solution corresponding to $\mathcal{T}_0$ is a good initial guess for any iterative solver for the discrete problem corresponding to $\mathcal{T}_1$.

These considerations lead to the following *nested iteration*. Here $\mathcal{T}_0$, ..., $\mathcal{T}_R$ denotes a sequence of successively (globally or locally) refined meshes with corresponding finite element problems

$$L_k u_k = f_k \qquad 0 \le k \le R.$$

ALGORITHM IV.1.1. (Nested iteration)

(1) *Compute*
$$\widetilde{u}_0 = u_0 = L_0^{-1} f_0.$$
(2) *For $k = 1, \ldots, R$ compute an approximate solution $\widetilde{u}_k$ for $u_k = L_k^{-1} f_k$ by applying m_k iterations of an iterative solver for the problem*
$$L_k u_k = f_k$$
with starting value $I_{k-1,k}\widetilde{u}_{k-1}$, where $I_{k-1,k}$ is a suitable interpolation operator from the mesh $\mathcal{T}_{k-1}$ to the mesh $\mathcal{T}_k$.

Usually, the number m_k of iterations in Algorithm IV.1.1 is determined by the stopping criterion

$$\boxed{\|f_k - L_k\widetilde{u}_k\| \le \varepsilon\|f_k - L_k(I_{k-1,k}\widetilde{u}_{k-1})\|.}$$

That is, the residual of the starting value measured in an appropriate norm should be reduced by a factor ε. Typically, $\|\cdot\|$ is a weighted Euclidean norm and ε is in the realm 0.05 to 0.1. If the iterative solver has the convergence rate δ_k, the number m_k of iterations is given by

$$\boxed{m_k = \left\lceil \frac{\ln \varepsilon}{\ln \delta_k} \right\rceil.}$$

Table IV.1.2 gives the number m_k of iterations that require the classical Gauß-Seidel algorithm, the conjugate gradient algorithm IV.3.1 (p. 116) and the preconditioned conjugate gradient algorithm IV.3.2 (p. 117) with SSOR-preconditioning IV.3.3 (p. 119) for reducing an initial residual by the factor $\varepsilon = 0.1$. These algorithms need the following number of operations per unknown:

$$\begin{aligned} &2d+1 \quad \text{(Gauß-Seidel)},\\ &2d+6 \quad \text{(CG)},\\ &5d+8 \quad \text{(SSOR-PCG)}. \end{aligned}$$

TABLE IV.1.2. Number of iterations required for reducing an initial residual by the factor 0.1

h	Gauß-Seidel	CG	SSOR-PCG
$\frac{1}{16}$	236	12	4
$\frac{1}{32}$	954	23	5
$\frac{1}{64}$	3820	47	7
$\frac{1}{128}$	15287	94	11

Table IV.1.2 shows that the preconditioned conjugate gradient algorithm with SSOR-preconditioning yields satisfactory results for problems that are not too large. Nevertheless, its computational work is not proportional to the number of unknowns; for a fixed tolerance ε it approximately is of the order $N_h^{1+\frac{1}{2d}}$. The multigrid algorithm IV.4.1 (p. 121) overcomes this drawback. Its convergence rate is independent of the mesh-size. Correspondingly, for a fixed tolerance ε, its computational work is proportional to the number of unknowns. The advantages of the multigrid algorithm are reflected by Table IV.1.3.

TABLE IV.1.3. Arithmetic operations required by the preconditioned conjugate gradient algorithm with SSOR-preconditioning and the V-cycle multigrid algorithm with one Gauß-Seidel step for pre- and post-smoothing applied to the model problem in $(0,1)^d$

d	h	PCG-SSOR	multigrid
2	$\frac{1}{16}$	16′200	11′700
	$\frac{1}{32}$	86′490	48′972
	$\frac{1}{64}$	500′094	206′988
	$\frac{1}{128}$	3′193′542	838′708
3	$\frac{1}{16}$	310′500	175′500
	$\frac{1}{32}$	3′425′965	1′549′132
	$\frac{1}{64}$	$4.0 \cdot 10^7$	$1.3 \cdot 10^7$
	$\frac{1}{128}$	$5.2 \cdot 10^8$	$1.1 \cdot 10^8$

IV.2. Classical iterative solvers

The setting of this and the following section is as follows: We want to solve a linear system of equations

$$Lu = f$$

with N unknowns and a symmetric positive definite matrix L. We denote by κ the *condition* of L, i.e. the ratio of its largest to its smallest eigenvalue. Moreover we assume that $\kappa \approx N^{\frac{2}{d}}$.

All methods of this section are so-called *stationary iterative solvers* and have the following structure:

ALGORITHM IV.2.1. (Stationary iterative solver)

(0) *Given: matrix L, right-hand side f, initial guess u_0 and tolerance ε.*
Sought: an approximate solution of the linear system of equations $Lu = f$.

(1) *Set $i = 0$.*

(2) *If*
$$|Lu_i - f| < \varepsilon,$$
return u_i as approximate solution; stop.

(3) *Compute*
$$u_{i+1} = F(u_i; L, f)$$
increase i by 1 and return to step (2).

Here, $u \mapsto F(u; L, f)$ is an affine mapping, the so-called *iteration method*, which characterises the particular iterative solver. $|\cdot|$ is any norm on $\mathbb{R}^N$, e.g., the Euclidean norm.

The simplest method is the *Richardson iteration*. The iteration method is given by

$$\boxed{u \mapsto u + \frac{1}{\omega}(f - Lu).}$$

Here, ω is a *damping parameter*, which has to be of the same order as the largest eigenvalue of L. The convergence rate of the Richardson iteration is $\frac{\kappa-1}{\kappa+1} \approx 1 - N^{-\frac{2}{d}}$.

The *Jacobi iteration* is closely related to the Richardson iteration. The iteration method is given by

$$\boxed{u \mapsto u + D^{-1}(f - Lu).}$$

Here, D is the diagonal of L. The convergence rate again is $\frac{\kappa-1}{\kappa+1} \approx 1 - N^{-\frac{2}{d}}$. Notice, the Jacobi iteration sweeps through all equations and exactly solves the current equation for the corresponding unknown without modifying subsequent equations.

The *Gauß-Seidel iteration* is a modification of the Jacobi iteration: Now every update of an unknown is immediately transferred to all subsequent equations. This modification gives rise to the following iteration method:

$$\boxed{u \mapsto u + \mathcal{L}^{-1}(f - Lu).}$$

Here, $\mathcal{L}$ is the lower diagonal part of L diagonal included. The convergence rate again is $\frac{\kappa-1}{\kappa+1} \approx 1 - N^{-\frac{2}{d}}$.

IV.3. Conjugate gradient algorithms

IV.3.1. The conjugate gradient algorithm. The conjugate gradient algorithm is based on the following ideas:

- For symmetric positive definite stiffness matrices L the solution of the linear system of equations

  $$Lu = f$$

  is equivalent to the minimization of the quadratic functional

  $$J(u) = \frac{1}{2} u \cdot (Lu) - f \cdot u.$$

- Given an approximation v to the solution u of the linear system, the negative gradient
$$-\nabla J(v) = f - Lv$$
of J at v gives the direction of the steepest descent.
- Given an approximation v and a search direction $d \neq 0$, J attains its minimum on the line $t \mapsto v + td$ at the point
$$t^* = \frac{d \cdot (f - Lv)}{d \cdot (Ld)}.$$
- When successively minimizing J in the directions of the negative gradients, the algorithm slows down since the search directions become nearly parallel.
- The algorithm speeds up when choosing the successive search directions L-orthogonal, i.e.
$$d_i \cdot (Ld_{i-1}) = 0$$
for the search directions of iterations $i - 1$ and i.
- These L-orthogonal search directions can be computed during the algorithm by a suitable three-term recursion.

ALGORITHM IV.3.1. (Conjugate gradient algorithm)

(0) *Given: a linear system of equations $Lu = f$ with a symmetric, positive definite matrix L, an initial guess u_0 for the solution, and a tolerance $\varepsilon > 0$.*
Sought: an approximate solution of the linear system.

(1) *Compute*
$$r_0 = f - Lu_0,$$
$$d_0 = r_0,$$
$$\gamma_0 = r_0 \cdot r_0.$$
Set $i = 0$.

(2) *If*
$$\gamma_i < \varepsilon^2$$
return u_i as approximate solution; stop. Otherwise go to step 3.

(3) *Compute*
$$s_i = Ld_i,$$
$$\alpha_i = \frac{\gamma_i}{d_i \cdot s_i},$$
$$u_{i+1} = u_i + \alpha_i d_i,$$
$$r_{i+1} = r_i - \alpha_i s_i,$$
$$\gamma_{i+1} = r_{i+1} \cdot r_{i+1},$$
$$\beta_i = \frac{\gamma_{i+1}}{\gamma_i},$$

$$d_{i+1} = r_{i+1} + \beta_i d_i.$$

Increase i by 1 and go to step 2.

The convergence rate of the CG-algorithm is given by

$$\boxed{\delta = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}}$$

where κ is the condition number of L and equals the ratio of the largest to the smallest eigenvalue of L. For finite element discretizations of elliptic equations of second order, we have $\kappa \approx h^{-2}$ and correspondingly $\delta \approx 1 - h$, where h is the mesh-size.

IV.3.2. The preconditioned conjugate gradient algorithm. The idea of the preconditioned conjugate gradient algorithm is the following:

- Instead of the original system

  $$Lu = f$$

  solve the equivalent system

  $$\widehat{L}\widehat{u} = \widehat{f}$$

  with

  $$\widehat{L} = H^{-1} L H^{-t}$$
  $$\widehat{f} = H^{-1} f$$
  $$\widehat{u} = H^t u$$

  and an invertible square matrix H.
- Choose the matrix H such that:
  - The condition number of $\widehat{L}$ is much smaller than the one of L.
  - Systems of the form $Cv = d$ with $C = HH^t$ are much easier to solve than the original system $Lu = f$.
- Apply the conjugate gradient algorithm to the new system $\widehat{L}\widehat{u} = \widehat{f}$ and express everything in terms of the original quantities L, f, and u.

ALGORITHM IV.3.2. (Preconditioned conjugate gradient algorithm)

(0) *Given: a linear system of equations $Lu = f$ with a symmetric, positive definite matrix L, an approximation C to L, an initial guess u_0 for the solution, and a tolerance $\varepsilon > 0$.*
Sought: an approximate solution of the linear system.

(1) *Compute*

$$r_0 = f - Lu_0,$$

solve

$$Cz_0 = r_0,$$

and compute

$$d_0 = z_0,$$
$$\gamma_0 = r_0 \cdot z_0.$$

Set $i = 0$.

(2) *If*

$$\gamma_i < \varepsilon^2$$

return u_i *as approximate solution; stop. Otherwise go to step 3.*

(3) *Compute*

$$s_i = Ld_i,$$
$$\alpha_i = \frac{\gamma_i}{d_i \cdot s_i},$$
$$u_{i+1} = u_i + \alpha_i d_i,$$
$$r_{i+1} = r_i - \alpha_i s_i,$$

solve

$$Cz_{i+1} = r_{i+1},$$

and compute

$$\gamma_{i+1} = r_{i+1} \cdot z_{i+1},$$
$$\beta_i = \frac{\gamma_{i+1}}{\gamma_i},$$
$$d_{i+1} = z_{i+1} + \beta_i d_i.$$

Increase i *by* 1 *and go to step 2.*

For the trivial choice $C = I$, the identity matrix, Algorithm IV.3.2 reduces to the conjugate gradient Algorithm IV.3.1. For the non-realistic choice $C = A$, Algorithm IV.3.2 stops after one iteration and produces the exact solution.

The convergence rate of the PCG-algorithm is given by

$$\boxed{\delta = \frac{\sqrt{\widehat{\kappa}} - 1}{\sqrt{\widehat{\kappa}} + 1}}$$

where $\widehat{\kappa}$ is the condition number of $\widehat{L}$ and equals the ratio of the largest to the smallest eigenvalue of $\widehat{L}$.

Obviously the efficiency of the PCG-algorithm hinges on the good choice of the preconditioning matrix C. It has to satisfy the contradictory goals that $\widehat{L}$ should have a small condition number and that problems of the form $Cz = d$ should be easy to solve. A good compromise is the SSOR-preconditioner. It corresponds to

$$C = \frac{1}{\omega(2-\omega)}(D - \omega U^t)D^{-1}(D - \omega U)$$

where D and U denote the diagonal of L and its strictly upper diagonal part, respectively and where $\omega \in (0,2)$ is a relaxation parameter.

The following algorithm realizes the SSOR-preconditioning.

ALGORITHM IV.3.3. (SSOR-preconditioning)

(0) *Given: r and a relaxation parameter $\omega \in (0,2)$.*
Sought: $z = C^{-1}r$.

(1) *Set*

$$z = 0.$$

(2) *For $i = 1, \ldots, N_h$ compute*

$$z_i = z_i + \omega L_{ii}^{-1}\left\{r_i - \sum_{j=1}^{N_h} L_{ij}z_j\right\}.$$

(3) *For $i = N_h, \ldots, 1$ compute*

$$z_i = z_i + \omega L_{ii}^{-1}\left\{r_i - \sum_{j=1}^{N_h} L_{ij}z_j\right\}.$$

For finite element discretizations of elliptic equations of second order and the SSOR-preconditioning of Algorithm IV.3.3, we have $\widehat{\kappa} \approx h^{-1}$ and correspondingly $\delta \approx 1 - h^{\frac{1}{2}}$, where h is the mesh-size.

IV.3.3. Non-symmetric and indefinite problems. The CG- and the PCG-algorithms IV.3.1 and IV.3.2 can only be applied to problems with a symmetric positive definite stiffness matrix, i.e., to scalar linear elliptic equations without convection and the displacement formulation of the equations of linearized elasticity. Scalar linear elliptic equations with convection – though possibly being small – and mixed formulations of the equations of linearized elasticity lead to non-symmetric or indefinite stiffness matrices. For these problems Algorithms IV.3.1 and IV.3.2 break down.

There are several possible remedies to this difficulty. An obvious one is to consider the equivalent normal equations

$$L^tLu = L^tf$$

which have a symmetric positive matrix. This simple device, however, cannot be recommended, since passing to the normal equations squares the condition number and thus doubles the number of iterations. A much better alternative is the bi-conjugate gradient algorithm. It tries to solve simultaneously the original problem $Lu = f$ and its adjoint or conjugate problem $L^t v = L^t f$.

ALGORITHM IV.3.4. (Stabilized bi-conjugate gradient algorithm Bi-CG-stab)

(0) *Given: a linear system of equations $Lx = b$, an initial guess x_0 for the solution, and a tolerance $\varepsilon > 0$.*
Sought: an approximate solution of the linear system.

(1) *Compute*

$$r_0 = b - Lx_0,$$

and set

$$\overline{r}_0 = r_0, \qquad v_{-1} = 0, \qquad p_{-1} = 0,$$
$$\alpha_{-1} = 1, \qquad \rho_{-1} = 1, \qquad \omega_{-1} = 1.$$

Set $i = 0$.

(2) *If*

$$r_i \cdot r_i < \varepsilon^2$$

return x_i as approximate solution; stop. Otherwise go to step 3.

(3) *Compute*

$$\rho_i = \overline{r}_i \cdot r_i,$$
$$\beta_{i-1} = \frac{\rho_i \alpha_{i-1}}{\rho_{i-1}\omega_{i-1}}.$$

If

$$|\beta_{i-1}| < \varepsilon$$

there is a possible break-down; stop. Otherwise compute

$$p_i = r_i + \beta_{i-1}\{p_{i-1} - \omega_{i-1} v_{i-1}\},$$
$$v_i = Lp_i,$$
$$\alpha_i = \frac{\rho_i}{\overline{r}_0 \cdot v_i}.$$

If

$$|\alpha_i| < \varepsilon$$

there is a possible break-down; stop. Otherwise compute

$$s_i = r_i - \alpha_i v_i,$$
$$t_i = Ls_i,$$
$$\omega_i = \frac{t_i \cdot s_i}{t_i \cdot t_i},$$
$$x_{i+1} = x_i + \alpha_i p_i + \omega_i s_i,$$
$$r_{i+1} = s_i - \omega_i t_i.$$

Increase i by 1 and go to step 2.

IV.4. Multigrid algorithms

Multigrid algorithms are based on the following observations:

- Classical iterative methods such as the Gauß-Seidel algorithm quickly reduce highly oscillatory error components.
- Classical iterative methods such as the Gauß-Seidel algorithm on the other hand are very poor in reducing slowly oscillatory error components.
- Slowly oscillating error components can well be resolved on coarser meshes with fewer unknowns.

IV.4.1. The multigrid algorithm. The multigrid algorithm is based on a sequence of meshes $\mathcal{T}_0, \ldots, \mathcal{T}_R$, which are obtained by successive local or global refinement, and associated discrete problems $L_k u_k = f_k$, $k = 0, \ldots, R$, corresponding to a partial differential equation. The finest mesh $\mathcal{T}_R$ corresponds to the problem that we actually want to solve.

The multigrid algorithm has three ingredients:

- a *smoothing operator* M_k, which should be easy to evaluate and which at the same time should give a reasonable approximation to L_k^{-1},
- a *restriction operator* $R_{k,k-1}$, which maps functions on a fine mesh $\mathcal{T}_k$ to the next coarser mesh $\mathcal{T}_{k-1}$,
- a *prolongation operator* $I_{k-1,k}$, which maps functions from a coarse mesh $\mathcal{T}_{k-1}$ to the next finer mesh $\mathcal{T}_k$.

For a concrete multigrid algorithm these ingredients must be specified. This will be done in the next sections. Here, we discuss the general form of the algorithm and its properties.

ALGORITHM IV.4.1. (*MG* $(k, \mu, \nu_1, \nu_2, L_k, f_k, u_k)$ one iteration of the multigrid algorithm on mesh $\mathcal{T}_k$)

(0) *Given: the level number k of the actual mesh, parameters μ, ν_1, and ν_2, the stiffness matrix L_k of the actual discrete problem, the actual right-hand side f_k, and an inital guess u_k.*
Sought: an improved approximate solution u_k.

(1) *If $k = 0$, compute*

$$u_0 = L_0^{-1} f_0;$$

stop. Otherwise go to step 2.

(2) (Pre-smoothing) *Perform ν_1 steps of the iterative procedure*

$$u_k = u_k + M_k(f_k - L_k u_k).$$

(3) (Coarse grid correction)

(a) *Compute*

$$f_{k-1} = R_{k,k-1}(f_k - L_k u_k)$$

and set

$$u_{k-1} = 0.$$

(b) *Perform* μ *iterations of* $MG(k-1, \mu, \nu_1, \nu_2, L_{k-1}, f_{k-1}, u_{k-1})$ *and denote the result by* u_{k-1}.

(c) *Compute*

$$u_k = u_k + I_{k-1,k} u_{k-1}.$$

(4) (Post-smoothing) *Perform* ν_2 *steps of the iterative procedure*

$$u_k = u_k + M_k(f_k - L_k u_k).$$

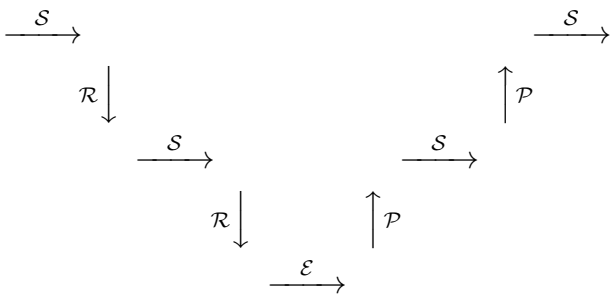

FIGURE IV.4.1. Schematic presentation of a multigrid algorithm with V-cycle and three grids. The labels have the following meaning: $\mathcal{S}$ smoothing, $\mathcal{R}$ restriction, $\mathcal{P}$ prolongation, $\mathcal{E}$ exact solution.

REMARK IV.4.2. (1) The parameter μ determines the complexity of the algorithm. Popular choices are $\mu = 1$ called *V-cycle* and $\mu = 2$ called *W-cycle*. Figure IV.4.1 gives a schematic presentation of the multigrid algorithm for the case $\mu = 1$ and $R = 2$ (three meshes). Here, $\mathcal{S}$ denotes smoothing, $\mathcal{R}$ restriction, $\mathcal{P}$ prolongation, and $\mathcal{E}$ exact solution.

(2) The number of smoothing steps per multigrid iteration, i.e. the parameters ν_1 and ν_2, should not be chosen too large. A good choice for positive definite problems such as the Poisson equation is $\nu_1 = \nu_2 = 1$. For indefinite problems such as mixed formulations of the equations of linearized elasticity, a good choice is $\nu_1 = \nu_2 = 2$.

(3) If $\mu \leq 2$, one can prove that the computational work of one multigrid iteration is proportional to the number of unknowns of the actual discrete problem.

(4) Under suitable conditions on the smoothing algorithm, which is determined by the matrix M_k, one can prove that the convergence rate

of the multigrid algorithm is independent of the mesh-size, i.e., it does not deteriorate when refining the mesh. These conditions will be discussed in the next section. In practice one observes convergence rates of 0.1 – 0.5 for positive definite problems such as the Poisson equation and of 0.3 – 0.7 for indefinite problems such as mixed formulations of the equations of linearized elasticity.

IV.4.2. Smoothing. The symmetric *Gauss-Seidel algorithm* is the most popular smoothing algorithm for positive definite problems such as the Poisson equation. It corresponds to the choice

$$M_k = (D_k - U_k^t)D_k^{-1}(D_k - U_k),$$

where D_k and U_k denote the diagonal and the strictly upper diagonal part of L_k respectively.

For non-symmetric or indefinite problems such as scalar linear elliptic equations with convection or mixed formulations of the equations of linearized elasticity, the most popular smoothing algorithm is the squared Jacobi iteration. This is the Jacobi iteration applied to the squared system $L_k^t L_k u_k = L_k^t f_k$ and corresponds to the choice

$$M_k = \omega^{-2} L_k^t$$

with a suitable damping parameter satisfying $\omega > 0$ and $\omega = O(h_K^{-2})$.

IV.4.3. Prolongation. Since the partition $\mathcal{T}_k$ of level k always is a refinement of the partition $\mathcal{T}_{k-1}$ of level $k-1$, the corresponding finite element spaces are nested, i.e., finite element functions corresponding to level $k-1$ are contained in the finite element space corresponding to level k. Therefore, the values of a coarse-grid function corresponding to level $k-1$ at the nodal points corresponding to level k are obtained by evaluating the nodal bases functions corresponding to $\mathcal{T}_{k-1}$ at the requested points. This defines the interpolation operator $I_{k-1,k}$.

Figures III.1.2 (p. 91) and III.1.3 (p. 92) show various partitions of a triangle and of a square, respectively. The numbers outside the element indicate the enumeration of the element vertices and edges. Thus, e.g., edge 2 of the triangle has the vertices 0 and 1 as its endpoints. The numbers +0, +1 etc. inside the elements indicate the enumeration of the child elements. The remaining numbers inside the elements give the enumeration of the vertices of the child elements.

Example IV.4.3. Consider a piecewise constant approximation, i.e. $S^{0,-1}(\mathcal{T})$. The nodal points are the barycentres of the elements. Every element in $\mathcal{T}_{k-1}$ is subdivided into several smaller elements in $\mathcal{T}_k$. The nodal value of a coarse-grid function at the barycentre of a child element in $\mathcal{T}_k$ then is its nodal value at the barycentre of the parent element in $\mathcal{T}_k$.

EXAMPLE IV.4.4. Consider a piecewise linear approximation, i.e. $S^{1,0}(\mathcal{T})$. The nodal points are the vertices of the elements. The refinement introduces new vertices at the midpoints of some edges of the parent element and possibly – when using quadrilaterals – at the barycentre of the parent element. The nodal value at the midpoint of an edge is the average of the nodal values at the endpoints of the edge. Thus, e.g., the value at vertex 1 of child +0 is the average of the values at vertices 0 and 1 of the parent element. Similarly, the nodal value at the barycentre of the parent element is the average of the nodal values at the four element vertices.

IV.4.4. Restriction. The restriction is computed by expressing the nodal bases functions corresponding to the coarse partition $\mathcal{T}_{k-1}$ in terms of the nodal bases functions corresponding to the fine partition $\mathcal{T}_k$ and inserting this expression in the variational formulation. This results in a lumping of the right-hand side vector which, in a certain sense, is the transpose of the interpolation.

EXAMPLE IV.4.5. Consider a piecewise constant approximation, i.e. $S^{0,-1}(\mathcal{T})$. The nodal shape function of a parent element is the sum of the nodal shape functions of the child elements. Correspondingly, the components of the right-hind side vector corresponding to the child elements are all added and associated with the parent element.

EXAMPLE IV.4.6. Consider a piecewise linear approximation, i.e. $S^{1,0}(\mathcal{T})$. The nodal shape function corresponding to a vertex of a parent *triangle* takes the value 1 at this vertex, the value $\frac{1}{2}$ at the midpoints of the two edges sharing the given vertex and the value 0 on the remaining edges. If we label the current vertex by a and the midpoints of the two edges emanating form a by m_1 and m_2, this results in the following formula for the restriction on a *triangle*

$$R_{k,k-1}\psi(a) = \psi(a) + \frac{1}{2}\{\psi(m_1) + \psi(m_2)\}.$$

When considering a *quadrilateral*, we must take into account that the nodal shape functions take the value $\frac{1}{4}$ at the barycentre b of the parent quadrilateral. Therefore the restriction on a *quadrilateral* is given by the formula

$$R_{k,k-1}\psi(a) = \psi(a) + \frac{1}{2}\{\psi(m_1) + \psi(m_2)\} + \frac{1}{4}\psi(b).$$

REMARK IV.4.7. An efficient implementation of the prolongation and restrictions loops through all elements and performs the prolongation or restriction element-wise. This process is similar to the usual element-wise assembly of the stiffness matrix and the load vector.

Bibliography

[1] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Wiley, New York, 2000.

[2] D. Braess, *Finite Elements*, second ed., Cambridge University Press, Cambridge, 2001, Theory, fast solvers, and applications in solid mechanics, Translated from the 1992 German edition by Larry L. Schumaker.

[3] R. Verfürth, *A Posteriori Error Estimation Techniques for Finite Element Methods*, Oxford University Press, Oxford, 2013.

Index