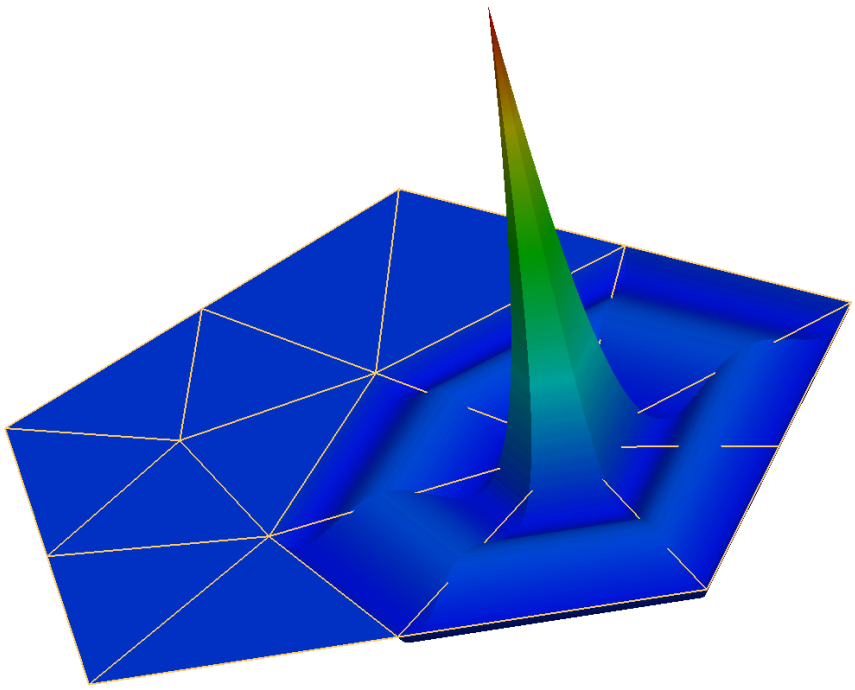
Lecture Notes on Scientific Computing with Partial Differential Equations

Peter Bastian
Interdisziplinäres Zentrum für Wissenschaftliches Rechnen
Universität Heidelberg, Im Neuenheimer Feld 205, 69120 Heidelberg
Peter.Bastian@iwr.uni-heidelberg.de



October 17, 2017

Contents

1. Gravity **7**
- 1.1. Newton's Law for Point Masses . . . 7
- 1.2. Distributed Mass . . . 8
- 1.3. Conservative Force . . . 9
- 1.4. Poisson's Equation . . . 10
- 1.5. Numerical Simulation . . . 10

2. Conservation Laws **15**
- 2.1. Continuum Hypothesis and Scales . . . 15
- 2.2. Conservation of Mass . . . 15
- 2.3. Conservation of Energy . . . 17
- 2.4. Conservation of Linear Momentum . . . 17
- 2.5. Heat Transfer . . . 18
- 2.6. Flow in Porous Media . . . 24
- 2.7. Inviscid Fluid Flow . . . 26
- 2.8. Propagation of Sound Waves . . . 30
- 2.9. Viscous Fluid Flow . . . 31

3. Calculus of Variations **37**
- 3.1. Equilibrium Principle . . . 37
- 3.2. Variational Approach . . . 42
- 3.3. Taut String Approximation . . . 45
- 3.4. Linear Elasticity and Plate Problem . . . 48
- 3.5. Hamilton's Principle . . . 49

4. Type Classification and Model Problems **51**
- 4.1. Basic Mathematical Questions . . . 51
- 4.2. Second-order Scalar Equations . . . 52
- 4.3. First-order Hyperbolic Systems . . . 61
- 4.4. Model Problems . . . 67

5. Elements of Functional Analysis **69**
- 5.1. Motivation . . . 69
- 5.2. Banach Spaces . . . 70
- 5.3. Hilbert Spaces . . . 72
- 5.4. Linear Mappings in Banach Spaces . . . 73
- 5.5. Abstract Existence Theory . . . 76

5.6. Lebesgue Spaces 81
5.7. Sobolev Spaces 85
5.8. Properties of Sobolev Spaces 89

6. Well-posed Scalar Elliptic PDEs 93
6.1. Dirichlet Problem 93
6.2. Neumann Problem 95
6.3. Mixed Problem 98
6.4. Convection-Diffusion Problem 99

7. Conforming Finite Element Methods 103
7.1. Abstract Galerkin Method 103
7.2. One-dimensional Finite Element Spaces 105
7.3. Mesh Construction in Arbitrary Dimensions 108
7.4. P_k Finite Elements 112
7.5. Q_k Finite Elements 115
7.6. Construction of the Finite Element Stiffness Matrix 116
7.7. Case Studies 119

8. Finite Element Convergence Theory 131
8.1. Bramble Hilbert Lemma 133
8.2. Approximation Results 135
8.3. Error Estimates 143
8.4. Loss of Coercivity 147

9. Adaptive Finite Element Methods 159
9.1. Introduction 159
9.2. Residual-based a-posteriori Error Estimator 160
9.3. Local Mesh Adaptation 163
9.4. Numerical Results 167

10. Multigrid Methods 175
10.1. Some Examples 176
10.2. Smoothing Property of Richardson Iteration 180
10.3. Variational Multigrid 183
10.4. Convergence Analysis 186

11. Finite Element Methods for Parabolic Problems 195
11.1. Method of Lines 195
11.2. Rothe Method 199
11.3. Space-time Method 200

12.Numerical Methods for First-order Hyperbolic Equations 201
12.1. Finite Difference Methods 201

A. Nabla and Friends 207
A.1. Notation for Derivatives 207
A.2. Vector Differential Calculus 208
A.2.1. Nabla Operator 208
A.2.2. Gradient 208
A.2.3. Divergence 209
A.2.4. Curl 210
A.2.5. Convection Term in Navier-Stokes Equations 210
A.2.6. Laplacian 210
A.3. Vector Integral Calculus 211
A.3.1. Matrix Product 211
A.3.2. Integration by Parts 211

Bibliography 213

"It is physics which gives us many important problems, which we would not have thought of without it. It is by the aid of physics that we can foresee the solutions."

Henri Poincaré,
cited from "Four Lectures on
Mathematics Delivered at Columbia University in 1911"
by Jacques Hadamard,
EBook #29788 in Project Gutenberg.

MANY practical applications involve the description of the state of a solid body, a fluid (in continuum mechanics no distinction is made between fluids and gases) or just any region of space. As examples consider the gravitational field within and outside of an inhomogeneous body, the temperature of a solid body, the flow of water in the subsurface, the flow of gases in a complicated duct, the propagation of sound or water waves or the mechanical stress in a bridge. In this part, we will derive the equations of mathematical physics that describe all these phenomena.

Chapter 1.

Gravity

1.1. Newton's Law for Point Masses

Newton's famous law of gravitation

$$F(x,y) = G\frac{mM}{\|y-x\|^2}\frac{y-x}{\|y-x\|} \qquad (x \neq y) \tag{1.1}$$

gives the force vector acting on a point mass m at position $x \in \mathbb{R}^3$ excerted by another point mass M located at a point $y \in \mathbb{R}^3$ and G is the gravitational constant with the approximate value $6.67 \cdot 10^{-11}$ N m^2 kg^{-2} (there is some debate about the value – it is difficult to measure). Newton's law is stated for point masses as it has first been applied to the sun and the planets in the solar system. But how does it act in a cloud of gas of varying density? Since there are so many atoms (or molecules) in the gas it would be overwhelmingly expensive to compute all the forces ($O(N^2)$ effort for N particles).

We now wish to derive a new form of Newton's law in the form of a partial differential equation (PDE) that is usable in this case. First we rewrite Newton's law a little bit by introducing the function

$$\psi(x,y) = -\frac{GM}{\|y-x\|} \tag{1.2}$$

which is called the gravitational *potential* of a point mass in physics. In mathematics $1/\|y-x\|$ is called *singularity function*. It has the following interesting properties:

$$\nabla_x \psi(x,y) = -\frac{GM(y-x)}{\|y-x\|^3}, \quad \Delta_x \psi(x,y) = \sum_{i=1}^{3} \partial^2_{x_i^2} \psi(x,y) = 0 \quad (x \neq y).$$

Using it we can rewrite Newton's law as

$$F(x,y) = ma(x,y), \qquad a(x,y) = -\nabla_x \psi(x,y).$$

Note that the acceleration $a(x,y)$ is independent of the mass m (equivalence principle).

Now consider an arbitrary domain $\omega \subset \mathbb{R}^3$ (open and connected set of points) with sufficiently smooth boundary $\partial\Omega$, a point $y \notin \partial\omega$ and compute the surface integral

$$\int_{\partial\omega} a(x,y) \cdot n(x)\, ds_x = -\int_{\partial\omega} \nabla_x \psi(x,y) \cdot n(x)\, ds_x \tag{1.3}$$

where $n(x)$ denotes the exterior unit outer normal vector to ω. By ds_x we indicate that the surface integral is done with respect to the variable x and not y. For the evaluation of the integral we need to consider two cases:

i) $y \notin \omega$. By applying Gauss' integral theorem $\int_\omega \nabla \cdot u\, dx = \int_{\partial\omega} u \cdot n\, ds$ we get

$$-\int_{\partial\omega} \nabla_x \psi(x,y) \cdot n(x)\, ds_x = -\int_{\omega} \Delta_x \psi(x,y)\, dx = 0$$

since $\Delta_x \psi(x,y) = 0$ for any $x \in \omega$ since y is outside ω.

ii) $y \in \omega$. Now the trick from case i) can not be done so easily because ψ has a singularity for $x = y$ but it can be modified. Let $B_\epsilon(y) = \{x \in \mathbb{R}^3 : \|x - y\| < \epsilon\}$ be the open ball of radius ϵ around y. Then again applying Gauss' theorem we get

$$0 = \int_{\omega \setminus B_\epsilon(y)} \Delta_x \psi(x,y) dx = \int_{\partial\omega} \nabla_x \psi(x,y) \cdot n(x) ds_x - \int_{\partial B_\epsilon(y)} \nabla_x \psi(x,y) \cdot n(x) ds_x$$

The left hand side integral is zero and the minus sign is due to the fact the normal to $\omega \setminus B_\epsilon(y)$ points *into* the ball $B_\epsilon(y)$. The second integral on the right hand side can be computed directly as

$$\int_{\partial B_\epsilon(y)} \nabla_x \psi(x,y) \cdot n(x)\, ds_x = 4\pi G M$$

independent of ϵ.

So we get the following result:

$$\int_{\partial\omega} a(x,y) \cdot n(x)\, ds_x = \begin{cases} -4\pi G M & y \in \omega \\ 0 & \text{else} \end{cases} . \tag{1.4}$$

1.2. Distributed Mass

Now we extend this to a distributed mass by introducing the density function $\rho : \mathbb{R}^3 \to \mathbb{R}$ with units kg m^{-3}. For any domain $\omega \subset \mathbb{R}^3$ then $M_\omega = \int_\omega \rho\, dx$ gives

the mass contained in ω. We further assume that the density distribution is such that the integral $\int_{\mathbb{R}^3} \rho \, dx$ exists. Then the acceleration experienced at a point x excerted by the mass distribution ρ can be computed by subdividing the mass into an infinite number of infinitesimal pieces V_i at position y_i (superposition principle):

$$a(x) = \lim_{N\to\infty} \sum_{i=1}^{N} G\rho(y_i) V_i \nabla_x \left(\frac{1}{\|y_i - x\|} \right) = G \int_{\mathbb{R}^3} \rho(y) \nabla_x \left(\frac{1}{\|y - x\|} \right) dy. \tag{1.5}$$

One can check that this integral is well defined despite the singularity, i.e. it holds also for a point x *inside* a body with mass distribution ρ (transform to spherical coordinates around x).

Now using (1.4) one finds that for this acceleration and any suitable $\omega \subset \mathbb{R}^3$

$$\begin{aligned} \int_{\partial\omega} a(x) \cdot n(x) \, ds &= \int_{\partial\omega} \lim_{N\to\infty} \sum_{i=1}^{N} \rho(y_i) V_i \nabla_x \left(\frac{G}{\|y_i - x\|} \right) \cdot n(x) \, dx \\ &= \lim_{N\to\infty} \sum_{i=1}^{N} \rho(y_i) V_i \int_{\partial\omega} \nabla_x \left(\frac{G}{\|y_i - x\|} \right) \cdot n(x) \, dx \\ &= -4\pi G \int_{\omega} \rho(x) \, dx \end{aligned}$$

(only the mass inside ω plays a role). This is a continuum version of (1.4).

Applying again Gauss' theorem to the integral on the left hand side we find

$$\int_{\omega} \nabla \cdot a(x) + 4\pi G\rho(x) \, dx = 0.$$

If a is sufficiently smooth the fact that ω can be chosen arbitrarily implies that the equality also holds for the integrand itself (see e.g. [Smirnow, 1981, Paragraph 74]) and we arrive at

$$\nabla \cdot a(x) = -4\pi G\rho(x) \qquad (x \in \mathbb{R}^3). \tag{1.6}$$

1.3. Conservative Force

The final piece is the observation that all fundamental forces in nature are conservative (a basic principle that is assumed to hold by physicists). In a conservative force field the path integral $w(a, b) = \int_a^b F(s) \cdot t(s) \, ds$ (t is the unit tangential vector) does only depend on the points a, b but not on the particular path taken

from a to b. Conservativity of the force is a consequence of conservation of energy because otherwise it would be possible to generate energy in a force field by taking different paths back and forth. With an arbitray reference point r_0 we then have $w(a,b) = w(a,r_0) + w(r_0,b) = w(r_0,b) - w(r_0,a) = w'(b) - w'(a)$ where $w'(x) = w(r_0,x)$ is now only a function with a single argument, called the *gravitational potential*. Invoking the main theorem of calculus in its multidimensional form

$$\int_a^b \nabla\Psi(s) \cdot t(s)\, ds = \Psi(b) - \Psi(a) \tag{1.7}$$

we see that a force is conservative if and only if it can be represented as the gradient of a potential. The potential is only unique up to a constant as can be seen from (1.7).

1.4. Poisson's Equation

Since $ma(x)$ with $a(x)$ from (1.6) is the gravitational force experienced by a point mass m at position x and the gravitational force is supposed to be conservative we conclude that there must exist a scalar function $\Psi(x)$ such that $a(x) = -\nabla\Psi(x)$. Inserting this into (1.6) we obtain

$$-\nabla \cdot \nabla\Psi(x) = -\Delta\Psi(x) = -4\pi G\rho(x) \qquad (x \in \mathbb{R}^3). \tag{1.8}$$

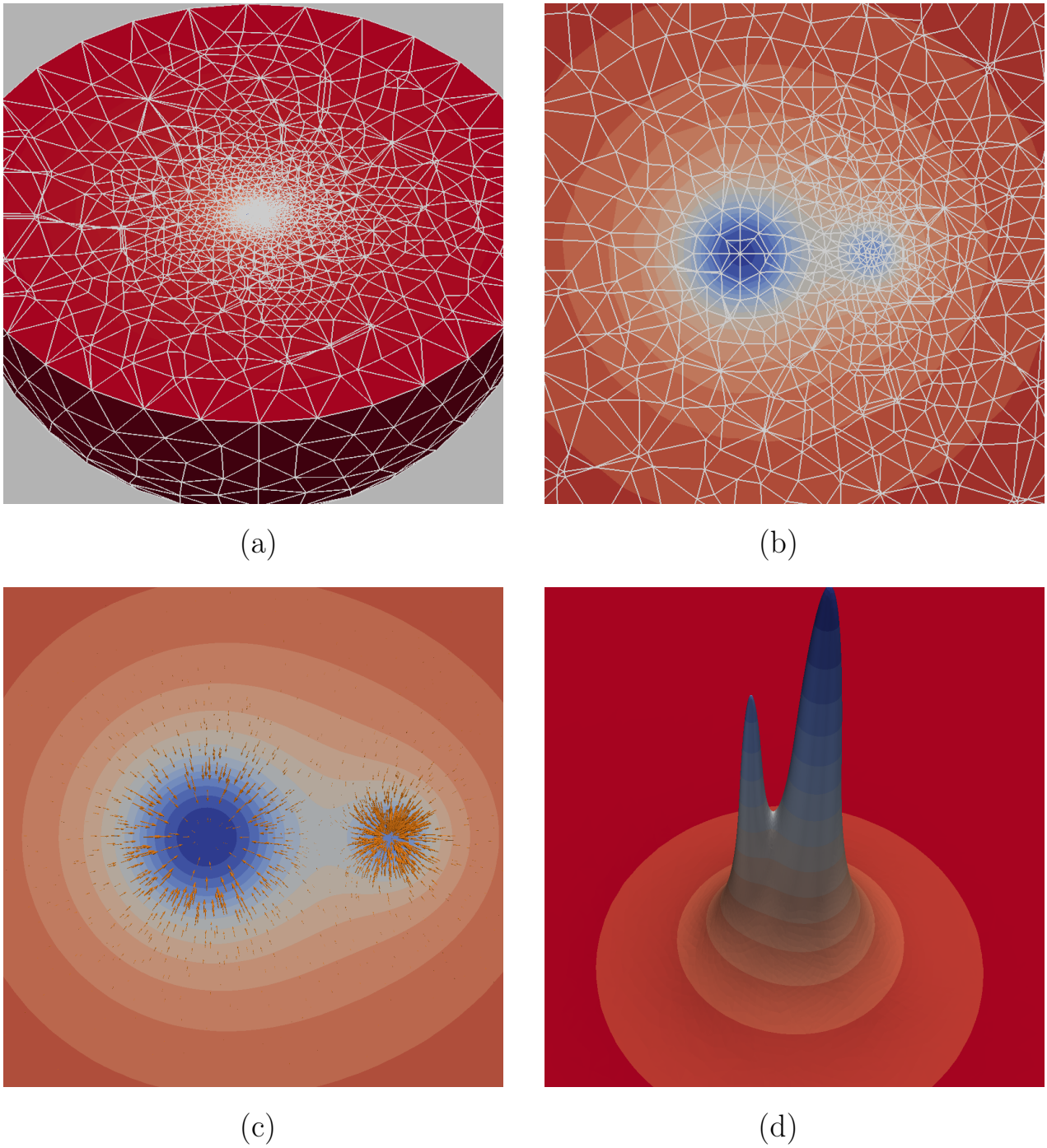
This equation is called *Poisson's* equation. As stated Ψ is assumed to be twice continuously differentiable, which requires ρ to be at least continuous. This is practically very restrictive since, for example, the density function of the moon (having no atmosphere) might be very well approximated by a discontinuous function. It is an important part of PDE theory to give equation (1.8) a precise mathematical meaning also in this sense. The potential is determined by Equation (1.8) up to a constant. To fix the constant, an additional condition for the behaviour of Ψ for $x \to \infty$ can be imposed.

1.5. Numerical Simulation

In these lecture notes we are concerned with the numerical solution of (1.8) among other equations. Let us demonstrate the typical workflow of a numerical simulation by way of this example.

Example 1.1. We consider the gravitational field of two solid, homogeneous spheres with the following radii, densities and positions:

$$\begin{aligned} R_1 &= 6 \cdot 10^6 \text{ m}, & \rho_1 &= 1000 \text{ kg m}^{-3}, & x_1 &= (-7 \cdot 10^6, 0, 0)^T, \\ R_2 &= 3 \cdot 10^6 \text{ m}, & \rho_2 &= 2500 \text{ kg m}^{-3}, & x_2 &= (1 \cdot 10^7, 0, 0)^T. \end{aligned}$$



(a) (b)

(c) (d)

Figure 1.1.: Numerical simulation of a gravity problem using a piecewise linear finite element method.

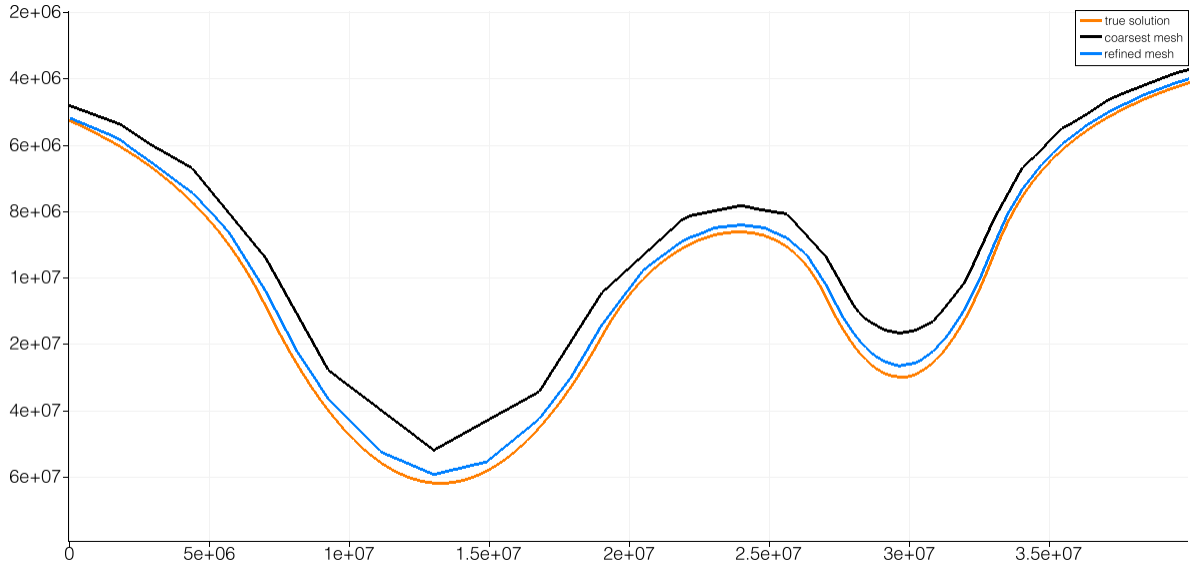


Figure 1.2.: Comparison of numerical and true solution.

The true potential then is

$$\Psi(x) = \sum_{i=1}^{2} \Psi_i(x), \quad \Psi_i(x) = \begin{cases} \frac{M_i G}{2R_i}\left(\frac{\|x-x_i\|^2}{R_i^2} - 3\right) & \|x - x_i\| \le R_i \\ -\frac{M_i G}{\|x-x_i\|} & \|x - x_i\| > R_i \end{cases} \qquad (i = 1, 2),$$

with $M_i = \frac{4}{3} R_i^3 \pi \rho_i$.

A first problem with numerically solving (1.8) is that it is posed in the whole space $\mathbb{R}^3$. Since the numerical solution (with the methods used in these lecture notes) requires the subdivision of the domain into a finite number of finite, simple-shaped volumes we have to restrict ourselves to a *finite* domain Ω, e.g. $\Omega = \{x \in \mathbb{R}^3 : \|x\| < R\}$ with R sufficiently large. On the boundary $\partial\Omega$ the values of the potential have to be prescribed. This can be done easily here, because we know the exact solution. In the general case when the exact potential is (of course) not known one makes R sufficiently large that it is far away from any mass inside Ω and takes the potential from concentrating all masses in the center of gravity. The gravity problem to be solved in a finite domain then reads

$$-\Delta u = -4\pi G\rho \qquad \text{in } \Omega, \tag{1.9a}$$

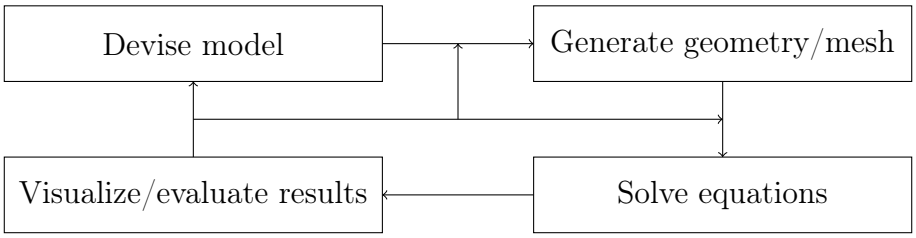
$$u = \Psi \qquad \text{on } \partial\Omega. \tag{1.9b}$$

with

$$\rho(x) = \begin{cases} \rho_1 & \|x - x_1\| \le R_1 \\ \rho_2 & \|x - x_2\| \le R_2 \\ 0 & \text{else} \end{cases}.$$

Note that in the general case $u \neq \Psi$ because only an approximation of the boundary values is available. Figure 1.1(a) shows the subdivision of Ω into tetrahedra, (b) shows a close up of the numerically determined potential, (c) shows in addition the gravitational field with arrows and (d) gives a "warped" view of the numerical solution. □

The typical simulation workflow consists of the following steps:



All simulations in these lecture notes are done with open source software: `Gmsh`[1] for geometry and mesh generation, `Dune`[2] for the actual solution of the discretized problem and `Paraview`[3] for visualization.

In general the numerical solution differs from the true solution. It is important to distinguish different sources of error:

- *Modelling error* comes from not taking the correct model. An example is the restriction to a finite domain in a gravity problem with artificial (and usually inexact) boundary conditions. In general modelling error can mean effects from physical processes that have not been taken into account.
- *Data error* means that coefficients in a given model (e.g. density here) might not be known with sufficient accuracy.
- *Discretization error* stems from the fact that in the computer the solution is represented by a finite number of degrees of freedom (e.g. coefficients of a polynomial).
- *Iteration error* arises from the solution of linear or nonlinear algebraic equations with iterative methods (such as e.g. Newton's method).
- *Floating-point error* comes from the fact that numbers are represented with finite precision.

Figure 1.2 illustrates the effect of discretization error for the gravity example. To that end we plot the solution along a segment of the x-axis which goes straight through the two bodies. The black curve shows the potential computed

[1] `http://geuz.org/gmsh/`
[2] `http://www.dune-project.org/`
[3] `http://www.paraview.org/`

numerically on a relatively coarse mesh. As the mesh is refined (blue curve) it approaches the true solution (orange curve).

Due to the different errors present in the numerical solution the simulation workflow is executed in a loop where, depending on the type of error, the model, mesh or solution process has to be refined.

Chapter 2.

Conservation Laws

2.1. Continuum Hypothesis and Scales

Materials such as solids or fluids are made up of atoms or molecules with void space in between (we do not consider quantum effects, although, also there, partial differential equations do play a role). Practical problems often involve excessively large numbers of atoms as we are interested in the behaviour of the material on a length scale that is very large compared to the average distance of the atoms. We call this scale of interest the macroscopic scale and the scale of the discrete particles the microscopic scale.

In continuum mechanics, the properties of the material are assumed to be (piecewise) continuous (or even differentiable) functions in the mathematical sense. The discrete particles are not considered, instead macroscopic properties (e.g. velocity) are defined as appropriate averages of the microscopic properties. By averaging, new quantities (such as density, temperature or pressure) arise that have no equivalent on the microscopic scale. The validity of this continuum hypothesis depends on the number of atoms (so that averages are representative) and whether the micro- and macroscale are sufficiently separated (a property called scale separation).

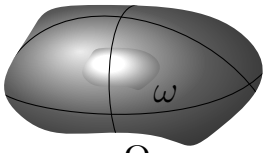
The laws on the microscale give now rise to new (effective) laws on the macroscale that connect the macroscopic variables. Current research is very much interested in so-called multiscale problems where the effective macroscopic laws (or coefficients in these laws) are not easily determined from the micoscopic scale (such as porous medium problems) or where there is no scale separation (e.g. turbulence).

In this chapter, fluids are considered while in the next chapter the deformation of elastic solid bodies is considered.

2.2. Conservation of Mass

Conservation of mass, linear and angular momentum as well as energy are basic empirical law of physics (throughout this text we consider only classical mechanics where mass and energy are distinct quantities). Conservation states that the total amount of such an extensive state variable in a closed system remains con-

stant over time. In an open system, the total amount of the quantity can vary through exchange with the environment. We are now about to first state the principle of conservation of mass in mathematical form. This is then extended to energy and linear momentum.



We consider a compressible fluid material that fills a domain $\Omega \subseteq \mathbb{R}^n$, $n = 1, 2, 3$, which is open and connected. The domain $\omega \subset \Omega$ is chosen arbitrarily within Ω (see figure). For the subsequent derivation, ω and Ω are fixed in space and do not depend on time (an assumption to be relaxed when solids are considered). This is called the *Eulerian point of view*. The function $\rho(x,t)$ gives the mass density in units[1] kg m^{-3} for any point $x \in \Omega$ at time t (other units, such as mol m^{-3} may be appropriate depending on the problem). The total mass $M_\omega(t)$ (in kg) contained in ω at time t is then given by

$$M_\omega(t) = \int_\omega \rho(x,t)\,dx \ .$$

The principle of mass conservation now states that over time the mass in ω can change only due to flow of material over the boundary $\partial\omega$ or due to injection or extraction of material into or from ω. To formulate this precisely, the velocity of the material $v(x,t)$ in m s^{-1} and the source function $f(x,t)$ in kg s^{-1} m^{-3} is given. For an arbitrary time interval, Δt the we can state:

$$M_\omega(t+\Delta t) - M_\omega(t) = \int_t^{t+\Delta t} \left\{ \int_\omega f(x,r)\,dx - \int_{\partial\omega} \rho(x,r)v(x,r)\cdot n(x)\,ds \right\} dr \ . \tag{2.1}$$

The volume integral gives the contribution from sources and sinks with $f > 0$ denoting a source and $f < 0$ denoting a sink. In the surface integral, $n(x)$ denotes the exterior unit normal vector at $x \in \partial\omega$ and therefore $v \cdot n > 0$ results in a reduction of the mass in ω.

Using $\int_t^{t+\Delta t} g(r)\,dr = \Delta t\, g(t) + O(\Delta t^2)$ for sufficiently smooth g, passing to the limit $\Delta t \to 0$ and applying Gauß' theorem $\int_\omega \nabla \cdot u\,dx = \int_{\partial\omega} u \cdot n\,ds$ we obtain from (2.1) the integro-differential form of the conservation law:

$$\partial_t \int_\omega \rho(x,t)\,dx + \int_\omega \nabla\cdot(\rho(x,t)v(x,t))\,dx = \int_\omega f(x,t)\,dx \qquad \text{(for any } \omega\text{)}. \tag{2.2}$$

For sufficiently smooth functions, the fact that (2.2) holds for any ω implies the final differential form of the mass conservation law (see e.g. [Smirnow, 1981, §

[1] We always state units in the MKS (meter kilogram second) system.

74], this is the same argument used in deriving eq.(1.6)):

$$\partial_t \rho(x,t) + \nabla \cdot (\rho(x,r)v(x,r)) = f(x,t), \qquad x \in \Omega. \tag{2.3}$$

If the fluid is incompressible then $\rho(x,t) = \text{const}$ implies

$$\nabla \cdot v(x,t) = f(x,t), \qquad x \in \Omega \tag{2.4}$$

which further reduces to $\nabla \cdot v = 0$ when there are no sources and sinks present (i.e the velocity field of an incompressible fluid without sources and sinks is divergence free).

2.3. Conservation of Energy

The other conserved quantities energy and momentum can be imagined as being attached to mass. In the case of energy we set $e(x,t) = \rho(x,t)u(x,t)$, where e is the energy density with units J m^{-3} and u is the specific energy with units J kg^{-1}. We can compute the energy stored in the material occupying the volume ω as

$$E_\omega(t) = \int_\omega e(x,t)\,dx = \int_\omega \rho(x,t)u(x,t)\,dx\ .$$

Repeating the reasoning given above with ρ replaced by ρu yields the energy conservation equation

$$\partial_t(\rho(x,t)u(x,t)) + \nabla \cdot q(x,t) = f(x,t), \qquad x \in \Omega, \tag{2.5}$$

where $q(x,t)$ is now the energy density flux vector. If energy is simply flowing with the fluid (e.g. no conductive heat transport) we have $q = \rho u v$.

2.4. Conservation of Linear Momentum

Similarly the (linear) momentum density (having units momentum per volume) is defined as ρv. Integration over an arbitrary volume ω gives the total momentum in ω:

$$P_\omega(t) = \int_\omega \rho(x,t)v(x,t)\,dx\ .$$

Note however, that $P(x,t)$ is a vector-valued function! For each component ρv_i of the momentum density vector we obtain the conservation equation

$$\partial_t(\rho(x,t)v_i(x,t)) + \nabla \cdot j_i = f_i(x,t), \qquad x \in \Omega,\ i = 1,\ldots,d,$$

where j_i is the momentum density flux vector for the given component. If momentum is only transported with the fluid (as in inviscid flow, see § 2.7) we have $j_i = \rho v_i v$.

By defining the j_i to be the rows of the matrix J and defining $\nabla \cdot J$ as applying the divergence to each row (yielding a vector, see § A.2.3) one can write the momentum conservation law in compact form as

$$\partial_t(\rho(x,t)v(x,t)) + \nabla \cdot J = f(x,t), \qquad x \in \Omega. \tag{2.6}$$

In the case of inviscid flow we then have $J = \rho v v^T$. The term $\partial_t(\rho v)$ on the left hand side is rate of change of momentum density which is a force density (units N m^{-3}). Equation (2.6) is Newton's second law generalized to spatially extended bodies.

2.5. Heat Transfer

As an application of conservation laws we consider the flow of heat in a solid or fluid filling the bounded domain $\Omega \subset \mathbb{R}^3$. The conserved quantity is the thermal energy. Its density e is assumed to be proportional to temperature

$$e = \rho c T$$

where c is the specific heat capacity in J kg^{-1} K^{-1}, ρ is the mass density of the material in kg m^{-3} and the absolute temperature T is given in Kelvin K.

In fluids and solids the flow of thermal energy is modelled as

$$q_d = -\lambda \nabla T$$

which is known as *Fourier's law* or *diffusive heat flux*. It states that flow is in direction of the steepest descent of temperature. The constant of proportionality is the heat conductivity $\lambda > 0$ with units J s^{-1} m^{-1} K^{-1}. Heat conductivity may depend on position and time (e.g. in a fluid with varying composition).

In a fluid thermal energy is also transported with the fluid velocity v which gives rise to a *convective heat flux*

$$q_c = ev = \rho c T v.$$

The total flux is then the sum of convective and diffusive flux. Inserting all this into the conservation law (2.5) (now with $u = cT$) we obtain the *convection-diffusion equation*

$$\partial_t(\rho c T) + \nabla \cdot (\rho c T v - \lambda \nabla T) = f \qquad \text{in } \Omega \tag{2.7}$$

which is a scalar linear second-order PDE. In order to fully determine the temperature $T(x,t)$ for $x \in \Omega$ and $t > 0$, boundary conditions

$$T(x,t) = g(x,t) \quad (x \in \Gamma \subseteq \partial\Omega,\, t > 0,\, \textit{Dirichlet}), \tag{2.8a}$$
$$(\rho c T v - \lambda \nabla T)(x,t) \cdot n(x) = j(x,t) \quad (x \in \partial\Omega \setminus \Gamma,\, t > 0,\, \textit{Neumann}) \tag{2.8b}$$

and the initial condition

$$T(x,0) = T_0(x) \qquad (x \in \Omega) \tag{2.9}$$

must be given.

Modeling sources and sinks The right hand side f with units J s^{-1} m^{-3} of equation (2.7) models sources and sinks. In a solid this rate is usually known. In a fluid the source/sink term depends on the temperature of the fluid going in or out of the domain. It can be modelled as $f = rcT$ where r in kg s^{-1} m^{-3} is the amount of fluid entering or leaving the domain. When $r > 0$ fluid (and with it thermal energy) is going in and the temperature of this fluid is assumed to be known. When $r < 0$ fluid is going out and the temperature of this fluid is unknown and must be computed. This leads to the final form, the so-called *convection-diffusion-reaction equation*:

$$\partial_t(\rho c T) + \nabla \cdot (\rho c T v - \lambda \nabla T) + rcT = f \qquad \text{in } \Omega. \tag{2.10}$$

Note that in this equation all coefficient functions may depend on position and time.

Special Cases Several important special cases of this equation can be stated:

a) No convective flux (*reaction-diffusion equation*):

$$\partial_t(\rho c T) - \nabla \cdot (\lambda \nabla T) + rcT = f \qquad \text{in } \Omega.$$

b) No diffusive flux (*first-order PDE*):

$$\partial_t(\rho c T) + \nabla \cdot (\rho c T v) + rcT = f \qquad \text{in } \Omega.$$

c) Stationary heat flow (all coefficients are independent of time):

$$\nabla \cdot (\rho c T v - \lambda \nabla T) + rcT = f \qquad \text{in } \Omega.$$

d) Stationary heat flow in a solid with a sink (this will be our model equation to introduce the finite element method):

$$-\nabla \cdot (\lambda \nabla T) + rcT = f \qquad \text{in } \Omega. \tag{2.11}$$

e) Stationary heat flow with constant conductivity and no sinks (again we obtain *Poisson's equation*):

$$-\nabla \cdot (\nabla T) = -\Delta T = f \qquad \text{in } \Omega.$$

Example 2.1. Figure 2.1 illustrates the solution for a three-dimensional heat transfer problem. The domain is $\Omega = (0,3) \times (0,3) \times (0,1)$ and the parameters were $v = 0$ (no convective flux), $\rho = 1$, $c = 1$, $\lambda = 1$, $r = 0$ and $f = 0$. The lateral boundaries and the region $(1,2) \times (1,2) \times \{1\}$ on the top boundary were isolated, i.e. $\nabla T \cdot n = 0$, the bottom boundary was held at constant temperature $T = 8$ and at the remaining part of the top boundary a Dirichlet condition oscillating in space and time was given. Practically, one can imagine a piece of subsurface that is heated periodically from the top and that is held at constant temperature from below. The Figure shows that the oscillations are quickly dampened by the diffusion, a fact that is also observed in nature. □

Another important feature of the solution of the heat transfer problem without sources and sinks and divergence free velocity field v is that the maximum (minimum) temperature in the interior of the domain Ω does not exceed (go below) the maximum (minimum) temperature at the boundary and initial condition. This is called a maximum principle. For details we refer to [Hackbusch, 1986] or [Evans, 2010].

Multiscale Problems Multiscale problems are problems with highly oscillating coefficient functions. Imagine a heterogeneous solid composed of two materials with different heat conductivity coefficient. The two materials occupy different regions of space and are arranged in a periodic fashion with periodicity ϵ in all directions:

$$\lambda_\epsilon(x) = \hat{\lambda}\left(\frac{x}{\epsilon}\right), \qquad \hat{\lambda}(x + e_i) = \hat{\lambda}(x) \quad (i = 1, \ldots, n) \tag{2.12}$$

(e_i being the ith cartesian unit vector). The 1-periodic coefficient function $\hat{\lambda}$ taken in $\Omega = (0,1)^n$ defines the "unit cell". Then we consider the family of stationary heat transfer problems

$$-\nabla \cdot (\lambda_\epsilon(x) \nabla T_\epsilon) = f \qquad \text{in } \Omega \tag{2.13}$$

depending on the parameter $\epsilon > 0$ together with appropriate boundary conditions.

Example 2.2. We consider an example of a multiscale problem in two space dimensions. Figure 2.2 on the left shows the setup of the macroscopic problem

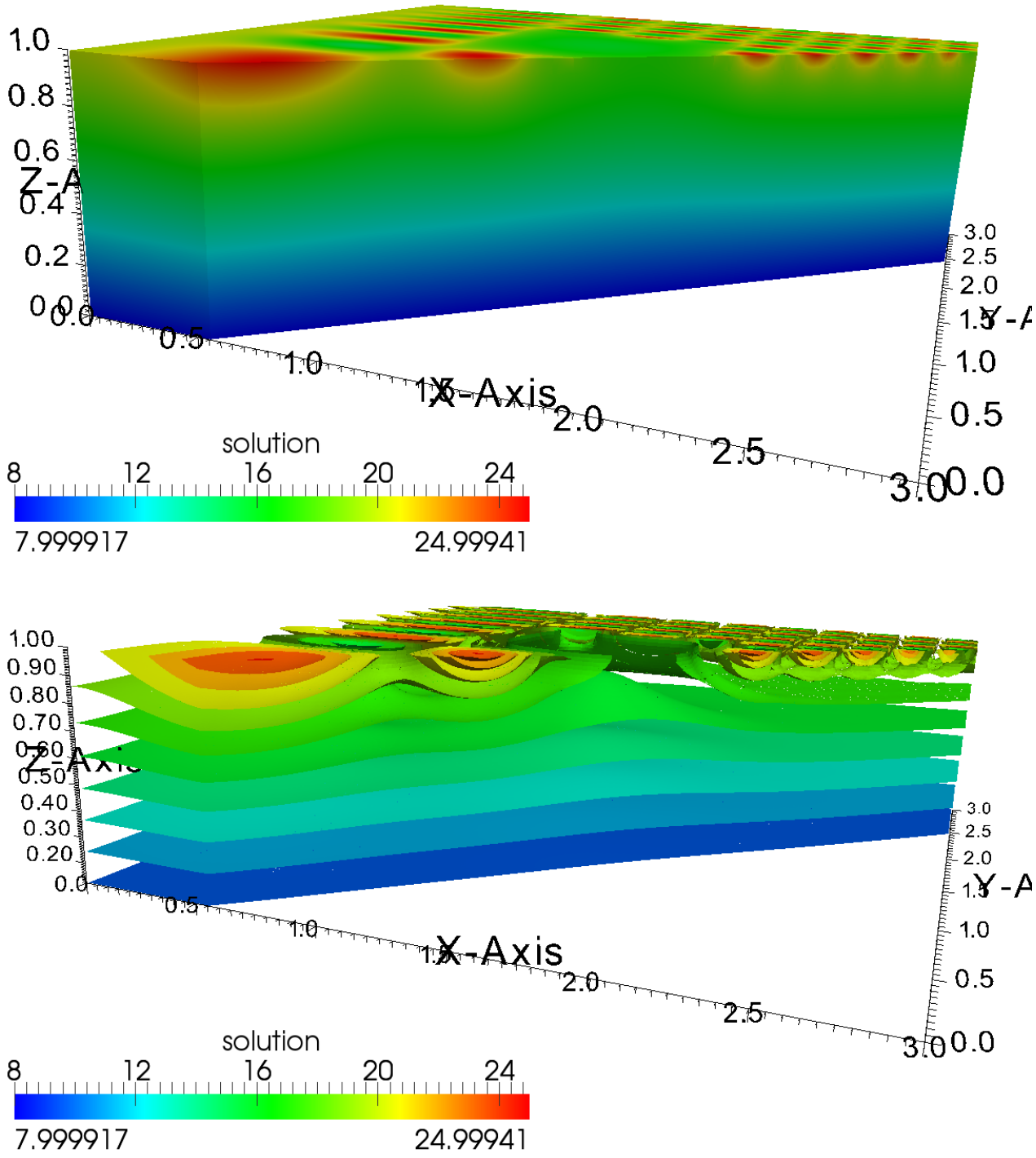


Figure 2.1.: Solution of a 3d heat transfer problem (details given in the text).

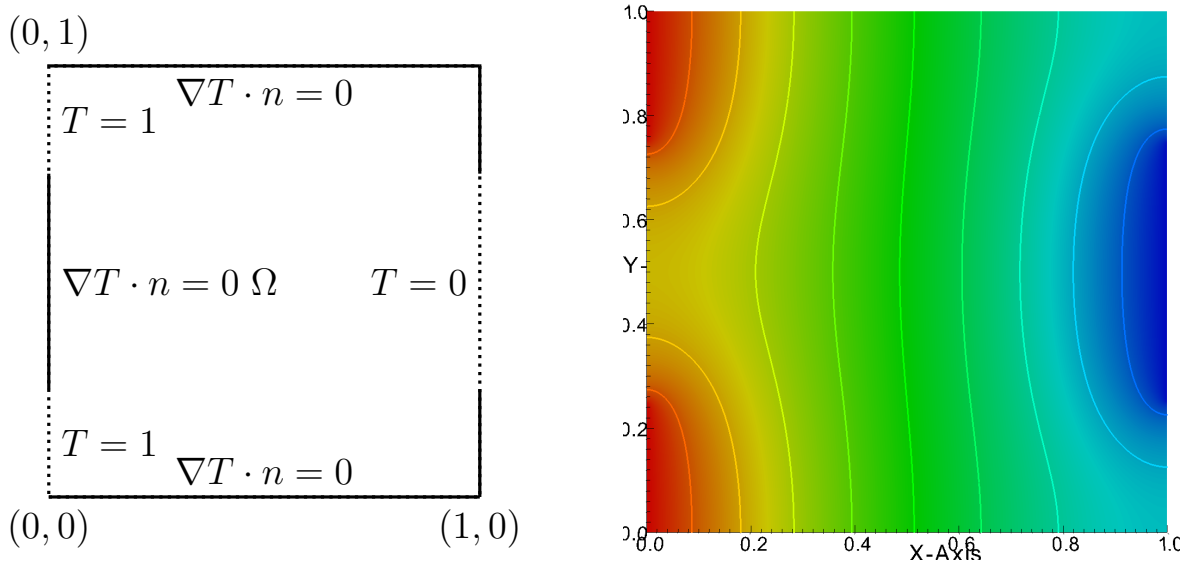


Figure 2.2.: Setup and solution for homogenous coefficient in the multiscale example.

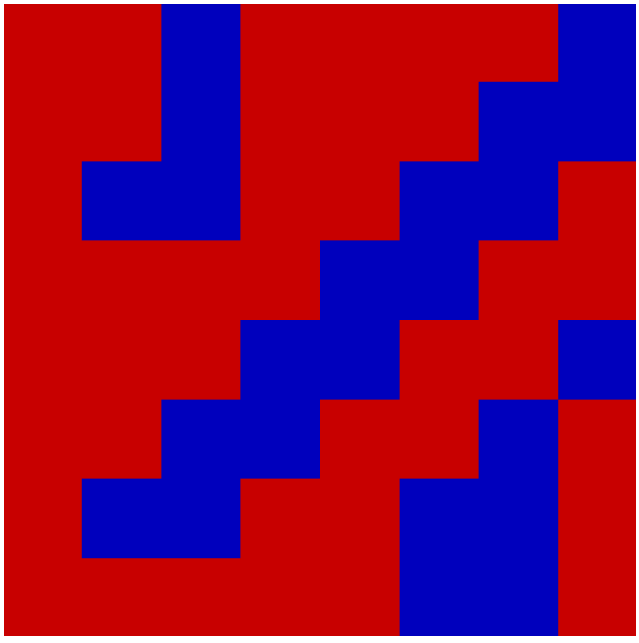


Figure 2.3.: Conductivity distribution in the unit cell.

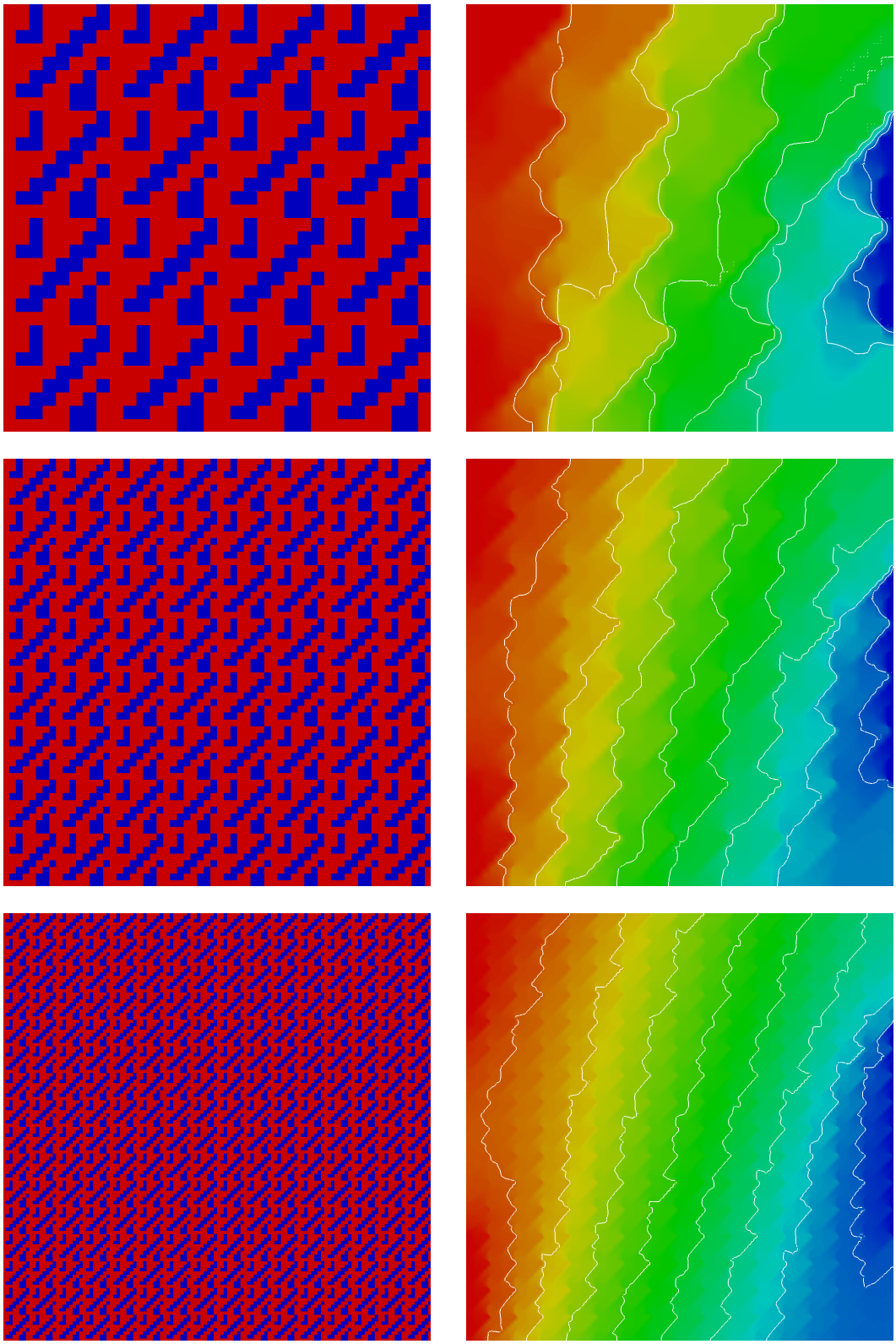


Figure 2.4.: Example of a multiscale problem in 2d (details given in the text).

and the image to the right shows the solution to this problem with a homogeneous conductivity coefficient. Now we solve a problem with the same boundary conditions and a heterogeneous periodic coefficient as defined above. The conductivity distribution in the unit cell is shown in Figure 2.3 and in Figure 2.4 the solution for $\epsilon = 1/4$, $\epsilon = 1/8$ and $\epsilon = 1/16$ is shown. The solutions suggest that for $\epsilon \to 0$ the solution T_ϵ converges to a smooth function. For finite $\epsilon > 0$ the solution has small oscillations of the order ϵ. □

In practical applications $\epsilon \ll 1$ and computing T_ϵ is prohibitively expensive. Moreover one is only interested in the macroscopic behaviour and not in the behaviour on the scale ϵ. Homogenization theory, see e.g. [Kozlov et al., 1994], shows that the limit solution $T = \lim_{\epsilon \to 0} T_\epsilon$ can be computed as the solution of a homogeneous heat transfer problem

$$-\nabla \cdot (\Lambda \nabla T) = f \qquad \text{in } \Omega$$

where the *effective coefficient* $\Lambda \in \mathbb{R}^{n \times n}$ is a symmetric and positive definite matrix that only depends on the conductivity distribution in the unit cell and is therefore cheap to compute. From example 2.2 it becomes clear that the effective coefficient cannot just be a scalar as in this case the solution would be symmetric around $y = 1/2$ as in Figure 2.2. Instead, the contour lines are tilted to the right because the material conducts better in the direction $(1, 1)$ than in the direction $(1, -1)$.

The discussion so far involved only two scales, the macroscopic scale of interest and the scale ϵ (actually there is a third scale, the atomistic scale that has has already been eliminated by deriving the heat transfer equation). In practice, there might be more than two scales involved. For an effective solution it is important that the macroscopic scale of interest and the small scales (one is not really interested in) are clearly separated.

Another situation arises when the precise arrangement of the materials is unknown, as is often the case with natural materials (such as e.g. rock). Then a stochastic approach may be appropriate leading to the field of stochastic partial differential equations.

2.6. Flow in Porous Media

Flows in porous media such as the subsurface, foams or biological tissue are highly relevant in practice. In such problems at least three different scales are involved as shown in Figure 2.5. On the microscopic scale (b), which is the scale of the sand grains, the flow can be described (under suitable assumptions) by a continuum approach (the Navier-Stokes equations introduced below) ignoring the individual molecules of the fluid. However, we are usually interested in the

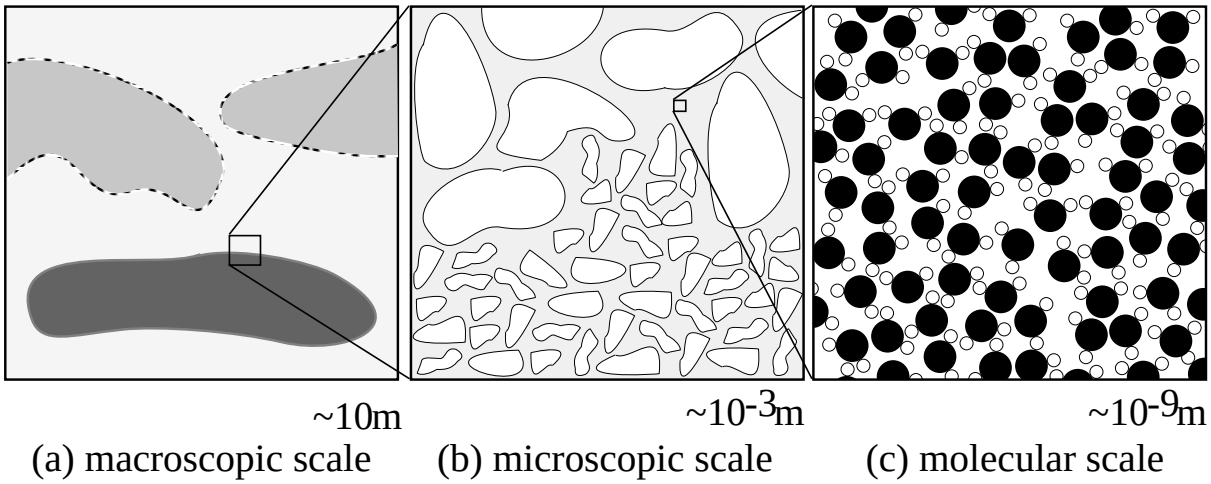


Figure 2.5.: Different scales involved in porous media flow.

flow on the macroscopic scale comprising a huge number of pores where the individual geometry is usually not available in detail.

Under certain assumptions equations describing the flow on the macroscopic scale can be derived. Conservation of fluid mass is expressed by

$$\partial_t(\Phi\rho) + \nabla \cdot \{\rho v\} = f \qquad \text{in } \Omega \tag{2.14}$$

where the scalar function $\Phi : \Omega \to (0, 1)$ describes the *porosity* of the porous medium which is the fraction of the volume available to fluid flow, ρ is the mass density of the fluid and v is the apparent velocity of the fluid on the macroscopic scale.

In 1856 Henry Darcy stated a phenomenological law describing the fluid flow in response to a given pressure drop which is since then known as *Darcy's law*:

$$v = -\frac{K}{\mu}(\nabla p - \rho g). \tag{2.15}$$

Here $p(x, t)$ is the fluid pressure with units Pa = N m^{-2}, K is the permeability tensor with units m^2 describing the pore structure of the porous medium, μ is the dynamic viscosity of the fluid with units Pa s and $g = (0, 0, -9.81)^T$ is the gravity vector pointing in negative z-direction and having units of acceleration m s^{-2}.

Inserting Darcy's law into the mass conservation equation yields the flow equation:

$$\partial_t(\Phi\rho) - \nabla \cdot \left\{ \rho \frac{K}{\mu}(\nabla p - \rho g) \right\} = f \qquad \text{in } \Omega \tag{2.16}$$

with ρ and p to be determined. In case of an incompressible fluid $\rho = const$ and

the equation reduces to the stationary equation

$$-\nabla \cdot \left\{ \frac{K}{\mu}(\nabla p - \rho g) \right\} = f/\rho \qquad \text{in } \Omega \tag{2.17}$$

which is a again a linear second-order PDE for pressure.

In case of an ideal gas $\rho = \rho(p)$ and we obtain a time-dependent PDE.

Geothermal Power Plant The flow of heat in a porous medium can be modeled by the modified heat transfer equation

$$\partial_t(s(T)T) + \nabla \cdot (\rho_f(T)c_f T v - \lambda \nabla T) = f \qquad \text{in } \Omega \tag{2.18}$$

where $s(T) = (1-\Phi)\rho_s c_s + \Phi \rho_f(T) c_f$ is now the effective volumetric heat capacity in J m^{-3} K^{-1} of the combined water/rock mixture (with ρ_s, c_s being mass density and specific heat capacity of the solid and $\rho_f(T)$, c_f being mass density and specific heat capacity of the fluid) and λ is the effective heat capacity of the water/rock mixture.

Together equations (2.16) (with $\rho = \rho_f(T)$) and (2.18) form a coupled time-dependent, nonlinear system of PDEs (also the dynamic viscosity depends on temperature) which can be used e.g. to model the performance of a geothermal power plant.

2.7. Inviscid Fluid Flow

The flow of a gas is a very interesting and important problem. It has applications e.g.in weather and climate prediction or in star formation and the development of galaxies in astronomy. Figure 2.6 shows an image of the Cone Nebula in the galaxy NGC 2264 which is just a pillar of gas and dust. It is supposed to be a region where new stars are formed. In this section, we consider the flow of a gas ignoring the effect of internal friction. Besides the conserved quantities density, linear momentum and energy an additional concept is needed to derive the governing equations.

Pressure In a gas that is macroscopically at rest the molecules still perform a random motion at the microscopic level. The molecules hitting the walls of the container excert a macroscopic force that must be counterbalanced by the rigid wall. This force per unit area is called pressure with units N m^{-2}. Through experiment one finds that the force per unit area excerted by the gas (at constant pressure) is always the same regardless of the shape of the wall. Therefore, the (scalar) pressure is the magnitude of a force (per unit area) that acts always perpendicular to the wall of the container (i.e. in the exterior normal direction).

http://www.spacetelescope.org/images/heic0206c/

Figure 2.6.: Cone Nebula (NASA/ESA image taken with the Hubble Space Telescope).

If we would suddenly introduce a new (infinitely thin) wall inside the container (imagine a test volume ω) a force (per unit area) would be exerted at every point from each side of the wall that has equal magnitude and opposite direction so that it cancels out. We can therefore imagine pressure to be a (scalar) quantity that is defined everywhere in the gas.

The effect of pressure (being a force per unit area) needs to be considered in the momentum balance equation (2.6). If we consider a small test volume ω then the total force (including the direction) acting on the surface is given by

$$-\int_{\partial\omega} pn\,ds = -\int_{\omega} \nabla\cdot(pI)\,dx = -\int_{\omega} \nabla p\,dx. \tag{2.19}$$

This term is part of the right hand side of the integral version of equation (2.6) and I denotes the identity matrix. Note how the force always acts in negative normal direction. The sign can be understood as follows. Imagine the test volume to be a cuboid and consider e.g. the x-direction with the two faces located at x_1, x_2 with $x_1 < x_2$ and corresponding normal directions $n_1 = (-1,0,0)^T$ and $n_2 = (1,0,0)^T$. Then x- momentum must increase when pressure acts at the face at x_1 and it must decrease when pressure acts at the face at x_2. Note also, that equal pressure at x_1 and x_2 does have a zero net effect for the x-momentum in the test volume (so it is pressure difference that does have an effect).

The pressure contribution is sometimes called an interior force to distinguish it from exterior forces (such as e.g. gravity) which are only present in open systems.

Energy In a macroscopic body of gas the total energy consists of two different forms of energy, the internal energy (translation, rotation and vibration of the molecules on the microscopic level) and the macroscopic kinetic energy due to the movement of the fluid that is macroscopically observed. Using the concept of densities we write this as

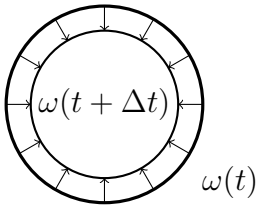
$$e = \rho u + \rho \|v\|^2/2 \tag{2.20}$$

with e the total energy density in J m^{-3} and u the specific internal energy in J kg^{-1}. According to the theory of gases an algebraic relation, called an "equation of state" (depending on the type of gas), of the form

$$u = u(\rho, p) \tag{2.21}$$

relating specific internal energy, density and pressure can be derived. A well-known example is the ideal gas law $p = \bar{R}\rho T = \rho u$ (here the internal energy $u = \bar{R}T$ is proportional to temperature). The specific gas constant $\bar{R}$ has the same units as the specific heat capacity. See below for another popular example of an equation of state.

Total energy e is a conserved quantity that is transported with the fluid with a flux $q = ev$. On the right hand side of the energy balance equation (2.5) internal work done in the fluid has to be considered. This internal work is known as "volume changing work" and can be experienced when using a bicycle pump: when a gas is compressed (i.e. its volume is decreased), it heats up.



We can derive the expression for volume changing work as follows: Imagine a set of molecules occupying the volume $\omega(t)$ at time t (see figure to the right). The same particles are contained in $\omega(t+\Delta t) \subset \omega(t)$ at small time interval Δt later. Subdividing $\partial\omega(t)$ into small surface elements Δs_i the work done against pressure of the gas in the time interval Δt is to first order

$$\begin{aligned}\Delta W_\omega(t) &= -\lim_{N\to\infty} \sum_{i=1}^{N} \underbrace{p(x_i,t)ds_i}_{\text{normal force}} \underbrace{v(x_i)\cdot n_i \Delta t}_{\text{distance}} \\ &= -\Delta t \int_{\partial\omega(t)} pv \cdot n \, ds = -\Delta t \int_{\omega(t)} \nabla \cdot (pv)\, dx.\end{aligned}$$

The sign is chosen such that compression ($v \cdot n < 0$) results in a positive value.

Euler Equations Considering the internal forces due to pressure in the momentum balance law and the volume change work in the energy balance law we

obtain the famous nonlinear system of partial differential equations known as the Euler equations of gas dynamics in conservative form

$$\partial_t \rho + \nabla \cdot (\rho v) = m, \tag{2.22a}$$
$$\partial_t (\rho v) + \nabla \cdot (\rho v v^T + pI) = f, \tag{2.22b}$$
$$\partial_t e + \nabla \cdot ((e+p)v) = w, \tag{2.22c}$$

which together with the thermodynamical relation

$$p = p(\rho, e) = (\gamma - 1)(e - \rho \|v\|^2/2) \tag{2.23}$$

and appropriate boundary and initial conditions describe the flow of a polytropic ideal gas. The functions m, f and w denote the mass source term, the external forces and the energy source term. Equation (2.23) is a consequence of the equation of state $u = p/((\gamma - 1)\rho)$ and the definition of total energy (2.20). The constant γ is the adiabatic exponent and depends on the type of gas. For more details, see [Leveque, 2002, § 14.4]. Pressure is considered a dependent variable in (2.22) which can be eliminated using (2.23) resulting in a system of five equations for the five unknown functions ρ, v_1, v_2, v_3 and e in three space dimensions. It is interesting to note that we can combine all the equations (2.22) into a single equation for the unknown vector function $w = (\rho, \rho v, e)^T$:

$$\partial_t w + \nabla \cdot F(w) = g \tag{2.24}$$

with

$$F(w) = \begin{pmatrix} \rho v_1 & \rho v_2 & \rho v_3 \\ \rho v_1 v_1 + p(\rho, e) & \rho v_1 v_2 & \rho v_1 v_3 \\ \rho v_2 v_1 & \rho v_2 v_2 + p(\rho, e) & \rho v_2 v_3 \\ \rho v_3 v_1 & \rho v_3 v_2 & \rho v_3 v_3 + p(\rho, e) \\ (e + p(\rho, e)) v_1 & (e + p(\rho, e)) v_2 & (e + p(\rho, e)) v_3 \end{pmatrix}. \tag{2.25}$$

An equation of the general form (2.24) is called a (nonlinear) conservation law. Yet another often encountered form is obtained by writing out the divergence:

$$\partial_t w + \sum_{j=1}^{n} \partial_{x_i} F_j(w) = g \tag{2.26}$$

where $F_j(w)$ is the j-th column of $F(w)$. Various other forms of the Euler equations can be found in the literature, most notably the nonconservative formulation. But (2.22) is the most general form that is also valid e.g. in the case of strong density contrasts.

2.8. Propagation of Sound Waves

Sound waves are small variations in pressure (and correspondingly density) that move through the gas. In order to derive an equation for the propagation of these variations we start with the Euler equations (2.22). We write all quantities as a constant background value (indicated by the bar) plus a small variation depending on space and time (indicated by the tilde):

$$\rho = \bar{\rho} + \tilde{\rho}, \qquad p = \bar{p} + \tilde{p}, \qquad v = \bar{v} + \tilde{v}.$$

The background velocity is actually assumed to be zero, $\bar{v} = 0$, and the temperature of the gas is assumed to be constant throughout the domain. From the ideal gas law we get $p = c^2\rho$ with $c = \sqrt{\bar{R}T}$ the speed of sound and therefore $p = c^2\rho = c^2(\bar{\rho} + \tilde{\rho}) = c^2\bar{\rho} + c^2\tilde{\rho} = \bar{p} + \tilde{p}$.

Linearizing mass and momentum equations around the background state, dropping all higher-order terms in fluctuations (note especially that $\tilde{v}\tilde{v}^T$ can be dropped) and assuming *constant background pressure* results in (no external sources)

$$\partial_t\tilde{\rho} + \nabla \cdot (\bar{\rho}\tilde{v}) = 0, \tag{2.27a}$$
$$\partial_t(\bar{\rho}\tilde{v}) + \nabla\tilde{p} = 0. \tag{2.27b}$$

Nonconservative Form of Linear Acoustics Using $\tilde{\rho} = \tilde{p}/c^2$ and assuming that c is constant throughout the domain the density variation is eliminated and we obtain the equations of linear acoustics:

$$\partial_t\tilde{p} + c^2\bar{\rho}\nabla \cdot \tilde{v} = 0, \tag{2.28a}$$
$$\bar{\rho}\partial_t\tilde{v} + \nabla\tilde{p} = 0. \tag{2.28b}$$

Taking the temporal derivative of the first equation and applying the divergence to the second the velocity variation can be eliminated from this system and we obtain the so-called wave equation:

$$\partial_t^2\tilde{p} - c^2\Delta\tilde{p} = 0. \tag{2.29}$$

In the analysis of the wave equation, (2.29) is often reduced to a first order system by setting $u = \partial_t\tilde{p}$ and $w = -\nabla\tilde{p}$. Together with the identities $\partial_{x_i}\partial_t\tilde{p} = \partial_t\partial_{x_i}\tilde{p}$ we obtain the system

$$\partial_t u + c^2\nabla \cdot w = 0,$$
$$\partial_t w + \nabla u = 0,$$

which is equivalent to (2.28) (simply use the transformation $w = \bar{\rho}\tilde{v}$). It should be noted that it is the first order system that is derived from the physics and not the scalar second order wave equation, see also [Leveque, 2002, § 2.7].

Conservative Form of Linear Acoustics We now consider the case that the speed of sound c is *piecewise constant* (e.g. due to temperature variations). Equation (2.27) is still valid in this case since only $\bar{p}$ being constant has been assumed. We conclude that pressure $\tilde{p}$ and normal momentum $\bar{\rho}\tilde{v}\cdot n$ are continuous at subdomain boundaries where c is discontinuous.

Due to $\rho = p/c^2 = (\bar{p}+\tilde{p})/c^2 = \bar{p}/c^2 + \tilde{p}/c^2 = \bar{\rho} + \tilde{\rho}$ also the background density $\bar{\rho}$ is piecewise constant. In case of varying speed of sound it is then more appropriate to use the conservative variables $(\tilde{\rho}, \bar{\rho}\tilde{v}) = (\tilde{\rho}, \tilde{q})$ resulting in the system

$$\partial_t \tilde{\rho} + \nabla \cdot \tilde{q} = 0, \tag{2.30a}$$

$$\partial_t \tilde{q} + \nabla(c^2\tilde{\rho}) = 0. \tag{2.30b}$$

At subdomain boundaries where c is discontinuous $c^2\tilde{\rho}$ and $\tilde{q}\cdot n$ are continuous. Figure 2.7 shows the results of a linear acoustics simulation. The conservative formulation has been used and only the density variations $\tilde{\rho}$ are shown. In the upper right part of the domain the speed of sound is smaller than in the rest of the domain. When the wave hits the internal boundary the velocity (not shown) becomes smaller and the density variations increase. An inverted wave is reflected at the internal boundary. Reflective boundary conditions at the out boundary have been used.

Waves in Solids Solid bodies are also able to support a propagation of waves, an example being earthquakes. In the one-dimensional situation we may imagine a string of beads connected by springs with each other. One type of wave consists of small displacements of a bead in the direction of the string resulting in displacements of the neighbouring beads. This type of wave is called a compression wave or P-wave and it is similar to the sound waves in a gas. Another type of wave results from displacements of a bead in a direction perpendicular to the string which also results in the propagation of a wave in the direction of the string. This is called S-wave which usually travels slower than a P-wave. In the one-dimensional situation both types of waves are described by the one-dimensional wave equation $\partial_t^2 u - c^2\partial_x^2 u = 0$ (A derivation of the P-wave is in [Eriksson et al., 1996, § 17.2] and the S-wave can be found in [Smirnow, 1981, § 176]). In a multi-dimensional solid both types of waves interact and more complicated equations result (see [Leveque, 2002, § 2.12] for some discussion). At the surface or at internal boundaries surface waves can be observed.

2.9. Viscous Fluid Flow

In many real fluids the effect of internal friction cannot be neglected. In a Newtonian fluid the stress tensor describing the additional flux of linear momentum

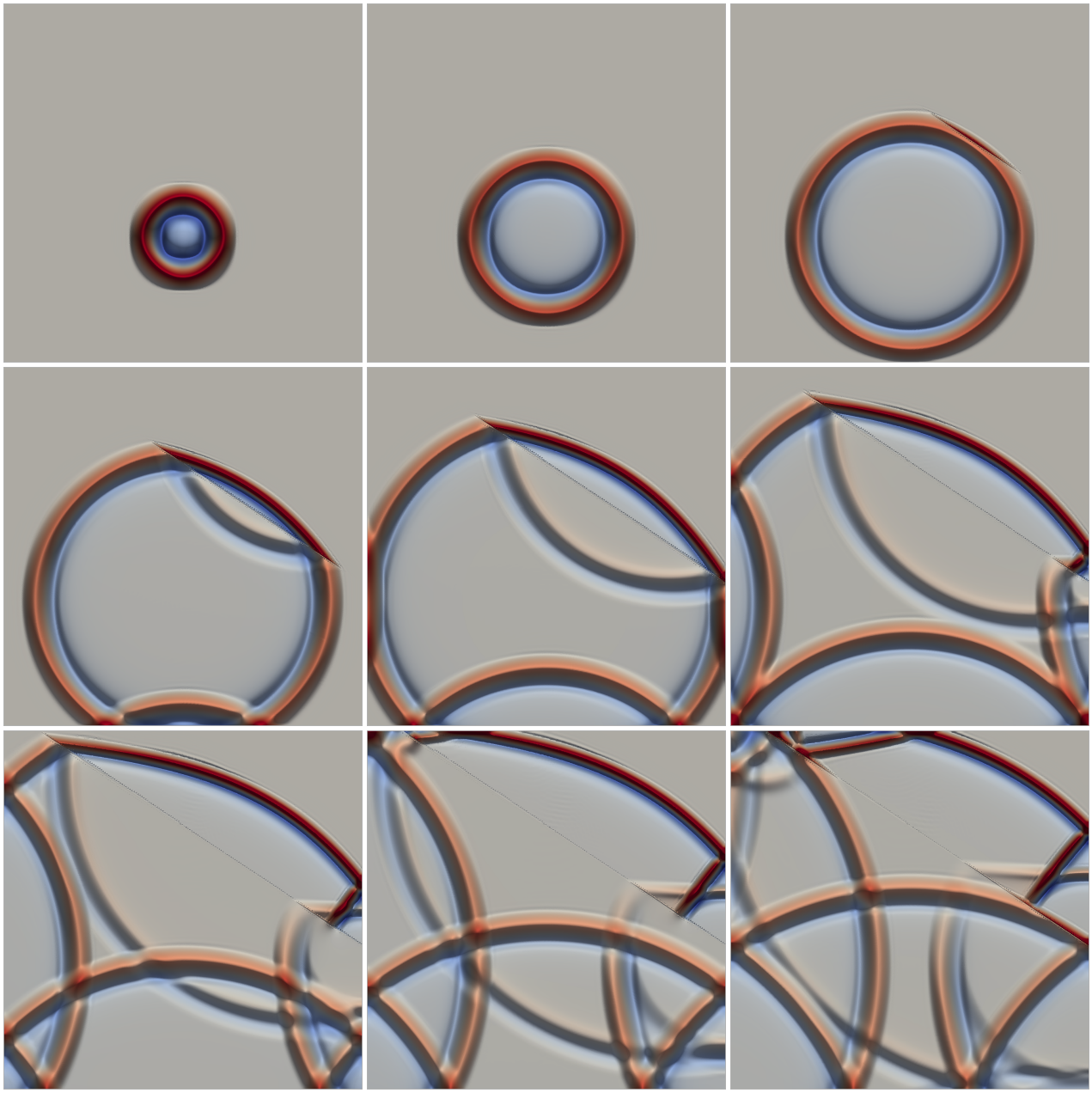


Figure 2.7.: Acoustic wave propagation in a heterogeneous medium with reflective boundary conditions. Fully third-order discontinuous Galerkin scheme. Time sequence goes from top left to bottom right.

is proportional to gradients of velocity. The result is the system of *compressible Navier-Stokes equations*:

$$\partial_t \rho + \nabla \cdot (\rho v) = m, \tag{2.31a}$$
$$\partial_t (\rho v) + \nabla \cdot (\rho v v^T + pI - \tau(v)) = f, \tag{2.31b}$$
$$\partial_t e + \nabla \cdot ((e+p)v - \tau(v)v - \lambda \nabla T(e, \rho, v)) = w, \tag{2.31c}$$

with the *stress tensor*

$$\tau(v) = 2\mu \left[D(v) - \frac{1}{3}(\nabla \cdot v) I \right] \tag{2.32}$$

where shear viscosity μ is a parameter of the fluid and the *rate of strain tensor*

$$D(v) = \frac{1}{2} \left(\nabla v + (\nabla v)^T \right). \tag{2.33}$$

There are three new terms in the Navier-Stokes equations (2.31) compared to the Euler equations (2.22). The last term on the right hand side of the momentum equation describes the forces due to internal friction. The term $\tau(v)v$ in the energy equation describes the energy flux due to internal friction and $-\lambda \nabla T$ describes the heat conduction (Temperature T is a function of the state variables). Depending on the application, e.g. in star formation, heat transfer might also include the effect of radiation.

A full derivation of the new terms in the Navier-Stokes equations is beyond the scope of these lecture notes, we refer e.g. to [Chung, 1996] for details. Note however, that all the new terms involve second derivatives, i.e. the Navier-Stokes equations are a second-order system of PDEs.

Incompressible Viscous Flow In many applications the fluid can be regarded as *incompressible* which means that density is independent of pressure. If temperature variations are also insignificant it is a constant. Neglecting also the energy equation (because temperature is assumed to have no effect on the fluid) and assuming that the fluid enters and leaves the domain only via the boundary ($m = 0$) results in the system of equations known as the *incompressible Navier-Stokes equations*:

$$\nabla \cdot v = 0, \tag{2.34a}$$
$$\partial_t v + \nabla \cdot (v v^T) - \nu \Delta v + \nabla p = f, \tag{2.34b}$$

with the kinematic viscosity $\nu = \mu/\rho$. Here p (which has been rescaled by $1/\rho$) is now an independent variable to be determined. In order to derive the momentum equation (2.34b) the incompressibility constraint $\nabla \cdot v = 0$ has been

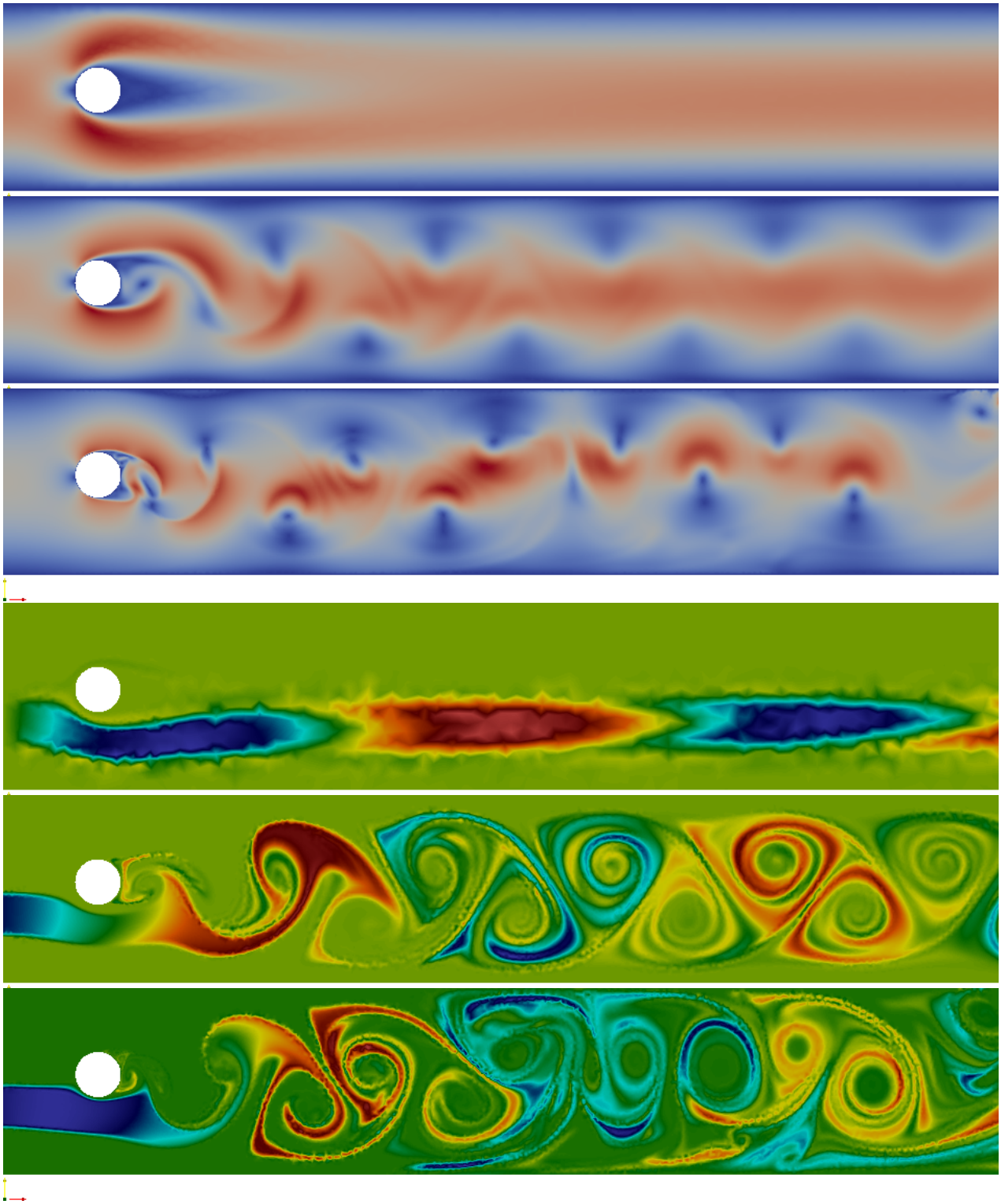


Figure 2.8.: Flow behind a cylinder for Reynolds numbers 20, 200 and 1500. Numerical computation with P_2/P_1 Taylor-Hood elements in space and Alexander scheme in time (results provided by Marian Piatkowski). Top three images shows velocity magnitude, bottom three images show concentration of a tracer. The flow is stationary for Re 20, periodic for 200 and turbulent for Re 1500.

applied twice: once to simplify the stress tensor τ and a second time to conclude $\nabla \cdot D(u) = \Delta u$ (i.e. the assumption $m = 0$ is essential in deriving the equations).

Figure 2.8 shows an example of incompressible flow around a cylindrical obstacle in a two-dimensional channel.

Chapter 3.

Calculus of Variations

In this chapter we present a general approach that is used to study many mechanical and geometrical problems. For simplicity it will be illustrated by modelling the deflection of an elastic string where it leads to a two-point boundary value problem in ordinary differential equations. The general principle, however, applies to the multi-dimensional situation and is essential to understand the finite element method for the numerical solution of partial differential equations. The motivating example in this chapter is taken from the lecture notes of Hiptmair [2010].

3.1. Equilibrium Principle

We are interested in modelling the deflection of an elastic string under a load. As an example consider a string where cloth is put on for drying. The property of elasticity means that after the load is removed the string returns exactly to its unloaded position without any lasting effect. In order to derive the model we will first consider systems of finitely many straight and ideal springs connected together. Then we will derive a continuum version by an appropriate limit process.

Discrete Spring System Figure 3.1 shows a system of $n \in \mathbb{N}$ point masses $m_1^{(n)}, \ldots, m_n^{(n)}$ located at the positions $u_1^{(n)}, \ldots, u_n^{(n)}$ and connected by springs. To each mass $m_i^{(n)}$ a constant force given by the vector $f_i^{(n)}$ is applied. Assuming all forces are applied in a plane we have $u_i^{(n)} \in \mathbb{R}^2$. Spring number i, $0 \leq i \leq n$, is elongated from position $u_i^{(n)}$ to position $u_{i+1}^{(n)}$ with the two endpoints

$$u_0^{(n)} = \begin{pmatrix} x_a \\ z_a \end{pmatrix}, \qquad u_{n+1}^{(n)} = \begin{pmatrix} x_b \\ z_b \end{pmatrix} \tag{3.1}$$

held fixed. All interior positions to be determined are collected in a big vector

$$u^{(n)} = (u_1^{(n)}, \ldots, u_n^{(n)})^T \in \mathbb{R}^{2n}$$

which completely describes the state of the system.

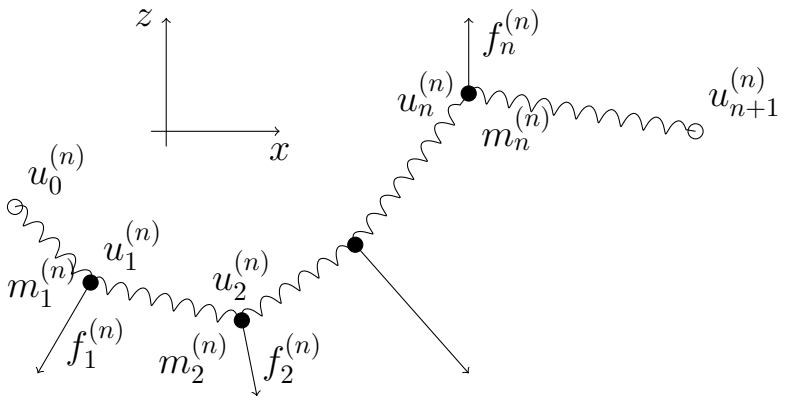


Figure 3.1.: Discrete mass-spring system.

In order to place the system in the state $u^{(n)}$ work has to be done against the forces excerted by the springs and the forces f_i. This work is stored as elastic energy $J_{\text{el}}^{(n)}$ and potential energy $J_{\text{f}}^{(n)}$ (in physics elastic energy is also a form of potential energy but for ease of writing we stick to these names). We now consider both energies separately.

The magnitude of the force excerted by a single spring extended to length l is given by *Hooke's law*

$$F(l) = \kappa(l - l_0)$$

where κ is the spring constant with units N m^{-1} and l_0 the length of the unloaded spring. The work done when extending the spring from length l_0 to l is then

$$W_{\text{el}}(l) = \int_{l_0}^{l} F(s)ds = \int_{l_0}^{l} \kappa(s - l_0)ds = \left[\frac{\kappa}{2}(s - l_0)^2\right]_{l_0}^{l} = \frac{\kappa}{2}(l - l_0)^2.$$

Then the total elastic energy in all springs in state $u^{(n)}$ is

$$J_{\text{el}}^{(n)}(u^{(n)}) = \frac{1}{2}\sum_{i=0}^{n} \kappa_i(\|u_{i+1}^{(n)} - u_i^{(n)}\| - l_i)^2 \tag{3.2}$$

where κ_i and l_i are the individual spring parameters and $\|.\|$ denotes the Euclidean norm.

The work done to bring a mass m to position u against the exterior force f is given by the path integral

$$W_{\text{f}}(u) = -\int_{0}^{u} f \cdot t\, ds = -\|u - 0\| \frac{u - 0}{\|u - 0\|} \cdot f = -f \cdot u.$$

Here we used 0 as the reference point but any other position is also in order. Note that when $u \cdot f$ is negative (e.g. the mass is lifted *up* in the gravity field $f = (0, -mg)^T$ pointing *down*) then the potential energy increases. The potential energy of all mass points is then

$$J_{\mathrm{f}}^{(n)}(u^{(n)}) = -\sum_{i=1}^{n} f_i^{(n)} \cdot u_i^{(n)} \tag{3.3}$$

and the total (potential) energy stored in the system at state $u^{(n)}$ is

$$J^{(n)}(u^{(n)}) = J_{\mathrm{el}}^{(n)}(u^{(n)}) + J_{\mathrm{f}}^{(n)}(u^{(n)}) = \frac{1}{2}\sum_{i=0}^{n} \kappa_i (\|u_{i+1}^{(n)} - u_i^{(n)}\| - l_i)^2 - \sum_{i=1}^{n} f_i^{(n)} \cdot u_i^{(n)}. \tag{3.4}$$

The *equilibrium principle* in mechanics says that the state $u_*^{(n)}$ attained by the system at equilibrium is the state of minimal (potential) energy:

$$J^{(n)}(u_*^{(n)}) \leq J^{(n)}(u) \quad \forall u \in \mathbb{R}^{2n}.$$

A short notation of the same statement is

$$u_*^{(n)} = \operatorname*{argmin}_{u \in \mathbb{R}^{2n}} J^{(n)}(u). \tag{3.5}$$

Note that problem (3.5) does in general not have a unique solution. An example for nonuniquess is the case $f_i^{(n)} = 0$ for all i and $\sum_{i=0}^{n} l_i > \|u_{n+1}^{(n)} - u_0^{(n)}\|$ where infinitely many solutions exist. When the endpoints of the string are sufficently far apart, however, one can prove that the functional $J^{(n)}(u)$ can be bounded from below, i.e.

$$J^{(n)}(u) \geq C \qquad \forall u \in \mathbb{R}^{2n} \tag{3.6}$$

and that it is convex, i.e.

$$J^{(n)}(\theta u + (1-\theta)v) \leq \theta J^{(n)}(u) + (1-\theta) J^{(n)}(v) \qquad \forall u, v \in \mathbb{R}^{2n}, \theta \in [0,1]. \tag{3.7}$$

By analogy with functions in one variable we may conclude that the problem has a unique global minimum. We will prove such a result later in a related context.

Continuum Limit We now aim at describing the position of the string by a continuous curve $u : I = [0,1] \to \mathbb{R}^2$. The parameter interval I is in principle arbitrary and the equations to be derived should not depend on the particular parametrization. A number $\xi \in I$ is used to "label" a point on the string and is

called *material coordinate*. The space $\mathbb{R}^2$ of positions is called the *configuration space* in this context.

To go from the discrete to the continuum model we introduce for every $n \in \mathbb{N}$ a discretization of the parameter interval

$$\xi_i^{(n)} = \frac{i}{n+1}$$

with the idea that $u(\xi_i^{(n)})$ corresponds to position $u_i^{(n)}$ of the discrete spring model. Furthermore we assume that the total length of the unloaded and unclamped string is given by L and set the lengths of the individual strings to

$$l_i^{(n)} = \frac{L}{n+1}.$$

With the abbreviation $\xi_{i\pm1/2}^{(n)} = \frac{1}{2}(\xi_i^{(n)} + \xi_{i\pm1}^{(n)})$ the other parameters of the discrete system are

$$\kappa_i^{(n)} = \kappa(\xi_{i+1/2}^{(n)}), \qquad f_i^{(n)} = \int_{\xi_{i-1/2}^{(n)}}^{\xi_{i+1/2}^{(n)}} f(\xi)\, d\xi$$

where $\kappa : I \to \mathbb{R}$ is a given continuous function describing the elastic properties of the string and $f : I \to \mathbb{R}^2$ is an integrable function giving the load density with units N m^{-1}. Inserting these definitions into Equation (3.2) for the discrete elastic energy yields (with slight abuse of notation):

$$\begin{aligned} J_{\text{el}}^{(n)}(u) &= \frac{1}{2}\sum_{i=0}^{n} \kappa_i (\|u(\xi_{i+1}^{(n)}) - u(\xi_i^{(n)})\| - l_i)^2 \\ &= \frac{1}{2}\sum_{i=0}^{n} \kappa_i \left(\frac{\|u(\xi_{i+1}^{(n)}) - u(\xi_i^{(n)})\|}{\xi_{i+1}^{(n)} - \xi_i^{(n)}} (\xi_{i+1}^{(n)} - \xi_i^{(n)}) - \frac{L}{n+1} \right)^2 \\ &= \frac{1}{2}\sum_{i=0}^{n} \kappa_i \, (\xi_{i+1}^{(n)} - \xi_i^{(n)})^2 \left(\left\| \frac{u(\xi_{i+1}^{(n)}) - u(\xi_i^{(n)})}{\xi_{i+1}^{(n)} - \xi_i^{(n)}} \right\| - L \right)^2 \end{aligned} \tag{3.8}$$

where we used $\xi_{i+1}^{(n)} - \xi_i^{(n)} = 1/(n+1)$. At this point we need to reconsider the spring "constant" κ. It has units N m^{-1} and depends on the length of the spring. This becomes important as the length of the individual springs now decreases as n increases. Mechanics tells us that a spring with cross-sectional area A_i, modulus of elasticity E_i and length l_i has a spring "constant"

$$\kappa_i = \frac{A_i E_i}{l_i} = \frac{A_i E_i}{L/(n+1)} = \frac{\tilde{\kappa}(\xi_{i+1/2}^{(n)})}{L(\xi_{i+1}^{(n)} - \xi_i^{(n)})}.$$

Note that the new material property function $\tilde{\kappa}(\xi)$ has units N and is now independent of the length of the string. Inserting this expression into Equation (3.8) yields

$$J_{\text{el}}^{(n)}(u) = \frac{1}{2}\sum_{i=0}^{n} \frac{\tilde{\kappa}_i}{L}\left(\left\|\frac{u(\xi_{i+1}^{(n)}) - u(\xi_i^{(n)})}{\xi_{i+1}^{(n)} - \xi_i^{(n)}}\right\| - L\right)^2 (\xi_{i+1}^{(n)} - \xi_i^{(n)})$$

where we can now pass to the limit

$$J_{\text{el}}(u) = \lim_{n\to\infty} J_{\text{el}}^{(n)}(u) = \int_0^1 \frac{\tilde{\kappa}(\xi)}{2L}\left(\|u'(\xi)\| - L\right)^2 \, d\xi. \tag{3.9}$$

Hereby we assumed that the derivative $u'(\xi)$ is well defined, i.e. $u \in \left(C^1([0,1])\right)^2$.

Now the potential energy is

$$J_{\text{f}}^{(n)}(u) = -\sum_{i=1}^{n} f_i^{(n)} \cdot u(\xi_i^{(n)}) = -\sum_{i=1}^{n} \int_{\xi_{i-1/2}^{(n)}}^{\xi_{i+1/2}^{(n)}} f(\xi) \cdot u(\xi_i^{(n)})\, d\xi$$

and passing to the limit gives

$$J_{\text{f}}(u) = \lim_{n\to\infty} J_{\text{f}}^{(n)}(u) = -\int_0^1 f(\xi) \cdot u(\xi)\, d\xi.$$

As in the discrete case we have

$$J(u) = J_{\text{el}}(u) + J_{\text{f}}(u) = \int_0^1 \frac{\tilde{\kappa}(\xi)}{2L}\left(\|u'(\xi)\| - L\right)^2 - f(\xi)\cdot u(\xi)\, d\xi. \tag{3.10}$$

Application of the equilibrium principle now results in a minimization problem in *function space*

$$u_* = \underset{u\in V}{\text{argmin}}\ J(u) \tag{3.11}$$

where the space of all admissible functions V is

$$V = \left\{ v \in \left(C^1([0,1])\right)^2 \,:\, v(0) = \begin{pmatrix} x_a \\ z_a \end{pmatrix}, v(1) = \begin{pmatrix} x_b \\ z_b \end{pmatrix} \right\} \tag{3.12}$$

since $u'(\xi)$ turns up in the energy functional. This now raises the question how to solve a minimization problem in function space?

3.2. Variational Approach

To find the minimum of a function $g(x)$ in one real variable one searches for stationary points $g'(x_*) = 0$ and then checks whether x_* really is a minimum. Transfering this idea to minimization problems in function space such as (3.11) is the central idea of the *calculus of variations*. As in the case of a function in one variable the search for stationary points of the functional $J(u)$ results only in a necessary condition for a minimum.

To start let us rewrite the minimization property as:

$$u_* = \underset{u\in V}{\operatorname{argmin}}\, J(u) \qquad \Leftrightarrow \qquad J(u_*) \le J(u_* + tv) \quad \forall t \in \mathbb{R}, \forall v \in V_0$$

where

$$V_0 = \left\{ v \in \left(C^1([0,1])\right)^2 \,:\, v(0) = v(1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\}. \tag{3.13}$$

The function v is called a *variation* or *test function* and the definition of V_0 ensures that the function $u_* + tv$ always satisfies the given boundary conditions which are already incorporated in u_*. The energy functional $J(u)$ to be minimized is called *Lagrangian* in the calculus of variations and the function spaces V and V_0 are called *trial space* and *test space* respectively.

Now the function $\phi_v(t) = J(u_* + tv)$ is an ordinary function in one variable for a fixed $v \in V_0$. If $\frac{d\phi_v}{dt}$ exists then we have

$$J(u_*) \le J(u_* + tv) \quad \forall t \in \mathbb{R}, \forall v \in V_0 \qquad \Rightarrow \quad \frac{d\phi_v}{dt}(0) = 0 \;\; \forall v \in V_0. \tag{3.14}$$

The reverse conclusion can also be shown if the minimizer exists. Now let us compute the *configurational derivative* $\frac{d\phi_v}{dt}$. For any given $u \in V$, $v \in V_0$ we get

$$\frac{d}{dt} J_{\mathrm{f}}(u + tv) = \frac{d}{dt}\left[-\int_0^1 f(\xi) \cdot (u(\xi) + tv(\xi))\, d\xi \right] = -\int_0^1 f(\xi) \cdot v(\xi)\, d\xi$$

and so

$$\frac{d}{dt} J_{\mathrm{f}}(u + tv)\Big|_{t=0} = -\int_0^1 f(\xi) \cdot v(\xi)\, d\xi.$$

For the more complicated elastic part we get

$$\begin{aligned} \frac{d}{dt} J_{\mathrm{el}}(u + tv) &= \frac{d}{dt} \int_0^1 \frac{\tilde\kappa(\xi)}{2L} \left(\|u'(\xi) + tv'(\xi)\| - L \right)^2 d\xi \\ &= \int_0^1 \frac{\tilde\kappa(\xi)}{L} \left(\|u'(\xi) + tv'(\xi)\| - L \right) \frac{[u'(\xi) + tv'(\xi)] \cdot v'(\xi)}{\|u'(\xi) + tv'(\xi)\|}\, d\xi \end{aligned}$$

where we have used $\frac{d}{dt}\|x+ty\| = (x+ty)\cdot y/\|x+ty\|$ for any two vectors $x, y \in \mathbb{R}^n$ and the Euclidean scalar product and norm. By setting $t=0$ we get

$$\frac{d}{dt} J_{\mathrm{el}}(u+tv)\Big|_{t=0} = \int_0^1 \frac{\tilde{\kappa}(\xi)}{L} \frac{\|u'(\xi)\| - L}{\|u'(\xi)\|} u'(\xi)\cdot v'(\xi)\, d\xi.$$

Putting both parts together results in the necessary condition for u (we refrain from writing u_* for the minimum from now on!) being a minimizer of the functional $J(u)$:

$$\int_0^1 \frac{\tilde{\kappa}(\xi)}{L} \frac{\|u'(\xi)\| - L}{\|u'(\xi)\|} u'(\xi)\cdot v'(\xi) - f(\xi)\cdot v(\xi)\, d\xi = 0 \qquad \forall v \in V_0. \tag{3.15}$$

This equation is called a (nonlinear) *variational equation.*

Abstract Variational Problem For a general Lagrangian of the form

$$J(u) = \int_0^1 F(u'(\xi), u(\xi))\, d\xi$$

we get by applying the chain rule the variational equation

$$\begin{aligned} \frac{d}{dt} J(u+tv)\Big|_{t=0} &= \frac{d}{dt}\left[\int_0^1 F(u'(\xi)+tv'(\xi), u(\xi)+tv(\xi))\, d\xi\right]\Big|_{t=0} \\ &= \int_0^1 \partial_1 F(u'(\xi), u(\xi))v'(\xi) + \partial_2 F(u'(\xi), u(\xi))v(\xi)\, d\xi = 0 \qquad \forall v \in V_0 \end{aligned} \tag{3.16}$$

where $\partial_1 F$, $\partial_2 F$ denote the partial derivatives of F with respect to the first and second argument. Note that the variation v always enters *linearly* in this equation! Therefore, the general variational equation has the abstract form:

$$\text{Find } u \in V: \quad r(u,v) = 0 \quad \forall v \in V_0 \tag{3.17}$$

where $r : V \times V_0 \to \mathbb{R}$ is linear in v, i.e.

$$r(u, v_1+v_2) = r(u,v_1) + r(u,v_2), \qquad r(u,kv) = kr(u,v)$$

but possibly nonlinear in u. In the applications the test space V_0 is a real vector space of functions and V is an affine space $V = u_0 + V_0 = \{u : u = u_0 + v, v \in V_0\}$ incorporating the boundary conditions.

A note on the requirement of the differentiability of u and v. The minimization problem as well as the variational problem were derived under the assumption that $u, v \in \left(C^1([0,1])\right)^2$. It will turn out that this function space is neither appropriate for proving the existence of a solution nor practical for the applications (consider for example a pointwise load on the string).

Differential Equation Integrating by parts the first term in Equation (3.16) gives

$$\begin{aligned}
&\int_0^1 \partial_1 F(u'(\xi), u(\xi))v'(\xi) + \partial_2 F(u'(\xi), u(\xi))v(\xi)\, d\xi \\
&\qquad = \int_0^1 -\frac{d}{d\xi}\left(\partial_1 F(u'(\xi), u(\xi))\right) v(\xi) + \partial_2 F(u'(\xi), u(\xi))v(\xi)\, d\xi \\
&\qquad\qquad + \left[\partial_1 F(u'(\xi), u(\xi))v(\xi)\right]_0^1
\end{aligned}$$

where the boundary term vanishes due to the boundary condition on v! In order to do the integration by parts it is necessary to assume that F is now twice differentiable with respect to each variable and also that $u \in \left(C^2([0,1])\right)^2$. This leads then to the following variant of the variational equation

$$\int_0^1 \left[-\frac{d}{d\xi}\partial_1 F(u'(\xi), u(\xi)) + \partial_2 F(u'(\xi), u(\xi))\right] v(\xi)\, d\xi = 0 \qquad \forall v \in V_0.$$

Now the fundamental lemma of the calculus of variation states that if this equation is true for all test functions v then the function in square brackets must vanish pointwise:

$$-\frac{d}{d\xi}\partial_1 F(u'(\xi), u(\xi)) + \partial_2 F(u'(\xi), u(\xi)) = 0 \qquad (\xi \text{ in } (0,1)). \tag{3.18}$$

Equation (3.18) is a nonlinear two-point boundary value problem called the *Euler-Lagrange equation* for the variational problem (3.16). Note the similarity to the reasoning when going from Equation (2.2) to (2.3). There we applied Gauss' theorem to the arbitrary domain $\omega \subseteq \Omega$ which can be interpreted as a special case of integration by parts with a piecewise constant function (the characteristic function of ω).

Setting up the Euler-Lagrange equation for our string example, i.e. applying integration by parts to Equation (3.15), results in the nonlinear second-order ordinary differential equation

$$-\frac{d}{d\xi}\left[\frac{\tilde{\kappa}(\xi)}{L}\frac{\|u'(\xi)\|-L}{\|u'(\xi)\|}u'(\xi)\right] = f(\xi) \qquad (\xi \text{ in } (0,1))$$

with boundary values

$$u(0) = \begin{pmatrix} x_a \\ z_a \end{pmatrix}, \qquad u(1) = \begin{pmatrix} x_b \\ z_b \end{pmatrix}.$$

Note that for this equation to make sense for us at the moment we require $\tilde{\kappa} \in C^1([0,1])$ and $u \in \left(C^2([0,1])\right)^2$.

In summary, we now have the following situation

(I) *Minimization problem*	$\Rightarrow$	(II) *Variational problem*	$\Rightarrow$	(III) *Differential equation*
$u = \underset{v \in V}{\operatorname{argmin}} J(v)$		Find $u \in V$ such that $r(u,v) = 0 \quad \forall v \in V_0$		Solve BVB $g(u, u', u'') = 0 \quad \text{in } \Omega$ u given on $\partial\Omega$

For step (I)→(II) we introduced the concept of the configurational derivative. We will later see that (I) follows also from (II) provided that the minimum exists. For the step (II)→(III) we applied integration by parts and had to assume additional smoothness for the solution and coefficient functions. In general, a solution of problem (II) need not be a solution of problem (III) therefore. By taking the perspective of the differential equation the variational problem (II) is called the *weak formulation* of the boundary value problem.

3.3. Taut String Approximation

In this paragraph we are interested in the situation where the length L of the string with zero elastic energy is much shorter than the distance of the two points where it is clamped to, i.e.

$$L \ll \|u(1) - u(0)\|.$$

Under this assumption we get

$$L \ll \|u(1) - u(0)\| = \left\| \int_0^1 u'(\xi)\, d\xi \right\| \leq \int_0^1 \|u'(\xi)\|\, d\xi.$$

With this the energy functional (3.10) simplifies to

$$\begin{aligned} J(u) &= \int_0^1 \frac{\tilde{\kappa}(\xi)}{2L} \left(\|u'(\xi)\| - L\right)^2 - f(\xi) \cdot u(\xi)\, d\xi \\ &\approx \int_0^1 \frac{\tilde{\kappa}(\xi)}{2L} \|u'(\xi)\|^2 - f(\xi) \cdot u(\xi)\, d\xi =: \tilde{J}(u). \end{aligned}$$

Now $\tilde{J}(u)$ is a *quadratic functional* in u. The associated variational problem is

$$u \in V: \quad \int_0^1 \frac{\tilde{\kappa}(\xi)}{L} u'(\xi) \cdot v'(\xi) - f(\xi) \cdot v(\xi)\, d\xi = 0 \qquad \forall v \in V_0 \tag{3.19}$$

which is now a *linear* variational problem in u. The related differential equation is then also linear and reads

$$-\frac{d}{d\xi}\left(\frac{\tilde{\kappa}(\xi)}{L}\frac{du}{d\xi}\right) = f \qquad \text{in } (0,1) \tag{3.20}$$

which decouples into two seperate equations for $x(\xi)$ and $z(\xi)$.

Let us assume now the special situation where there is only a vertical load $f(\xi) = (0, f_z(\xi))^T$. Naming the components $u(\xi) = (x(\xi), z(\xi))^T$ and $v(\xi) = (\phi(\xi), \psi(\xi))^T$ the variational problem (3.19) reads

$$\begin{aligned} u \in V: \quad & \int_0^1 \frac{\tilde{\kappa}(\xi)}{L}(x'(\xi)\phi'(\xi) + z'(\xi)\psi'(\xi)) - f_z(\xi)\psi(\xi)\, d\xi \\ = & \int_0^1 \frac{\tilde{\kappa}(\xi)}{L} x'(\xi)\phi'(\xi)\, d\xi + \int_0^1 \frac{\tilde{\kappa}(\xi)}{L} z'(\xi)\psi'(\xi)) - f_z(\xi)\psi(\xi)\, d\xi = 0 \quad \forall \left(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix}\right) \in V_0. \end{aligned}$$

The equation for $x(\xi)$ can be solved analytically by solving the corresponding differential equation and we find:

$$x(\xi) = x_a + (x_b - x_a)\frac{\int_0^\xi \frac{1}{\tilde{\kappa}(s)}\, ds}{\int_0^1 \frac{1}{\tilde{\kappa}(s)}\, ds}.$$

Since L and $\tilde{\kappa}$ are strictly positive quantities the function $\xi \to x(\xi)$ is strictly increasing and therefore has an inverse $x \to x^{-1}(x)$.

We now want to write the second component $z(\xi)$ as a function of $x(\xi)$ instead of ξ. Therefore we define the new function $\hat{z}(x)$ and use the chain rule:

$$z(\xi) = \hat{z}(x(\xi)) \qquad \Rightarrow \qquad \frac{dz}{d\xi}(\xi) = \frac{d\hat{z}}{dx}(x(\xi))\frac{dx}{d\xi}(\xi).$$

The same applies for the test function $\psi(x) = \hat{\psi}(x(\xi))$. Recalling the transformation theorem for integrals $\int_a^b g(s)ds = \int_{a'}^{b'} g(\mu(t))|\frac{d\mu}{dt}(t)|\,dt$ with $\mu : [a', b'] \to [a, b]$ a differentiable map, we obtain for the variational problem for the second component $z(\xi)$:

$$\begin{aligned}
&\int_0^1 \frac{\tilde{\kappa}(\xi)}{L}\frac{d\hat{z}}{dx}(x(\xi))\frac{dx}{d\xi}(\xi)\frac{d\hat{\psi}}{dx}(x(\xi))\frac{dx}{d\xi}(\xi) - f_z(\xi)\hat{\psi}(x(\xi))\,d\xi \\
&= \int_{x_a}^{x_b} \left[\frac{\tilde{\kappa}(x^{-1}(x))}{L}\frac{d\hat{z}}{dx}(x)\left(\frac{dx}{d\xi}(x^{-1}(x))\right)^2\frac{d\hat{\psi}}{dx}(x) - f_z(x^{-1}(x))\hat{\psi}(x)\right]\frac{dx}{\left|\frac{dx}{d\xi}(x^{-1}(x))\right|} \\
&= \int_{x_a}^{x_b} \underbrace{\frac{\tilde{\kappa}(x^{-1}(x))}{L}\left|\frac{dx}{d\xi}(x^{-1}(x))\right|}_{\hat{\sigma}(x)}\frac{d\hat{z}}{dx}(x)\frac{d\hat{\psi}}{dx}(x) - \underbrace{\frac{f_z(x^{-1}(x))}{\left|\frac{dx}{d\xi}(x^{-1}(x))\right|}}_{\hat{f}(x)}\hat{\psi}(x)\,dx = 0, \quad \forall\psi \in C_0^1([x_a, x_b]
\end{aligned}$$

The corresponding linear second-order *scalar* differential equation for the function $\hat{z}$ now in "physical coordinates" reads:

$$-\frac{d}{dx}\left(\hat{\sigma}(x)\frac{d\hat{z}}{dx}\right) = \hat{f}(x) \qquad \text{in } (x_a, x_b)$$

with boundary conditions

$$\hat{z}(x_a) = z_a, \qquad \hat{z}(x_b) = z_b.$$

In two space dimensions the equation

$$-\nabla \cdot (\sigma(x)\nabla u) = f \qquad \text{in } \Omega \subset \mathbb{R}^2$$

with boundary conditions

$$u = g \qquad \text{on } \partial\Omega \tag{3.21}$$

is a model for the vertical position of a thin sheet of rubber under vertical load that is clamped at the boundary. Figure 3.2 shows an example for the two-dimensional case. Note that $\|\nabla u\|$ can become very large near so-called "reentrant corners" of the domain.

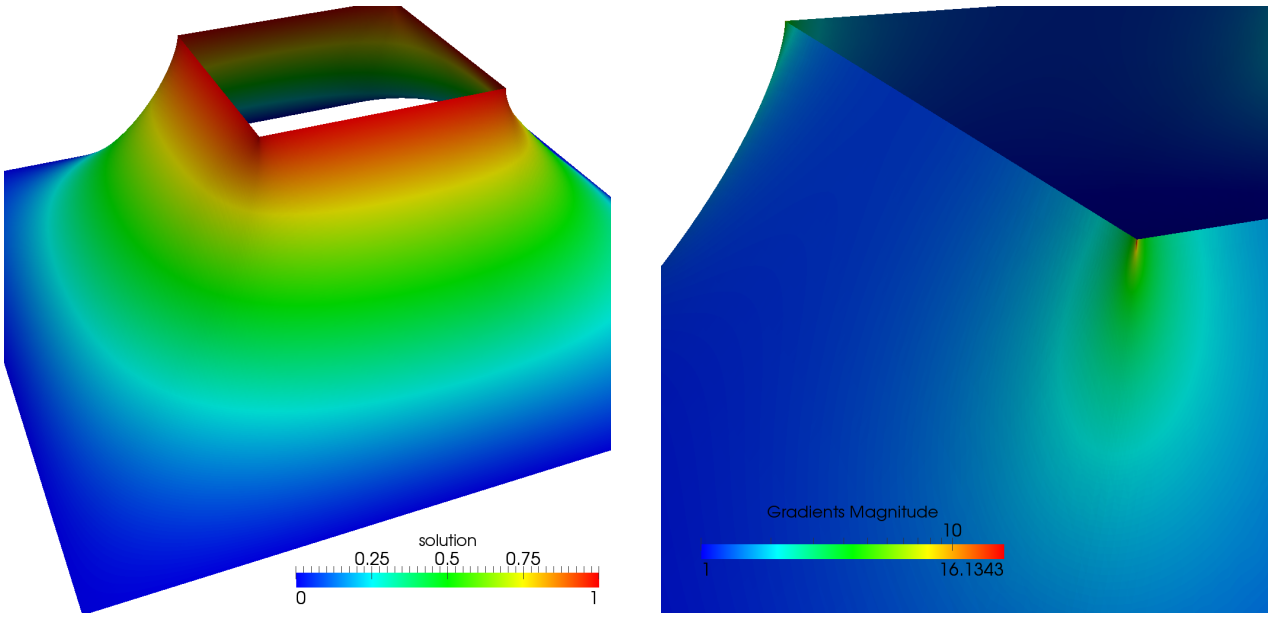


Figure 3.2.: A thin rubber sheet over the region $\Omega = (0,3)^2 \setminus [1,2]^2$ clamped to height 1 at the inner boundary and to 0 at the outer boundary. Left image shows rubber sheet colored by height and right image shows rubber sheet colored by the norm of the gradient in logarithmic (!) scale.

3.4. Linear Elasticity and Plate Problem

The considerations of this chapter can be generalized to small deformations of a three-dimensional elastic material experiencing both tension and compression. The resulting energy functional for the *linear elasticity problem* is

$$J(u) = \int_\Omega \frac{1}{2} \left\{ \lambda (\nabla \cdot u)^2 + 2\mu D(u) : D(u) \right\} - f \cdot u \, dx \tag{3.22}$$

where $u \in \left(C^1(\Omega)\right)^3$ is the unknown *displacement* of the material from its unloaded configuration (i.e. $x + u(x)$ is the position of the material point $x \in \Omega$ under load). Then $D(u) = \frac{1}{2}(\nabla u + (\nabla u)^T)$ is the strain tensor from (2.33), λ, μ are the *Lamé coefficients* of the material and f are volume forces. The boundary condition

$$u(x) = g(x) \qquad \text{on } \partial\Omega$$

models clamping of the material at the boundary. The Euler-Lagrange equation corresponding to the variational formulation of (3.22) is now a linear second-order *system* of partial differential equations. For details we refer to [Braess, 2003, §3] and [Ciarlet, 2002, §1.2].

If we are interested in the deflection of a thin plate of elastic material with constant thickness, a well established model is the *plate problem* with the energy functional

$$J(u) = \int_\Omega \frac{1}{2} \left\{ |\Delta u|^2 + 2(1-\sigma)((\partial_{x_1 x_2} u)^2 - \partial_{x_1}^2 u \, \partial_{x_2}^2 u) \right\} - f u \, dx \tag{3.23}$$

with $\sigma = \lambda/(2(\lambda+\mu))$ the *Poisson coefficent* computed from the Lamé coefficients. Here u is again a scalar function giving the vertical displacement of the plate out of the planar reference configuration. As boundary conditions we consider $u = 0$ on $\partial\Omega$. Note that the functional (3.23) involves second derivatives of u! The corresponding Euler-Lagrange equation of the variational formulation is now a *fourth-order* partial differential equation

$$\partial_{x_1}^4 u + \partial_{x_2}^4 = \Delta^2 u = f \qquad \text{in } \Omega \tag{3.24}$$

with boundary conditions $u = \partial_n u = 0$ on $\partial\Omega$. We refer to [Ciarlet, 2002, §1.2] or [Hackbusch, 1986, §5.3].

3.5. Hamilton's Principle

So far we considered stationary problems where the state attained by the system is a minimizer of potential energy (equilibrium principle). The energy functional (Lagrangian) is convex and bounded from below which ensures that a solution of the corresponding variational problem (which determines stationary points of the Lagrangian) is the globally unique minimizer.

In the dynamic case the energy functional involves kinetic and potential energy and is typically not convex any more. It turns out, that for certain systems the state can still be determined by finding stationary points of the energy functional, i.e. solving the corresponding variational problem. This principle is called *Hamilton's principle* and the energy functional is called a *Hamiltonian* in this case. For more information and some examples we refer to [Eriksson et al., 1996, §11.2].

Chapter 4.

Type Classification and Model Problems

4.1. Basic Mathematical Questions

So far we have derived several different PDEs by physical reasoning without mentioning a word about their solvability. From other types of mathematical equations, most notably ordinary differential equations, it is clear that we have to ask the following questions:

a) *Existence*: Does a given PDE problem have a solution?

b) *Uniqueness*: Is this solution the only one?

c) *Stability*: How does the data influence the solution?

A PDE problem is informally called *well-posed* (in the sense of Hadamard) if

a) it has a solution,

b) this solution is unique and

c) it depends continuously on the data.

If any of these conditions is not fullfilled it is called *ill-posed*. The discovery of chaos theory in the 20th century tells us that a problem need not be well-posed to be physically meaningful. Even the existence of solutions to many practically relevant problems such as the Euler or Navier-Stokes equations is open. In fact, a proof of the existence (and regularity, see below) of a solution to the incompressible Navier-Stokes equations in three space dimensions (this is important) is one of the millenium prize problems[1].

The informal definition of a well-posed problem needs to be made mathematically precise. The first question is what a solution should be. A natural assumption for an equations such as

$$\partial_x^2 u + \partial_y^2 u + \partial_z^2 u = f \qquad (x \in \Omega)$$

[1]The precise description of the task can be found here: `http://www.claymath.org/millennium/Navier-Stokes_Equations/`

would be that u is twice continuously differentiable with respect to each variable. We say that a function is a *classical solution* of a PDE problem if it has continuous derivatives up to the required order and it satisfies the PDE for every $x \in \Omega$.

The number of derivatives of a function is called its *regularity*. If a solution to a PDE problem possesses higher derivatives than required by the PDE it is said to have "additional regularity". This is important to assess the (speed of) convergence of numerical schemes. It turns out that also functions that *do not have* the derivatives required by the PDE may be called "solutions" in an appropriate sense (so-called weak solutions). A particular example is the conservation law

$$\partial_t u + \partial_x u = 0 \qquad (x \in \mathbb{R}, t > 0)$$

where also discontinuous functions $u(x,t)$ do make sense, as we will show below.

Unfortunately there is no theory that covers the solvability of PDEs in general and it is unlikely that such a theory exists. Instead techniques have been developed that can be used to analyze certain classes of PDEs. Similarly, there are no general numerical methods that can be applied successfully to any PDE but the development of numerical schemes follows the different classes introduced for the analysis. Below we will introduce the following important classes of PDEs:

a) Second-order scalar elliptic equation.

b) Second-order scalar hyperbolic equation.

c) Second-order scalar parabolic equation.

d) First-order hyperbolic systems.

4.2. Second-order Scalar Equations

Type Classification Our aim is now to sort second-order scalar equations into different classes. We restrict ourselves to linear equations (nonlinear equations are classified after linearization). From the viewpoint of physical applications the independent variables are characterized as "time" and "space". In the following definition this distinction is not made, i.e. $x = (x_1, \ldots, x_n)^T$ denotes just a vector of n independent variables where one of them may be time.

The general linear second-order scalar PDE has one of the two forms

$$Lu = -\sum_{i,j=1}^{n} \partial_{x_j}(a_{ij}(x)\partial_{x_i} u) + \sum_{i=1}^{n} \partial_{x_i}(b_i(x)u) + c(x)u = f \qquad \text{in } U \qquad (4.1)$$

or

$$Lu = -\sum_{i,j=1}^{n} a_{ij}(x)\partial_{x_j}\partial_{x_i}u + \sum_{i=1}^{n} b_i(x)\partial_{x_i}u + c(x)u = f \qquad \text{in } U, \tag{4.2}$$

where L is called a linear *differential operator* and U is some domain. The specificiation of boundary conditions to obtain a well-posed problem is intentionally omitted as it depends on the given coefficent functions.

We say that the PDE is in *divergence form* if it is given by (4.1). This is the form that arises when deriving the equation from a conservation principle. This also explains the minus sign in the second-order terms which reminds us that some flow in direction of the negative gradient is modelled. If the coefficients are continuously differentiable we can rewrite the form (4.1) into the form (4.2) and vice versa using the product rule. Doing so results in new coefficients $\tilde{b}_i(x)$ and $\tilde{c}(x)$ but the coefficients $a_{ij}(x)$ remain the same. Moreover, since $\partial_{x_i}\partial_{x_j}u = \partial_{x_j}\partial_{x_i}u$ we may assume without loss of generality that

$$a_{ij}(x) = a_{ji}(x) \qquad (i, j = 1, \ldots, n, x \in U).$$

Either form (4.1) or (4.2) may be more appropriate for different purposes. In the following we will consider only form (4.2), tacitly assuming that both forms are equivalent.

Definition 4.1. For every point $x \in U$ define the real symmetric $n \times n$ matrix $(A(x))_{ij} = a_{ij}(x)$ and the column vector $b(x) = (b_1(x), \ldots, b_n(x))^T$. Then the *partial differential operator* L (or Equation (4.2)) is called

a) *elliptic* in x if all eigenvalues of $A(x)$ are nonzero and have the same sign,

b) *hyperbolic* in x if all eigenvalues are nonzero, $n-1$ eigenvalues have the same sign and the remaining eigenvalue has the opposite sign,

c) *parabolic* in x if one eigenvalue is zero, the remaining eigenvalues have the same sign and the $n \times (n+1)$ matrix $(A(x), b(x))$ has full rank.

The operator (or the equation) is called elliptic (hyperbolic, parabolic) if it is elliptic (hyperbolic, parabolic) in every point $x \in U$. □

The names elliptic, hyperbolic and parabolic are taken from the special case $n = 2$ where a level set of the quadratic form $q(x_1, x_2) = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2$ is either an ellipse, a hyperbola or a parabola. In the case $n = 2$ the classification is complete, i.e. every linear second-order PDE is either elliptic, hyperbolic or parabolic or it is not a PDE (this case is excluded by the rank condition in the parabolic case). For $n > 2$ there are PDEs that are neither elliptic, parabolic or hyperbolic. Moreover, there are useful PDEs that have different types in different parts of the domain. Note also that the type of the operator depends only on the coefficients of the second-order terms, the so-called *leading part*.

Characteristics In the initial value problem for a second-order ordinary differential equation the solution u and its derivative du/dt are prescribed at some t_0 in order to determine the solution at later times $t > t_0$. We can transfer the idea of an initial value problem to partial differential equations in the following way: Given u on a surface Γ in $\mathbb{R}^n$ together with its derivative in direction normal to the surface, can we determine the second derivative in normal direction, and with it the solution in the neighborhood of the surface, from the given data and the PDE (4.2)? This problem is generally called the *Cauchy problem* for (4.2). Points $x \in \Gamma$ where the Cauchy problem can *not* be solved are called *characteristic points* and if it is not solvable in any point of a given surface the surface itself is called a *characteristic surface*. On a characteristic surface the solution of the PDE or its normal derivative may not be continuous although the coefficients (and the surface) are smooth. Therefore the existence of characteristic surfaces for a differential operator gives important information.

In order to answer this question about characteristic points and surfaces we assume the surface Γ to be smooth and denote by $q_1(x), \dots, q_n(x)$ a system of orthonormal vectors in $x \in \Gamma$ such that $q_1(x), \dots, q_{n-1}(x)$ are tangential to the surface and $q_n(x)$ points in direction normal to the surface. Since Γ is smooth and u as well as its normal derivative ∂_{q_n} are given on all of Γ also $\partial_{q_i} u$ and $\partial_{q_i}\partial_{q_n} u$ for $1 \le i < n$ are given. So the task is to compute the single derivative $\partial^2_{q_n} u$ from the given data and the PDE.

In order to do that we introduce the coordinate transformation $x(s) = Q(y)s + y$ for an arbitrary $y \in \Gamma$ and $Q(y) = [q_1(y), \dots, q_n(y)]$ the column matrix of tangential and normal vectors. We now want to derive a PDE for the new function

$$v(s) = u(x(s)) = u(Q(y)s + y)$$

locally around $y \in \Gamma$. Employing the chain rule we obtain for the gradient and the Hessian

$$\begin{aligned} \partial_{s_j} v(s) &= q_j^T \nabla_x u(x(s)) && \Rightarrow && \nabla_s v(s) = Q^T \nabla_x u(x(s)) \\ \partial_{s_i}\partial_{s_j} v(s) &= q_j^T \nabla_x^2 u(x(s)) q_i && \Rightarrow && \nabla_s^2 v(s) = Q^T \nabla_x^2 u(x(s)) Q. \end{aligned}$$

Note that $\partial^2_{s_n} v(s) = q_n^T \nabla_x^2 u(x(s)) q_n = q_n^T \nabla_x (q_n^T \nabla_x u(x(s))) = \partial^2_{q_n} u(x(s))$ is the second derivative in normal direction. Since Q is orthogonal we get

$$\nabla_x u(x(s)) = Q \nabla_s v(s) \qquad \nabla_x^2 u(x(s)) = Q \nabla_s^2 v(s) Q^T$$

which we insert into the PDE:

$$-A(x(s)) : (Q \nabla_s^2 v(s) Q^T) + b(x(s)) \cdot (Q \nabla_s v(s)) + c(x(s)) = f(x(s)).$$

Here we used the notation $A : B = \sum_{i,j=1}^n (A)_{i,j} (B)_{i,j}$. Using the identity $A : (QBQ^T) = (Q^T A Q) : B$, see appendix A.3.1, we obtain the transformed

PDE for $v(s)$:

$$-(Q^T A(x(s))Q) : \nabla_s^2 v(s) + (Q^T b(x(s)))^T \cdot \nabla_s v(s) + c(x(s)) = f(x(s)).$$

Writing out components in the leading order part we find

$$-q_n^T A q_n \partial_{s_n}^2 v - \sum_{i,j=1(i\neq j)}^{n} q_i^T A q_j \partial_{s_i}\partial_{s_j} v + (Q^T b)^T \cdot \nabla_s v + c = f. \tag{4.3}$$

From this we see that the missing derivative $\partial_{s_n}^2 v(s) = \partial_{q_n}^2 u(x(s))$ can be computed from the given data if and only if

$$q_n^T A q_n \neq 0. \tag{4.4}$$

The matrix A is always symmetric and therefore diagonalizable, i.e. it has n real eigenvalues $\lambda_1, \ldots, \lambda_n$ together with a set of orthonormal eigenvectors $r_1, \ldots, r_n$. With the column matrix $R = [r_1, \ldots, r_n]$ we have $R^T A R = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. We now discuss the condition (4.4) depending on the type of equation:

i) The PDE (4.2) is of elliptic type. Then all eigenvalues of A are either positive or negative, i.e. $q_n^T A q_n \neq 0$ for any $q_n \neq 0$. Therefore the desired derivative $\partial_{q_n}^2 u(x(s))$ can always be computed for any surface Γ. We conclude that an elliptic PDE does not have characteristic surfaces. This also means that the solution and its gradient are smooth as long as the coefficients of the equation are smooth enough.

ii) The PDE (4.2) is of parabolic type. Then A has one zero eigenvalue and all others are nonzero and have the same sign. Without loss of generality let $\lambda_n = 0$ with corresponding eigenvector r_n. Now

$$q_n A q_n = 0 \quad \Leftrightarrow \quad q_n = \alpha_n r_n$$

for any α_n, meaning that $\partial_{q_n}^2 u(x(s))$ can *not* be computed at a point on the surface when the normal to points in direction of the eigenvector r_n. Characteristic surfaces have normal direction $r_n(y)$ in every $y \in \Gamma$. In fact this does not explain why all nonzero eigenvalues need to have the same sign.

iii) The PDE (4.2) is of hyperbolic type. Then A has $n-1$ eigenvalues of the same sign and one with the opposite sign. Without loss of generality let λ_n be this eigenvalue. Decomposing $q_n = \alpha_n r_n + \sum_{i=1}^{n-1} \alpha_i r_i$ we get

$$q_n A q_n = \alpha_n^2 \lambda_n + \sum_{i=1}^{n-1} \alpha_i^2 \lambda_i = 0 \quad \Leftrightarrow \quad \alpha_n = \pm \sqrt{\sum_{i=1}^{n-1} -\frac{\lambda_i}{\lambda_n} \alpha_i^2}.$$

Note that the radicand is always nonnegative due to the sign condition in the definition of hyperbolicity. Now

$$q_n = \pm\sqrt{\sum_{i=1}^{n-1} -\frac{\lambda_i}{\lambda_n}\alpha_i^2 r_n + \sum_{i=1}^{n-1}\alpha_i r_i}$$

are all the surface normal directions for which the derivative $\partial_{q_n}^2 u(x(s))$ can not be computed. For any choice of $\alpha_1, \ldots, \alpha_{n-1}$ we get *two* possible directions. Since q_n is a direction it can be scaled arbitrarily. Therefore in the case $n = 2$ we can fix $\alpha_1 = 1$ and there are exactly two directions:

$$q_n = r_1 \pm \sqrt{-\frac{\lambda_1}{\lambda_2}}\, r_2.$$

If $n > 2$, there is an $n - 2$ dimensional set of directions.

Characteristic surfaces are closely related to the Cauchy-Kovalevskaya theorem, see [Renardy and Rogers, 1993, §2.2], which asserts the local existence of solutions of a system of PDEs in the neighborhood of noncharacteristic surfaces. It is not of much practical use because the data and the surface are required to be analytic and it is indifferent to well-posed and ill-posed problems. Although it turns out that the choice of boundary and initial conditions that lead to a well-posed problem strongly depends on the type of the equation this question can not be answered with the techniques given so far. In the following we will give boundary and initial conditions that lead to well-posed problems for the different types with proofs given later in the text or in the literature. Ill-posedness of certain problems will be shown by the way of counter examples.

Elliptic Equations The condition for ellipticity results in $A(x)$ being either positive or negative definite. Since the sign can be changed arbitrarily by multiplying the equation by -1 the convention is to require that $A(x)$ is positive definite (then $A(x)$ models a permeability tensor). From §2.5 we learn that elliptic equations model e.g. the stationary flow of heat in a solid or fluid material.

The "simplest" elliptic equation is obtained by setting $A = I$, $b = 0$, $c = 0$:

$$-\Delta u = f \qquad \text{in } \Omega \tag{4.5}$$

and is called *Poisson equation*. If also $f = 0$ the equation is called *Laplace equation* or *potential equation*. In §1 we have seen that the Poisson equation describes e.g. the gravitational potential.

Now we turn to the question of boundary conditions on $\partial\Omega$. The analysis of the Cauchy problem above suggests that the solution in the neighborhood of

the boundary can be determined from u and $\partial_n u$ given on the boundary, where n denotes by convention the direction of the unit outer normal to $\partial\Omega$. The following counter example shows that such boundary data may not lead to a well-posed problem.

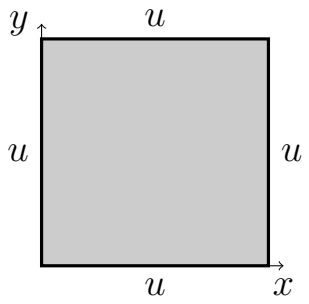
Example 4.2. [Rannacher, 2006, §1.2]. Consider $n = 2$ and $\Omega = \{(x, y) \in \mathbb{R}^2 : x > 0\}$. On the curve $\Gamma = \{(0, y) \in \mathbb{R}^2\}$ we prescribe the Cauchy data $u(0, y) = u_0^0(y) = 0$ and $\partial_x u(0, y) = u_0^1 = 0$. Clearly the function $u(x, y) = 0$ solves the Laplace equation $\Delta u = 0$ with this boundary data. Now we chose $\epsilon > 0$ and set the boundary data to

$$u_\epsilon^0(y) = 0, \qquad\qquad u_\epsilon^1(y) = \epsilon \sin(y/\epsilon).$$

One verifies that

$$u_\epsilon(x, y) = \epsilon^2 \sinh(x/\epsilon) \sin(y/\epsilon), \qquad \sinh(z) = \frac{1}{2}(e^z - e^{-z})$$

solves $\Delta u = 0$ in Ω and satisfies the given boundary data. Now we have on the one hand $\lim_{\epsilon\to 0} u_\epsilon^1 = u_0^1 = 0$ but on the other hand for any fixed point $(x, y) \in \Omega$: $\lim_{\epsilon\to 0} u_\epsilon(x, y) \to \infty$. Therefore it is not possible to bound $u_\epsilon(x, y)$ by the data uniformly in ϵ and the problem is not well posed. □



It turns out that on every point of the boundary only *one* of the following conditions

$$u = g \qquad \text{(Dirichlet)}, \tag{4.6a}$$
$$\partial_n u = g \qquad \text{(Neumann)}, \tag{4.6b}$$
$$\partial_n u + \alpha u = g \qquad \text{(Robin)}, \tag{4.6c}$$

can be prescribed. That these prescriptions actually lead to well-posed problems for the general elliptic equation will be shown later as part of the convergence theory. Note also, that the Neumann condition requires a compatibility condition with the right hand side f in order to be solvable and the solution is only unique up to a constant. The boundary conditions can be mixed, i.e. on different parts of the boundary, different conditions can be given.

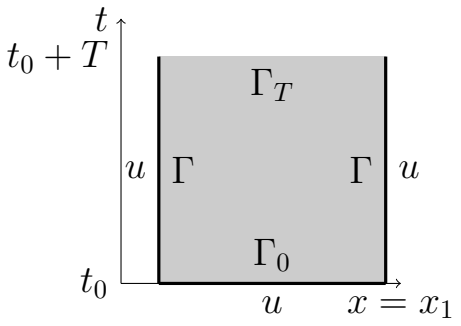
Parabolic Equations In the parabolic case one eigenvalue is zero. We assume that this eigenvalue is λ_n and rename the variable x_n to t. The "simplest" parabolic equation is obtained by setting $A = \left(\begin{smallmatrix} \lambda I & 0 \\ 0 & 0 \end{smallmatrix}\right)$ with some constant $\lambda > 0$, $b = (0, \ldots, 0, 1)^T$ and $c = 0$ which leads to the equation

$$\partial_t u - \lambda \sum_{i=1}^{n-1} \partial_{x_i}^2 u = \partial_t u - \lambda \Delta u = f \qquad \text{in } U. \tag{4.7}$$

Note that second derivatives are taken only with respect to the "spatial" variables $x_1, \ldots, x_{n-1}$. From §2.5 we learn that this equation models instationary heat flow in a homogeneous solid body with heat conductivity λ. The equation with $\lambda = 1$ and $f = 0$

$$\partial_t u = \Delta u \qquad \text{in } U \tag{4.8}$$

is generally referred to as *heat equation* in the mathematical literature.



Parabolic equations are typically solved in a space-time cylinder $U = \Omega \times \Sigma$ with Ω the spatial domain and $\Sigma = (t_0, t_0 + T)$ a time interval. For the boundary conditions we identify the boundaries $\Gamma = \{(x,t) \in U : x \in \partial\Omega, t \in \Sigma\}$, $\Gamma_0 = \{(x,t) \in U : x \in \Omega, t = t_0\}$ and $\Gamma_T = \{(x,t) \in U : x \in \Omega, t = t_0 + T\}$. This is illustrated for one spatial dimension in the figure to the left. Since $-\Delta$ is an elliptic operator on the $n-1$ spatial variables the same boundary conditions as in the elliptic case (with g now depending on time as well), i.e. those in (4.6), can be applied on Γ. The surfaces Γ_0 and Γ_T are characteristic since the normal direction $(0, \ldots, 0, 1)^T$ points in direction of the eigenvector corresponding to the zero eigenvalue. But the equation is only first order in that direction, so the prescription of the solution on either Γ_0 or Γ_T is sufficient. We now show by way of a counter example that a prescription of u on Γ_T ("at the end of the time interval") may lead to an ill-posed problem.

Example 4.3. [Braess, 2003]. Consider $n = 2$, $U = \Omega \times \Sigma = (0,1) \times [-1, 0)$. We prescribe the condition $u(0,t) = u(1,t) = 0$ on Γ and the condition

$$u(x,0) = \frac{1}{k} \sin(k\pi x) \qquad (k \in \mathbb{N})$$

for $t = 0$ which corresponds to Γ_T above. One verifies that

$$u(x,t) = \frac{1}{k} e^{-k^2\pi^2 t} \sin(k\pi x)$$

solves the heat equation $\partial_t u = \partial_x^2 u$ in U and satisfies the boundary data. We observe that for $k \to \infty$: $\sup_{x\in\Omega} |u(x,0)| \to 0$ but $\sup_{(x,t)\in U} |u(x,t)| \to \infty$. On the other hand $u(x,t) = 0$ solves the heat equation for $u(x,0) = 0$. So again the solution does not depend continuously on the data. Observe, however, that changing U to $U = (0,1) \times (0,1]$ and the same data at $t = 0$ which now corresponds to Γ_0 (!) in the notation introduced above. we get $|u(x,t)| \to 0$ for $k \to \infty$. □

From the example we motivate that conditions on Γ and Γ_0 do lead to a well-posed problem. On Γ_T no condition is necessary. The conditions on Γ are

referred to as boundary conditions while the condition on Γ_0 is referred to as initial condition.

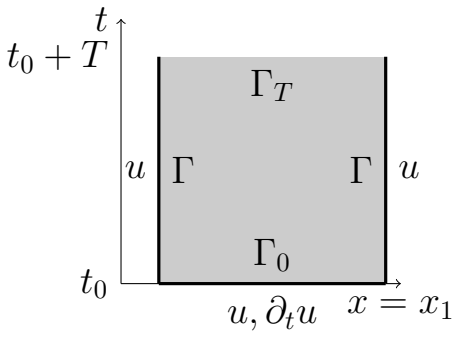
Hyperbolic Equations In the hyperbolic case $n-1$ eigenvalues have the same sign and one eigenvalue has the opposite sign. We assume without loss of generality that $\lambda_n < 0$, $\lambda_i > 0$, $1 \le i \le n-1$ and rename the variable x_n to t. The "simplest" hyperbolic equation is obtained by setting $A = \left(\begin{smallmatrix} \kappa^2 I & 0 \\ 0 & -1 \end{smallmatrix}\right)$ with some constant $\kappa > 0$, $b = 0$ and $c = 0$ which leads to the equation

$$\partial_t^2 u - \kappa^2 \sum_{i=1}^{n-1} \partial_{x_i}^2 u = \partial_t^2 u - \kappa^2 \Delta u = f \qquad \text{in } U. \tag{4.9}$$

As in the heat equation the Laplace operator acts only on the "spatial" variables $x_1, \ldots, x_{n-1}$. From §2.8 we learn that this equation models the propagation of sound waves in a gas with κ the speed of sound. The equation with $\kappa = 1$ and $f = 0$

$$\partial_t^2 u = \Delta u \qquad \text{in } U \tag{4.10}$$

is generally referred to as *wave equation* in the mathematical literature.



Hyperbolic equations are typically solved in a space-time cylinder $U = \Omega \times \Sigma$ with Ω the spatial domain and $\Sigma = (t_0, t_0 + T)$ a time interval. For the boundary conditions we identify the boundaries $\Gamma = \{(x,t) \in U : x \in \partial\Omega, t \in \Sigma\}$, $\Gamma_0 = \{(x,t) \in U : x \in \Omega, t = t_0\}$ and $\Gamma_T = \{(x,t) \in U : x \in \Omega, t = t_0 + T\}$. This is illustrated for one spatial dimension in the figure to the left. In the hyperbolic case no boundary surface is characteristic. Since the operator $-\Delta u$ is elliptic an elliptic operator with respect to the $n-1$ spatial variables the same boundary conditions (4.6) as in the elliptic case, with g now depending as well on time, can be applied on Γ. On Γ_0, it turns out, the prescription of the initial conditions u and $\partial_t u$ does lead to a well-posed problem. If such a condition is prescribed on Γ_0 no condition on Γ_T can be given.

However, in contrast to the parabolic case, the role of Γ_0 and Γ_T can be reversed without leading to an ill-posed problem. This can be seen as follows: Suppose $u^+(x,t)$ is a solution of the wave equation in $U^+ = \Omega \times (0, T]$ with initial data prescribed at $t = 0$. Then the function $u^-(x,t) = u^+(x,-t)$ is a solution of the wave equation in $U^- = \Omega \times [-T, 0)$ with "final data" prescribed at $t = 0$ and boundary data $u^-(x,t) = u^+(x,-t), x \in \partial\Omega, t \in [-T, 0)$. So when the problem in U^+ is well posed also the "reflected problem" in U^- is well-posed.

We now turn to the question whether supplying u on Γ_0 *and* Γ_T instead of $u, \partial_t u$ on either Γ_0 *or* Γ_T leads to a well-posed problem. The following counter example shows that this is in general not the case.

Example 4.4. [Hackbusch, 1986, §1.4]. Consider $U = \Omega \times \Sigma = (0,1) \times (0, 1/\pi)$. We prescribe the data $u(x,0) = u(0,t) = u(1,t) = 0$ and $u(x, 1/\pi) = \sin(k\pi x)$ for $k \in \mathbb{N}$. One verifies that

$$u(x,t) = \sin(k\pi x) \frac{\sin(k\pi t)}{\sin k}$$

solves the wave equation in U and satisfies the given data. Now for $k \to \infty$ we have $|u(x, 1/\pi)| \leq 1$ but $\sup\{1/\sin\nu \,:\, \nu \in \mathbb{N}\} = \infty$. Again, the solution does not depend continuously on the data. □

In the following, we consider therefore the wave equation with boundary conditions (4.6) on Γ and initial conditions $u, \partial_t u$ on Γ_0.

Extensions Sometimes the partial differential operator $L(\epsilon)$ depends on a parameter $\epsilon \geq 0$. Then the PDE is called *singularly perturbed* if the type of $L(0)$ is different from the type $L(\epsilon)$ for $\epsilon > 0$. As an example consider the equation

$$\partial_t u - \epsilon \Delta u + b \cdot \nabla u = 0.$$

For any $\epsilon > 0$ the equation is second-order parabolic but for $\epsilon = 0$ the equation is first-order (hyperbolic).

We now turn to the type classification of nonlinear partial differential equations. The most general form of a scalar second-order nonlinear PDE in n variables is

$$F(\partial^2_{x_1} u, \partial_{x_1}\partial_{x_2} u, \ldots \partial^2_{x_n} u, \partial_{x_1} u, \ldots, \partial_{x_n} u, u, x) = 0 \qquad (x \in U)$$

with F a function in $n(n-1)/2 + 2n + 1$ variables which we may name

$$z = (z_{11}, z_{12}, \ldots, z_{nn}, z_1, \ldots, z_n, z_0)^T.$$

A linearization of the PDE around the state $\bar{u}(x)$ is obtained by decomposing $u(x) = \bar{u}(x) + \tilde{u}(x)$ and applying Taylor expansion:

$$F(\partial^2_{x_1} u, \ldots, u, x) \doteq \sum_{j=1}^{n} \sum_{i \geq j} \partial_{z_{ij}} F \, \partial_{x_j}\partial_{x_i} \tilde{u} + \sum_{i=1}^{n} \partial_{z_i} F \, \partial_{x_i} \tilde{u} + \partial_{z_0} F \, \tilde{u} + \bar{F} = 0.$$

This is a linear PDE in $\tilde{u}$ with coefficients $a_{ij}(x) = \partial_{z_{ij}} F(\partial^2_{x_1} \bar{u}, \ldots, \bar{u}, x)$. The type classification is then applied to this linear a PDE for a given state $\bar{u}$.

As an example consider the *porous medium equation*

$$\partial_t u - \Delta u^p = \partial_t u - \nabla \cdot (p u^{p-1} \nabla u) = 0 \tag{4.11}$$

with $p > 1$ and where we assume that boundary and initial conditions are such that $u \geq 0$ is ensured. This equation is called *degenerate parabolic* as it is parabolic at points where $u(x) > 0$ and it degenerates into an ordinary differential equation at points where $u(x) = 0$.

4.3. First-order Hyperbolic Systems

Method of Characteristics We will consider the linear scalar conservation law

$$\partial_t u + \nabla \cdot (vu) = 0 \qquad \text{in } \mathbb{R}^n \times \mathbb{R}^+ \tag{4.12}$$

with initial conditions

$$u(x,0) = u^0(x) \qquad \text{on } \mathbb{R}^n. \tag{4.13}$$

Here $v : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n$ is a given time-dependent velocity field assumed sufficiently smooth. The extension to a bounded domain Ω is straightforward. The nonlinear case is treated in [Evans, 2010, § 3.2]. The following Theorem gives an explicit solution formula for this equation.

Proposition 4.5. Let $u \in C^1(\mathbb{R}^n \times \mathbb{R}_0^+)$ solve (4.12) for a smooth velocity field $v \in [C^1(\mathbb{R}^n \times \mathbb{R}^+)]^n$. For any point $x^0 \in \mathbb{R}^n$ define the *characteristic curve* $(\hat{x}(t), t)$ by

$$\frac{d\hat{x}}{dt}(t) = v(\hat{x}(t), t), \ t > 0, \qquad\qquad \hat{x}(0) = x^0 \ . \tag{4.14}$$

Then the solution of (4.12) at any point along the curve (4.14) is given by

$$u(\hat{x}(t), t) = u^0(x^0) \exp\left(- \int_0^t (\nabla \cdot v)(\hat{x}(s), s)\, ds \right) \quad . \tag{4.15}$$

Proof. Differentiating u along the curve gives

$$\begin{aligned} \frac{d}{ds} u(\hat{x}(s), s) &= \partial_t u(\hat{x}(s), s) + \sum_{i=1}^n \partial_{x_i} u(\hat{x}(s), s) \frac{d\hat{x}_i}{ds}(s) \\ &= \partial_t u(\hat{x}(s), s) + \nabla u(\hat{x}(s), s) \cdot v(\hat{x}(s), s) \\ &= \partial_t u(\hat{x}(s), s) + \nabla \cdot (v(\hat{x}(s), s) u(\hat{x}(s), s)) - (\nabla \cdot v(\hat{x}(s), s)) u(\hat{x}(s), s) \end{aligned}$$

where we have used the definition of the characteristic curve and the product rule $\nabla \cdot (vu) = v \cdot \nabla u + (\nabla \cdot v)u$. The first two terms vanish since u solves (4.12) and we are left with the linear ordinary differential equation

$$\frac{d}{ds} u(\hat{x}(s), s) = -(\nabla \cdot v(\hat{x}(s), s)) u(\hat{x}(s), s) \tag{4.16}$$

which has the solution stated in the Theorem. □

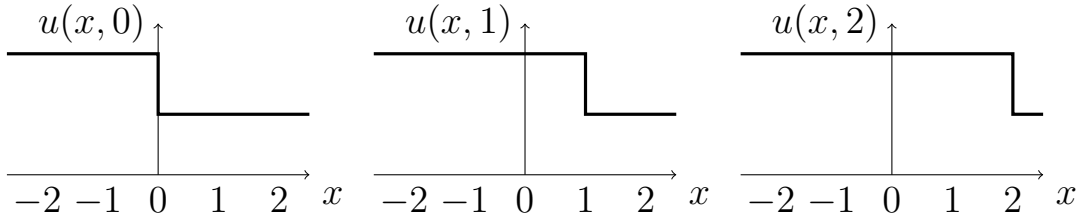


Figure 4.1.: Solution of the equation $\partial_t u + \partial_x u = 0$ with a step initial condition.

The characteristic curves defined in the Theorem are the paths followed by particles in the flow. A special case is given when v is divergence free, $\nabla \cdot v = 0$. Then Theorem 4.5 states that the solution is constant along any characteristic curve. In particular $v = \text{const}$ is a divergence free velocity field and the corresponding characteristic curves $(\hat{x}(t), t) = (x^0 + vt, t)$ are straight lines in space-time. The solution at any point is then given by $u(x,t) = u^0(x - vt)$.

As a further specialization consider now the special initial data $u^0(x) = \phi(y \cdot x)$ where y is an arbitrary vector of modulus 1 and ϕ is a strictly monotone scalar function. Since $\nabla\phi(y \cdot x) = \phi'(y \cdot x)y$ the level sets of $u^0(x)$ are hyperplanes in $\mathbb{R}^n$ which are perpendicular to the given direction y. For this special initial data the solution of (4.12) with $v = \text{const}$ is given by $u(x,t) = \phi(y \cdot (x - vt)) = \phi(y \cdot x - ty \cdot v)$. Thus the level contours move with velocity $y \cdot v$ in the direction y. This explains the fact that solutions of (4.12) support the propagation of *waves* (without a formal definition of a wave). More specifically, solutions of the form $\phi(y \cdot x - ty \cdot v)$ are called *plane waves* and they play a role in the generalization to systems of equations.

Using the explicit solution formula from Theorem 4.5 we can analyze the regularity of the solution. Clearly, when the velocity field is smooth enough (e.g. Lipschitz continuous in x) and the initial data u^0 is continuously differentiable then also $u(x,t)$ will be continuously differentiable. However, the solution formula makes also sense when the initial data is discontinuous (which might be a perfectly good approximation for a density or a concentration)!

Example 4.6. Consider the one-dimensional case with $v = 1$ and the step initial condition

$$u^0(x) = \begin{cases} 2 & x \leq 0 \\ 1 & x > 0 \end{cases} .$$

According to the method of characteristics we get the discontinuous solution $u(x,t) = u^0(x - t)$ illustrated in Figure 4.1 for the times $t = 0, 1, 2$. □.

This example motivates that requiring in general a solution of a kth order PDE to be k times continuously differentiable might be too restrictive. The question is then to give such "generalized" solutions lacking the required regularity a precise mathematical sense.

One-dimensional Systems We now turn, as an intermediate step, to the case of a vector-valued function $u(x,t) = (u_1(x,t), \ldots, u_m(x,t))^T$ in one spatial dimension and consider the system of equations

$$\partial_t u + B\partial_x u = 0 \qquad \text{in } \mathbb{R}^n \times \mathbb{R}^+ \tag{4.17}$$

where B is a constant $m \times m$ matrix. One idea to solve this equation is to require B to be real diagonalizable, i.e. B has m real eigenvalues $\lambda_1, \ldots, \lambda_m$ and a corresponding set of eigenvectors $r_1, \ldots, r_m$ that form a basis of $\mathbb{R}^m$. In that case there exists an orthogonal matrix Q with $QBQ^T = D$, $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$. Using the transformation $w = Qu$ we can transform the system (4.17) into the equivalent system

$$\partial_t w + D\partial_x w = 0 \qquad \text{in } \mathbb{R}^n \times \mathbb{R}^+ \quad .$$

In the transformed system all components decouple and each component can be solved independently using the method of characteristics. Note that the velocities are the eigenvalues λ_k which might be different for each component. Each component of the solution of the original system $u = Q^T w$ is then a linear combination of these "simple" waves w_j.

We can also ask whether a system of the form (4.17) can have plane wave solutions. To find them we make the ansatz

$$u(x,t) = \phi(yx - \sigma t)$$

where the "direction" y is now reduced to a scalar, $\phi : \mathbb{R} \to \mathbb{R}^m$ is now a vector-valued function in one argument and the scalar factor σ is to be determined. Inserting this ansatz into the PDE (4.17) results in

$$\frac{d\phi}{ds}(yx - \sigma t)(-\sigma) + A\frac{d\phi}{ds}(yx - \sigma t)y = (-\sigma I + yA)\frac{d\phi}{ds}(yx - \sigma t) = 0 \quad .$$

This equation can not be satisfied by any profile function ϕ in contrast to the scalar case. However, if we assume A to be diagonalizable we can require that $\frac{d\phi}{ds}$ is equal to an eigenvector:

$$\frac{d\phi}{ds} = r_k \qquad \Rightarrow \qquad \phi(s) = sr_k + \phi_k \quad (k = 1, \ldots, m).$$

Then our equation reduces to

$$(-\sigma + y\lambda_k)r_k = 0 \qquad \Leftrightarrow \qquad \sigma = \lambda_k \quad (k = 1, \ldots, m).$$

Thus we can conclude that equation (4.17) supports m plane wave solutions that have the form

$$u_k(x,t) = (yx - y\lambda_k t)r_k + \phi_k \qquad (k = 1, \ldots, m)$$

with arbitrary ϕ_k provided A is diagonalizable. Since y has the meaning of a direction we may assume $y = 1$ and so the possible velocities $\sigma = \lambda_k$ are just the eigenvalues of A.

It turns out that the diagonalizability of A is not just a nice mathematical structure that allows one to solve the system (4.17) but that such systems are also practically relevant!

Multi-dimensional Systems We now turn to the linear system of m equations in n space dimensions of the form

$$\partial_t u + \sum_{j=1}^{n} B_j \partial_{x_j} u = f \qquad \text{in } \mathbb{R}^n \times \mathbb{R}^+ \tag{4.18}$$

with the initial condition

$$u(x,0) = u^0(x) \qquad \text{on } \mathbb{R}^n \tag{4.19}$$

where $u : \mathbb{R}^n \times \mathbb{R}_0^+ \to \mathbb{R}^m$ is the unknown function and $B_j : \mathbb{R}^n \times \mathbb{R}_0^+ \to \mathbb{R}^{m\times m}$, $f : \mathbb{R}^n \times \mathbb{R}_0^+ \to \mathbb{R}^m$ are given functions that may depend on position and time.

This system can not be transformed into a decoupled system for the individual components by a transformation $w = Qu$ because the matrices B_j are, in general, not simultaneously diagonalizable. However, we can still ask for the existence of plane wave solutions. In analogy we make the ansatz $u(x,t) = \phi(y \cdot x - \sigma t)$ where $y \in \mathbb{R}^n$ is now a given direction, ϕ is a vector-valued function in one variable and the B_j are assumed to be constant matrices. Inserting this ansatz into the PDE (4.18) yields

$$\left(-\sigma I + \sum_{j=1}^{n} y_j B_j \right) \frac{d\phi}{ds}(y \cdot x - \sigma t) = 0 \quad .$$

If we now assume that the matrix $B(y) = \sum_{j=1}^n y_j B_j$ is diagonalizable for any $y \in \mathbb{R}^n$ with m eigenvalues $\lambda_k(y)$ and corresponding eigenvectors $r_k(y)$we can set again $\frac{d\phi}{ds} = r_k(y)$ and the system reduces to

$$\left(-\sigma + \lambda_k(y)\right) r_k(y) = 0 \quad .$$

Consequently we will have the m plane wave solutions of the form

$$u_k(x,t) = (y \cdot x - \lambda_k(y)t) r_k(y) + \phi_k \qquad (k = 1, \ldots, m)$$

with arbitrary ϕ_k. This motivates now the following definition.

Definition 4.7 (Hyperbolic linear first-order systems)**.** The system of equations (4.18) is called *hyperbolic* if for each $x, y \in \mathbb{R}^n$ and $t \geq 0$ the $m \times m$ matrix

$$B(x,t;y) = \sum_{j=1}^{n} y_j B_j(x,t) \tag{4.20}$$

is real diagonalizable, i.e. it has m real eigenvalues $\lambda_1(x,t;y), \ldots, \lambda_m(x,t;y)$ and its corresponding eigenvectors $r_1(x,t;y), \ldots, r_m(x,t;y)$ form a basis of $\mathbb{R}^m$. In addition there are two special cases:

i) The system is called *symmetric hyperbolic* if $B_j(x,t)$ is symmetric for every $x \in \mathbb{R}^n$, $t \geq 0$ and $j = 1, \ldots, m$.

ii) The system is called *strictly hyperbolic* if for $x, y \in \mathbb{R}^n$, $y \neq 0$ and $t \geq 0$ the matrix $B(x,t;y)$ has m distinct real eigenvalues. □

Example 4.8 (Linear Acoustics)**.** We consider the system of linear acoustics in three space dimensions given in Equation (2.28). Setting $u = (\tilde{p}, \tilde{v}_1, \tilde{v}_2, \tilde{v}_3)$ this system can be written as

$$\partial_t u + \sum_{j=1}^{n} B_j \partial_{x_j} u = 0$$

with

$$B_1 = \begin{pmatrix} 0 & c^2\bar{\rho} & 0 & 0 \\ 1/\bar{\rho} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \; B_2 = \begin{pmatrix} 0 & 0 & c^2\bar{\rho} & 0 \\ 0 & 0 & 0 & 0 \\ 1/\bar{\rho} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \; B_3 = \begin{pmatrix} 0 & 0 & 0 & c^2\bar{\rho} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1/\bar{\rho} & 0 & 0 & 0 \end{pmatrix}.$$

For any $y \in \mathbb{R}^3$ we therefore have

$$B(y) = \sum_{j=1}^{3} y_j B_j = \begin{pmatrix} 0 & y_1 c^2\bar{\rho} & y_2 c^2\bar{\rho} & y_3 c^2\bar{\rho} \\ y_1/\bar{\rho} & 0 & 0 & 0 \\ y_2/\bar{\rho} & 0 & 0 & 0 \\ y_3/\bar{\rho} & 0 & 0 & 0 \end{pmatrix}.$$

With the transformation matrix $T = \mathrm{diag}(\bar{\rho}c, 1, 1, 1)$ we see that $B(y)$ is similar to the symmetric matrix

$$T^{-1}B(y)T = \begin{pmatrix} 0 & y_1 c & y_2 c & y_3 c \\ y_1 c & 0 & 0 & 0 \\ y_2 c & 0 & 0 & 0 \\ y_3 c & 0 & 0 & 0 \end{pmatrix}$$

and therefore is diagonalizable with eigenvalues

$$\lambda_{1,2} = \pm c\|y\| \qquad \text{and} \qquad \lambda_{3,4} = 0.$$

Since y is a direction vector we may assume $\|y\| = 1$ and therefore the system supports two wave solutions with velocities $\pm c$ (explaining that c is the speed of sound). □

Definition 4.7 can be extended to the slightly more general system

$$B_0 \partial_t u + \sum_{j=1}^{n} B_j \partial_{x_j} u = 0 \qquad \text{in } \mathbb{R}^n \times \mathbb{R}^+ \tag{4.21}$$

where B_0 is a constant symmetric positive definite matrix. This system is also called hyperbolic provided the matrix $B(x,t;y)$ defined in (4.20) is diagonalizable. This can be shown as follows. By assumption there exists an orthogonal matrix Q such that $QBQ^T = D = \operatorname{diag}(\mu_1, \ldots, \mu_m)$, $\mu_k > 0$. With the transformation $w = D^{1/2}Qu$ the system (4.21) is equivalent to

$$\partial_t w + \sum_{j=1}^{n} D^{1/2} Q B_j Q^T D^{-1/2} \partial_{x_j} w = 0 \qquad \text{in } \mathbb{R}^n \times \mathbb{R}^+.$$

For this transformed system and any $y \in \mathbb{R}^n$ we have

$$\begin{aligned} \sum_{j=1}^{n} y_j D^{1/2} Q B_j Q^T D^{-1/2} &= D^{1/2} Q \left(\sum_{j=1}^{n} y_j B_j \right) Q^T D^{-1/2} \\ &= D^{1/2} Q B(x,t;y) Q^T D^{-1/2}, \end{aligned}$$

so the diagonalizability of $B(x,t;y)$ also implies the hyperbolicity of the transformed system.

First- and Second-order Hyperbolic Equations We now establish a connection between first-order hyperbolic systems and second-order scalar hyperbolic equations. We only consider the case $b \equiv 0$, $c \equiv 0$ in (4.2). With the vector-valued function $v = (v_1, \ldots, v_n, v_{n+1})^T = (\partial_{x_1} u, \ldots, \partial_{x_n} u, \partial_t u)^T$ we obtain the system of $n+1$ equations

$$\begin{aligned} \sum_{j=1}^{n} a_{ij} \partial_t v_j - \sum_{j=1}^{n} a_{ij} \partial_t v_{n+1} &= 0 \qquad (i = 1, \ldots, n), \\ \partial_t v_{n+1} - \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} \partial_{x_j} v_i &= f. \end{aligned}$$

Here the first n equations are a consequence of the n identities $\partial_t \partial_{x_i} u = \partial_{x_i} \partial_t u$ and the fact that the rows of $A(x,t)$ are linearly independent. The last equation is our second-order hyperbolic PDE. Now this system can be written as first-order system of the form (4.21) with the matrices

$$B_0 = \begin{pmatrix} a_{11} & \dots & a_{1n} & 0 \\ \vdots & & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}, \qquad B_j = \begin{pmatrix} 0 & \dots & 0 & -a_{1j} \\ \vdots & & \vdots & \vdots \\ 0 & \dots & 0 & -a_{nj} \\ -a_{1j} & \dots & -a_{nj} & 0 \end{pmatrix}.$$

The positive definiteness of $A(x,t)$ ensures the positive definiteness of B_0 and since the B_j are symmetric any combination $\sum_{j=1}^n y_j B_j$ is diagonalizable. Thus we have shown that a scalar second-order hyperbolic PDE can be written as a (symmetric) hyperbolic first-order system. This also implies that (appropriately generalized) solutions of a scalar second-order hyperbolic PDE may be discontinuous.

Nonlinear Hyperbolic Conservation Laws Definition 4.7 can be extended to the case of a nonlinear conservation law (2.24) (such as the Euler equations) as follows. We rewrite the conservation law using the chain rule (assuming F to be sufficiently smooth):

$$\partial_t u + \sum_{j=1}^n \partial_{x_j} F_j(u) = \partial_t u + \sum_{j=1}^n \nabla F_j(u) \partial_{x_j} u$$

(∇F_j denotes the Jacobian matrix). Then the nonlinear system is called hyperbolic if the matrix

$$B(x,t;y) = \sum_{j=1}^n y_j \nabla F_j(u(x,t))$$

is diagonalizable for any $y \in \mathbb{R}^n$ and possible state $u(x,t)$.

4.4. Model Problems

In order to summarize this chapter we give a list of problems (including boundary and initial conditions) which will serve as model problems in the rest of the text and which we will be able to solve numerically in the course of the lecture. In the following $\Omega \subset \mathbb{R}^n$ is a (spatial) domain, $\Sigma = (0,T]$ is a time interval and $u : \Omega \to \mathbb{R}$ or $u : \Omega \times \Sigma \to \mathbb{R}$ denotes the unknown scalar function.

a) Transport problem (first-order hyperbolic).

$$\begin{aligned} \partial_t u + \nabla \cdot (vu) &= f && \text{in } \Omega \\ u &= g && \text{on } \Gamma = \{x \in \partial\Omega \,:\, v(x) \cdot n(x) < 0\}. \end{aligned}$$

b) Laplace equation with Dirichlet boundary conditions (second-order elliptic).

$$\begin{aligned} -\Delta u &= 0 && \text{in } \Omega \\ u &= g && \text{on } \partial\Omega. \end{aligned}$$

c) Poisson equation with Neumann boundary conditions (second-order elliptic).

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega \\ -\nabla u \cdot n &= g && \text{on } \partial\Omega. \end{aligned}$$

In order for a solution to exist the compatibility condition $\int_\Omega f\,dx = \int_{\partial\Omega} g\,ds$ is required. The solution is unique up to a constant.

d) Heat equation (second-order parabolic).

$$\begin{aligned} \partial_t u - \Delta u &= f && \text{in } \Omega \times \Sigma \\ u &= g && \text{on } \partial\Omega \times \Sigma \\ u &= u_0 && \text{on } \Omega \times \{0\}. \end{aligned}$$

e) Wave equation (second-order hyperbolic).

$$\begin{aligned} \partial_t^2 u - \Delta u &= f && \text{in } \Omega \times \Sigma \\ u &= g && \text{on } \partial\Omega \times \Sigma \\ u = u_0, \partial_t u &= u_1 && \text{on } \Omega \times \{0\}. \end{aligned}$$

f) Convection-diffusion-reaction equation

$$\begin{aligned} \partial_t u + \nabla \cdot (vu - D\nabla u) + cu &= f && \text{in } \Omega \times \Sigma \\ u &= g && \text{on } \partial\Omega \times \Sigma \\ u &= u_0 && \text{on } \Omega \times \{0\}. \end{aligned}$$

Chapter 5.

Elements of Functional Analysis

5.1. Motivation

In this chapter we study the existence and uniqueness of linear variational problems that arise from elliptic partial differential equations.

As a motivation consider the elliptic boundary value problem

$$-\nabla \cdot (K\nabla u) + cu = f \qquad \text{in } \Omega, \tag{5.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{5.1b}$$

Multiplying both sides with a suitable test function (variation) v, $v = 0$ on $\partial\Omega$ (because no variation is needed on the Dirichlet boundary, see Section 3.2) and integrating we obtain through integration by parts:

$$-\int_\Omega \nabla \cdot (K\nabla u)v + cuv\,dx = \int_\Omega (K\nabla u)\cdot\nabla v + cuv\,dx = \int_\Omega fv\,dx$$

where we have observed that the boundary term vanishes due to $v = 0$ on $\partial\Omega$. Defining the so-called bilinear and linear forms

$$a(u,v) = \int_\Omega (K\nabla u)\cdot\nabla v + cuv\,dx, \qquad l(v) = \int_\Omega fv\,dx,$$

we obtain a problem of the abstract form

$$\boxed{\text{Find } u \in U: \qquad a(u,v) = l(v) \quad \forall v \in V}. \tag{5.2}$$

where U and V are appropriate function spaces for the solution u and the test functions (variations) v. The purpose of this chapter is to define the appropriate function spaces and to give necessary and sufficient conditions for the existence of a unique solution of problem (5.2).

Provided the solution of (5.2) is in $C^2(\Omega)\cap C^0(\Omega)$ the integration by parts can be reversed and the solution satisfies also (5.1). Since it will turn out that problem (5.2) has a solution under weaker assumptions on the data and coefficients it is called the *weak formulation* and (5.1) correspondingly is called *strong formulation.*

The abstract setting (5.2) is not restricted to scalar problems. As an example for a system consider the Stokes problem

$$\begin{aligned} -\Delta u + \nabla p &= s && \text{in } \Omega, \\ \nabla \cdot u &= g && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

Scalar multiplication with a vector-valued test function $v \in V_v$ and integration by parts applied to the momentum equation results in

$$\int_\Omega \left(\sum_{i=1}^{n} \nabla u_i \cdot \nabla v_i \right) - p(\nabla \cdot v)\, dx = \int_\Omega s \cdot v\, dx \qquad \forall v \in V_v.$$

Multiplication with a test function $q \in V_p$ in the mass conservation equation and integration results in

$$\int_\Omega (\nabla \cdot u) q\, dx = \int_\Omega gq\, dx \qquad \forall q \in V_q.$$

Combining both variational equations results in a problem in the abstract form (5.2) with $U = U_v \times U_p$, $V = V_v \times V_p$ and the bilinear and linear forms

$$\begin{aligned} a((u,p),(v,q)) &= \int_\Omega \left(\sum_{i=1}^{n} \nabla u_i \cdot \nabla v_i \right) - p(\nabla \cdot v) + q(\nabla \cdot u)\, dx, \\ l((v,q)) &= \int_\Omega s \cdot v + gq\, dx. \end{aligned}$$

5.2. Banach Spaces

In the following V, W are vector spaces (also called linear spaces) over the field $\mathbb{R}$, i.e. a set that is closed under addition and scalar multiplication. (The generalization to the field $\mathbb{C}$ is possible but not needed here).

Definition 5.1 (Norm)**.** A mapping $\|.\|_V : V \to \mathbb{R}$ is called a *norm* if it satisfies the following three properties:

i) $\|v\| = 0$ if and only if $v = 0$ (definiteness).

ii) $\forall c \in \mathbb{R}, \forall v \in V : \|cv\|_V = |c|\|v\|_V$ (homogeneity).

iii) $\forall v, w \in V : \|v + w\|_V \leq \|v\|_V + \|w\|_V$ (triangle inequality).

A mapping that only satisfies ii) and iii) is called a *semi-norm.* □

The pair $(V, \|.\|_V)$ is called a *normed (vector) space.* Examples for normed spaces are

a) $\mathbb{R}^n$ with the Euclidean norm $\|x\| = \left(\sum_{i=1}^n x_i^2\right)^{\frac{1}{2}}$.

b) Let $\Omega \subset \mathbb{R}^n$ be a bounded domain. $C^k(\Omega)$ denotes the vector space of functions u with continuous partial derivatives $\partial^\alpha u$ up to order k. Since Ω is open these functions may be unbounded. If all $\partial^\alpha u$ up to order k are bounded in Ω and uniformly continuous then there exists a unique, bounded, continuous extension to the closure $\overline{\Omega}$. The space $C^k(\overline{\Omega})$ of all functions with bounded, continuous partial derivatives up to order k is a normed vector space with the norm

$$\|u\|_{C^k(\overline{\Omega})} = \max_{0 \le |\alpha| \le k} \sup_{x \in \Omega} |\partial^\alpha u(x)|.$$

Sometimes we will write $\|u\|_\infty = \|u\|_{C^k(\overline{\Omega})}$.

The norm defines a topology (i.e. the notion of open sets) on V. A subset $X \subset V$ is called *open*, if

$$\forall x \in X, \exists \epsilon > 0 : B_\epsilon(x) \subset X$$

where $B_\epsilon(x) = \{y \in V \,:\, \|x - y\|_V < \epsilon\}$ is the open ball with radius ϵ around x.

Norms are important in connection with limit processes. We say that a sequence $(v_k)_{k \in \mathbb{N}} \subset V$ converges to $v \in V$ if and only if $\lim_{k \to \infty} \|v_k - v\|_V = 0$. Convergence is denoted by $v_k \to v$ or $v = \lim_{k \to \infty} v_k$. A consequence of the triangle inequality is the inverse triangle inequality

$$|\|v\| - \|w\|| \le \|v - w\| \quad \forall v, w \in V$$

which implies the continuity of the norm

$$(v_k \to v) \quad \Rightarrow \quad (\|v_k\| \to \|v\|).$$

Definition 5.2 (Equivalent norms)**.** Two norms $\|.\|$ and $|||.|||$ on V are called *equivalent* if

$$c_1 \,|||\, v \,|||_V \le \|v\|_V \le c_2 \,|||\, v \,|||_V \quad \forall v \in V.$$

with two constants $c_1, c_2 > 0$. □

In finite-dimensional spaces all norms are equivalent

Definition 5.3 (Cauchy sequence)**.** A sequence $\{v_k : k \in \mathbb{N}\}$ with

$$\sup\{\|v_n - v_m\|_V \, : \, n, m \geq k\} \to 0 \quad \text{for} \quad k \to \infty.$$

is called *Cauchy sequence*. □

A space V is called *complete*, if every Cauchy sequence has a limit in V.

Definition 5.4 (Banach space)**.** A complete, normed vector space is called *Banach space*. □

Examples for Banach spaces are

a) $\mathbb{R}, \mathbb{C}$ with modulus $|.|$ as a norm are complete, $\mathbb{Q}$ is not.

b) $(\mathbb{R}^n, \|.\|)$ is Banach space.

c) $\Omega \subset \mathbb{R}^n$ bounded domain, $(C^0(\overline{\Omega}), \|.\|_\infty)$ is Banach space.

Banach spaces can be constructed via the process of *completion*: Let $(X, \|.\|_X)$ be a normed vector space that is not complete. Then $(V = \overline{X}, \|.\|_V)$ is called the completion of X. Every element $v \in V$ is limit of a Cauchy sequence $\{x_k\} \subseteq X$ and $\|v\|_V = \lim_{k\to\infty} \|x_k\|_X$. Especially, for $x \in X \cap V$ we have $\|x\|_X = \|x\|_V$. Conversely, X is called *dense* in $(V, \|.\|_V)$ if $X \subset V$ and $\overline{X} = V$.

5.3. Hilbert Spaces

Definition 5.5 (Scalar Product)**.** A mapping $(.,.)_V : V \times V \to \mathbb{R}$ is called *scalar product* if

i) $(v, v)_V > 0$ if and only if $v \neq 0$ (definiteness).

ii) $(\alpha v + w, z)_V = \alpha(v, z)_V + (w, z)_V \; \forall v, w, z \in V, \forall \alpha \in \mathbb{R}$ (linearity).

iii) $(v, w)_V = (w, v)_V \; \forall v, w \in V$ (symmetry). □

For any scalar product $(.,.)_V$ on a space V,

$$\|v\|_V = \sqrt{(v, v)_V}$$

defines a norm, the so-called induced norm, on V.

Definition 5.6 (Hilbert Space)**.** A *Hilbert space* is a vector space with scalar product that is complete with respect to the norm induced by the scalar product. □

By definition, a Hilbert space is also a Banach space but not necessarily vice versa.

Lemma 5.7 (Cauchy Schwarz Inequality)**.** In a Hilbert space the following inequality holds true:

$$\forall v, w \in V : \qquad |(v, w)_V| \leq \|v\|_V \|w\|_V.$$

Since

$$\|v\|_V\|w\|_V - (v, w)_V = \frac{\|v\|_V\|w\|_V}{2} \left\| \frac{v}{\|v\|_V} - \frac{w}{\|w\|_V} \right\|_V^2$$

equality holds true if and only if v and w are colinear, see [Ern and Guermond, 2004, A.6] □

We call $v, w \in V$ orthogonal if $(v, w)_V = 0$. Given $X \subseteq V$ we denote by $X^\perp = \{w \in V : (w, v)_V = 0, \forall v \in X\}$ the space orthogonal to X with respect to V. $X^\perp$ is a closed subspace of V.

5.4. Linear Mappings in Banach Spaces

Definition 5.8. Let V, W be normed spaces. Then $\mathcal{L}(V; W)$ denotes the set of all linear and continuous mappings from V to W. An element $A \in \mathcal{L}(V; W)$ is called operator. □

We recall that a mapping $A : V \to W$ is called *linear* if the following two conditions hold:

i) $A(v_1 + v_2) = A(v_1) + A(v_2)$, $\forall v_1, v_2 \in V$.

ii) $A(\alpha v) = \alpha A(v)$, $\forall v \in V, \forall \alpha \in \mathbb{R}$.

Instead of $A(v)$ we often write Av when A is a linear operator. With

$$\|A\|_{\mathcal{L}(V;W)} = \sup_{v \in V, v \neq 0} \frac{\|Av\|_W}{\|v\|_V}$$

a norm is declared on $\mathcal{L}(V; W)$. The condition $v \neq 0$ when taking the supremum is implicitly understood in the following. The norm $\|A\|_{\mathcal{L}(V;W)}$ of the operator A exists (i.e. the operator is bounded) if and only if A is continuous, see [Hackbusch, 1986, Excercise 6.1.5].

$(\mathcal{L}(V; W), \|.\|_{\mathcal{L}(V;W)})$ is a normed vector space. Since $\frac{\|Av\|_W}{\|v\|_V} \leq \|A\|_{\mathcal{L}(V;W)}$ for any $v \neq 0$, the inequality

$$\|Av\|_W \leq \|A\|_{\mathcal{L}(V;W)} \|v\|_V$$

holds for all $v \in V$.

If V is a normed space and W is a Banach space then $\mathcal{L}(V;W)$ is a Banach space, see [Ern and Guermond, 2004, Prop. A.10].

The following proposition shows that the continuity of the operator implies convergence of the image of a convergent sequence.

Proposition 5.9. Let V, W be Banach spaces, $A \in \mathcal{L}(V;W)$ a bounded linear operator and $v_k \to v$ a convergent sequence in V. Then Av_k converges to Av.
Proof. $\{v_k : k \in \mathbb{N}\}$ is Cauchy sequence. $\|Av_n - Av_m\|_W \leq \|A\|_{\mathcal{L}(V;W)}\|v_n - v_m\|_V$ shows that $\{Av_k : k \in \mathbb{N}\}$ is a Cauchy sequence in W. Since W is complete this sequence has the limit $w = \lim_{k\to\infty} Av_k$. It remains to show that $w = Av$, which follows from $\|Av - W\|_W = \lim_{k\to\infty} \|Av - Av_k\|_W \leq \|A\|_{\mathcal{L}(V;W)}\|v - v_k\|_V$ and $v_k \to v$. □

Often an operator is defined only on a dense subspace. The following proposition states how the operator can be extended to the whole space.

Proposition 5.10. Let V_0 be a dense subspace of the normed space V and W is Banach space.

i) A bounded linear operator $A_0 \in \mathcal{L}(V_0;W)$ defined on the subspace V_0 has a unique extension $A \in \mathcal{L}(V;W)$ with $Av = A_0 v$ for all $v \in V_0$.

ii) For any sequence $v_k \to v$, $(v_k \in V_0, v \in V)$ there holds $Av = \lim_{k\to\infty} A_0 v_k$.

iii) $\|A\|_{\mathcal{L}(V;W)} = \|A_0\|_{\mathcal{L}(V_0;W)}$.

Proof. [Hackbusch, 1986, Satz 6.1.11] □

Definition 5.11. Let V, W be Banach spaces. $A \in \mathcal{L}(V;W)$ is called *compact* if for every bounded sequence $\{v_n : n \in \mathbb{N}\} \subset V$ there exists a subsequence $\{v_{n_k} : k \in \mathbb{N}\}$ such that Av_{n_k} converges in W. □

After these more general definitions and properties we turn to some special operators.

Definition 5.12 (Dual Space)**.** Let V be a normed vector space. $V' = \mathcal{L}(V;\mathbb{R})$ is called the *dual space* of V. An element $A \in V'$ is called a *continuous (or bounded) linear form.* Instead of Av we will write $\langle A, v\rangle_{V',V}$. □

Since $\mathbb{R}$ is a Banach space, V' is also Banach space with the canonical norm

$$\|A\|_{V'} = \sup_{v\in V} \frac{Av}{\|v\|_V} = \sup_{v\in V} \frac{\langle A, v\rangle_{V',V}}{\|v\|_V}.$$

Theorem 5.13 (Riesz Representation Theorem)**.** Let $(V, (.,.)_V)$ be a Hilbert space. Then for every $v' \in V'$ there exists a unique $u \in V$ such that

$$\forall w \in V : \qquad \langle v', w\rangle_{V',V} = (u, w)_V.$$

The map $\tau : V' \to V$ mapping $v' \in V'$ to the corresponding $u \in V$ is linear and an isometry, i.e. $\|\tau v'\|_V = \|v'\|_{V'}$. For a constructive proof of the Riesz Theorem see [Brenner and Scott, 1994, §2.4]. The property that τ is an isometry will be important below and can be easily seen as follows: For any $u, w \in V$, $w \neq 0$ it follows from the Cauchy-Schwarz inequality that $\frac{(u,w)_V}{\|w\|_V} \leq \frac{\|u\|_V\|w\|_V}{\|w\|_V} = \|u\|_V$. Equality holds only for u and w colinear and therefore $\sup_{w\in V} \frac{(u,w)_V}{\|w\|_V} = \|u\|_V$. But this means that $\|v'\|_{V'} = \sup_{w\in V} \frac{\langle v',w\rangle_{V',V}}{\|w\|_V} = \sup_{w\in V} \frac{(\tau v',w)_V}{\|w\|_V} = \|\tau v'\|_V$. □

Definition 5.14 (Dual Operator)**.** Let V, W be normed vector spaces and $A \in \mathcal{L}(V; W)$. Then $A^T : W' \to V'$ given by

$$\forall v \in V, \forall w' \in W' : \qquad \langle A^T w', v\rangle_{V',V} = \langle w', Av\rangle_{W',W}$$

is called *dual operator*. The dual operator is a generalization of matrix transposition in $\mathbb{R}^n$. □

Definition 5.15 (Bilinear Forms)**.** Let Z_1, Z_2 be normed spaces. Then $\mathcal{L}(Z_1 \times Z_2; \mathbb{R})$ is the vector space of continuous *bilinear forms* on $Z_1 \times Z_2$. With the norm

$$\|a\|_{Z_1,Z_2} = \sup_{z_1\in Z_1, z_2\in Z_2} \frac{a(z_1, z_2)}{\|z_1\|_{Z_1}\|z_2\|_{Z_2}}$$

$\mathcal{L}(Z_1 \times Z_2; \mathbb{R})$ is a Banach space. □

The following observation associates a special operator with a bilinear form that will later play an important role in connection with the variational formulation of partial differential equations.

Proposition 5.16. Let Z_1, Z_2 be Banach spaces and $a \in \mathcal{L}(Z_1 \times Z_2; \mathbb{R})$ a bilinear form. The map $A : Z_1 \to Z_2'$ given by

$$\forall z_1 \in Z_1, \forall z_2 \in Z_2 : \qquad \langle Az_1, z_2\rangle_{Z_2',Z_2} = a(z_1, z_2)$$

is an element of $\mathcal{L}(Z_1; Z_2')$ and $\|A\|_{\mathcal{L}(Z_1;Z_2')} = \|a\|_{Z_1,Z_2}$.
Proof. For fixed $z_1 \in Z_1$, $l(z_2) = a(z_1, z_2)$ is a continuous linear map. Moreover,

$$\begin{aligned}\|A\|_{\mathcal{L}(Z_1;Z_2')} &= \sup_{z_1\in Z_1} \frac{\|Az_1\|_{Z_2'}}{\|z_1\|_{Z_1}} = \sup_{z_1\in Z_1} \sup_{z_2\in Z_2} \frac{\langle Az_1, z_2\rangle_{Z_2',Z_2}}{\|z_1\|_{Z_1}\|z_2\|_{Z_2}} \\ &= \sup_{z_1\in Z_1} \sup_{z_2\in Z_2} \frac{a(z_1, z_2)}{\|z_1\|_{Z_1}\|z_2\|_{Z_2}} = \|a\|_{Z_1,Z_2}\end{aligned}$$

shows that the norms of the operator A and the bilinear form a coincide. □

Definition 5.17 (Double Dual)**.** Let V be a Banach space. The dual space of V' is called the *double dual* of V and is denoted by V''. V'' is also a Banach space. □

Proposition 5.18. Let V be a Banach space. Define the map $J_V : V \to V''$ as

$$\forall u \in V, \forall v' \in V' : \qquad \langle J_V u, v' \rangle_{V'',V'} = \langle v', u \rangle_{V',V}.$$

Then J_V is an isometry, i.e. $\|J_V u\|_{V''} = \|u\|_V$.
Proof. See [Ern and Guermond, 2004, Prop. A.24]. □

Isometric maps in normed spaces are always injective, since when $u_1 \neq u_2$ we have $\|J_V u_1 - J_V u_2\| = \|J_V(u_1 - u_2)\| = \|u_1 - u_2\| \neq 0$, i.e. $J_V u_1 \neq J_V u_2$. But isometric maps need not be surjective in general. The isometry J_V being surjective defines a special kind of Banach space.

Definition 5.19. A Banach space is called *reflexive* if J_V is an isomorphism. □

5.5. Abstract Existence Theory

With the definitions of function spaces and linear mappings in place we can now turn back to the abstract problem (5.2). The following theorem states existence and uniqueness of the solution.

Theorem 5.20 (Banach-Nečas-Babuška)**.** Let U be a Banach space and V a reflexive Banach space, $a \in \mathcal{L}(U \times V, \mathbb{R})$ and $l \in V'$. Then Problem (5.2) is well posed if and only if

$$\exists \alpha > 0 : \qquad \inf_{u \in U} \sup_{v \in V} \frac{a(u,v)}{\|u\|_U \|v\|_V} \geq \alpha \tag{BNB1}$$

and

$$\forall v \in V : \qquad (\forall u \in U : \ a(u,v) = 0) \Rightarrow (v = 0). \tag{BNB2}$$

Moreover, the following a-priori estimate holds:

$$\forall l \in V' : \qquad \|u\|_U \leq \frac{1}{\alpha} \|l\|_{V'}.$$

Proof. See [Ern and Guermond, 2004, Theorem 2.6]. □

For the proof of this theorem we refer to the literature. Instead we provide only a motivation of the result here.

a) As in Proposition 5.16 we define the map $A : U \to V'$ via the bilinear form: $\forall u \in U, \forall v \in V : \langle Au, v \rangle_{V',V} = a(u,v)$, so $Au \in V'$ is the continuous linear

form that is obtained from $a(u,v)$ by fixing the argument u. Then Problem (5.2) can be understood as a linear operator equation

$$\text{Find } u \in U: \qquad Au = l. \tag{5.3}$$

Thus, Problem (5.2) has a unique solution if and only if the corresponding operator A is invertible, i.e. it is injective and surjective.

b) In order to characterize injective and surjective operators we go back to the $\mathbb{R}^n$. So set $Y = \mathbb{R}^m$, $Z = \mathbb{R}^n$ and $B \in \mathcal{L}(Y;Z)$. We recall the definitions

- $\text{range}(B) = \{z \in Z : By = z \text{ for some } y \in Y\}$.
- $\ker(B) = \{y \in Y : By = 0\}$.

Then the following holds true:

- B is injective if and only if $\ker(B) = \{0\}$. *Proof.* Suppose B is injective. Then there exist $y_1 \neq y_2$ such that for any $y \neq 0$ we have $0 \neq y = y_1 - y_2 \Rightarrow By = B(y_1 - y_2) = By_1 - By_2 \neq 0$. So 0 is the only element in $\ker(B)$. Now suppose $\ker(B) = 0$. Then for $y_1 \neq y_2$ we have $By_1 - By_2 = B(y_1 - y_2) \neq 0$ so B is injective.
- $\text{range}(B)^\perp = \ker(B^T)$ where $.^\perp$ is the orthogonal complement with respect to the Euclidean scalar product and B^T denotes the transposed matrix. *Proof.* $z \in \text{range}(B)^\perp \Leftrightarrow (z,v) = 0 \ \forall v \in \text{range}(B) \Leftrightarrow (z, By) = 0 \ \forall y \in Y \Leftrightarrow (B^T z, y) = 0 \ \forall y \in Y \Leftrightarrow B^T z = 0 \Leftrightarrow z \in \ker(B^T)$.
- B is surjective if and only if $\ker(B^T) = \{0\}$. *Proof.* B is surjective $\Leftrightarrow$ $\text{range}(B) = Z \Leftrightarrow \text{range}(B)^\perp = \{0\} \Leftrightarrow \ker(B^T) = \{0\}$.

The first and third equivalence illustrate that the invertability of B can be characterized by its kernel and range (respectively the kernel of B^T). In general Banach spaces the role of the transposed is taken over by the dual operator given in Definition 5.14.

c) Now we go back to the operator A from a). Clearly, if $\ker(A) = \{0\}$ then $\|Au\|_{V'}/\|u\|_U > 0$ for any $u \neq 0$. In infinite-dimensional spaces the latter condition is not sufficient to imply that A is injective. The precise condition is

$$A \text{ is injective} \wedge \text{range}(A) \text{ closed} \quad \Leftrightarrow \quad \exists \alpha > 0, \forall u \in U, u \neq 0 : \frac{\|Au\|_{V'}}{\|u\|_U} \geq \alpha$$

[Ern and Guermond, 2004, Lemma A.39]. With this result we obtain

$$\begin{aligned} \inf_{u \in U, u \neq 0} \frac{\|Au\|_{V'}}{\|u\|_U} &= \inf_{u \in U, u \neq 0} \sup_{v \in V, v \neq 0} \frac{\langle Au, v \rangle_{V',V}}{\|u\|_U \|v\|_V} && \text{(Definition of } \|.\|_{V'}) \\ &= \inf_{u \in U, u \neq 0} \sup_{v \in V, v \neq 0} \frac{a(u,v)}{\|u\|_U \|v\|_V} \geq \alpha. && \text{(Definition of } A) \end{aligned}$$

Thus (BNB1) ensures the injectivity of A.

d) The condition (BNB2) ensures the surjectivity of A, see [Ern and Guermond, 2004, A.2.2, A.2.3].

Theorem 5.20 gives necessary and sufficient conditions for the existence and uniqueness of a solution of Problem (5.2). We now consider a theorem that only gives a sufficient condition for existence and uniqueness. An advantage of this theorem will be that the assumptions on the bilinear form a are easier to check. In this theorem the spaces U and V are the same, i.e. $U = V$.

Theorem 5.21 (Lax-Milgram)**.** Let V be a Hilbert space, $a \in \mathcal{L}(V \times V, \mathbb{R})$ and $l \in V'$. Then Problem (5.2) is well posed provided a is *coercive*, i.e. it satisfies the condition

$$\exists \alpha > 0, \forall u \in V : \qquad a(u,u) \geq \alpha \|u\|_V^2. \tag{5.4}$$

Moreover, the following a-priori estimate holds:

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

Proof. We show that (5.4) implies (BNB1) and (BNB2). For $0 \neq w \in V$ coercivity implies:

$$\alpha \|w\|_V \leq \frac{a(w,w)}{\|w\|_V} \leq \sup_{v \in V} \frac{a(w,v)}{\|v\|_V}.$$

Since $0 \neq w \in V$ was chosen arbitrarily this implies

$$\inf_{w \in V} \sup_{v \in V} \frac{a(w,v)}{\|w\|_V \|v\|_V} \geq \alpha$$

which is (BNB1). Now assume that $v \in V$ is chosen such that $\forall w \in V : a(w,v) = 0$. Using again coercivity we have $\alpha \|v\|_V^2 \leq a(v,v) = 0$ implying $v = 0$, i.e. (BNB2). □

Remark 5.22. a) Lax-Milgram implies BNB but not vice-versa. Coercivity is only a sufficient condition for existence and uniqueness.

b) The condition $U = V$ is inherent in the condition of coercivity.

c) Coercivity also implies that V is a Hilbert space. The symmetrized bilinear form $\overline{a}(u,v) = a(u,v) + a(v,u)$ is a scalar product on V and implies the norm $\|v\|_{\overline{a}} = \sqrt{\overline{a}(v,v)}$. This norm is equivalent to the norm $\|.\|_V$ in V which follows from coercivity and continuity of a. Even if V is only assumed to be a Banach space, $(V, \overline{a}(.,.))$ is a Hilbert space and $\|.\|_{\overline{a}}$ induces the same topology as $\|.\|_V$ due to the equivalence of the norms.

d) The expression *ellipticity* of the bilinear form is used synonymously with coercivity.

e) The symmetry of the bilinear form a is *not* required for the proof of the Lax-Milgram theorem although in some books it is proven only under this assumption. □

We have proven the Lax-Milgram theorem by reducing it to the BNB-thoerem which we did not prove. We will now present a stand-alone proof of the Lax-Milgram theorem, taken from [Brenner and Scott, 1994, Theorem 2.7.7].

As stated in Equation (5.3) we can write the variational problem as an operator equation with an operator $A \in \mathcal{L}(V; V')$. Due to the Riesz representation theorem 5.13 there is a linear and isometric map $\tau \in \mathcal{L}(V', V)$ such that $\phi(v) = \langle \phi, v \rangle_{V',V} = (\tau\phi, v)_V$ for any $\phi \in V'$. Therefore we have the following reformulations of the variational problem:

$$\begin{aligned}
&\text{Find } u \in V: && a(u,v) = l(v)\ \forall v \in V && \text{(Variational formulation)}\\
\Leftrightarrow\ &\text{Find } u \in V: && Au = l && \text{(Operator formulation)}\\
\Leftrightarrow\ &\text{Find } u \in V: && \tau Au = \tau l && (\tau \text{ is bijective}).
\end{aligned}$$

Now the last equation is solved using the Banach fixed point theorem. To that end consider the map $T : V \to V$ given as

$$Tv = v - \rho(\tau Av - \tau l), \qquad \rho \in \mathbb{R}, \rho \neq 0.$$

If T is a contraction, i.e. $\|Tv_1 - Tv_2\|_V \leq q\|v_1 - v_2\|_V$ with $0 \leq q < 1$, then the fixed point theorem states that there exists a unique fixed point $u \in V$ such that $Tu = u$, i.e. $\tau Au - \tau l = 0$ since $\rho \neq 0$.

We now show that ρ can always be chosen such that T is a contraction.

$$\begin{aligned}
\|Tv_1 - Tv_2\|_V^2 &= \|v_1 - \rho(\tau Av_1 - \tau l) - (v_2 - \rho(\tau Av_2 - \tau l))\|_V^2 && \text{(Def. of } T)\\
&= \|v_1 - v_2 - \rho(\tau Av_1 - \tau Av_2)\|_V^2\\
&= \|v - \rho\tau Av\|_V^2 && (v = v_1 - v_2)\\
&= (v - \rho\tau Av, v - \rho\tau Av)_V\\
&= \|v\|_V^2 - 2\rho(v, \tau Av)_V + \rho^2(\tau Av, \tau Av)_V\\
&= \|v\|_V^2 - 2\rho\langle Av, v\rangle_{V',V} + \rho^2\langle Av, \tau Av\rangle_{V',V} && \text{(Def. of } \tau)\\
&= \|v\|_V^2 - 2\rho a(v,v) + \rho^2 a(v, \tau Av) && \text{(Def. of } A)\\
&\leq \|v\|_V^2 - 2\rho\alpha\|v\|_V^2 + \rho^2\|a\|_{U,V}\|v\|_V\|\tau Av\|_V && \text{(coerc., cont.)}\\
&\leq (1 - 2\rho\alpha + \rho^2\|a\|_{U,V}^2)\|v\|_V^2 && (\tau \text{ isom.})\\
&= q^2\|v_1 - v_2\|_V^2
\end{aligned}$$

Now it remains to show that ρ can always be chosen such that $q < 1$. From

$$q^2 = 1 - 2\rho\alpha + \rho^2\|a\|_{U,V}^2 = 1 - \rho(2\alpha - \rho\|a\|_{U,V}^2) < 1 \;\Leftrightarrow\; \rho(\rho\|a\|_{U,V}^2 - 2\alpha) < 0$$

we conclude that

$$\rho \in \left(0, \frac{2\alpha}{\|a\|_{U,V}^2}\right) \quad \Rightarrow \quad q^2 < 1.$$

The stability estimate follows from coercivity:

$$\alpha\|u\|_V \leq \frac{a(u,u)}{\|u\|_V} = \frac{\langle l, u\rangle_{V',V}}{\|u\|_V} \leq \sup_{v\in V} \frac{\langle l, v\rangle_{V',V}}{\|v\|_V} = \|l\|_{V'}.$$

We now turn to the special case when the bilinear form is symmetric. Then the variational problem is equivalent to a minimization problem as is shown in the following theorem.

Theorem 5.23 (Characterization Theorem)**.** Let V be a normed vector space (completeness is not necessary), a a *symmetric* and coercive bilinear form and l a linear functional. Then the following two assertions are equivalent:

i) u is a minimizer of the functional $J(v) = \frac{1}{2}a(v,v) - l(v)$.

ii) u solves the variational problem $a(u,v) = l(v)\ \forall v \in V$.

Moreover, if the minimizer exists then it is unique.
Proof. [Braess, 2003, Satz 2.2]. For arbitrary $u, v \in V$ and $t \in \mathbb{R}$ we have

$$\begin{aligned} J(u+tv) &= \frac{1}{2}a(u+tv, u+tv) - l(u+tv) \\ &= \frac{1}{2}(a(u,u) + 2ta(u,v) + t^2a(v,v)) - l(u) - tl(v) \\ &= J(u) + t[a(u,v) - l(v)] + \frac{t^2}{2}a(v,v). \end{aligned}$$

i) $\Rightarrow$ *ii)*. Let $u \in V$ be a minimizer of $J(v)$ and set $\phi_v(t) = J(u+tv)$. Since u is a minimizer we have $\frac{d\phi_v}{dt}(0) = 0$ for all $v \in V$. This implies

$$\frac{d\phi_v}{dt}(t)\bigg|_{t=0} = [a(u,v) - l(v)] + ta(v,v)\bigg|_{t=0} = a(u,v) - l(v) = 0 \qquad \forall v \in V.$$

ii) $\Rightarrow$ *i)*. Now $a(u,v) = l(v)\ \forall v \in V$. This implies (set $t = 1$ above) for any $v \neq 0$

$$J(u+v) = J(u) + \frac{1}{2}a(v,v) > J(u)$$

since a is coercive and $v \neq 0$. This shows that u is a minimizer of $J(v)$. The last argument also shows that the minimizer is unique when it exists. Suppose that $u_1 \neq u_2$ are both minimizers, then we have $J(u_1) = J(u_2 + (u_1 - u_2)) = J(u_2) + \frac{1}{2}a(u_1 - u_2, u_1 - u_2) > J(u_2)$ which is a contradiction. □

Note, that the previous theorem does not claim that a minimizer always exists. It only states that both formulations are equivalent when a minimizer exists.

5.6. Lebesgue Spaces

We now turn to the question: What are the appropriate function spaces to be used for solving the variational problem (5.2) with either the Banach-Nečas-Babuška Theorem or the Lax-Milgram Theorem? Both theorems require establishing certain properties of the bilinear form a which in the case of a scalar elliptic boundary value problem is given by $a(u, v) = \int_\Omega (K\nabla u) \cdot \nabla v \, dx$. The function spaces in the theorems need to be either Banach spaces or Hilbert spaces. Assuming homogeneous Dirichlet boundary conditions a candidate for a suitable function space would be $V = \{v \in C^1(\overline{\Omega}) \, : \, v|_{\partial\Omega} = 0\}$. Equipped with the norm $\|v\|_V = \max_{|\alpha|\leq 1} \sup_{x\in\Omega} |\partial^\alpha v(x)|$ V is a Banach space provided Ω is bounded.

The Lax-Milgram Theorem requires V to be a Hilbert space. This is not the case and therefore the theorem is not applicable.

So let us turn to the Banach-Nečas-Babuška Theorem. A first observation is that functions of the form $u(x) = \|x - y\|^\beta$ with $0 < \beta < 1$ and $y \in \partial\Omega$ are not in $(V, \|.\|_V)$ since the derivative is unbounded. However, functions of this form do occur as solutions of the Poisson problem in domains with reentrant corners as has been illustrated in the example in Figure 3.2.

Accepting that the theory can not cover some interesting cases one could still try to verify the inf-sup condition. To make things more simple consider the one-dimensional case $\Omega = (-1/2, 1/2)$ and the two-point boundary value problem

$$-\frac{d^2u}{dx^2} = f \text{ in } \Omega, \qquad \text{u(-1/2)=u(1/2)=0.}$$

a) In this case we have $\|u\|_V = \max(\sup_{x\in\Omega} |u(x)|, \sup_{x\in\Omega} |u'(x)|)$. For any $x \in \Omega$ we have

$$u(x) = u(x) - u(-1/2) = \int_{-1/2}^{x} u'(\xi) \, d\xi \leq \sup_{\xi\in\Omega} |u'(\xi)| \int_{-1/2}^{x} 1 \, dx \leq \sup_{\xi\in\Omega} |u'(\xi)|$$

and therefore $\sup_{x\in\Omega} |u(x)| \leq \sup_{x\in\Omega} |u'(x)|$ and $\|u\|_V = \sup_{x\in\Omega} |u'(x)|$.

Note that $\sup_{x\in\Omega}|u'(x)|$ is a norm on V due to the homogeneous Dirichlet boundary conditions.

b) For any $u \in V$ consider now

$$\sup_{v\in V}\frac{a(u,v)}{\|v\|_V} = \sup_{v\in V,\|v\|_V=1}\int_{-1/2}^{1/2} u'v'\,dx.$$

The integral is maximized for the function v_* defined by

$$v_*'(x) = \begin{cases} 1 & u'(x) \geq 0 \\ -1 & u'(x) < 0 \end{cases}.$$

v_* (obviously) can have a discontinuous derivative but it can be approximated arbitrarily close with functions from V which have the property $\|v\|_V = 1$. Thus we can conclude

$$\sup_{v\in V}\frac{a(u,v)}{\|v\|_V} = \int_{-1/2}^{1/2}|u'|\,dx.$$

c) Now finally the inf-sup condition reads

$$\inf_{u\in V}\sup_{v\in V}\frac{a(u,v)}{\|u\|_V\|v\|_V} = \inf_{u\in V,\|u\|_V=1}\int_{-1/2}^{1/2}|u'|\,dx.$$

Consider the sequence of functions $u_k(x) = (1-(2x)^{2k})/(4k)$ with $k \in \mathbb{N}$. On $\Omega = (-1/2, 1/2)$ we have $\|u_k\|_V = 1$ and we get for the integral $\int_{-1/2}^{1/2}|u_k'|\,dx = 2\int_0^{1/2}(2x)^{2k-1}\,dx = 2\left[(2x)^{2k}/(4k)\right]_0^{1/2} = 1/(2k)$. So the inf-sup condition does *not* hold.

These considerations show that the classical Banach spaces $C^k(\Omega)$ equipped with the supremum norm are not suitable in connection with the arising bilinear forms containing integrals of function values and derivatives.

We are instead interested in Banach and Hilbert spaces where the norm is defined via an integral over the function. A problem of the classical Riemann integral is that a sequence of integrable functions may not converge to a limit function that is again integrable as is shown by the following example.

Example 5.24. [Brenner and Scott, 1994, §1.1] The function $\log(x)$ has an improper integral on the interval $[0,1]$. Now let $\{r_n : n \in \mathbb{N}_0\}$ be a set of

numbers that are dense in $[0,1]$, e.g. $\mathbb{Q} \cap [0,1]$. Then define the sequence of functions $f_k(x) = \sum_{n=0}^{k} 2^{-n} \log|x - r_n|$. Every f_k is Riemann-integrable and we have

$$\left| \int_0^1 f_k(x)\,dx \right| \leq 2 \int_0^1 |\log x|\,dx.$$

Since $\lim_{k\to\infty} \left|\int_0^1 f_k(x)\,dx\right|$ exists we would like to conclude that also the integral of the limit function $f(x) = \sum_{n=0}^{\infty} 2^{-n} \log|x - r_n|$ exists. This, however, is not the case: on any interval (a,b) the function f is infinite at some point and therefore the Riemann integral is not defined. □

As a consequence the Riemann integral needs to be replaced by a more general notion of integral given by Lebesgue. In the following $\int_\Omega f(x)\,dx$ denotes the Lebesgue integral of the function f (the symbol is the same because the Lebesgue integral coincides with the Riemann integral if it exists).

Definition 5.25. With the Lebesgue integral being defined consider the scalar product

$$(u,v)_{0,\Omega} = \int_\Omega u(x)v(x)\,dx \tag{5.5}$$

and the corresponding norm

$$\|u\|_{0,\Omega} = \left(\int_\Omega u^2(x)\,dx \right)^{\frac{1}{2}}. \tag{5.6}$$

We may omit the subscript Ω if the domain is not ambiguous. The normed linear space $L^2(\Omega)$ then is the completion of $C^0(\overline{\Omega})$ with respect to the norm $\|.\|_{0,\Omega}$. □

For every $v \in L^2(\Omega)$ therefore exists a sequence $(v_k)_{k\in\mathbb{N}} \subset C^0(\overline{\Omega})$ which is a Cauchy sequence w.r.t. $\|.\|_{0,\Omega}$ and which converges to v "almost everywhere" in the sense of Lebesgue. In $L^2(\Omega)$ two functions f and g are identified if they differ at most on a set of zero measure. In one space dimension, e.g., a set of measure zero can consist of countably infinitely many points. In 2D even lines and in 3D surfaces are sets of measure zero.

Definition 5.26. For $1 \leq p < \infty$ set

$$\|u\|_{L^p(\Omega)} = \left(\int_\Omega |u(x)|^p\,dx \right)^{\frac{1}{p}}.$$

The normed linear space $L^p(\Omega)$ is obtained as the completion of $C^0(\overline{\Omega})$ with respect to the norm $\|.\|_{L^p(\Omega)}$. For $p = \infty$ we call $L^\infty(\Omega)$ the space of Lebesgue measurable and essentially bounded functions equipped with the norm

$$\|u\|_{L^\infty(\Omega)} = \operatorname{ess\,sup}_{x\in\Omega} |u(x)| = \inf\{M \geq 0 \,:\, |u(x)| \leq M \text{ almost everywhere}\}.$$

□

Proposition 5.27. For $1 \leq p \leq \infty$, $L^p(\Omega)$ is a Banach space. For $p = 2$, $L^2(\Omega)$ is a Hilbert space.
Proof. A proof can be found in [Adams, 1978]. □

Alternatively, the space L^2 can be defined via the completion of the set of infinitely differentiable functions with compact support:

$$C_0^\infty(\Omega) = \left\{ v \in C^\infty(\Omega) \,:\, \operatorname{supp}(v) = \overline{\{x \in \Omega : v(x) \neq 0\}} \subset \Omega \right\} \tag{5.7}$$

as is stated by the following theorem:

Proposition 5.28. The set $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$. See [Hackbusch, 1986, Lemma 6.2.2]. □

Example 5.29. In this example we consider when functions of the form $f(x) = \|x\|^\alpha$, $\alpha < 0$ are in L^2. These functions do occur as solution of the Poisson equation in domains with reentrant corners.

a) $n = 1$, i.e. $\Omega = (0, 1)$. In order to show that $f(x) = x^\alpha \in L^2(\Omega)$ we have to show that f is the limit of a Cauchy sequence in $C^0(\overline{\Omega})$. Therefore consider the functions $f_k \in C^0(\overline{\Omega})$, $k \in \mathbb{N}$:

$$f_k(x) = \begin{cases} x^\alpha & \frac{1}{k} < x \leq 1 \\ \left(\frac{1}{k}\right)^\alpha & 0 \leq x \leq \frac{1}{k} \end{cases} .$$

For $m < n$ we have $1/n < 1/m$ and we obtain

$$\begin{aligned} \|f_n - f_m\|_0^2 = \int_0^1 (f_n - f_m)^2 \, dx &\leq \int_0^{1/m} x^{2\alpha} \, dx + \left(\frac{1}{m}\right)^{2\alpha+1} \\ &= \left[\frac{x^{2\alpha+1}}{2\alpha+1}\right]_0^{1/m} + \left(\frac{1}{m}\right)^{2\alpha+1} . \end{aligned}$$

The improper integral exists for $2\alpha + 1 > 0$ i.e. $\alpha > -1/2$, and we get

$$\|v_n - v_m\|_0^2 \leq \frac{2\alpha+2}{2\alpha+1} \left(\frac{1}{m}\right)^{2\alpha+1} \qquad \to 0 \text{ for } m \to \infty,$$

i.e. (v_k) is a Cauchy sequence and hence x^α in $L^2(\Omega)$ provided $\alpha > -1/2$. The estimate is sharp since for fixed m and $n \to \infty$ the estimate is arbitrarily close. Moreover, we observe that in this example the existence of the improper integral is equivalent to the convergence of the Cauchy sequence.

b) $n = 2$, $\Omega = \{x \in \mathbb{R}^2 : \|x\| < 1\}$. Transform the integral $\int_\Omega \|x\|^{2\alpha}\, dx$ to polar coordinates through $\mu : [0,1) \times [-\pi, \pi) \to \Omega$ given by

$$\mu(r,\phi) = \begin{pmatrix} r\cos\phi \\ r\sin\phi \end{pmatrix}, \qquad \nabla\mu(r,\phi) = \begin{pmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{pmatrix} \tag{5.8}$$

and $|\det \nabla\mu(r,\phi)| = r$. Transforming the integral then yields

$$\int_\Omega \|x\|^{2\alpha}\, dx = \int_{-\pi}^{\pi}\int_0^1 \|\mu(r,\phi\|^{2\alpha} r\, dr d\phi = \int_{-\pi}^{\pi}\int_0^1 r^{2\alpha+1}\, dr d\phi = 2\pi\left[\frac{r^{2\alpha+2}}{2\alpha+2}\right]_0^1.$$

This integral exists for $2\alpha + 2 > 0$, i.e. $\alpha > -1$.

c) $n = 3$, $\Omega = \{x \in \mathbb{R}^3 : \|x\| < 1\}$. The Transformation to spherical coordinates $\mu : [0,1) \times [-\pi,\pi)^2 \to \Omega$ is now given by

$$\begin{aligned} \mu(r,\phi,\theta) &= \begin{pmatrix} r\cos\phi\cos\theta \\ r\sin\phi\cos\theta \\ r\sin\theta \end{pmatrix}, \\ \nabla\mu(r,\phi) &= \begin{pmatrix} \cos\phi\cos\theta & -r\sin\phi\cos\theta & -r\cos\phi\sin\theta \\ \sin\phi\cos\theta & r\cos\phi\cos\theta & -r\sin\phi\sin\theta \\ \sin\theta & 0 & r\cos\theta \end{pmatrix} \end{aligned} \tag{5.9}$$

and $|\det \nabla\mu(r,\phi,\theta)| = r^2|\cos\theta|$ resulting in the transformed integral

$$\int_\Omega \|x\|^{2\alpha}\, dx = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_0^1 r^{2\alpha} r^2 |cos\theta|\, dr d\phi d\theta$$

which is finite for $2\alpha + 3 > 0$, i.e. $\alpha > -3/2$.

d) One can show that in n dimensions the singularity function $\|x\|^\alpha$ is in L^2 provided $\alpha > -n/2$. □

5.7. Sobolev Spaces

Our applications require derivatives of functions. Derivatives of L^2-functions are defined as follows.

Definition 5.30. A function $f \in L^2(\Omega)$ has a *weak partial derivative* $g \in L^2(\Omega)$ with respect to x_i if

$$(g, \phi)_{0,\Omega} = -(f, \partial_{x_i}\phi)_{0,\Omega} \qquad \forall \phi \in C_0^\infty(\Omega).$$

As in the classical case we write $g = \partial_{x_i} f$. It can be shown that if f has a classical derivative then it is also a weak derivative. Moreover, the definition can be iterated to define derivatives $\partial^\alpha f$ of arbitrary order $k = |\alpha|$. □

Example 5.31. [Brenner and Scott, 1994, 1.2.5] Consider the function $f(x) = 1 - |x|$ on $\Omega = (-1, 1)$. f has no classical derivative at $x = 0$. We show that f has the weak derivative

$$g(x) = \begin{cases} 1 & x < 0 \\ -1 & x > 0 \end{cases}.$$

Since $g \in L^2(\Omega)$ the value at $x = 0$ is not relevant.

In order to show that g is the weak derivative of f we need to verify that

$$\int_{-1}^{1} g(x)\phi(x)\,dx = -\int_{-1}^{1} f(x)\phi'(x)\,dx \qquad \phi \in C_0^\infty((-1, 1)).$$

So,

$$\begin{aligned}
\int_{-1}^{1} f(x)\phi'(x)\,dx &= \int_{-1}^{0} f(x)\phi'(x)\,dx + \int_{0}^{1} f(x)\phi'(x)\,dx \\
&= -\int_{-1}^{0} (+1)\phi(x)\,dx + [f\phi]_{-1}^{0} - \int_{0}^{1} (-1)\phi(x)\,dx + [f\phi]_{0}^{1} \\
&= -\int_{-1}^{1} g(x)\phi(x)\,dx + \underbrace{(f\phi)(0-) - (f\phi)(0+)}_{=0 \text{ since } f,\phi \text{ continuous}}
\end{aligned}$$

Which is the result. Note that the continuity of f is crucial and that the argument would fail for f discontinuous at $x = 0$. □

Definition 5.32. All functions $v \in L^2(\Omega)$ with weak square integrable derivatives up to order 1 form the function space $H^1(\Omega)$ equipped with the scalar product

$$(u, v)_{1,\Omega} = \int_\Omega uv + \nabla u \cdot \nabla v\,dx \tag{5.10}$$

and the norm

$$\|u\|_{1,\Omega} = \sqrt{(u,u)_{1,\Omega}} = \left(\int_\Omega u^2 + \|\nabla u\|^2 \, dx \right)^{\frac{1}{2}}. \tag{5.11}$$

$H^1(\Omega) \subset L^2(\Omega)$ is called *Sobolev space of order 1* and the symbol H is chosen in honor of David Hilbert. Alternatively, $H^1(\Omega)$ can also be defined as the completion of $C^1(\overline{\Omega})$ with respect to the norm $\|.\|_{1,\Omega}$. □

Definition 5.33. The completion of $C_0^\infty(\Omega)$ with respect to the norm $\|.\|_{1,\Omega}$ is called $H_0^1(\Omega)$. Functions in $H_0^1(\Omega)$ are zero on $\partial\Omega$ "almost everywhere" and $H_0^1(\Omega)$ is a proper subspace of $H^1(\Omega)$. □

Proposition 5.34. $H^1(\Omega)$ and $H_0^1(\Omega)$ are Hilbert spaces. □

Example 5.35. In this example we consider singularity functions in H^1. From example 5.29 we learned that it suffices in this case to check the existence of the improper integral.

a) Let $\Omega = \{x \in \mathbb{R}^n : \|x\| < 1\}$ for $n \geq 1$ and set $f(x) = \|x\|^\alpha$. Then

$$\partial_{x_i} f(x) = \partial_{x_i} \|x\|^\alpha = \partial_{x_i} \left(\sum_{i=1}^n x_i^2 \right)^{\frac{\alpha}{2}} = \alpha \left(\sum_{i=1}^n x_i^2 \right)^{\frac{\alpha}{2}-1} = \alpha x_i \|x\|^{\alpha-2}$$

and therefore $\nabla f(x) = \alpha x \|x\|^{\alpha-2}$. For the H^1-norm we get by transformation to (generalized) polar coordinates in n dimensions:

$$\begin{aligned}
\|f\|_{1,\Omega}^2 &= \int_\Omega \|x\|^{2\alpha} + \alpha^2 \|x\|^{2\alpha-4} \|x\|^2 \, dx \\
&= \int_{-\pi}^{\pi} \dots \int_{-\pi}^{\pi} \int_0^1 (r^{2\alpha} + \alpha^2 r^{2\alpha-2}) r^{n-1} \, dr d\theta_1 \dots d\theta_{n-1} \\
&= \int_{-\pi}^{\pi} \dots \int_{-\pi}^{\pi} \int_0^1 r^{2\alpha+n-1} + \alpha^2 r^{2\alpha+n-3} \, dr d\theta_1 \dots d\theta_{n-1}.
\end{aligned}$$

This integral exists if $2\alpha + n - 3 > -1$, i.e.

$$\alpha > 1 - n/2 = \begin{cases} \frac{1}{2} & n = 1 \\ 0 & n = 2 \\ -\frac{1}{2} & n = 3 \end{cases}$$

b) Now consider $f(x) = \ln \|x\|$. Then

$$\partial_{x_i} f(x) = \partial_{x_i} \ln \|x\| = \partial_{x_i} \ln \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}} = \frac{1}{\|x\|} \left(\sum_{i=1}^{n} x_i^2 \right)^{-\frac{1}{2}} x_i = \frac{x_i}{\|x\|^2}$$

and therefore $\nabla f(x) = x\|x\|^{-2}$ (note that this formally coincides with $\alpha = 0$ in case a).

We then get

$$\begin{aligned} \|f\|_{1,\Omega}^2 &= \int_\Omega (\ln \|x\|)^2 + \|x\|^{-4} x \cdot x \, dx = \int_\Omega (\ln \|x\|)^2 + \|x\|^{-2} \, dx \\ &= \int_{-\pi}^{\pi} \dots \int_{-\pi}^{\pi} \int_0^1 ((\ln r)^2 + r^{-2}) r^{n-1} \, dr d\theta_1 \dots d\theta_{n-1} \\ &= \int_{-\pi}^{\pi} \dots \int_{-\pi}^{\pi} \int_0^1 (\ln r)^2 r^{n-1} + r^{n-1} \, dr d\theta_1 \dots d\theta_{n-1}. \end{aligned}$$

The integral of the first term exists for $n > 1$ while the second term requires $n - 3 > -1$, i.e. $n > 2$.

It turns out that H^1-functions in $n = 1$ do not have singularities ($\alpha > 1/2$) and are always continuous (Sobolov embedding theorem below). This is not the case for $n = 2$ where one can give examples of singular functions in H^1, see [Rannacher, 2006, Beispiel 1.1]. We observe that the functions $\ln \|x\|$ for $n = 2$ as well as $\frac{1}{\|x\|}$ for $n = 3$, which are classical solutions of the Poisson equation for point sources (Dirac delta function on the right hand side) are *not* in $H^1(\Omega)$. □

Sobolev spaces are extended to arbitrary order in the following definition.

Definition 5.36. For $k \geq 1$ the completion of $C^k(\overline{\Omega})$ with respect to the norm

$$\|u\|_{k,\Omega} = \sqrt{(u,u)_{k,\Omega}} \tag{5.12}$$

induced by the scalar product

$$(u,v)_{k,\Omega} = \sum_{0 \leq |\alpha| \leq k} \int_\Omega (\partial^\alpha u)(\partial^\alpha v) \, dx \tag{5.13}$$

is called *Sobolev space of order k* and is denoted by $H^k(\Omega)$. The completion of $C_0^\infty(\Omega)$ with respect to the norm $\|.\|_{k,\Omega}$ is denoted by $H_0^k(\Omega)$.

$$|u|_{k,\Omega} = \left(\sum_{|\alpha|=k} \int_\Omega (\partial^\alpha u)^2 \right)^{\frac{1}{2}}$$

denotes the so-called H^k-*seminorm*. □

The spaces $H^k(\Omega)$ and $H_0^k(\Omega)$ are Hilbert spaces and we have the following inclusions:

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

Example 5.37. With respect to singularity functions of the form $f(x) = \|x\|^\alpha$ in $H^k(\Omega)$, $\Omega = \{x \in \mathbb{R}^n : \|x\| < 1\}$, one should verify that $f \in H^k(\Omega)$ provided $\alpha > k - n/2$. □

Remark 5.38. There are additional function spaces of importance in numerical analysis which we do not handle in detail:

a) $H^s(\Omega)$, $s \in \mathbb{R}^+$, are Sobolev spaces of real order which can be defined e.g. via Fourier transformation, see [Hackbusch, 1986]. For $s \in \mathbb{N}$ they coincide with the spaces introduced above. One property of these spaces is that they include precisely functions $\|x\|^\alpha$ provided $\alpha > s - n/2$.

b) $H^{-s}(\Omega) = (H^s(\Omega))' = \mathcal{L}(H^s(\Omega); \mathbb{R})$ denotes the dual space of $H^s(\Omega)$ for any $s \in \mathbb{R}^+$.

c) $W_p^k(\Omega)$, $1 \le p \le \infty$ are the Sobolev spaces based on L^p equipped with the norm

$$\|v\|_{W_p^k(\Omega)} = \left(\sum_{0 \le |\alpha| \le k} \|\partial^\alpha v\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}} \qquad 1 \le p < \infty$$

$$\|v\|_{W_p^\infty(\Omega)} = \max_{0 \le |\alpha| \le k} \|\partial^\alpha v\|_{L^\infty(\Omega)} \qquad p = \infty$$

□

5.8. Properties of Sobolev Spaces

Proposition 5.39 (Friedrich inequality I)**.** Let Ω be enclosed in a n-dimensional cube with side length s. Then the following inequality holds:

$$\|v\|_{0,\Omega} \le s|v|_{1,\Omega} \qquad \forall v \in H_0^1(\Omega). \tag{5.14}$$

Proof. Since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$ it suffices to show the result for C_0^∞-functions. The completeness of the function spaces then ensures the result for all $v \in H_0^1(\Omega)$. The main ingredient is the central theorem of calculus in connection

with the fact that C_0^∞-functions are zero at the boundary (and can be extended by zero to all of $\mathbb{R}^n$):

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

Applying the Cauchy-Schwarz inequality then gives

$$\begin{aligned} |v(x)|^2 &= \left| \int_0^{x_1} 1 \, \partial_1 v(t, x_2, \ldots, x_n)\, dt \right|^2 \\ &\leq \int_0^{x_1} 1^2\, dt \int_0^{x_1} |\partial_1 v(t, x_2, \ldots, x_n)|^2\, dt \\ &\leq s \int_0^{s} |\partial_1 v(t, x_2, \ldots, x_n)|^2\, dt. \end{aligned}$$

Note that the last integral is independent of x_1. Using this results we obtain

$$\begin{aligned} \|v(x)\|_{0,\Omega}^2 &= \int_\Omega |v(x)|^2\, dx = \int_{x_n=0}^{s} \ldots \int_{x_1=0}^{s} |v(x)|^2\, dx_1 \ldots dx_n \\ &\leq \int_{x_n=0}^{s} \ldots \int_{x_1=0}^{s} s \int_{t=0}^{s} |\partial_1 v(t, x_2, \ldots, x_n)|^2\, dt\, dx_1 \ldots dx_n \\ &= s \int_{x_n=0}^{s} \ldots \int_{x_2=0}^{s} \int_{t=0}^{s} |\partial_1 v(t, x_2, \ldots, x_n)|^2\, dt \int_{x_1=0}^{s} 1\, dx_1\; dx_2 \ldots dx_n \\ &= s^2 \int_\Omega |\partial_1 v(x)|^2\, dx \\ &\leq s^2 |v|_{1,\Omega}^2. \end{aligned}$$

In the last step the remaining terms $\int_\Omega \sum_{i=2}^n |\partial_i v(x)|^2\, dx$ have been added. □

The proof of the Friedrich inequality shows that the zero boundary conditions are not needed on the whole boundary. In fact more general results are given below.

Definition 5.40. We state two properties on the regularity of a domain Ω:

1) $\Omega \subset \mathbb{R}^n$, bounded, is a *Lipschitz domain* if for every point $x_0 \in \partial\Omega$ there exists $\epsilon > 0$ and a map $\mu_{x_0} : B_\epsilon(x_0) \to B_1(0)$ such that
   i) μ_{x_0} is bijective and both μ_{x_0} and $\mu_{x_0}^{-1}$ are Lipschitz-continuous,
   ii) $\mu_{x_0}(\partial\Omega) = \{(x_1, \ldots, x_{n-1}, x_n) \in B_1(0) \,:\, x_n = 0\}$and
   iii) $\mu_{x_0}(\Omega \cap B_\epsilon(x_0)) = \{(x_1, \ldots, x_{n-1}, x_n) \in B_1(0) \,:\, x_n > 0\}$.

2) Ω satisfies a *cone condition*, if a cone of finite size and with a finite opening angle can be positioned at any point on the boundary $\partial\Omega$ such that the cone is completely inside Ω. □

Proposition 5.41 (Friedrich inequality II)**.** Let Ω be a bounded Lipschitz domain and let $\Gamma \subseteq \partial\Omega$ be part of the boundary with non-vanishing $(n-1)$-dimensional measure. Then there exist constants $c_1, c_2 > 0$ only depending on Ω and Γ such that

$$\|v\|_{0,\Omega}^2 \leq c_1 |v|_{1,\Omega}^2 + c_2 \|v\|_{0,\Gamma}^2 \qquad \forall v \in H^1(\Omega). \tag{5.15}$$

See [Toselli and Widlund, 2005, Lemma A.14]. □

This second version of the Friedrich inequality bounds the L^2-norm on the whole domain by the derivatives within the domain and the L^2-norm on part of the boundary. That the L_2-norm of an H^1-function on part of the boundary does make sense is not obvious and is established by the trace theorem given below.

Instead of the values at the boundary one can also use the average of the function.

Proposition 5.42 (Poincaré inequality)**.** Let Ω be a bounded Lipschitz domain, then there exist constants $c_1, c_2 > 0$ only depending on Ω such that

$$\|v\|_{0,\Omega}^2 \leq c_1 |v|_{1,\Omega}^2 + c_2 \left(\int_\Omega v \, dx \right)^2 \qquad \forall v \in H^1(\Omega). \tag{5.16}$$

See [Toselli and Widlund, 2005, Lemma A.13]. □

Remark 5.43. The naming of the inequalities given above is ambiguous. Often, Proposition 5.39 is called Poincaré-Friedrich inequality, e.g. in [Braess, 2003]. Inequalities bounding the L_2-norm of a function by its derivatives are referred to as Poincaré type inequalities. □

As an application of the Friedrich and Poincaré inequalities we prove

Corollary 5.44. In $H_0^1(\Omega)$ and $\bar{H}(\Omega) = \{v \in H^1(\Omega) : \int_\Omega v\,dx = 0\}$ the semi-norm $|.|_{1,\Omega}$ is a norm which is equivalent to $\|.\|_{1,\Omega}$.
Proof. From either Proposition 5.41 or Proposition 5.42 we have $\|v\|_{0,\Omega}^2 \leq s^2|v|_{1,\Omega}^2$ and therefore $\|v\|_{1,\Omega}^2 = \|v\|_{0,\Omega}^2 + |v|_{1,\Omega}^2 \leq (1+s^2)|v|_{1,\Omega}^2$. Trivially, we have $|v|_{1,\Omega}^2 \leq \|v\|_{0,\Omega}^2 + |v|_{1,\Omega}^2 = \|v\|_{1,\Omega}^2$ and therefore

$$\frac{1}{\sqrt{1+s^2}}\|v\|_{1,\Omega} \;\leq\; |v|_{1,\Omega} \;\leq\; \|v\|_{1,\Omega} \tag{5.17}$$

which was to be shown. □

Theorem 5.45 (Trace theorem)**.** Let Ω be bounded, have piecewise smooth boundary and satisfy a cone condition. Then there exists a continuous linear map

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

called the "trace operator", with the following properties:

i) $(\gamma v)(x) = v(x)$ almost everywhere on $\partial\Omega$.

ii) $\|\gamma v\|_{0,\partial\Omega} \leq c\|v\|_{1,\Omega}$.

Proof. See [Braess, 2003, Satz 3.1] □

The trace theorem ensures that evaluation of H^1 functions on the boundary makes sense. The following theorem shows that H^k functions are continuous and bounded if k is large enough.

Theorem 5.46 (Embedding theorem)**.** Let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$. For $k > n/2$ we have that $H^k(\Omega) \subset C^0(\overline{\Omega})$ and, moreover, the embedding is continuous, i.e. there exists a constant c such that $\|v\|_{k,\Omega} \leq c\sup_{x\in\overline{\Omega}}|u|$.
Proof. See[Adams, 1978, Lemma 5.17, p. 108] □

A consequence of Theorem 5.46 is that pointwise evaluation of H^1 functions is well defined in one space dimension and pointwise evaluations of H^2 functions is well defined for $n = 1, 2, 3$.

Chapter 6.

Well-posedness of Scalar Elliptic Partial Differential Equations

With the theory of the last chapter we are now in a position to prove the well-posedness of the weak fromulation of linear scalar elliptic partial differential equations with a variety of boundary conditions

6.1. Dirichlet Problem

We begin with the Dirichlet problem

$$-\nabla \cdot (K\nabla u) = f \qquad \text{in } \Omega, \tag{6.1a}$$
$$u = g \qquad \text{on } \partial\Omega. \tag{6.1b}$$

According to Section 5.1 the problem has the corresponding weak formulation

$$\text{Find } u \in U: \qquad a(u,v) = l(v) \quad \forall v \in V \tag{6.2}$$

with

$$a(u,v) = \int_\Omega (K\nabla u)\cdot \nabla v \, dx, \qquad l(v) = \int_\Omega fv \, dx. \tag{6.3}$$

Definition 6.1. Let $\Omega \subset \mathbb{R}^n$. Problem (6.1) and the matrix $K(x)$ are called *uniformly elliptic* if there exists a constant $k_0 \in \mathbb{R}$, $k_0 > 0$ such that

$$\xi^T K(x)\xi \geq k_0 \|\xi\|^2 \qquad \forall x \in \Omega, \forall \xi \in \mathbb{R}^n,$$

Since $K(x)$ is symmetric positive definite, k_0 can be chosen as the infimum over the smallest eigenvalue of $K(x)$ in Ω. □

Proposition 6.2. Assume that

i) problem (6.1) is uniformly elliptic

ii) the coefficient matrix $K(x)$ is bounded: $\forall x \in \Omega, 1 \leq i,j \leq n : |k_{ij}(x)| \leq M$,

iii) $f \in L^2(\Omega)$ and

iv) there exists $u_g \in H^1(\Omega)$ such that $\gamma u_g = g$ (for this to hold one requires $u \in H^s(\partial\Omega)$ with $s \geq 1/2$).

Then the problem

$$\text{Find } u \in u_g + H_0^1(\Omega): \qquad a(u,v) = l(v) \qquad \forall v \in H_0^1(\Omega)$$

with a, l from (6.3) has a unique solution.

Proof. Let us first reformulate the problem. Using assumption iv) we can write $u = u_g + u_0$ with $u_0 \in H_0^1(\Omega)$. Then, due to linearity, we have $a(u,v) = a(u_g + u_0, v) = a(u_g, v) + a(u_0, v)$ and the problem reads

$$\text{Find } u_0 \in H_0^1(\Omega): \qquad a(u_0,v) = l(v) - a(u_g,v) = l_0(v) \qquad \forall v \in H_0^1(\Omega).$$

Now the Lax-Milgram Theorem can be applied with $V = H_0^1(\Omega)$ after its assumptions have been verified.

Continuity of a.

$$\begin{aligned}
|a(u,v)| &= \left| \sum_{i,j=1}^{n} \int_\Omega k_{ij} \partial_j u \partial_i v \, dx \right| \leq \sum_{i,j=1}^{n} \int_\Omega |k_{ij}| \, |\partial_j u| \, |\partial_i v| \, dx && \text{(tria. ineq.)} \\
&\leq M \sum_{i,j=1}^{n} \int_\Omega |\partial_j u| \, |\partial_i v| \, dx && (K \text{ bounded}) \\
&\leq M \sum_{i,j=1}^{n} \left(\int_\Omega |\partial_j u|^2 \, dx \int_\Omega |\partial_i v|^2 \, dx \right)^{\frac{1}{2}} && \text{(C.S.)} \\
&= M \sum_{j=1}^{n} \left(\int_\Omega |\partial_j u|^2 \, dx \right)^{\frac{1}{2}} \sum_{i=1}^{n} \left(\int_\Omega |\partial_i v|^2 \, dx \right)^{\frac{1}{2}} && \text{(reorganize)} \\
&\leq M \left(\sum_{j=1}^{n} \int_\Omega |\partial_j u|^2 \, dx \right)^{\frac{1}{2}} n^{\frac{1}{2}} \left(\sum_{i=1}^{n} \int_\Omega |\partial_i v|^2 \, dx \right)^{\frac{1}{2}} n^{\frac{1}{2}} && \text{(C.S. in } \mathbb{R}^n) \\
&= Mn |u|_{1,\Omega} |v|_{1,\Omega} \leq Mn \|u\|_{1,\Omega} \|v\|_{1,\Omega}.
\end{aligned}$$

Coercivity of a. From the uniform ellipticity we conclude

$$(K(x)\nabla v(x)) \cdot \nabla v(x) \geq k_0 \|\nabla v(x)\|^2 \qquad \forall x \in \Omega$$

and therefore

$$k_0 |v|_{1,\Omega}^2 = \int_\Omega k_0 \nabla v \cdot \nabla v \, dx \leq \int_\Omega (K(x)\nabla v(x)) \cdot \nabla v(x) \, dx = a(v,v).$$

Using Corollary 5.44 we conclude

$$a(v,v) \geq k_0 |v|_{1,\Omega}^2 \geq \frac{k_0}{\sqrt{1+s^2}} \|v\|_{1,\Omega}^2$$

where s is the diameter of the domain Ω. Finally, the continuity of the linear form l

$$|l(v)| = \left| \int_\Omega f v \, dx \right| \leq \|f\|_{0,\Omega} \|v\|_{0,\Omega} \leq \|f\|_{0,\Omega} \|v\|_{1,\Omega}$$

follows from the Cauchy-Schwarz inequality in L^2 and $\|v\|_{0,\Omega} \leq \|v\|_{1,\Omega}$.
Uniqueness of u. The choice of u_g is not unique. So assume that for two different choices u_g^1, u_g^2 we obtain two different solutions $u^1 = u_g^1 + u_0^1$ and $u^2 = u_g^2 + u_0^2$. Then we have:

$$\begin{aligned} a(u^1, v) &= l(v) && \forall v \in H_0^1(\Omega), \\ a(u^2, v) &= l(v) && \forall v \in H_0^1(\Omega) \end{aligned}$$

and consequently $a(u^1 - u^2, v) = 0 \ \forall v \in H_0^1(\Omega)$. Since $u^1|_{\partial\Omega} = g = u^2|_{\partial\Omega}$ we have $u^1 - u^2 \in H_0^1(\Omega)$ and we conclude with coercivity that $\alpha \|u^1 - u^2\|_{1,\Omega} \leq a(u^1 - u^2, u^1 - u^2) = 0$ and therefore $u^1 = u^2$. □

6.2. Neumann Problem

Let us now consider the pure Neumann problem

$$\begin{aligned} -\nabla \cdot (K \nabla u) + cu &= f && \text{in } \Omega, && (6.4a) \\ -(K \nabla u) \cdot n &= j && \text{on } \partial\Omega && (6.4b) \end{aligned}$$

with the function $c(x)$ uniformly positive, i.e. $\forall x \in \Omega : c(x) \geq c_0 > 0$. (The case $c(x) = 0$ will be treated below).

Multiplying with a test function $v \in C^1(\Omega) \cap C^0(\overline{\Omega})$ and integrating over Ω we obtain

$$\begin{aligned} &\int_\Omega (-\nabla \cdot (K \nabla u) + cu) v \, dx \\ &\qquad = \int_\Omega (K \nabla u) \cdot \nabla v + cuv \, dx + \int_{\partial\Omega} -(K \nabla u) \cdot n v \, ds \\ &\qquad = \int_\Omega (K \nabla u) \cdot \nabla v + cuv \, dx + \int_{\partial\Omega} j v \, ds \end{aligned}$$

Now the weak formulation reads

$$\text{Find } u \in U: \qquad a(u,v) = l(v) \quad \forall v \in V \tag{6.5}$$

with

$$a(u,v) = \int_\Omega (K\nabla u)\cdot\nabla v + cuv\,dx, \qquad l(v) = \int_\Omega fv - \int_{\partial\Omega} jv\,dx. \tag{6.6}$$

Here the functions u and v are *not* constrained at the boundary, it will turn out that $U = V = H^1(\Omega)$ is the appropriate function space. The boundary condition (**??**) is built into the linear form l and $j = 0$ would result in the same linear form as in the Dirichlet problem. Therefore the Neumann boundary condition is called the *natural boundary condition* in the context of the weak formulation. In contrast, the Dirichlet boundary condition is called *essential boundary condition* as it needs to be built into the function space itself (there are also ways to enforce Dirichlet boundary conditions weakly which is not treated here).

Proposition 6.3. Assume that

i) $K(x)$ is uniformly elliptic and bounded,

ii) $c(x)$ is uniformly positive and bounded and

iii) $f \in L^2(\Omega)$ as well as $j \in L^2(\partial\Omega)$.

Then the problem

$$\text{Find } u \in H^1(\Omega): \qquad a(u,v) = l(v) \qquad \forall v \in H^1(\Omega)$$

with a, l from (6.6) has a unique solution.

Proof. The proof of continuity of a is the same as in Proposition 6.2, except the additional term

$$\left|\int_\Omega cuv\,dx\right| \le C\int_\Omega |u|\,|v|\,dx \le C\|u\|_{0,\Omega}\|v\|_{0,\Omega} \le C\|u\|_{1,\Omega}\|v\|_{1,\Omega}.$$

Showing the coercivity, we cannot use the Friedrich inequality since $v \in H^1(\Omega)$. Here, the uniform positivity of $c(x)$ with constant c_0 is crucial. Using

$$c_0\|v\|_{0,\Omega}^2 = c_0\int_\Omega v^2\,dx \le \int_\Omega c(x)v^2\,dx$$

and $k_0|v|^2_{1,\Omega} \leq \int_\Omega (K\nabla v)\cdot\nabla v\,dx$ following from uniform ellipticity (see proof of Proposition 6.2) we obtain

$$\|v\|^2_{1,\Omega} = \|v\|^2_{0,\Omega} + |v|^2_{1,\Omega} \leq \frac{1}{\min(c_0,k_0)} \int_\Omega (K\nabla v)\cdot\nabla v + c(x)v^2\,dx = a(v,v)$$

and we obtain coercivity with $\alpha = \min(c_0, k_0)$.

The continuity of the linear form l now requires the trace theorem 5.45 to estimate the boundary term:

$$\begin{aligned} |l(v)| &= \left| \int_\Omega fv - \int_{\partial\Omega} jv\,dx \right| \leq \|f\|_{0,\Omega}\|v\|_{1,\Omega} + \|j\|_{0,\partial\Omega}\|\gamma v\|_{0,\partial\Omega} \\ &= \|f\|_{0,\Omega}\|v\|_{1,\Omega} + \|j\|_{0,\partial\Omega} c\|v\|_{1,\Omega} = (\|f\|_{0,\Omega} + c\|j\|_{0,\partial\Omega})\|v\|_{1,\Omega} \end{aligned}$$

where c is the constant from the trace theorem. □

Let us now consider the case $c(x) = 0$. If $u(x)$ is a solution of 6.4, then $u(x) + C$ is also a solution for any $C \in \mathbb{R}$. This ambiguity needs to be fixed by selecting the right function space.

Proposition 6.4. Assume that

i) $K(x)$ is uniformly elliptic and bounded,

ii) $c(x) = 0$ and

iii) $f \in L^2(\Omega)$, $j \in L^2(\partial\Omega)$ with $\int_\Omega f\,dx = \int_{\partial\Omega} j\,ds$.

Then with $V = \{v \in H^1(\Omega) : \int_\Omega v\,dx = 0\}$ the problem

$$\text{Find } u \in V: \qquad a(u,v) = l(v) \qquad \forall v \in V$$

with a, l from (6.6) has a unique solution.

Proof. The compatibility condition on the data iii) is a consequence of Gauß' theorem:

$$\int_\Omega f\,dx = -\int_\Omega \nabla\cdot(K\nabla u)\,dx = -\int_{\partial\Omega} (K\nabla u)\cdot n\,ds = \int_{\partial\Omega} j\,ds.$$

Continuity of a was already established in proposition 6.2. Coercivity uses $k_0|v|^2_{1,\Omega} \leq a(v,v)$ (shown in proposition 6.2 and the Poincaré inequality propo-

sition 5.42:

$$\begin{aligned}\|v\|_{1,\Omega}^2 &= \|v\|_{0,\Omega}^2 + |v|_{1,\Omega}^2 \\ &\leq c_1|v|_{1,\Omega}^2 + c_2 \underbrace{\left(\int_\Omega v\,dx\right)^2}_{=0} + |v|_{1,\Omega}^2 = (1+c_1)|v|_{1,\Omega}^2 \\ &\leq (1+c_1)k_0^{-1}a(v,v).\end{aligned}$$

Therefore coercivity holds with $\alpha = k_0/(1+c_1)$ with c_1 the constant from Poincaré's inequality. □

6.3. Mixed Problem

We now turn to the mixed problem

$$-\nabla\cdot(K\nabla u) + cu = f \qquad \text{in } \Omega, \tag{6.7a}$$
$$u = g \qquad \text{on } \Gamma_D \text{ with non-vanishing measure}, \tag{6.7b}$$
$$-(K\nabla u)\cdot n = j \qquad \text{on } \Gamma_N = \partial\Omega \setminus \Gamma_D, \tag{6.7c}$$

with the function $c(x) \geq 0$ for all $x \in \Omega$. The weak formulation then reads

$$\text{Find } u \in U: \qquad a(u,v) = l(v) \quad \forall v \in V \tag{6.8}$$

with

$$a(u,v) = \int_\Omega (K\nabla u)\cdot\nabla v + cuv\,dx, \qquad l(v) = \int_\Omega fv - \int_{\Gamma_N} jv\,dx. \tag{6.9}$$

The only change compared to (6.6) is that the integration of the boundary term is only with respect to the Neumann boundary Γ_N. The functions u and v are now assumed to be zero at the Dirichlet boundary $\Gamma_D \subseteq \partial\Omega$.

Proposition 6.5. Assume that

i) $K(x)$ is uniformly elliptic and bounded,

ii) $c(x) \geq 0$ for all $x \in \Omega$,

iii) $f \in L^2(\Omega)$, $j \in L^2(\Gamma_N)$,

iv) Γ_D has nonvanishing measure and there exists $u_g \in H^1(\Omega)$ such that $\gamma_D u_g = g$ ($\gamma_D v$ being the trace of v on Γ_D).

Then with $V_D = \{v \in H^1(\Omega) : \gamma_D v = 0\}$, the problem

$$\text{Find } u \in u_g + V_D : \qquad a(u,v) = l(v) \qquad \forall v \in V_D$$

with a, l from (6.9) has a unique solution.
Proof. As in the pure Dirichlet case the solution is written as $u = u_g + u_0$ with $u_0 \in V_D$. Then the coercivity of a on V_D can be established using the second variant of Friedrich's inequality proposition 5.41. For the continuity of the linear form l a generalization of the trace theorem 5.45 to part of the boundary is required. □

6.4. Convection-Diffusion Problem

As a further example let us consider the stationary convection-diffusion problem

$$\begin{aligned} \nabla \cdot (bu - K\nabla u) &= f && \text{in } \Omega, && (6.10a)\\ u &= g && \text{on } \Gamma_D \text{ with non-vanishing measure}, && (6.10b)\\ -(K\nabla u)\cdot n &= j && \text{on } \Gamma_N = \partial\Omega \setminus \Gamma_D, && (6.10c) \end{aligned}$$

with $b(x) : \overline{\Omega} \to \mathbb{R}^n$ a given velocity field. We subdivide the boundary into the following parts:

$$\begin{aligned} \partial\Omega_+ &= \{x \in \partial\Omega : b(x)\cdot n(x) > 0\} && \text{outflow boundary},\\ \partial\Omega_0 &= \{x \in \partial\Omega : b(x)\cdot n(x) = 0\} && \text{characteristic boundary},\\ \partial\Omega_- &= \{x \in \partial\Omega : b(x)\cdot n(x) < 0\} && \text{inflow boundary}. \end{aligned}$$

Assuming that $\nabla \cdot b = 0$ the conservative and non-conservative forms of the equation are equivalent

$$\nabla \cdot (bu - K\nabla u) = b \cdot \nabla u - \nabla \cdot (K\nabla u) = f.$$

and the weak formulation of (6.10) then reads: Find $u \in u_g + V_D$ such that

$$\int_\Omega (K\nabla u)\cdot\nabla v + (b\cdot\nabla u)v \, dx = \int_\Omega fv \, dx - \int_{\Gamma_N} jv \, ds \quad \forall v \in V_D. \qquad (6.11)$$

Let us define the bilinear forms and the usual linear form of this problem:

$$a(u,v) = \int_\Omega (K\nabla u)\cdot\nabla v + (b\cdot\nabla u)v\,dx, \tag{6.12a}$$

$$c(u,v) = \int_\Omega (b\cdot\nabla u)v\,dx, \tag{6.12b}$$

$$l(v) = \int_\Omega fv\,dx - \int_{\Gamma_N} gv\,ds. \tag{6.12c}$$

Proposition 6.6. Assume that

i) $K(x)$ is uniformly elliptic and bounded,

ii) b is bounded and $\nabla\cdot b(x) = 0$ for all $x\in\Omega$,

iii) $f\in L^2(\Omega)$, $j\in L^2(\Gamma_N)$,

iv) $\Gamma_N\cap\partial\Omega_- = \emptyset$ (inflow boundary must be Dirichlet),

v) Γ_D has nonvanishing measure and there exists $u_g\in H^1(\Omega)$ s. t. $\gamma_D u_g = g$.

Then with $V_D = \{v\in H^1(\Omega)\,:\,\gamma_D v = 0\}$, the problem

$$\text{Find } u\in u_g + V_D: \qquad a(u,v) = l(v) \qquad \forall v\in V_D$$

with a, l from (6.12) has a unique solution.

Proof. This proof follows the one in [Elman et al., 2005]. The crucial part is to show the coercivity of a. For the convective part of the bilinear form we obtain:

$$\begin{aligned}
c(u,v) &= \int_\Omega (vb)\cdot\nabla u\,dx \\
&= -\int_\Omega \nabla\cdot(vb)u\,dx + \int_{\Gamma_N} uvb\cdot n\,ds && \text{(integration by parts)} \\
&= -\int_\Omega (v\underbrace{\nabla\cdot b}_{=0} + \nabla v\cdot b)u\,dx + \int_{\Gamma_N} uvb\cdot n\,ds && \text{(product rule)} \\
&= -\int_\Omega (ub)\cdot\nabla v\,dx + \int_{\Gamma_N} uvb\cdot n\,ds \\
&= -c(v,u) + \int_{\Gamma_N} uvb\cdot n\,ds.
\end{aligned}$$

From this it follows that

$$c(u,v) + c(v,u) = \int_{\Gamma_N} uvb \cdot n \, ds$$

and therefore

$$c(v,v) = \frac{1}{2}(c(v,v) + c(v,v)) = \frac{1}{2} \int_{\Gamma_N} v^2 \underbrace{b \cdot n}_{\geq 0} \, ds \geq 0$$

since we required Γ_N not to be part of the inflow boundary. Coercivity then follows in the usual way from

$$k_0 |v|^2_{1,\Omega} \leq \int_\Omega (K \nabla u) \cdot \nabla v \, dx \leq \int_\Omega (K \nabla u) \cdot \nabla v \, dx + c(v,v) = a(v,v).$$

□

Chapter 7.

Conforming Finite Element Methods

7.1. Abstract Galerkin Method

In order to solve an elliptic PDE *numerically* the idea is to solve its *weak formulation in finite-dimensional function spaces.*

Suppose the function space underlying the weak formulation is V and let $V_h \subset V$ be of finite dimension. (h denotes a parameter later to be identified as the "*mesh size*"). If the function spaces are chosen appropriately, a minimum requirement is that $\dim V_h \to \infty$ as $h \to 0$ unless $u \in V_h$ for some h, then we will later prove that

$$\inf_{v_h \in V_h} \|u - v_h\| \to 0 \quad \text{as} \quad h \to 0 \ .$$

Provided the bilinear form is coercive in V, the Lax-Milgram theorem ensures also the solvability of the problem in the subspace V_h

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

We now consider how the variational problem in V_h can be solved practically. Since V_h has finite dimension $\dim V_h = N_h$ we can find a basis

$$\Phi_h = \{\varphi_1^h, \ldots, \varphi_{N_h}^h\} \ .$$

Inserting the basis representation

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

yields a linear system of equations:

$$\begin{aligned}
& a(u_h, v) = l(v) \quad \forall v \in V_h \\
\Leftrightarrow \qquad & a\left(\sum_{j=1}^{N_h} z_j \varphi_j^h, \varphi_i^h\right) = l(\varphi_i^h) \quad \forall i = 1, \ldots, N_h \\
\Leftrightarrow \qquad & \sum_{j=1}^{N_h} z_j a(\varphi_j^h, \varphi_i^h) = l(\varphi_i^h) \\
\Leftrightarrow \qquad & Az = b \qquad (A)_{ij} = a(\varphi_j^h, \varphi_i^h), \quad (b)_i = l(\varphi_i^h)
\end{aligned}$$

for the coefficient vector $z \in \mathbb{R}^{N_h}$. The matrix A is symmetric if the bilinear form a is symmetric:

$$(A)_{ij} = a(\varphi_j, \varphi_i) = a(\varphi_i, \varphi_j) = (A)_{ji}$$

and it is positive definite since the bilinear form is coercive:

$$\begin{aligned} \forall z \neq 0: \quad z^T A z &= \sum_{i=1}^{N_h} z_i \left(\sum_{k=1}^{N_h} (A)_{ik}\, z_k \right) = \sum_{i=1}^{N_h} z_i \left(\sum_{k=1}^{N_h} a(\varphi_k, \varphi_i)\, z_k \right) \\ &= a\left(\underbrace{\sum_{k=1}^{N_h} z_k \varphi_k}_{v}, \underbrace{\sum_{i=1}^{N_h} z_i \varphi_i}_{v} \right) = a(v, v) > 0 \ . \end{aligned}$$

Remark 7.1. a) In the engineering literature A is called *stiffness matrix* and b is called *load vector*.

b) In the more general case $u \in U_h, v \in V_h,\ U_h \neq V_h$ the method is called Petrov-Galerkin method. □

The Galerkin method requires finite dimensional subspaces of Sobolev spaces. Such spaces are called *conforming* and are characterized in the following lemma.

Lemma 7.2. Assume $k \geq 1$ and let Ω be a bounded domain. A piecewise C^∞-function $v : \overline{\Omega} \to \mathbb{R}$ is in $H^k(\Omega)$ if and only if $v \in C^{k-1}(\overline{\Omega})$.

Proof [Braess, 2003, Satz 5.2]. It is sufficient to consider $k = 1$, the result for $k > 1$ follows by induction. Moreover, we restrict ourselves to $\Omega \subset \mathbb{R}^2$.

"$\Leftarrow$". Assume $v \in C(\overline{\Omega})$ and let $\mathcal{T} = \{t_j : j = 1, \ldots, m\}$ be a decomposition of Ω into open subdomains such that

$$\bigcup_{i=1}^{m} \overline{t}_i = \overline{\Omega} \quad \text{and} \quad t_i \cap t_j = \emptyset \quad \forall i \neq j.$$

For $i = 1, 2$ define $w_i \in L^2(\Omega)$ as

$$w_i(x) = \begin{cases} \partial_{x_i} v(x) & x \in t_j \text{ for some } j \ (t_j \text{ is open}), \\ \text{arbitrary} & \text{else.} \end{cases}$$

We show that w_i is a weak derivative of v (the argument is the same as in

example 5.31). For any $\varphi \in C_0^\infty$ we have

$$\begin{aligned}\int_\Omega w_i\varphi\,dx &= \sum_{j=1}^m \int_{t_j} \partial_{x_i} v\varphi\,dx = \sum_{j=1}^m \left[-\int_{t_j} v\partial_{x_i}\varphi\,dx + \int_{\partial t_j} v\varphi n_i\,ds \right] \\ &= -\sum_{j=1}^m \int_{t_j} v\partial_{x_i}\varphi\,dx + \sum_{j=1}^m \sum_{l>j} \int_{\partial t_j\cap\partial t_l} \underbrace{\left[(v\varphi)|_{t_j} - (v\varphi)|_{t_l}\right]}_{=0 \text{ because } v \text{ and } \varphi \text{ continuous}} n_i\,ds \\ &= -\int_\Omega v\partial_{x_i}\varphi\,dx\ .\end{aligned}$$

Here the integral over $\partial t_j \cap \partial t_l$ is only taken if $\partial t_j \cap \partial t_l$ is a one-dimensional measure (i.e. an edge in the decomposition). If $\partial t_j \cap \partial t_l$ is a single point, it is not considered. The minus sign in the boundary integral comes from the fact that n_i is the i-th component of the *outer* unit normal.

"$\Rightarrow$" Now assume $v \in H^1(\Omega)$. This means that v has a weak derivative $w_i = \partial_{x_i} v$. The argument above shows that this is only possible if v is continuous since ϕ is arbitrary. □

7.2. One-dimensional Finite Element Spaces

Let $\Omega = (a, b)$ be subdivided into

$$a = x_0 < x_1 < \ldots < x_m = b$$

(not necessarily equidistant) and set for $j = 0, \ldots, m-1$

$$t_j = (x_j, x_{j+1}),\ h(t_j) = x_{j+1} - x_j \text{ and } h = \max_{t\in\mathcal{T}} h(t)\ .$$

The set $\mathcal{T} = \{t_j : j = 0, \ldots, m-1\}$ is in general called a *mesh* or a *triangulation* and each individual $t \in \mathcal{T}$ is called an *element* or a *cell*.

For convergence we will need to consider sequences of meshes $\{\mathcal{T}_\nu : \nu \in \mathbb{N}\}$ such that $h_\nu \to 0$. The size of $t \in \mathcal{T}_\nu$ is denoted by $h_\nu(t)$.

By

$$\mathbb{P}_k^1 = \left\{ u \in C^\infty(\mathbb{R}) : u(x) = \sum_{i=0}^k c_i x^i \right\}$$

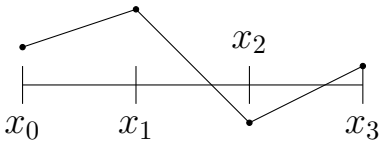
we denote the $k+1$-dimensional vector space of polynomials of at most degree k in $\mathbb{R}$ and for a given triangulation $\mathcal{T}$ we denote by

$$P_k(\mathcal{T}) = \left\{ u \in C^0(\overline{\Omega}) : u|_{\bar{t}} \in \mathbb{P}_k^1 \quad \forall t \in \mathcal{T} \right\} \tag{7.1}$$

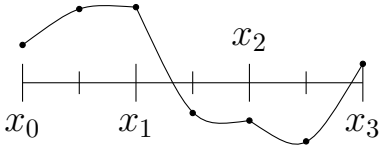
the space of piecewise polynomials of degree at most k. According to Lemma 7.2 we have $P_k(\mathcal{T}) \subset H^1(\Omega)$.

In (7.1) the space $P_k(\mathcal{T})$ is characterized without reference to a basis. In the following we show how to construct a basis of $P_k(\mathcal{T})$ that is needed to carry out the Galerkin procedure.

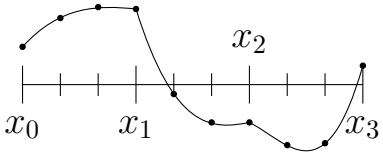
Example 7.3. A polynomial of degree k is defined uniquely by prescribing values at $k+1$ distinct points. Global continuity of the piecewise polynomial function is ensured by including at least the points $x_i, i = 0, \ldots, m$. This is illustrated for $m = 3$ and $k = 1, 2, 3$:

$k = 1$ 

$\dim P_k(\mathcal{T}) = m + 1$

$k = 2$ 

$\dim P_k(\mathcal{T}) = 2m + 1$

$k = 3$ 

$\dim P_k(\mathcal{T}) = 3m + 1$

For arbitrary k the dimension of $P_k(\mathcal{T})$ is $km + 1$. Note also that the $k - 1$ additional points within each element need not be chosen equidistantly (as it is done in the following proposition). □

Proposition 7.4. The functions $\varphi_i,\ 0 \le i \le km$, given by

$$\varphi_i \in P_k(\mathcal{T}), \quad \varphi_i(x'_j) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{else} \end{cases}$$

with $x'_j = x_{j/k} + \frac{j \bmod k}{k}(x_{(j/k)+1} - x_{j/k})$ for $0 \le j \le km$ are a basis of $P_k(\mathcal{T})$ (j/k denotes integer division without remainder). The φ_i are called *Lagrange basis functions*.

Proof. The property $\varphi_i(x'_i) = \delta_{ij}$ ensures that the φ_i are linearly independent: Since φ_i is 1 at x'_i and all $\varphi_j,\ j \ne i$ are 0 at x'_i, φ_i cannot be a linear combination of the $\varphi_j,\ j \ne i$. On the other hand $\varphi_i \in P_k(\mathcal{T})$ by construction. Since $\dim P_k(\mathcal{T}) = km + 1$ the φ_i are a basis. □

Affine Construction Let us recall the definition of the Lagrange polynomials.

Definition 7.5. For a given set of $k+1$ points $s = (s_0, s_1, \ldots, s_k)$, $s_i \neq s_j$ for $i \neq j$, the Lagrange polynomials $L_i^s(x)$ are given by

$$L_i^s(x) = \frac{\prod_{j \neq i}(x - s_j)}{\prod_{j \neq i}(s_i - s_j)}, \quad 0 \leq i \leq k \ .$$

Note that $L_i^s(s_j) = \delta_{ij}$ and therefore the L_i^s are a basis of $\mathbb{P}_k^1$. □

The following construction, known as *affine finite elements*, gives an efficient way to construct the basis polynomials φ_i. By $\hat{\Omega} = [0,1]$ we denote the *reference element*. For every $t_l = (x_l, x_{l+1}) \in \mathcal{T}$ we define the bijective map

$$\mu_{t_l} : \hat{\Omega} \to \overline{t_l}, \qquad \mu_{t_l}(\hat{x}) = x_l + \hat{x} h(t_l) \ .$$

For $\hat{s} = (0, 1/k, 2/k, \ldots, 1)$, the $L_i^{\hat{s}}(\hat{x})$ are the Lagrange polynomials for an equidistant subdivision of the reference element and for each $t_l \in \mathcal{T}$ and $s^l = (x_l, x_l + h(t_l)/k, x_l + 2h(t_l)/k, \ldots, x_{l+1})$ the $L_i^{s^l}$ are Lagrange polynomials for an equidistant subdivision of t_l.

We observe that that

$$\forall \hat{x} \in \hat{\Omega}, \ \forall t_l \in \mathcal{T} : L_i^{s^l}(\mu_{t_l}(\hat{x})) = L_i^{\hat{s}}(\hat{x})$$

since

$$\frac{\prod_{j \neq i}(x_l + \hat{x} h(t_l) - (x_l + j h(t_l)/k))}{\prod_{j \neq i}(x_l + i h(t_l)/k - (x_l + j h(t_l)/k))} = \frac{\prod_{j \neq i}(\hat{x} - j/k)}{\prod_{j \neq i}(i/k - j/k)} \ .$$

This means that the Lagrange polynomials on the individual elements can be generated from the Lagrange polynomials on the reference element and the transformations μ_{t_l}.

With the characteristic function

$$\chi_{t_j}(x) = \begin{cases} 1 & x \in t_j \\ 0 & \text{else} \end{cases}$$

we can write the basis functions φ_i in the form

$$\varphi_i(x) = \sum_{t_j \in \mathcal{T}} \sum_{l=0}^{k} \delta_{jk+l,i} L_l^{\hat{s}}(\mu_{t_j}^{-1}(x)) \chi_{t_j}(x) \ . \tag{7.2}$$

Formally, the function φ_i is only defined in the interior of each $t_j \in \mathcal{T}$ (due to χ_{t_j}). However, for $x \in \partial t \cap \partial t' \neq \emptyset$ the limit $\lim_{y \to x} \varphi_i(y)$ yields the same value for $y \in t$ and $y \in t'$. Therefore there exists a unique extension of φ_i to all points in $\overline{\Omega}$.

The map $g : \mathcal{T} \times \{0, \ldots, k\}$ with $g(t_j, l) = jk + l$ is called the *local-to-global map*. It determines that local basis function l in element t_j contributes to the global basis function $\varphi_{g(t_j,l)}$.

Remark 7.6. The positions $\hat{s}$ on the reference element need not be equidistant. The construction can be generalized to arbitrary positions within $\hat{\Omega}$. This may be advantageous in connection with certain quadrature formulae. □

Definition 7.7. The map $\mathcal{I}_k[\mathcal{T}] : C^0(\overline{\Omega}) \to P_k(\mathcal{T})$ given by

$$\mathcal{I}_k[\mathcal{T}](v) = \sum_{i=0}^{km} v(x_i')\varphi_i = \sum_{i=0}^{km} \gamma_i(v)\varphi_i$$

is called *Lagrange interpolation operator*.

The linear forms $\{\gamma_0, \ldots, \gamma_{km}\} \subset \mathcal{L}(C^0(\overline{\Omega}); \mathbb{R})$ given by $\gamma_i(v) = v(x_i')$ are called *global degrees of freedom*. □

When $\{\mathcal{T}_\nu : \nu \in \mathbb{N}\}$ is a sequence of triangulations with $h_\nu \to 0$ we write $P_{k,h}$, $\varphi_{i,h}$, $\mathcal{I}_{k,h}$ as short hand notation for $P_k(\mathcal{T}_\nu)$, $\varphi_i \in P_k(\mathcal{T}_\nu)$ and $\mathcal{I}_k[\mathcal{T}_\nu]$ with $h_\nu = h$.

These ideas are now extended to the case $n > 1$. The construction is essentially the same, only more technical.

7.3. Mesh Construction in Arbitrary Dimensions

In the following, Ω is a bounded domain in $\mathbb{R}^n$ with Lipschitz-continuous boundary.

Definition 7.8 (Polyhedron)**.** In $\mathbb{R}^2$ a polyhedron (or polygon) is a domain with a boundary that is a finite set of straight line segments. In dimension $n > 2$ a polyhedron is a domain with a boundary that is a finite set of polyhedra in $\mathbb{R}^{n-1}$. □

Definition 7.9 (Mesh)**.** Let Ω be a domain in $\mathbb{R}^n$. A mesh in its most general form is a finite set $\mathcal{T} = \{t_0, \ldots, t_{m-1}\}$ of bounded domains t_i with Lipschitz boundary that form a partitioning of Ω:

$$\overline{\Omega} = \bigcup_{i=0}^{m-1} \overline{t_i}, \quad t_i \cap t_j = \emptyset \ \forall i \neq j \ .$$

The domains t_i are called elements (or cells). Moreover, for all $t \in \mathcal{T}$ we set $h(t) := \operatorname{diam} t = \max_{x,y\in\bar{t}} \|x - y\|$ and $h = \max_{t\in\mathcal{T}} h(t)$ is called mesh size. □

Often we write $\mathcal{T}_h$ to indicate that h is the mesh size in $\mathcal{T}$. In order to study the convergence of functions $v_h \to v \in H^1(\Omega)$ we consider sequences of successively refined meshes $\{\mathcal{T}_\nu : \nu \in \mathbb{N}\}$ with mesh size $h_\nu = \max_{t\in\mathcal{T}_\nu} h_\nu(t) \to 0$.

Typically, we will consider meshes that are less general than those possible in Defintion 7.9. A typical assumption is that all mesh elements $t_i \in \mathcal{T}_h$ are generated by a geometric transformation from a reference element $\hat{\Omega}$, i.e.

$$\forall i = 0, \dots, m-1 : \bar{t}_i = \mu_{t_i}(\hat{\Omega}),$$

where μ_{t_i} is a C^1-diffeomorphism (μ is bijective and μ, μ^{-1} are C^1-functions). The reference element $\hat{\Omega}$ is typically either the reference simplex

$$\hat{S}_n = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n \, : \, 0 \le x_i \le 1, 0 \le \sum_{i=1}^{n} x_i \le 1 \right\}$$

or the reference cube

$$\hat{Q}_n = \{ (x_1, \dots, x_n) : 0 \le x_i \le 1 \} \ .$$

If $\hat{\Omega} = \hat{S}_n$ the mesh is called simplicial, if $\hat{\Omega} = \hat{Q}_n$ the mesh is called cuboid.

Definition 7.10. A mesh is called *affine* if $\bar{t}_i = \mu_{t_i}(\hat{\Omega})$ and all transformations are affine, i.e. $\mu_{t_i}(x) = B_{t_i} x + z_{t_i}$. If $\hat{\Omega} = \hat{S}_n$ ($\hat{Q}_n$), the mesh is called affine simplicial (cuboid). □

The smoothness of μ_t implies that the corners of the reference element $\hat{\Omega}$ and the transformed element t are numbered in a compatible way. In addition, one often requires that $\det \nabla \mu_t = \det B_t > 0$.

If a mesh is affine simplicial or affine cuboid the domain Ω needs to be a polyhedron. Therefore this is assumed in the following. The treatment of domains with curved boundaries requires transformations μ_t which are e.g. polynomials of degree greater than 1. Conforming spaces are most easily defined on the following type of mesh.

Definition 7.11 (Geometrically conforming mesh)**.** Let $\mathcal{T}$ be an affine simplicial (or cuboid) mesh. $\mathcal{T}$ is called geometrically conforming if the intersection of two elements $t, t' \in \mathcal{T}$ is either empty or a common *face*, i.e. a simplex (or cube) of smaller dimension. □

The following figure gives an example for a conforming triangular mesh (left) and a nonconforming triangular mesh (right):

For the error estimates we need assumptions on the *quality* of a mesh.

Definition 7.12. Let $\mathcal{T}$ be an affine mesh and $\rho(t)$ the diameter of the largest ball that can be inscribed in $t \in \mathcal{T}$. A sequence of meshes $\{\mathcal{T}_\nu : \nu \in \mathbb{N}\}$ is called

a) *uniform* if there exists a number $\kappa_1 > 0$ independent of ν such that

$$\forall \nu \in \mathbb{N},\ \forall t \in \mathcal{T}_\nu : \frac{h_\nu}{h_\nu(t)} \leq \kappa_1$$

b) *shape-regular* if there exists a number $\kappa_2 > 0$ independent of ν such that

$$\forall \nu \in \mathbb{N},\ \forall t \in \mathcal{T}_\nu : \frac{h_\nu(t)}{\rho_\nu(t)} \leq \kappa_2$$

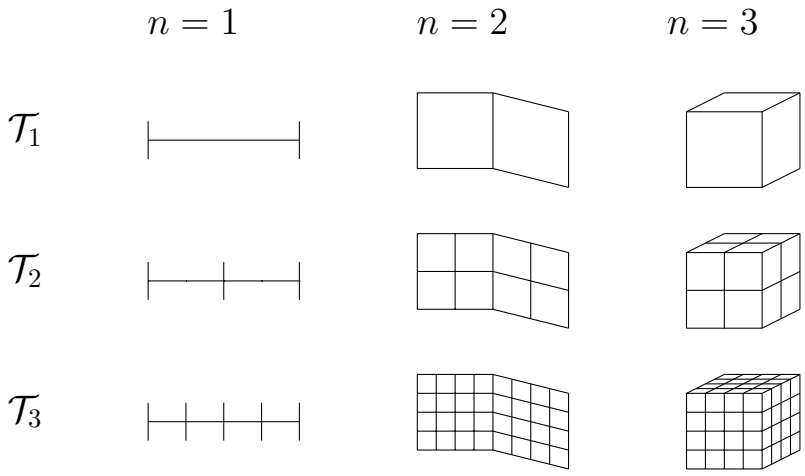
Both conditions are independent. □

If the mesh is uniform, we have $h_\nu/\kappa_1 \leq h_v(t) \leq h_\nu$ for all $t \in \mathcal{T}_\nu$, i.e. all elements have the same diameter up to a constant. If the mesh is shape-regular then $h_\nu(t)/\kappa_2 \leq \rho_v(t) \leq h_\nu(t)$ for all $t \in \mathcal{T}_\nu$ which means that the interior angles of the element can not degenerate.

A *mesh refinement algorithm* generates a sequence of successively refined meshes $\mathcal{T}_\nu,\ \nu = 2, 3, \ldots$, with mesh size h_ν from a given initial mesh $\mathcal{T}_1$.

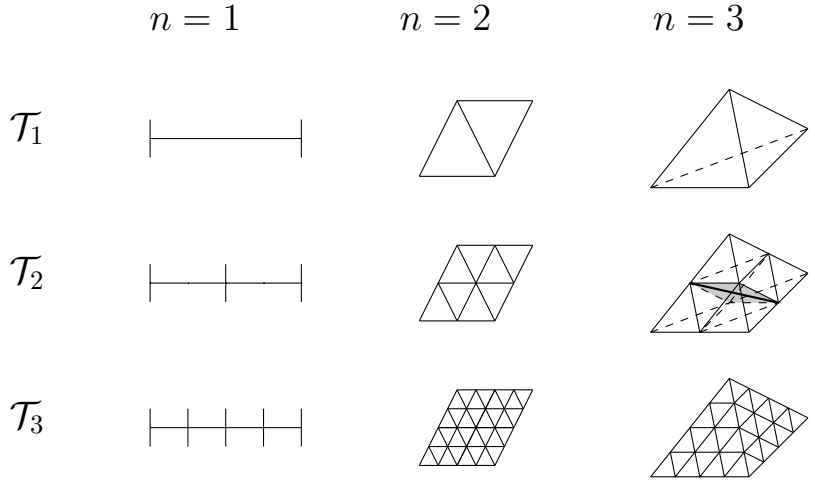
There are several possibilities to do this, depending on the type of element

Cuboid Mesh Refinement



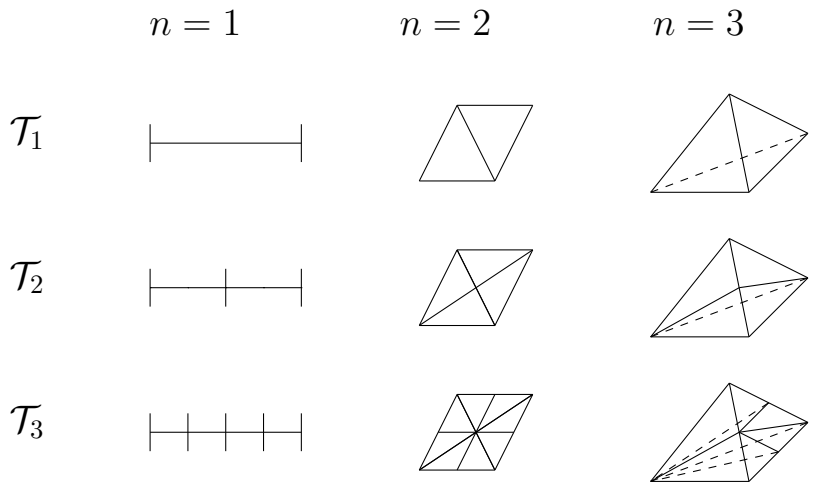
The meshes are conforming and uniform. Moreover, each element of the refined mesh is congruent to an element of $\mathcal{T}_1$.

Regular Subdivision of Simplices



It turns out that for $n \leq 3$ the regular refinement of a tetrahedron does not result in all refined tetrahedra being congruent to the initial one. But it can be shown that the refinement can be done such that the number of equivalence classes is finite [Bey, 2000].

Bisection Refinement



The selection of the edge to be refined is not unique. Possible choices are:

- longest edge bisection [Bänsch, 1991]
- (opposite) nearest vertex bisection [Mitchell and McClain, 2011]

Initial Mesh Generation Generating the initial mesh for a given domain Ω is called the *mesh generation problem*. There are two approaches in general use:

- advancing front method: add one element at a time
- Delauney triangulation: First place the vertices, then connect those by simplices

This problem is very hard, especially for complex geometries. Moreover, cube meshes are harder to generate than simplicial meshes. For details see [Ern and Guermond, 2004].

7.4. P_k Finite Elements

The space of polynomials of degree at most k in n space dimensions is

$$\mathbb{P}_k^n = \{u \in C^\infty(\mathbb{R}^n) : u(x) = \sum_{0 \le |\alpha| \le k} c_\alpha x^\alpha\} \ . \tag{7.3}$$

In the case $n = 2$ we have

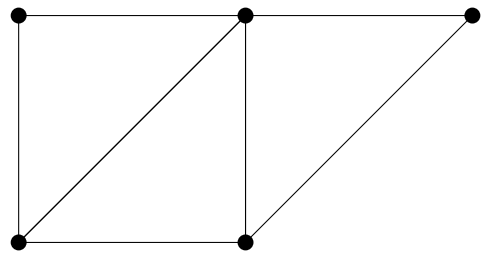
$$\dim \mathbb{P}_k^2 = \frac{(k+1)(k+2)}{2} \ .$$

As in the one-dimensional case the finite element space of piecewise polynomials of degree k on a conforming, simplicial mesh $\mathcal{T}$ is given by

$$P_k(\mathcal{T}) = \{u \in C^0(\overline{\Omega}) : u|_{\bar{t}} \in \mathbb{P}_k^n \ \forall t \in \mathcal{T}\} \ . \tag{7.4}$$

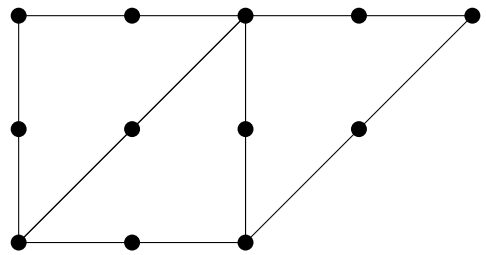
A Lagrange basis for $P_k(\mathcal{T})$ can be defined as follows. For the ease of drawing we illustrate only the case $n = 2$ but the construction can easily be extended to arbitrary dimension.

a) $k = 1$. Let $x_0, \ldots, x_{N-1} \in \overline{\Omega}$ denote the vertices of the conforming, simplicial mesh and define the basis functions $\varphi_i, \ i = 0, \ldots, N-1$ via $\varphi_i(x_j) = \delta_{ij}$.



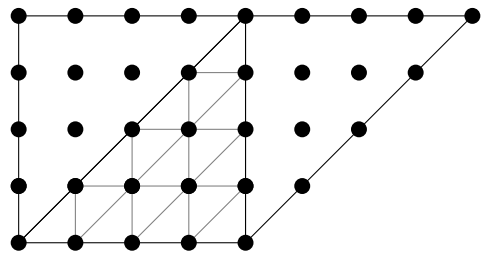
On each $t \in \mathcal{T}$ a polynomial $p_t \in \mathbb{P}_1^2$ is defined by the 3 values at its vertices. Each φ_i and therefore also $u = \sum z_i \varphi_i$ is conforming since p_t is linear on an edge and it is defined uniquely by the two values at its end points. p_t on an edge does *not* depend on the value of the vertex opposite of the edge!

b) $k = 2$. Additionally introduce the edge midpoints $x_N, \ldots, x_{N+E-1}$ and consider $\varphi_i(x_j) = \delta_{ij}, \ 0 \le i < N + E$.



On each $t \in \mathcal{T}$, $p_t \in \mathbb{P}_2^2$ is determined uniquely by the 6 point values. On an edge p_t is quadratic and is defined uniquely by the values of the three points lying on the edge.

c) General k (figure illustrates the case $k = 4$).



Each edge is subdivided into k equidistant intervals by introducing $k-1$ points on an edge. Within each $t \in \mathcal{T}$ these points are connected by lines parallel to the edges resulting in $\dim \mathbb{P}_k^2$ points associated with each triangle. On each $t \in \mathcal{T}$, $p_t \in \mathbb{P}_k^2$ is defined uniquely by the prescription of values at these points.

On each edge p_t is a polynomial of degree k in one variable. Let $a, b \in \overline{\Omega}$ be the two end points, then $x(\xi) = a + \xi(b-a)$ parametrizes the edge. Inserting this into p_t results in:

$$\begin{aligned} p_t(x(\xi)) &= \sum_{0\le|\alpha|\le k} c_\alpha (x(\xi))^\alpha \\ &= \sum_{0\le|\alpha|\le k} c_\alpha (a_1 + \xi(b_1 - a_1))^{\alpha_1} (a_2 + \xi(b_2 - a_2))^{\alpha_2} \\ &= \sum_{0\le|\alpha|\le k} c_\alpha (a_1^{\alpha_1} + \ldots + (b_1 - a_1)^{\alpha_1}\xi^{\alpha_1})(a_2^{\alpha_2} + \ldots + (b_2 - a_2)^{\alpha_2}\xi^{\alpha_2}) \\ &= \sum_{i=0}^{k} \tilde{c}_i \xi^i \quad \text{(after reorganization of the sum).} \end{aligned}$$

A polynomial of degree k is determined uniquely by $k+1$ values, so the p_t defined by local Lagrange interpolation form a continuous function u.

Figure 7.1 illustrates the global P_k basis functions for $k = 1, 2, 3, 4$ in two space dimensions. Note how some of the basis functions for $k > 1$ may exhibit undershoots (go below zero).

Let us now turn to the corresponding affine construction. As in the one-dimensional case the global basis function φ_i can be constructed from a basis on the reference simplex $\hat{S}_n$. We illustrate the case $n = 2$.

The basis functions $\hat{\varphi}_{i,j}$, $0 \le i + j \le k$ are then given by

$$\hat{\varphi}_{i,j}(\xi,\eta) = \prod_{\alpha=0}^{i-1} \underbrace{\frac{\xi - \alpha/k}{i/k - \alpha/k}}_{=0 \text{ for } \xi=\alpha/k} \prod_{\beta=0}^{j-1} \underbrace{\frac{\eta - \beta/k}{j/k - \beta/k}}_{=0 \text{ for } \eta=\beta/k} \prod_{\gamma=i+j+1}^{k} \underbrace{\frac{\gamma/k - \xi - \eta}{\gamma/k - i/k - j/k}}_{=0 \text{ for } \xi+\eta=\gamma/k}$$

and correspondingly $\varphi_{i,j}(\hat{x}_{l,m}) = \delta_{(i,j),(l,m)}$ for $\hat{x}_{l,m} = (l/k, m/k)$, $0 \le l+m \le k$. In addition, one can check that on an edge all basis functions *not* associated with a point on the edge are identically zero.

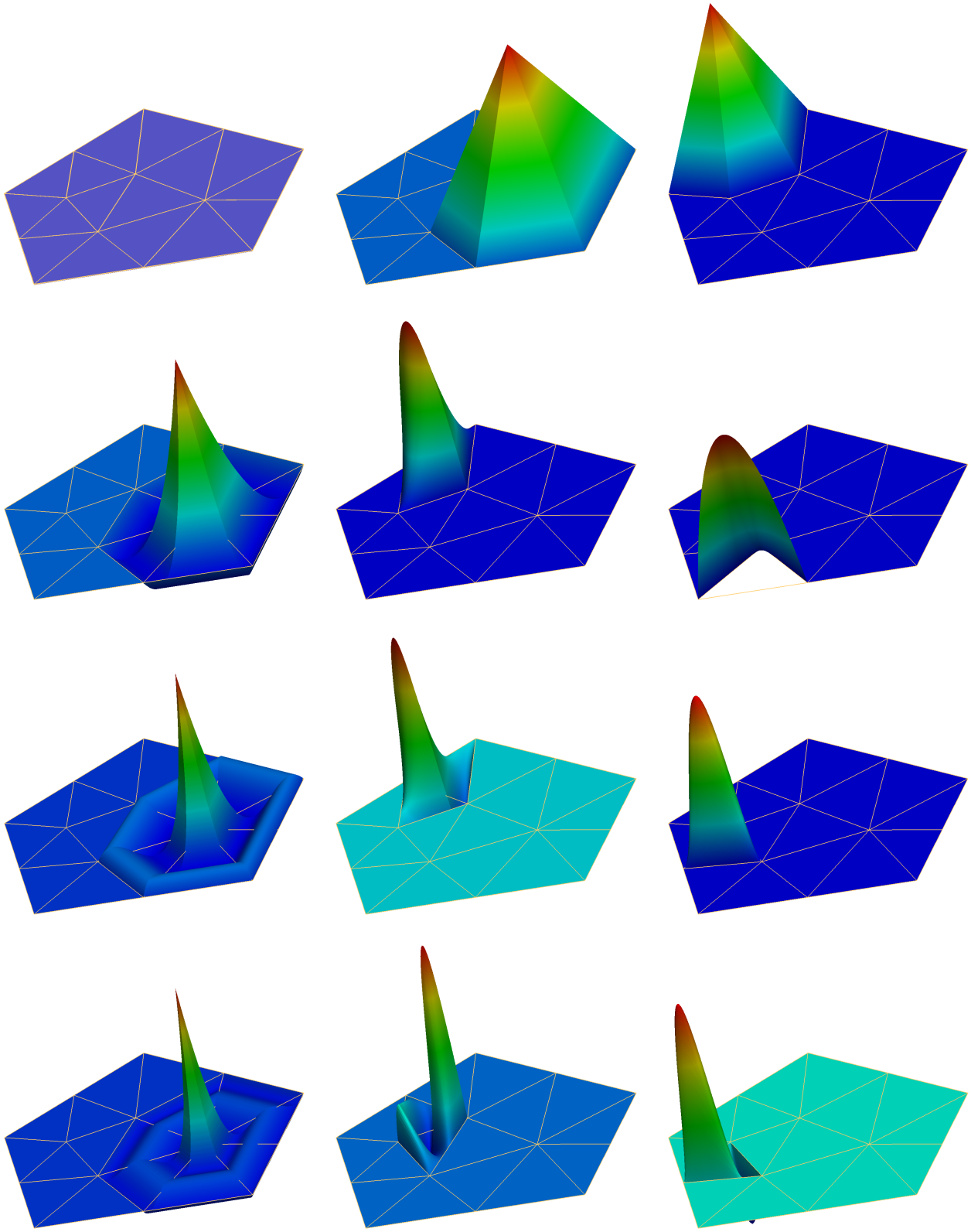


Figure 7.1.: Illustration of P_k finite element functions in two space dimensions for $k = 1, 2, 3, 4$ (top to bottom).

The transformation $\mu_t : \hat{S}_n \to \bar{t}$ for a simplex $t = \{x_0^t, \ldots, x_n^t\}$ is given by

$$\mu_t(\xi) = x_0^t + \sum_{i=1}^{n} \xi_i (x_i^t - x_0^t) = B_t \xi + x_0^t \tag{7.5}$$

where $B_t = [x_1^t - x_0, \ldots, x_n^t - x_0]$ is the $n \times n$ matrix built column-wise by the vectors $x_i^t - x_0^t$.

Then any $u \in P_k^2$ can be written in the form

$$u(x) = \sum_{t \in \mathcal{T}} \sum_{l=0}^{\dim \mathbb{P}_k^2 - 1} z_{g(t,l)} \hat{\varphi}_l(\mu^{-1}(x)) \chi_t(x) \tag{7.6}$$

where $\hat{\varphi}_l$, $0 \le l < \dim \mathbb{P}_k^2$ are the basis polynomials on the reference simplex, $g : \mathcal{T} \times \{0, \ldots, \dim \mathbb{P}_k^2 - 1\} \to \dim V_h - 1$ maps local degrees of freedom to global degrees of freedom and z is the vector of global degrees of freedom.

7.5. Q_k Finite Elements

Finite element spaces on cubes are based on the polynomials

$$\mathbb{Q}_k^n = \{u \in C^\infty(\mathbb{R}^n) : u(x) = \sum_{0 \le |\alpha|_\infty \le k} c_\alpha x^\alpha\} \tag{7.7}$$

with $|\alpha|_\infty = \max_i \alpha_i$, $\alpha_i \in \mathbb{N}_0$ and $\dim \mathbb{Q}_k^n = (k+1)^n$.

If all cubes in the mesh $\mathcal{T}$ are axi-parallel the space $\tilde{\mathbb{Q}}_k(\mathcal{T}) = \{u \in C^0(\overline{\Omega}) : u|_{\bar{t}} \in \mathbb{Q}_k \ \forall t \in \mathcal{T}\}$ can be defined as before. However this construction fails for general cube elements as can be shown by example.

Consider $n = 2, k = 1$. Then $u_t(x, y) = a_t xy + b_t x + c_t y + d_t$. On a straight edge given by $(x, y) = (\alpha + \xi\beta, \gamma + \xi\delta)$ u_t takes the values

$$\begin{aligned} u(\alpha + \xi\beta, \gamma + \xi\delta) &= a_t(\alpha + \xi\beta)(\gamma + \xi\delta) + b_t(\alpha + \xi\beta) + c_t(\gamma + \xi\delta) + d_t \\ &= a_t \beta\delta \xi^2 + \ldots \end{aligned}$$

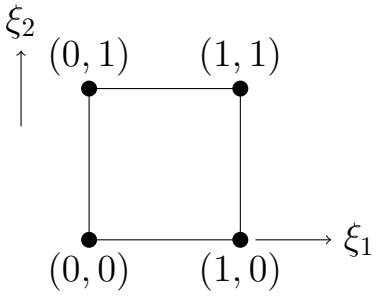
So in general, u_t is quadratic on an edge and therefore its values on an edge are *not* determined uniquely by the two values at the end points.

This problem is solved by the affine construction which is crucial in the case of general cube elements. First define the Lagrange basis $\tilde{\mathbb{Q}}_k^n$ on the reference cube $\hat{Q}_n$:

$$\hat{\varphi}_\alpha(\xi) = \prod_{i=1}^{n} \frac{\prod_{j=0, j \ne \alpha_i}^{k} (\xi_i - j/k)}{\prod_{j=0, j \ne \alpha_i}^{k} (\alpha_i/k - j/k)}, \quad 0 \le |\alpha|_\infty \le k \ .$$

For the points $\hat{x}_\beta = (\beta_1/k, \ldots, \beta_n/k)$, $0 \le |\beta|_\infty \le k$, we have $\hat{\varphi}_\alpha(\hat{x}_\beta) = \delta_{\alpha\beta}$. Also note that $\hat{\varphi}_\alpha \equiv 0$ on every face if $\hat{x} : \alpha$ is *not* on that face.

For $n = 2$, $k = 1$ we obtain for example:



$$\begin{aligned}
\hat{\varphi}_{(0,0)}(\xi_1,\xi_2) &= (1-\xi_1)(1-\xi_2)\\
\hat{\varphi}_{(1,0)}(\xi_1,\xi_2) &= \quad\ \xi_1\ \ (1-\xi_2)\\
\hat{\varphi}_{(0,1)}(\xi_1,\xi_2) &= (1-\xi_1)\ \ \xi_2\\
\hat{\varphi}_{(1,1)}(\xi_1,\xi_2) &= \quad\ \xi_1\quad \xi_2
\end{aligned}$$

For any cube $t \in \mathcal{T}$ given by the corners $\{x_\alpha : 0 \le |\alpha|_\infty \le 1\}$ the function

$$\mu_t(\xi) = \sum_{0 \le |\alpha| \le 1} \hat{\varphi}(\xi) x_\alpha \tag{7.8}$$

defines a map $\hat{Q}_n \to \bar{t}$ (the $\hat{\varphi}_\alpha$ are the Lagrange basis functions for $\mathbb{Q}_1^n$). In the following we assume that this map in bijective on $\bar{t}$. Note that the faces of the reference cube are mapped to the faces of $\bar{t}$ and vice versa.

Now the space of piecewise polynomials of degree k on a general cube mesh $\mathcal{T}$ is defined as

$$Q_k(\mathcal{T}) = \{u \in C^0(\overline{\Omega}) : u|_t = \hat{u}_t \circ \mu_t^{-1},\ \hat{u}_t \in \mathbb{Q}_k^n,\ t \in \mathcal{T}\}\ . \tag{7.9}$$

We show that Q_k is conforming. On $\bar{t}$, u_t is given by

$$u_t(x) = \hat{u}_t(\mu_t^{-1}(x)) \quad \Leftrightarrow \quad u_t(\mu_t(\hat{x})) = \hat{u}_t(\hat{x}) \qquad \text{for } x = \mu_t(\hat{x})\,.$$

On a face $\hat{f}$, $\hat{u}_t$ is defined uniquely by the $(k+1)^n$ values on that face and by construction the values on the transformed face $f = \mu_t(\hat{f})$ are those of $\hat{u}_t$.

Again, we have the representation

$$u(x) = \sum_{t \in \mathcal{T}} \sum_{l=0}^{\dim \mathbb{Q}_k^n - 1} z_{g(t,l)} \hat{\varphi}_l(\mu^{-1}(x)) \chi_t(x)\ . \tag{7.10}$$

7.6. Construction of the Finite Element Stiffness Matrix

We demonstrate how to construct the stiffness matrix and right hand side for the following problem:

$$u_h \in V_h : \quad a(u_h, v) = l(v) \qquad \forall v \in V_h \tag{7.11}$$

with

$$a(u,v) = \int_\Omega (K\nabla u)\cdot\nabla v\,dx\,, \quad l(v) = \int_\Omega f(v)\,dx\,,$$

and the simplest case of homogeneous Dirichlet boundary conditions. Using the basis representation $V_h = \mathrm{span}\{\varphi_0^h,\ldots,\varphi_{N-1}^h\}$ (7.11) is equivalent to the linear system

$$Az = b$$

with

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

How can the entries of A and b be efficiently computed on a given mesh $\mathcal{T}_h$? Exploiting (7.6) or (7.10), every global basis function φ_i has a local representation:

$$\varphi_i(x) = \sum_{t\in\mathcal{T}}\sum_{l=0}^{L-1}\delta_{g(t,l),i}\hat{\varphi}_l(\mu_t^{-1}(x))\chi_t(x)\ .$$

(L is the number of basis functions on the reference element). So we have

$$\begin{aligned}
(A)_{ij} &= \int_\Omega (K\nabla\varphi_j)\cdot\nabla\varphi_i\,dx \\
&= \sum_{t\in\mathcal{T}}\sum_{l,m=0}^{L-1}\delta_{g(t,l),i}\delta_{g(t,m),j}\int_t (K\nabla_x\hat{\varphi}_m(\mu_t^{-1}(x))\cdot\nabla_x\hat{\varphi}_l(\mu_t^{-1}(x))\,dx \\
&= \sum_{t\in\mathcal{T}}\sum_{l,m=0}^{L-1}\delta_{g(t,l),i}\delta_{g(t,m),j}\int_t \big(K(\nabla_x\mu_t^{-1}(x))^T\nabla_{\hat{x}}\hat{\varphi}_m(\mu_t^{-1}(x))\big)\cdot(\nabla_x\mu_t^{-1}(x))^T\nabla_{\hat{x}}\hat{\varphi}_l(\mu_t^{-1}(x))\,dx \\
&= \sum_{t\in\mathcal{T}}\sum_{l,m=0}^{L-1}\delta_{g(t,l),i}\delta_{g(t,m),j}\int_{\hat{\Omega}} \big(K(\nabla_{\hat{x}}\mu_t(\hat{x}))^{-T}\nabla_{\hat{x}}\hat{\varphi}_m(\hat{x})\big)\cdot(\nabla_{\hat{x}}\mu_t(\hat{x}))^{-T}\nabla_{\hat{x}}\hat{\varphi}_l(\hat{x})\ |\det\nabla_{\hat{x}}\mu_t(\hat{x})|\,d\hat{x} \\
&= \sum_{t\in\mathcal{T}}\sum_{l,m=0}^{L-1}\delta_{g(t,l),i}\delta_{g(t,m),j}(A_t)_{l,m}
\end{aligned} \tag{7.12}$$

where we have first transformed the gradient to the reference element and then transformed the integral to the reference element. The $L\times L$ matrix A_t is called *local (or element-wise) stiffness matrix.*

In order to show the transformation of the gradient, let $\mu_t : \hat{\Omega}\to\bar{t}$ be the transformation and $u:\bar{t}\to\mathbb{R}$ and $\hat{u}:\hat{\Omega}\to\mathbb{R}$ two functions linked by

$$u(x) = \hat{u}(\mu_t^{-1}(x))\ .$$

Application of the chain rule results in

$$\nabla_x u(x) = \left(\nabla_x \mu_t^{-1}(x)\right)^T \nabla_{\hat{x}} \hat{u}(\mu_t^{-1}(x)) = \left(\nabla_{\hat{x}} \mu_t(\mu_t^{-1}(x))\right)^{-T} \nabla_{\hat{x}} \hat{u}(\mu_t^{-1}(x))$$

because $I = \nabla_x x = \nabla_x \mu_t(\mu_t^{-1}(x)) = (\nabla_x \mu_t^{-1}(x))^T (\nabla_{\hat{x}} \mu_t(\mu_t^{-1}(x)))^T$.

Note that the computation in (7.12) involves

- the gradients of the basis functions on the reference element which can be precomputed and
- the Jacobian of the transformation which is in general different for each element.

In computer implementations the global stiffness matrix A is computed by looping over all elements $t \in \mathcal{T}$ and computing all entries of A_t at once. With the rectangular $L \times N_h$ matrix R_t defined by

$$(R_t)_{l,i} = \delta_{g(t,l),i}$$

we can write A as a sum of the local stiffness matrices

$$A = \sum_{t \in \mathcal{T}} R_t^T A_t R_t \,.$$

The matrix A is sparse due to the locality of the basis functions. Element t only contributes to the entry $(A)_{i,j}$ if $\operatorname{supp} \varphi_i \cap \operatorname{supp} \varphi_j \cap t \neq \emptyset$. The number of entries per row in the matrix A is bounded by a constant depending on the polynomial degree and the maximum number of elements sharing a vertex.

For the load vector we get through similar arguments

$$\begin{aligned}
(b)_i = \int_\Omega f \varphi_i^h \, dx &= \sum_{t \in \mathcal{T}} \sum_{l=0}^{L-1} \delta_{g(t,l),i} \int_t f(x) \hat{\varphi}_l(\mu^{-1}(x)) \, dx \\
&= \sum_{t \in \mathcal{T}} \sum_{l=0}^{L-1} \delta_{g(t,l),i} \int_{\hat{\Omega}} f(\mu(\hat{x})) \hat{\varphi}_l(\hat{x}) \\
&= \sum_{t \in \mathcal{T}} \sum_{l=0}^{L-1} \delta_{g(t,l),i} (b_t)_l
\end{aligned}$$

and

$$b = \sum_{t \in \mathcal{T}} R_t^T b_t \,.$$

Residual Formulation The formulation of the finite element method in DUNE is based on the so-called *residual formulation*. With the residual form $r(u,v) = a(u,v) - l(v)$ the discrete problem reads:

$$\text{Find } u_h \in V_h\text{:} \qquad r(u_h, v) = 0 \qquad \forall v \in V_h\,.$$

With the map $R : \mathbb{R}^{N_h} \to \mathbb{R}^{N_h}$ given by

$$(R(z))_i = r\left(\sum_{j=1}^{N_h} z_j \varphi_j, \varphi_i\right)$$

the discrete problem is equivalent to the algebraic problem

$$R(z) = 0\,.$$

An advantage of the residual formulation is that it can be readily used for nonlinear PDEs. In that case the algebraic problem is solved, e.g., by Newton's method. In case of a linear PDE we have $R(z) = Az - b$ and Newton's method converges in one step with the Jacobian matrix $\nabla R(z) = A$.

7.7. Case Studies

In this section we illustrate the convergence behavior of the finite element method, i.e.

$$\|u - u_h\|_{j,\Omega} \to 0 \qquad \text{for } h \to 0$$

for two different test problems. As norms we investigate $j = 0$ (the L^2-norm) and $j = 1$ (the H^1-norm). We will prove in the next chapter that the convergence is of the form

$$\|u - u_h\|_{j,\Omega} \leq Ch^\beta$$

where the exponent β depends on the polynomial degree, the norm in which the error is measured and the regularity of the solution. The exponent β is called the *convergence rate* of the method (with respect to a given norm).

Fully Regular Problem The first test problem uses a solution that is in $H^k(\Omega)$ for any $k \geq 1$.

Example 7.13 (Full Regularity Problem)**.** We consider the Poisson problem in two space dimensions

$$-\Delta u = 0 \qquad \text{in } \Omega = (0,2)^2$$

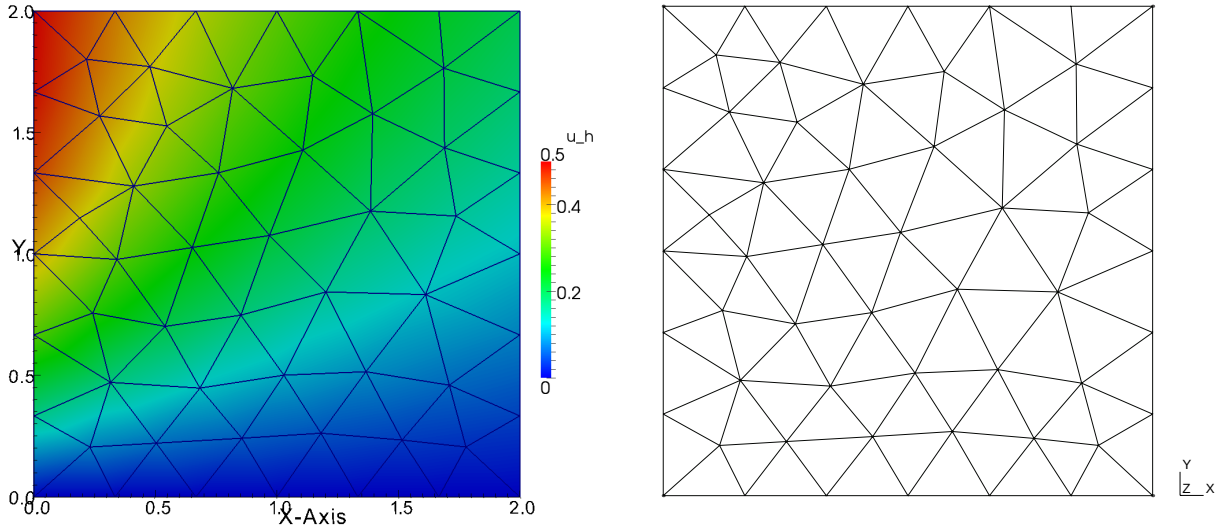


Figure 7.2.: Solution and simplicial coarse mesh for the full regularity example.

with the exact solution

$$u(x,y) = \frac{2y}{(2+x)^2 + y^2}$$

which is taken from [Elman et al., 2005, Example 1.1.3] The exact solution is taken as Dirichlet boundary data. Figure 7.2 shows the coarsest unstructured simplicial mesh generated with `Gmsh`[1], [Geuzaine and Remacle, 2009], used in the computations and the solution as a color plot.

Table 7.1 gives the error in the L^2-norm and the H^1-seminorm for polynomial degree $k = 1, \ldots, 5$ on different meshes. We observe that the convergence rate is $\beta = k + 1$ in L^2 and $\beta = k$ in the H^1-seminorm. We will prove that these convergence factors are optimal.

Figure 7.3 shows plots of the L^2-error $\|u - u_h\|_{0,\Omega}$ in the solution on the coarsest mesh using P_1, P_2, P_3 and P_4. It can be seen that the error has the same overall structure and is just scaled (note the legend!).

Figure 7.4 shows the errors $\|u - u_h\|_{0,\Omega}$ and $\|\nabla(u - u_h)\|_{0,\Omega}$ with respect to (inverse) mesh size h^{-1} for different polynomial degrees $k = 1, \ldots, 5$.

Figure 7.5 compares the errors $\|u - u_h\|_{0,\Omega}$ and $\|\nabla(u - u_h)\|_{0,\Omega}$ with respect to the number of degrees of freedom for P_1, P_2, Q_1 and Q_2. For a given number of degrees of freedom the plot shows that the Q_k solution is slightly more accurate than the P_k solution. □

Reentrant Corner Problem In this subsection we consider a problem where the solution is less regular.

[1] http://geuz.org/gmsh/

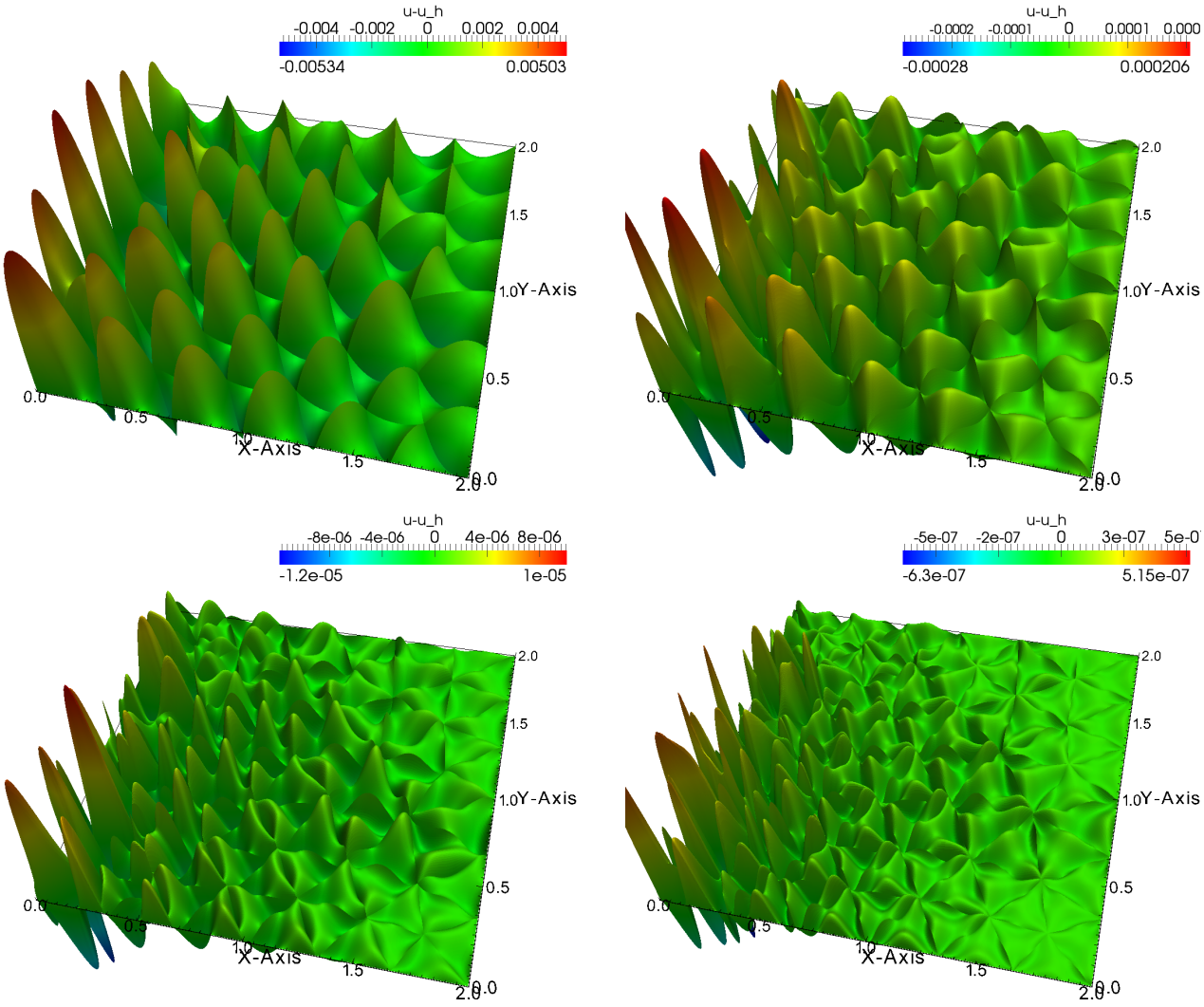


Figure 7.3.: L^2-error in the solution on the coarsest mesh using polynomial degree $k = 1, 2, 3, 4$ (top to bottom , left to right).

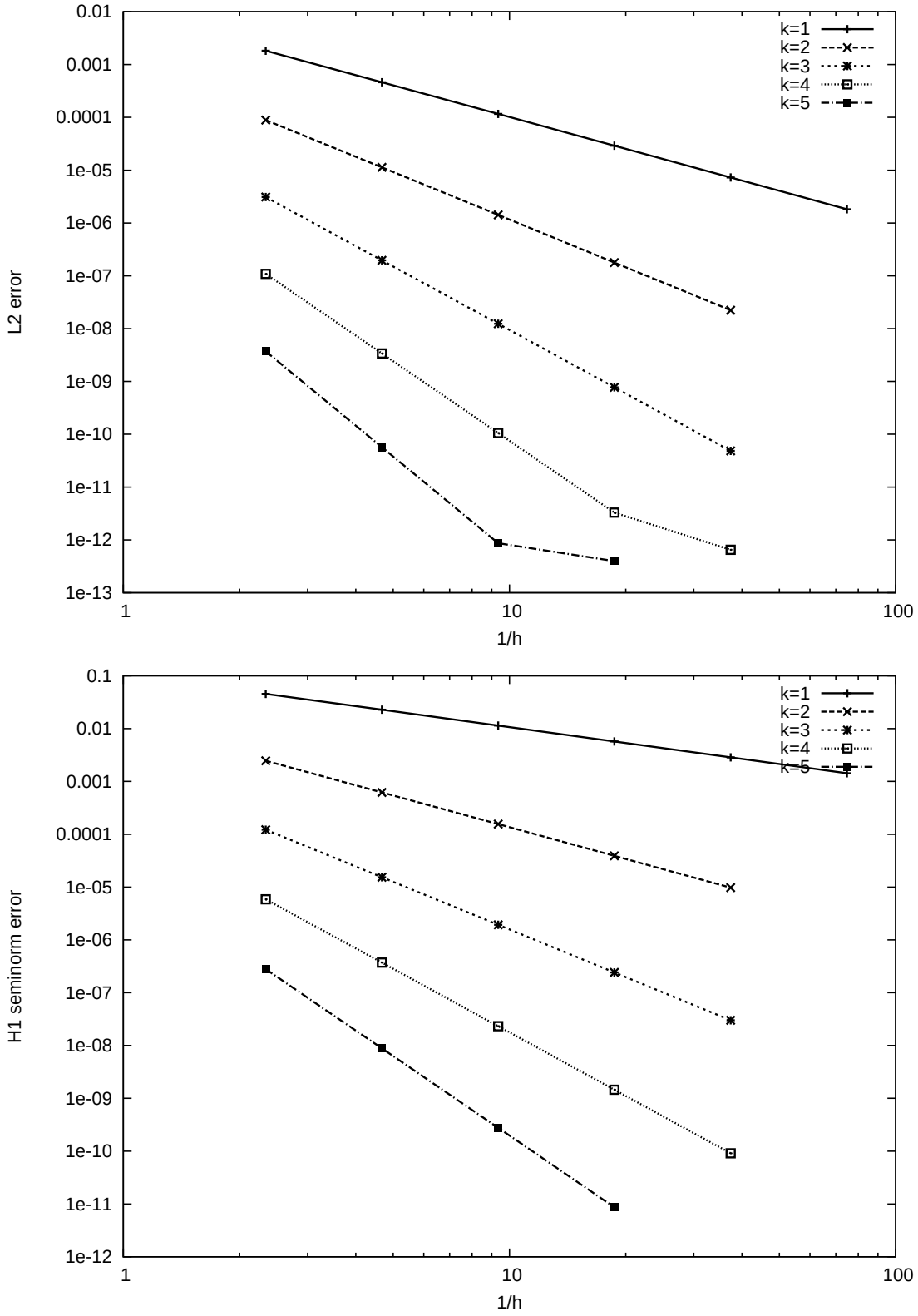


Figure 7.4.: The errors $\|u - u_h\|_{0,\Omega}$ (top) and $\|\nabla(u - u_h)\|_{0,\Omega}$ (bottom) for h- and p-refinement.

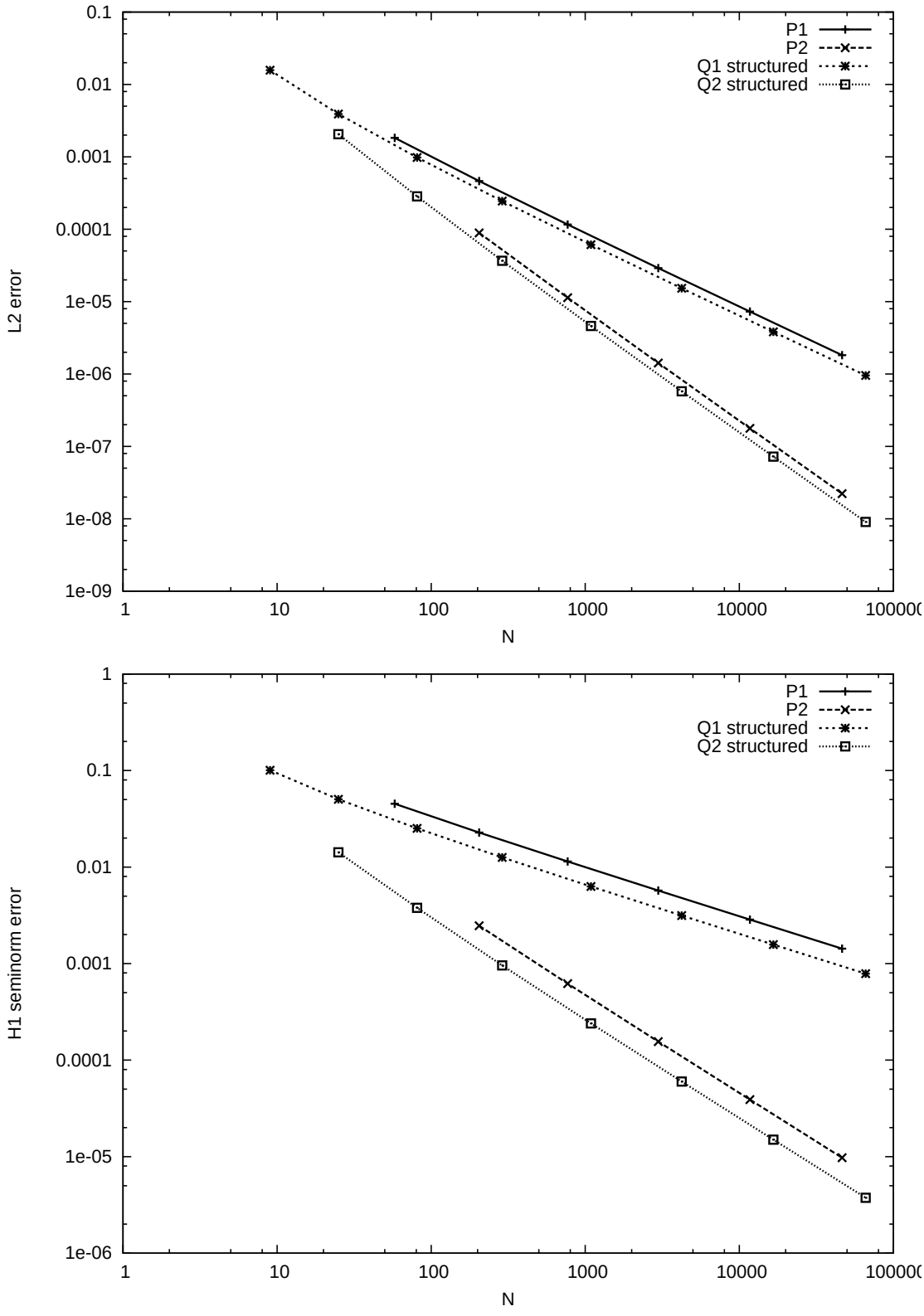


Figure 7.5.: Comparison of $\|u-u_h\|_{0,\Omega}$ (top) and $\|\nabla(u-u_h)\|_{0,\Omega}$ (bottom) using P_1, P_2 and Q_1, Q_2. Note that errors are shown with respect to the number of degrees of freedom.

Table 7.1.: Convergence rates for the example with full regularity using P_k finite elements.

N	$\|u-u_h\|_{0,\Omega}$	L^2-rate	$\|u-u_h\|_{1,\Omega}$	H^1-rate
		$k=1$		
58	1.8313e-03		4.5348e-02	
205	4.6209e-04	1.9866	2.2805e-02	0.99173
769	1.1610e-04	1.9928	1.1423e-02	0.99734
2977	2.9084e-05	1.9970	5.7149e-03	0.99919
11713	7.2765e-06	1.9989	2.8579e-03	0.99976
46465	1.8196e-06	1.9996	1.4290e-03	0.99993
		$k=2$		
205	8.9060e-05		2.4620e-03	
769	1.1318e-05	2.9762	6.2027e-04	1.9889
2977	1.4212e-06	2.9934	1.5556e-04	1.9955
11713	1.7793e-07	2.9978	3.8943e-05	1.9980
46465	2.2255e-08	2.9991	9.7419e-06	1.9991
		$k=3$		
442	3.1084e-06		1.2183e-04	
1693	1.9638e-07	3.9845	1.5378e-05	2.9859
6625	1.2346e-08	3.9915	1.9268e-06	2.9967
26209	7.7398e-10	3.9956	2.4097e-07	2.9993
104257	4.8446e-11	3.9979	3.0124e-08	2.9999
		$k=4$		
769	1.0864e-07		5.8722e-06	
2977	3.3959e-09	4.9996	3.7145e-07	3.9826
11713	1.0548e-10	5.0087	2.3283e-08	3.9958
46465	3.2843e-12	5.0053	1.4561e-09	3.9991
185089	6.4845e-13	2.3405	9.1024e-11	3.9997
		$k=5$		
1186	3.7143e-09		2.7910e-07	
4621	5.6760e-11	6.0321	8.8259e-09	4.9829
18241	8.6560e-13	6.0350	2.7575e-10	5.0003
72481	3.9778e-13	1.1217	8.6568e-12	4.9934

Example 7.14 (Reentrant Corner Problem)**.** We consider the Poisson problem in two space dimensions

$$-\Delta u = 0 \qquad \text{in } \Omega = (-1,1)^2 \setminus [0,1) \times (-1,0]$$

with the exact solution in polar coordinates

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

The exact solution is taken as Dirichlet boundary data. Figure 7.6 shows a color plot of the solution with contour lines and Figure 7.7 shows the two initial meshes generated with `Gmsh` used in the computations.

The exact solution has the regularity $u \in H^{1+\frac{2}{3}}$, see [Hackbusch, 1986, Example 9.7.2]. Table 7.2 shows that the convergence rate in H^1 is $1+2/3-1 = 2/3$ and in L^2 it is $2(1+2/3-1) = 4/3$.

Figure 7.8 shows the L^2-error on a fixed mesh with varying polynomial degree. It illustrates that the error is concentrated near the reentrant corner. Moreover, the error is reduced at a greater rate away from the corner with increasing polynomial degree. This suggests that the local convergence depends on the local regularity of the problem.

Figure 7.9 shows the error in L^2 and H^1 norms plotted against the mesh size for various polynomial degrees. Clearly, the convergence rate is independent of the polynomial degree. Here the uniform initial mesh has been used.

Figure 7.10 shows the error in L^2 and H^1 norms now plotted against the number of degrees of freedom for various polynomial degrees using the uniform mesh and the locally refined mesh. Theses results show that with higher polynomial degree the solution is still more accurate for the same number of degrees of freedom. The second observation is that the locally refined mesh is much more efficient in terms of error per degrees of freedom. Together with the observation from 7.8 this suggests that the most efficient method would be a small mesh size in the vicinity of the singularity and a high polynomial degree away from it. The goal of *hp*-methods is to automatically choose the appropriate mesh size and polynomial degree to reach a prescribed error tolerance with the minimum number of degrees of freedom. □

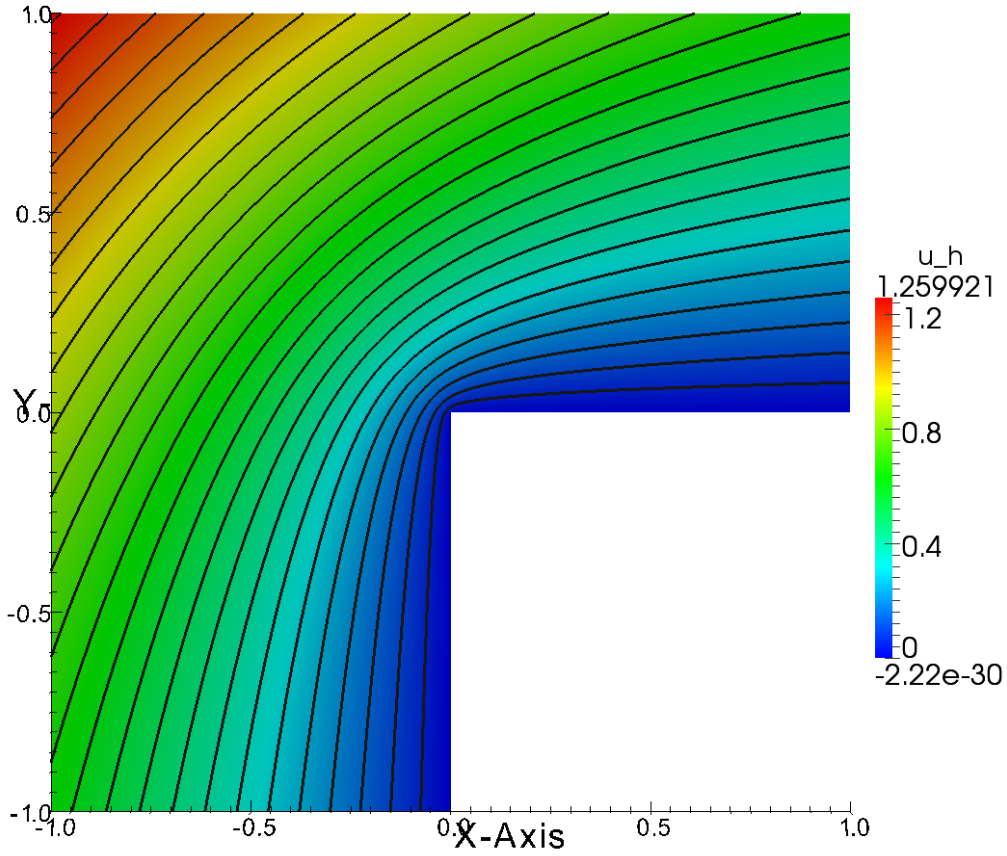


Figure 7.6.: Solution of the L-domain example.

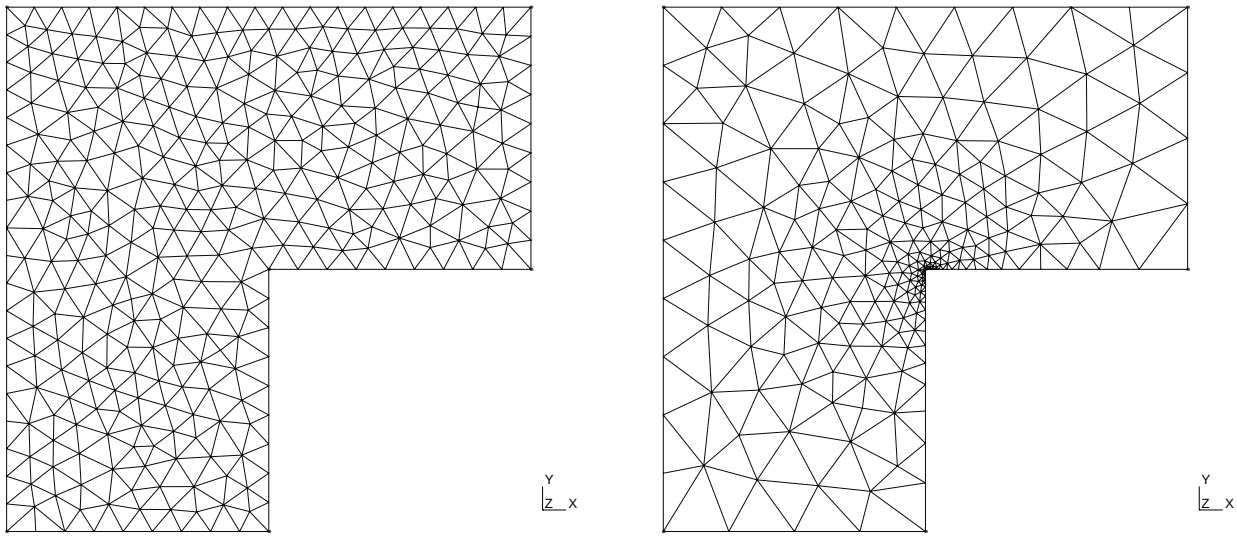


Figure 7.7.: Uniform mesh of the L-shaped domain and mesh with local refinement towards the reentrant corner.

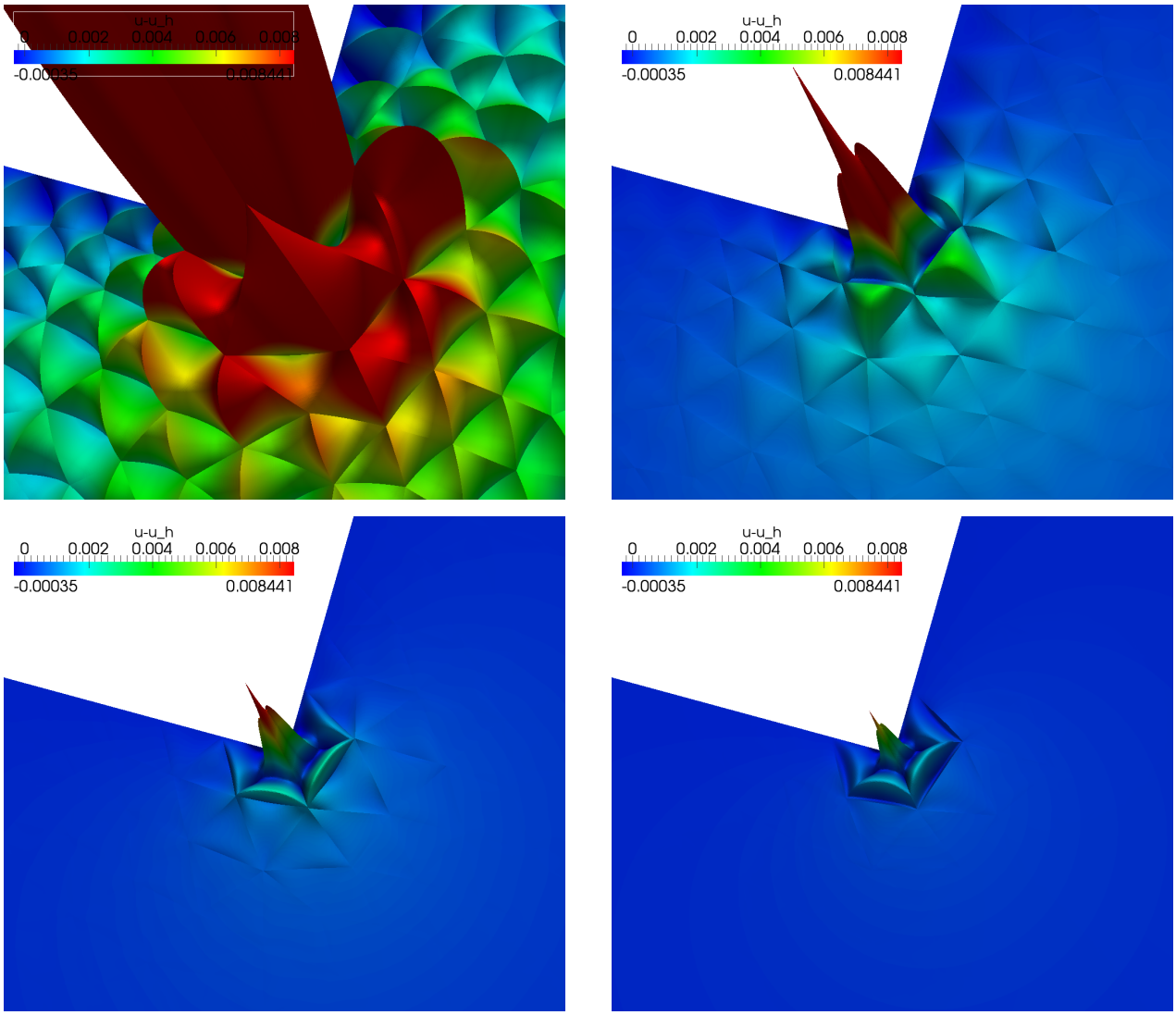


Figure 7.8.: L^2-error in the solution on the coarsest mesh using polynomial degree $k = 1, 2, 3, 4$ (top to bottom , left to right). Note that scaling is the same in all plots!

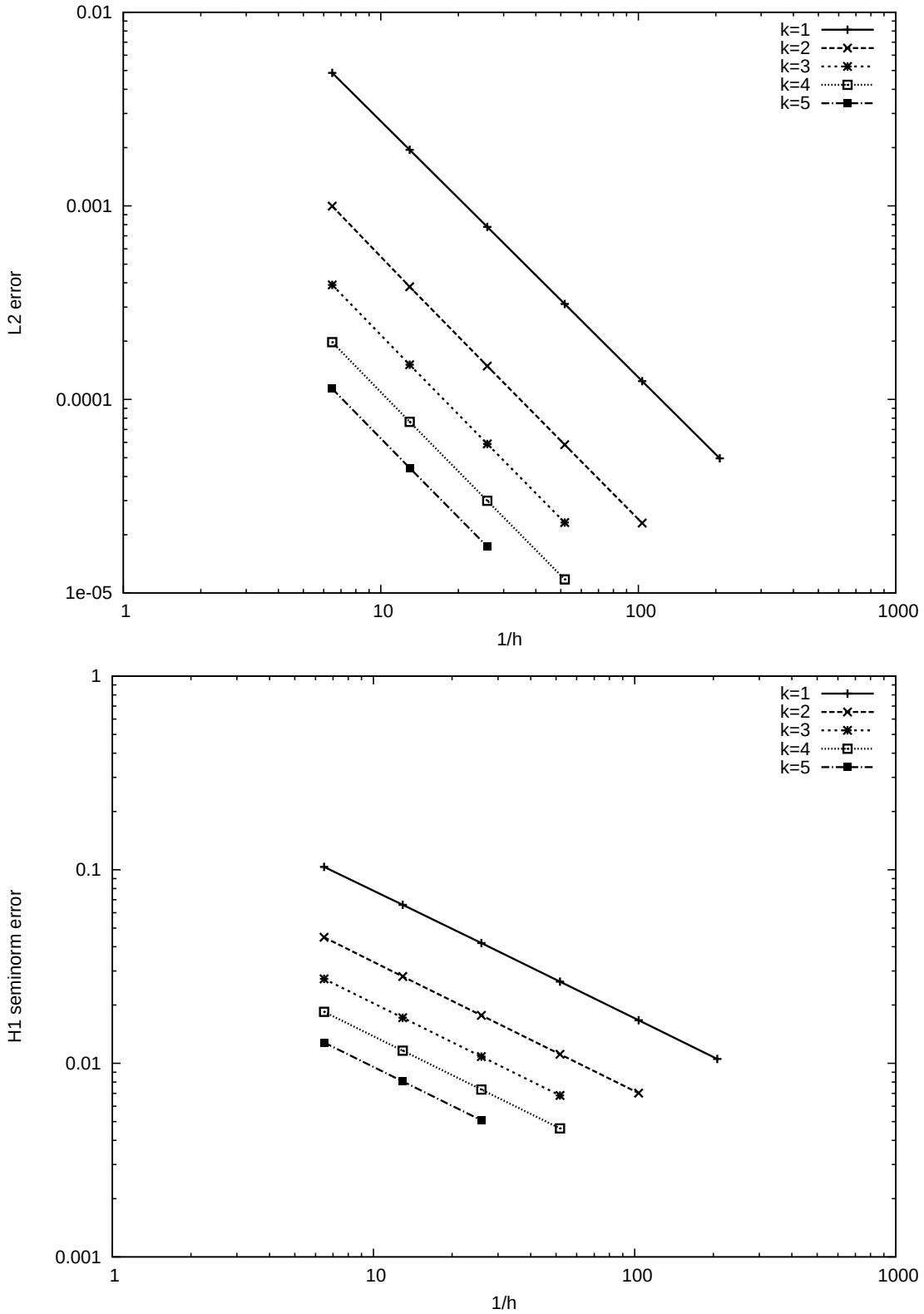


Figure 7.9.: Comparison of $\|u - u_h\|_{0,\Omega}$ (top) and $\|\nabla(u - u_h)\|_{0,\Omega}$ (bottom).

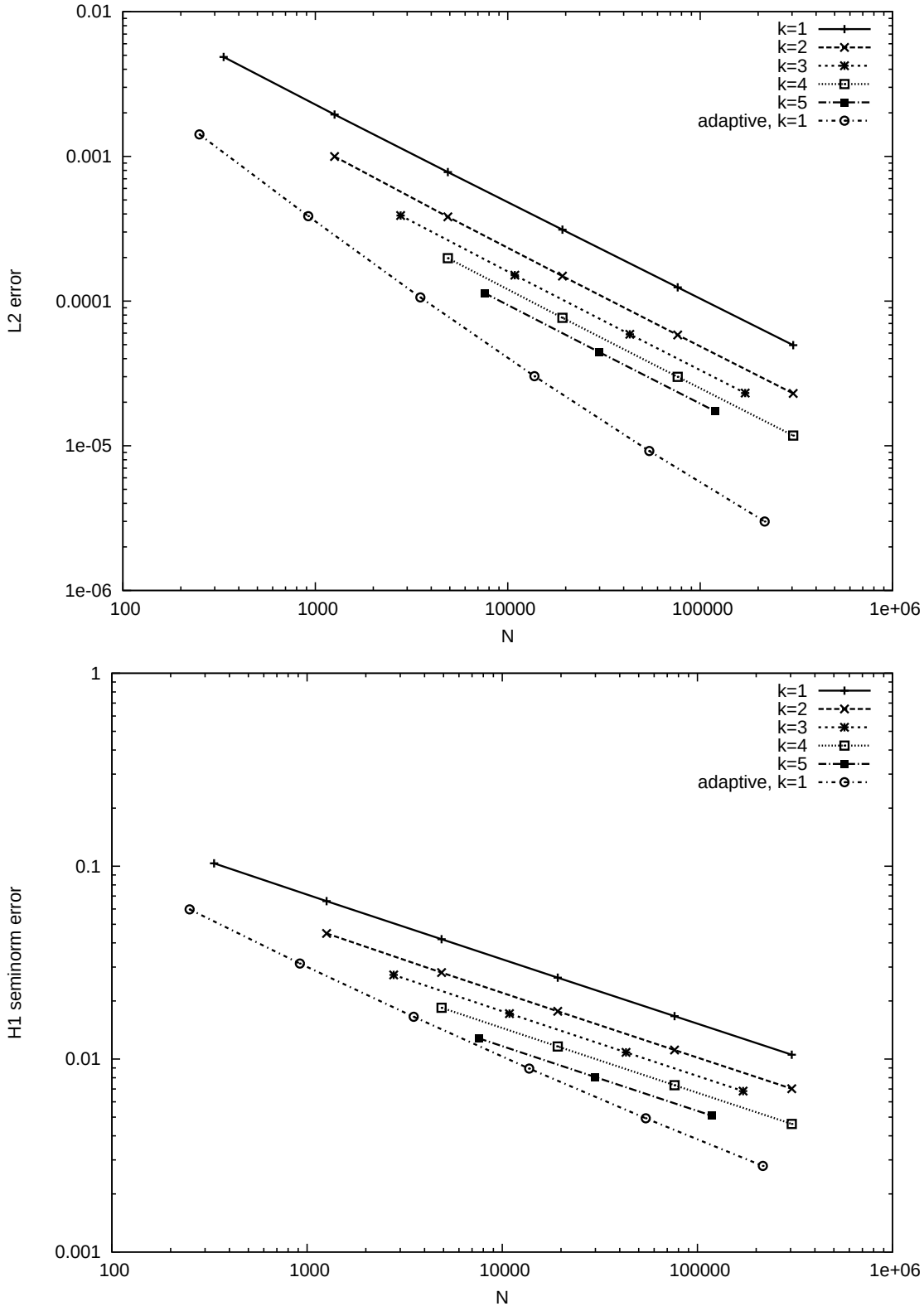


Figure 7.10.: Errors $\|u-u_h\|_{0,\Omega}$ (top) and $\|\nabla(u-u_h)\|_{0,\Omega}$ (bottom) with respect to degrees of freedom.

Table 7.2.: Convergence rates for the L-domain example.

N	$\|u-u_h\|_{0,\Omega}$	L^2-rate	$\|u-u_h\|_{1,\Omega}$	H^1-rate
$k=1$, uniform mesh				
334	4.8614e-03		1.0345e-01	
1259	1.9469e-03	1.3202	6.5916e-02	0.65024
4885	7.7909e-04	1.3213	4.1791e-02	0.65743
19241	3.1151e-04	1.3225	2.6431e-02	0.66099
76369	1.2438e-04	1.3245	1.6692e-02	0.66307
304289	4.9596e-05	1.3265	1.0532e-02	0.66439
$k=1$, adapted mesh				
250	1.4174e-03		5.9661e-02	
919	3.8599e-04	1.8766	3.1300e-02	0.93063
3517	1.0591e-04	1.8657	1.6561e-02	0.91836
13753	3.0263e-05	1.8072	8.9265e-03	0.89164
54385	9.1965e-06	1.7184	4.9287e-03	0.85688
216289	2.9978e-06	1.6172	2.7940e-03	0.81891

Chapter 8.

Finite Element Convergence Theory

Now we turn to the question how good $u_h \in V_h$ approximates $u \in V$.

Observation 8.1 (Galerkin Orthogonality)**.** Let $V_h \subset V$ and let u, u_h denote the solution of the problems

$$u \in V: \qquad a(u,v) = l(v) \qquad \forall v \in V\ , \tag{8.1}$$

$$u_h \in V_h: \qquad a(u_h,v) = l(v) \qquad \forall v \in V_h\ . \tag{8.2}$$

Then we have

$$a(u-u_h, v) = 0 \qquad \forall v \in V_h\,. \tag{8.3}$$

Proof. Substract (8.1) from (8.2). □

Property (8.3) is called *Galerkin orthogonality*. Though very simple to proof this property is an important ingredient in many proofs.

Lemma 8.2 (Céa)**.** Let $V_h \subset V$ and let u, u_h denote the solution of problems (8.1) and (8.2), respectively. Then we have

$$\|u-u_h\|_V \leq \frac{C}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V$$

where α, C are the coercivity and stability constants of the bilinear form a.

Proof. For any $v_h \in V_h$ we have

$$\begin{aligned}
\alpha\|u-u_h\|_V^2 &\leq a(u-u_h, u-u_h) = a(u-u_h, u-v_h+v_h-u_h) \\
&= a(u-u_h, u-v_h) + \underbrace{a(u-u_h, v_h-u_h)}_{=0 \text{ because of Galerkin}} \\
&\leq C\|u-u_h\|_V\|u-v_h\|_V \\
\Leftrightarrow \|u-u_h\|_V &\leq \frac{C}{\alpha}\|u-v_h\|_V \\
\Rightarrow \|u-u_h\|_V &\leq \frac{C}{\alpha} \inf_{v_h \in V_h} \|u-v_h\|_V
\end{aligned}$$

since $v_h \in V_h$ was arbitrary. □

The Lemma of Céa reduces the error estimate $\|u-u_h\|_V$ to the question of approximation of u in the space V_h.

Nonhomogeneous Dirichlet Boundary Conditions This case needs some extra care. Consider the setting of section 6.1 and set $W = H^1(\Omega)$, $V = H_0^1(\Omega)$, i.e. $V \subset W$, and denote by $V_h \subset W_h$ the corresponding finite element spaces $V_h \subset V$ and $W_h \subset W$.

By $\mathcal{I}_h : W \to W_h$ we denote the Lagrange interpolation operator. For any $w \in C^0(\overline{\Omega}) \cap H^1(\Omega)$ the Lagrange interpolation is given by

$$\mathcal{I}_h w = \sum_{i=1}^{N} w(a_i)\varphi_i$$

where φ_i is the nodal basis function corresponding to the point $a_i \in \overline{\Omega}$.

Furthermore we need to assume that the boundary condition g is such that there exists a sufficiently smooth extension $u_g \in C^0(\overline{\Omega}) \cap H^1(\Omega)$ of the boundary condition to the interior. Then we define the boundary interpolation

$$\mathcal{I}_h^{\partial\Omega} g = \sum_{a_i \in \partial\Omega} g(a_i)\gamma(\varphi_i)$$

where γ is the trace operator. Clearly, we then have for any $w \in C^0(\overline{\Omega}) \cap H^1(\Omega)$:

$$\gamma(\mathcal{I}_h w) = \gamma\left(\sum_{i=1}^{N} w(a_i)\varphi_i\right) = \sum_{i=1}^{N} w(a_i)\gamma(\varphi_i) = \sum_{a_i \in \partial\Omega} w(a_i)\gamma(\varphi_i) = \mathcal{I}_h^{\partial\Omega}\gamma(w)$$

since $a_i \notin \partial\Omega \Rightarrow \varphi_i|_{\partial\Omega} = 0$.

Now we are in a position to state the discrete problem with inhomogeneous Dirichlet boundary conditions: For an arbitrary extension $u_g \in C^0(\overline{\Omega}) \cap H^1(\Omega)$ choose its finite element interpolation $u_{gh} = \mathcal{I}_h u_g$ and find

$$u_h \in u_{gh} + V_h : \qquad a(u_h, v) = l(v) \qquad \forall v \in V_h.$$

This problem is well-posed due to the Lax-Milgram Theorem by setting $u_h = u_{gh} + u_{0h}$. Moreover, due to $\gamma(u_h) = \gamma(u_{gh}) + \gamma(u_{0h}) = \gamma(\mathcal{I}_h u_g) = \mathcal{I}_h^{\partial\Omega}\gamma(u_g) = \mathcal{I}_h^{\partial\Omega} g$ the boundary conditions are satisfied. Interestingly, Galerkin orthogonality is still valid (only $V_h \subset V$) is required:

$$\begin{aligned} u \in u_g + V : &\qquad a(u, v) = l(v) &&\forall v \in V\ , \\ u_h \in u_{gh} + V_h : &\qquad a(u_h, v) = l(v) &&\forall v \in V_h\ , \end{aligned}$$

from which we conclude $a(u - u_h, v) = 0$ for all $v \in V_h$.

Finally, we can state a relation corresponding to the Céa-Lemma in the case of inhomogenous Dirichlet conditions.

Lemma 8.3. With the notation from above there holds

$$\|u - u_h\|_{1,\Omega} \leq \left(1 + \frac{C}{\alpha}\right) \|u - \mathcal{I}_h u\|_{1,\Omega}. \tag{8.4}$$

Proof. From Galerkin orthogonality we obtain by adding and subtracting $\mathcal{I}_h u$:

$$a(\mathcal{I}_h u - u_h, v) = a(\mathcal{I}_h u - u, v) \quad \forall v \in V_h.$$

Using coercivity we get

$$\begin{aligned} \alpha \|\mathcal{I}_h u - u_h\|_{1,\Omega}^2 &\leq a(\mathcal{I}_h u - u_h, \mathcal{I}_h u - u_h) = a(\mathcal{I}_h u - u, \mathcal{I}_h u - u_h) \\ &\leq C \|\mathcal{I}_h u - u\|_{1,\Omega} \|\mathcal{I}_h u - u_h\|_{1,\Omega} \end{aligned}$$

which is equivalent to

$$\|\mathcal{I}_h u - u_h\|_{1,\Omega} \leq \frac{C}{\alpha} \|\mathcal{I}_h u - u\|_{1,\Omega}.$$

Finally, we conclude using the triangle inequality

$$\begin{aligned} \|u - u_h\|_{1,\Omega} &= \|u - \mathcal{I}_h u + \mathcal{I}_h u - u_h\|_{1,\Omega} \\ &\leq \|u - \mathcal{I}_h u\|_{1,\Omega} + \|\mathcal{I}_h u - u_h\|_{1,\Omega} \\ &\leq \|u - \mathcal{I}_h u\|_{1,\Omega} + \frac{C}{\alpha} \|\mathcal{I}_h u - u\|_{1,\Omega} \\ &= \left(1 + \frac{C}{\alpha}\right) \|u - \mathcal{I}_h u\|_{1,\Omega}. \end{aligned}$$

Note that in the case of inhomogeneous Dirichlet boundary conditions we have to make explicit use of the Lagrange interpolation operator in contrast to the homogeneous case.

□

8.1. Bramble Hilbert Lemma

We now study the approximation error

$$\inf_{v_h \in V_h} \|u - v_h\|_V$$

where $V_h \subset V$ is some finite dimensional subspace of a Sobolev space V. In order to obtain estimates in terms of some power of the mesh size h, additional regularity, i.e. $u \in W$ with $W \subset V$ has to be assumed. In our concrete application a typical setting would be

$$P_k(\mathcal{T}_h) \subset V \subseteq H^1(\Omega) \supset W = H^m(\Omega) \ .$$

Proposition 8.4. Let $\Omega \subset \mathbb{R}^n$ be a subdomain with Lipschitz boundary satisfying a cone condition. Moreover, for $m \in \mathbb{N}$, $m > \frac{n}{2}$ and let $\mathcal{I}_{m-1} : H^m(\Omega) \to \mathbb{P}^n_{m-1}$ be the Lagrange interpolation operator corresponding to certain points $s_1, \ldots, s_N \in \overline{\Omega}$, $N = \dim \mathbb{P}^n_{m-1}$. Then there exists a constant $c = c(\Omega, s_1, \ldots, s_N)$ such that

$$\|u - \mathcal{I}_{m-1}u\|_m \leq c|u|_m \qquad \forall u \in H^m(\Omega) \ . \tag{8.5}$$

Proof. See [Braess, 2003, Hilfssatz 6.2].

First observe that Lagrange interpolation is well-defined for $m > \frac{n}{2}$ due to the Sobolev embedding theorem 5.46. Then define the special norm

$$|||v||| = |v|_m + \sum_{i=1}^{N} |v(s_i)| \ .$$

We would like to prove that there exists $c \in \mathbb{R}$ such that

$$\|v\|_m \leq c\,|||v||| \qquad \forall v \in H^m(\Omega) \ . \tag{8.6}$$

The proof is by contradiction. Assume that (8.6) does not hold. Then there exists a sequence $\{v_k : k \in \mathbb{N}\}$ such that

$$\|v_k\|_m = 1, \quad |||v_k||| \leq \frac{1}{k}, \qquad k \in \mathbb{N} \tag{8.7}$$

(since then $\frac{|||v_k|||}{\|v_k\|_m} \leq \frac{1}{k} \Leftrightarrow \|v_k\|_m \geq k|||v_k|||$.)

Clearly, $H^m(\Omega) \subset H^{m-1}(\Omega)$ and this embedding is compact (Rellich-Kondrachov theorem [Adams, 1978, chapter VI]). This means that a subsequence $\{v_{k_i}\}$ can be selected that converges to (a possibly different) $v \in H^{m-1}(\Omega)$. Without loss of generality let us assume that $\{v_k\}$ already is that subsequence, i.e. $v = \lim_{k\to\infty} v_k \in H^{m-1}(\Omega)$. Then

$$\|v_k - v_l\|_m^2 = \underbrace{\|v_k - v_l\|_{m-1}^2}_{\substack{\to 0 \text{ since } \{v_k\} \to v \text{ in } H^{m-1} \text{ is Cauchy seq.}}} + \underbrace{|v_k - v_l|_m^2}_{\substack{\to 0 \text{ since } |v_k - v_l|_m \leq |||v_k - v_l||| \\ \leq |||v_k||| + |||v_l||| \leq \frac{1}{k} + \frac{1}{l} \to 0 \text{ for } l,k \to \infty}}$$

So $\{v_k\}$ is a Cauchy sequence in $H^m(\Omega)$ and so we even have $v \in H^m(\Omega)$. Due to the continuity of the norms $\|.\|$ and $|||.|||$ we conclude from (8.7) that

$$\|v\|_m = 1 \qquad \text{and} \qquad |||v||| = 0 \ .$$

From $|||v||| = |v|_m + \sum_{i=1}^{N} |v(s_i)| = 0$ it follows that

a) $|v|_m = 0 \Rightarrow v \in \mathbb{P}^n_{m-1}$ and

b) $v(s_i) = 0 \; \forall i = 1, \ldots, N \overset{\text{a)}}{\Rightarrow} v \equiv 0.$

This is a contradiction to $\|v\|_m = 1$. □

Lemma 8.5 (Bramble-Hilbert)**.** $\Omega \subset \mathbb{R}^n$ fulfills the conditions from Proposition 8.4. For $m > \frac{n}{2}$ let $L : H^m(\Omega) \to Y$ be a bounded linear operator into the normed vector space Y with the canonical norm $\|L\| = \sup_{v \neq 0} \frac{\|Lv\|_Y}{\|v\|_m}$. Furthermore, assume that $\mathbb{P}^n_{m-1} \subseteq \ker L$. Then there exists a constant $c \in \mathbb{R},\ c > 0$ such that

$$\|Lv\|_Y \leq c|v|_m \qquad \forall v \in H^m(\Omega).$$

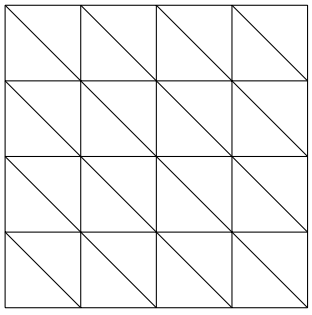
Proof. [Braess, 2003, Lemma 6.3].

$$\begin{aligned}
\|Lv\|_Y &= \|Lv - L\mathcal{I}_{m-1}v\|_Y && (L\mathcal{I}_{m-1}v = 0 \text{ since } \mathcal{I}_{m-1}v \in \mathbb{P}^n_{m-1}) \\
&= \|L(v - \mathcal{I}_{m-1}v)\|_Y && (\text{linearity}) \\
&\leq \|L\|\|v - \mathcal{I}_{m-1}v\|_m && (L \text{ bounded linear operator}) \\
&\leq c\|L\||v|_m && (\text{Proposition 8.4})
\end{aligned}$$

with c the constant from Proposition 8.4. □

In our application we will set $L = I - \mathcal{I}_k$ (I the identity) for some polynomial degree k. For any $v \in \mathbb{P}^n_k$ we then have $Lv = Iv - \mathcal{I}_k v = v - v = 0$, i.e. $\mathbb{P}^n_k \in \ker L$. Assuming that $n \leq 3$, Bramble-Hilbert holds with $m = k+1 \geq 2 \Leftrightarrow k \geq 1$, i.e. the smallest possible choice would be $\mathbb{P}^n_1$ and $H^2(\Omega)$.

8.2. Approximation Results



We first consider the special case of a uniform grid in $n = 2$ with the special form shown to the left.

Proposition 8.6. For $\Omega = (0,1)^2$ consider the uniform, conforming triangulations $\mathcal{T}_h,\ h = \frac{1}{\nu},\ \nu \in \mathbb{N}$ where each $t \in \mathcal{T}_h$ is generated by a mapping μ_t from the reference triangle $\hat{S}_2$ of the form

$$\mu_t(\hat{x}) = h d_t \hat{x} + b_t \qquad \text{with} \qquad d_t = \pm 1 \ .$$

For $k \in \mathbb{N}$, $k > \min(1, n/2)$, let $P_{k-1}(\mathcal{T}_h)$ be the conforming finite element space of piecewise polynomials of degree $k-1$ and $\mathcal{I}_h$ the corresponding Lagrange interpolation operator for the mesh $\mathcal{T}_h$. Then for any $u \in H^k(\Omega)$ the interpolation error measured in the $H^m(\Omega)$-norm, $0 \le m \le 1$, can be bounded by

$$\|u - \mathcal{I}_h u\|_{m,\Omega} \le c h^{k-m} |u|_{k,\Omega} \qquad \forall u \in H^k(\Omega) \; .$$

Proof.

a) The transformation μ_t has the form $\mu_t(\hat{x}) = h d_t \hat{x} + b_t$, so $\hat{x} = \mu_t^{-1}(x) = h^{-1} d_t^{-1}(x - b_t)$.

b) Transformation of derivatives. For $v \in H^k(\Omega)$ and any $t \in \mathcal{T}_h$ set $\hat{v}(\hat{x}) = v(\mu_t(\hat{x}))$. By the chain rule one gets for any multiindex $|\alpha| \le k$:

$$\hat{\partial}^\alpha \hat{v}(\hat{x}) = h^{|\alpha|} d_t^{|\alpha|} \partial^\alpha v(\mu_t(\hat{x})) \, .$$

c) On a single triangle $t \in \mathcal{T}_h$ we obtain for $0 \le l \le k$:

$$\begin{aligned}
|\hat{v}|^2_{l,\hat{S}_2} &= \sum_{|\alpha|=l} \int_{\hat{S}_2} (\hat{\partial}^\alpha \hat{v}(\hat{x}))^2 d\hat{x} \\
&= \sum_{|\alpha|=l} \int_{\hat{S}_2} h^{2l} |d_t|^{2l} (\partial^\alpha v(\mu_t(\hat{x}))^2 \, d\hat{x} \qquad \text{(insert transformation rule)} \\
&= h^{2l} |d_t|^{2l} \sum_{|\alpha|=l} \int_t (\partial^\alpha v[\mu_t(\mu_t^{-1}(x))]^2 \underbrace{|\nabla \mu^{-1}(x)|}_{h^{-n}|d_t|^{-n}} \, dx \\
&= h^{2l-n} |v|^2_{l,t} \; . \qquad \text{(since } |d_t|^{2l-n} = 1\text{)}
\end{aligned}$$

With the same argument one obtains for $v(x) = \hat{v}(\mu_t^{-1}(x))$

$$|v|^2_{l,t} = h^{n-2l} |\hat{v}|^2_{l,\hat{S}_2} \; .$$

d) With these preparations we obtain for the interpolation error on a single element $t \in \mathcal{T}_h$ for any $0 \le l \le m$

$$\begin{aligned}
|u - \mathcal{I}_t u|^2_{l,t} &= h^{n-2l} |\hat{u} - \mathcal{I}_{\hat{S}_2} \hat{u}|^2_{l,\hat{S}_2} && \text{(Transformation to } \hat{S}_2\text{)} \\
&\le h^{n-2l} \|\hat{u} - \mathcal{I}_{\hat{S}_2} \hat{u}\|^2_{l,\hat{S}_2} && \text{(extend to full norm)} \\
&\le h^{n-2l} c |\hat{u}|^2_{k,\hat{S}_2} && \text{(Bramble-Hilbert, } L : H^k(\hat{S}_2) \to H^l(\hat{S}_2)\text{)} \\
&\le h^{n-2l} c h^{2k-n} |u|^2_{k,t} && (\hat{S}_2 \to t) \\
&= c h^{2(k-l)} |u|^2_{k,t} \; .
\end{aligned}$$

Note c depends on Ω, k, but not on l. This argument is known as a "*scaling argument*".

e) H^m-norm on a single triangle

$$\begin{aligned}
\|u - \mathcal{I}_t u\|_{m,t}^2 &= \sum_{l=0}^{m} |u - \mathcal{I}_t u|_{l,t}^2 \\
&\leq \sum_{l=0}^{m} c h^{2(k-l)} |u|_{k,t}^2 \\
&= c|u|_{k,t}^2 h^{2(k-m)} \sum_{l=0}^{m} h^{2(m-l)} \quad (h^{2(k-l)} = h^{2(k-m+m-l)}) \\
&\leq c(m+1)|u|_{k,t}^2 h^{2(k-m)} \qquad (1 + h^2 + \ldots + h^{2m}) \leq m+1,\ h \leq 1.
\end{aligned}$$

f) Now on the whole domain

$$\begin{aligned}
\|u - \mathcal{I}_h u\|_{m,\Omega}^2 &= \sum_{t \in \mathcal{T}_h} \|u - \mathcal{I}_t u\|_{m,t}^2 \\
&\leq C h^{2(k-m)} \sum_{t \in \mathcal{T}_h} |u|_{k,t}^2 \\
&= C h^{2(k-m)} |u|_{k,\Omega}^2
\end{aligned}$$

Taking the square root proves the result. □

In the case of a general triangulation $\mathcal{T}_h$ of a polyhedral domain $\Omega \subset \mathbb{R}^n$ only steps b) and c) of the proof above get more complicated and technical (the factor $d_t^{|\alpha|}$ will be replaced), but Proposition 8.6 remains true with a different constant c.

Tensor Products We prepare the general approximation result with a few technical lemmata.

Definition 8.7 (Tensor product)**.** Given m vectors $y_k \in \mathbb{R}^{n_k}$, $1 \leq k \leq m$ (the dimension n_k may be different for each vector), then

$$y = \bigotimes_{k=1}^{m} y_k \in \mathbb{R}^N, \qquad N = \prod_{k=1}^{m} n_k$$

with

$$(y)_{i_1,\ldots,i_m} = \prod_{k=1}^{m} (y_k)_{i_k}$$

is called the tensor product of the y_k. The indices of components of y are from the set $\mathcal{I} = \{1,\ldots,n_1\} \times \ldots \times \{1,\ldots,n_m\}$ (which is the Cartesian product of the individual index sets). □

Example: Given two vectors $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, then $x \otimes y \in \mathbb{R}^{nm}$ and $(x \otimes y)_{i,j} = x_i y_j$. In other words, the entries of the matrix xy^T are written into a single vector.

For the Euclidean norm of a tensor product we have

$$\begin{aligned}\left\| \bigotimes_{k=1}^{m} y_k \right\|_{\mathbb{R}^N}^2 &= \sum_{(i_1,\ldots,i_m)\in\mathcal{I}} \left| \prod_{k=1}^{m} (y_k)_{i_k} \right|^2 = \sum_{i_1=1}^{n_1} \cdots \sum_{i_m=1}^{n_m} (y_1)_{i_1}^2 \cdot \ldots \cdot (y_m)_{i_m}^2 \\ &= \sum_{i_1=1}^{n_1} (y_1)_{i_1}^2 \left(\sum_{i_2=1}^{n_2} \cdots \sum_{i_m=1}^{n_m} (y_2)_{i_2}^2 \ldots (y_m)_{i_m}^2 \right) \\ &= \left(\sum_{i_2=1}^{n_2} \cdots \sum_{i_m=1}^{n_m} (y_2)_{i_2}^2 \ldots (y_m)_{i_m}^2 \right) \|y_1\|^2 \\ &= \|y_1\|^2 \cdot \|y_2\|^2 \cdot \ldots \cdot \|y_m\|^2 = \prod_{k=1}^{m} \|y_k\|^2 \ .\end{aligned}$$

Taking the square root we obtain

$$\left\| \bigotimes_{k=1}^{m} y_k \right\|_{\mathbb{R}^N} = \prod_{k=1}^{m} \|y_k\| \ .$$

We now use tensor products to represent derivatives. For $v \in H^1(\Omega)$ we define the first order differential operator

$$L[y] = \sum_{i=1}^{n} (y)_i \partial_{x_i}$$

taking linear combinations of partial derivatives. So we have

$$L[y]v(x) = \sum_{i=1}^{n} (y)_i \partial_{x_i} v(x).$$

Especially for $y = e_i$ (the ith unit vector) we get $L[e_i]v(x) = \partial_{x_i} v(x)$. We now extend this to higher order derivatives.

Definition 8.8 (Derivatives as Multilinear Form)**.** For $y_1, \ldots, y_m \in \mathbb{R}^n$

$$L[y_1, \ldots, y_m] = \prod_{k=1}^{m} \left(\sum_{i_k=1}^{n} (y_k)_{i_k} \partial_{i_k} \right).$$

is called a *multilinear form* (we have set $\partial_{i_k} = \partial_{x_{i_k}}$ for ease of writing). □

Example: For $m = n = 2$ and $v \in H^2(\Omega)$ we have

$$\begin{aligned} L[y_1, y_2]v(x) &= \left((y_1)_1\partial_1 + (y_1)_2\partial_2\right)\left((y_2)_1\partial_1 + (y_2)_2\partial_2\right)v(x) \\ &= (y_1)_1(y_2)_1\partial_1^2 v(x) + (y_1)_1(y_2)_2\partial_1\partial_2 v(x) \\ &\quad + (y_1)_2(y_2)_1\partial_2\partial_1 v(x) + (y_1)_2(y_2)_2\partial_2^2 v(x) \ . \end{aligned}$$

For $y_k = e_{i_k}$ we get $L[e_{i_1}, \ldots, e_{i_m}]v(x) = \partial_{i_1} \ldots \partial_{i_m} v(x)$. So we are able to represent arbitrary higher order derivatives with multilinear forms (although the representation is not unique).

We then have the following important estimate:

$$\begin{aligned} |L[y_1, \ldots, y_m]v(x)| &= \left| \left[\prod_{k=1}^{m} \left(\sum_{i_k=1}^{n} (y_k)_{i_k} \partial_{i_k} \right) \right] v(x) \right| \\ &= |\left((y_1)_1\partial_1 + \ldots + (y_1)_n\partial_n\right) \cdot \\ &\qquad \ldots \cdot \left((y_m)_1\partial_1 + \ldots + (y_m)_n\partial_n\right) v(x))| \\ &= \left| \sum_{(i_1,\ldots,i_m)\in\mathcal{I}} (y_1)_{i_1} \ldots (y_m)_{i_m} \partial_{i_1}\partial_{i_2} \ldots \partial_{i_m} v(x) \right| \qquad (8.8) \\ &\leq \left\| \bigotimes_{k=1}^{m} y_k \right\| \|D^m v(x)\| \qquad \text{(Cauchy-Schwarz in } \mathbb{R}^n) \\ &= \|D^m v(x)\| \prod_{k=1}^{m} \|y_k\| \end{aligned}$$

where $D^m v(x)$ denotes the vector (ordered set) of all n^m partial derivatives $\partial^\alpha v(x)$ with $|\alpha| = m$ where the different permutations of a multiindex $(\alpha_1, \ldots, \alpha_n)$ are distinguished.

The Chain Rule for Affine Transformations For $\mu_t : \hat{S}_n \to \bar{t},\ \mu_t(\hat{x}) = B_t\hat{x} + b_t$ and $v : \Omega \to \mathbb{R},\ \hat{v} : \hat{S}_n \to \mathbb{R}$ set $\hat{v}(\hat{x}) = v(\mu_t(\hat{x}))$. The chain rule then yields

$$\hat{\partial}_i \hat{v}(\hat{x}) = \hat{\partial}_i v(\mu_t(\hat{x})) = \sum_{l=1}^{n} \partial_l v(\mu_t(\hat{x}))\, \hat{\partial}_i \mu_{t,l}(\hat{x}) = \sum_{l=1}^{n} \partial_l v(\mu_t(\hat{x}))\, B_{li} \ .$$

Putting all components together we obtain

$$\nabla_{\hat{x}} \hat{v}(\hat{x}) = B^T \nabla_x v(\mu_t(\hat{x})) \ .$$

In order to extend this to higher order derivatives we use the multi-linear forms:

$$\begin{aligned}\hat{L}[y_1,\dots,y_m] &= \prod_{k=1}^{m}\left(\sum_{i_k=1}^{n}(y_k)_{i_k}\hat{\partial}_{i_k}\right) = \prod_{k=1}^{m}\left[\sum_{i_k=1}^{n}(y_k)_{i_k}\underbrace{\left(\sum_{l_k=1}^{n}B_{l_k i_k}\partial_{l_k}\right)}_{\text{chain rule applied to }\hat{\partial}_{i_k}}\right] \\ &= \prod_{k=1}^{m}\left[\sum_{l_k=1}^{n}\partial_{l_k}\underbrace{\sum_{i_k=1}^{n}B_{l_k i_k}(y_k)_{i_k}}_{=(By_k)_{l_k}}\right] = L[By_1,\dots,By_m]\end{aligned} \tag{8.9}$$

Together with 8.8 we obtain the following the estimate:

$$\begin{aligned}|\hat{L}[y_1,\dots,y_m]\hat{v}(\hat{x})| &= |L[By_1,\dots,By_m]v(\mu_t(\hat{x}))| \\ &\leq \left(\prod_{k=1}^{m}\|By_k\|\right)\|D^m v(\mu_t(\hat{x}))\| \\ &\leq \|B\|^m\left(\prod_{k=1}^{m}\|y_k\|\right)\|D^m v(\mu_t(\hat{x}))\| \ .\end{aligned} \tag{8.10}$$

Now we can prove the following transformation formulae.

Proposition 8.9. Let $\mu_t : \hat{S}_n \to \Omega$, $\Omega \subset \mathbb{R}^n$, be an affine linear transformation with Jacobian B_t and $\hat{v}(\hat{x}) = v(\mu_t(\hat{x}))$ for $v \in H^m(t)$ and $\hat{v} \in H^m(\hat{S}_n)$. Then we have

$$|\hat{v}|_{m,\hat{S}_n} \leq n^m\|B_t\|^m|\det B_t|^{-\frac{1}{2}}|v|_{m,t} \ , \tag{8.11a}$$

$$|v|_{m,t} \leq n^m\|B_t^{-1}\|^m|\det B_t|^{\frac{1}{2}}|\hat{v}|_{m,\hat{S}_n} \ . \tag{8.11b}$$

Proof. The proof is a generalization of Proposition 8.6 b).

$$
\begin{aligned}
|\hat{v}|^2_{m,\hat{S}_n} &= \int_{\hat{S}_n} \sum_{|\alpha|=m} |\hat{\partial}^\alpha \hat{v}(\hat{x})|^2 d\hat{x} \\
&= \int_{\hat{S}_n} \sum_{|\alpha|=m} |\hat{L}[e_{\alpha_1}, \ldots, e_{\alpha_m}]\hat{v}(\hat{x})|^2 d\hat{x} && \text{(multilinear form)} \\
&\leq \int_{\hat{S}_n} \sum_{|\alpha|=m} \|D^m v(\mu_t(\hat{x}))\|^2 \|B_t\|^{2m} \prod_{k=1}^{m} \|e_{i_k}\|^2 d\hat{x} && \text{(use estimate 8.10)} \\
&= \|B_t\|^{2m} n^m \int_{\hat{S}_n} \underbrace{\|D^m v(\mu(\hat{x}))\|^2}_{\sum_{(i_1,\ldots,i_m)\in\mathcal{I}} |\partial_{i_1}\ldots\partial_{i_m} v(\mu(\hat{x}))|^2} d\hat{x} && (\sum_{|\alpha|=m} 1 \leq n^m) \\
&\leq \|B_t\|^{2m} n^m n^m \int_{\hat{S}_n} \sum_{|\alpha|=m} |\partial^\alpha v(\mu(\hat{x}))|^2 d\hat{x} && (\#\text{ permut.} \leq n^m) \\
&= \|B_t\|^{2m} n^{2m} \int_{t} \sum_{|\alpha|=m} |\partial^\alpha v(\mu(\mu^{-1}(x)))|^2 |\det B_t^{-1}| dx && \text{(transform to } t) \\
&= \|B_t\|^{2m} n^{2m} |\det B_t|^{-1} |v|^2_{m,t} \ .
\end{aligned}
$$

Taking the square root proves the result. The second estimate (8.11b) is shown in the same way. □

Shape Regular Affine Transformations Let $\mu_t : B_t\hat{x} + b_t$ be the affine linear transformation from a reference simplex $\hat{S}_n$ to an element $t \in \mathcal{T}_h$.

We now prove estimates of the spectral norms (associated matrix norm for the Euclidean norm) $\|B\|$ and $\|B^{-1}\|$. Let $\rho(t)$ and $h(t)$ denote the diameter of largest ball inscribed in t and the maximum distance of any two points in t, respectively. $\rho(\hat{S}_n)$ and $h(\hat{S}_n)$ denote the same quantities for the reference simplex.

Then we have

$$
\begin{aligned}
\|B\| &= \sup_{\hat{x}\neq 0} \frac{\|B\hat{x}\|}{\|\hat{x}\|} = \sup_{\|\hat{x}-\hat{y}\|=\rho(\hat{S}_n)} \frac{\|B\hat{x} + b_t - (B\hat{y} + b_t)\|}{\|\hat{x} - \hat{y}\|} \\
&= \sup_{\|\hat{x}-\hat{y}\|=\rho(\hat{S}_n)} \frac{\|\mu_t(\hat{x}) - \mu_t(\hat{y})\|}{\rho(\hat{S}_n)} \leq \frac{h(t)}{\rho(\hat{S}_n)} \ .
\end{aligned}
$$

With the same argument we obtain

$$\begin{aligned}\|B^{-1}\| &= \sup_{x\neq 0} \frac{\|B^{-1}x\|}{\|x\|} \\ &= \sup_{\|x-y\|=\rho(t)} \frac{\|B^{-1}(x-b_t) - B^{-1}(y-b_t)\|}{\rho(t)} \\ &= \sup_{\|x-y\|=\rho(t)} \frac{\|\mu_t^{-1}(x) - \mu_t^{-1}(y)\|}{\rho(t)} \\ &\leq \frac{h(\hat{S}_n)}{\rho(t)} \leq \frac{\kappa_2 h(\hat{S}_n)}{h(t)} ,\end{aligned}$$

where the last step uses the shape regularity of the mesh (Def. 7.12 b). Thus we obtain for the condition number

$$\mathrm{cond}_2(B) = \|B\|\|B^{-1}\| \leq \frac{h(t)}{\rho(\hat{S}_n)} \frac{\kappa_2 h(\hat{S}_n)}{h(t)} = \kappa_2 \frac{h(\hat{S}_n)}{\rho(\hat{S}_n)}$$

which only depends on the reference element $\hat{S}_n$ and the mesh.

Approximation Result Now we extend Proposition 8.6 to the general case.

Theorem 8.10. Let $\{\mathcal{T}_\nu\}$ be a family of affine and shape regular triangulations of Ω with $h_\nu \to 0$ and shape regularity constant κ_2. For $k \in \mathbb{N}$, $k > \min(1, n/2)$ and continuous finite element functions of of degree $k-1$ the interpolation error in the H^m-norm, $0 \leq m \leq 1$, can be bounded by

$$\|u - \mathcal{I}_h u\|_{m,\Omega} \leq C h^{k-m} |u|_{k,\Omega} \qquad \forall u \in H^k(\Omega)$$

and the constant C depends only on the shape regularity of the mesh and the space dimension.

Proof. The proof is identical to that in Proposition 8.6 except step d) where we now use the general transformation formula on an individual element t:

$$\begin{aligned}|u - \mathcal{I}_t u|_{l,t} &\leq n^l \|B_t^{-1}\|^l |\mathrm{det} B_t|^{\frac{1}{2}} |\hat{u} - \mathcal{I}_{\hat{t}}\hat{u}|_{l,\hat{t}} && \text{(use (8.11b))} \\ &\leq n^l \|B_t^{-1}\|^l |\mathrm{det} B_t|^{\frac{1}{2}} \|\hat{u} - \mathcal{I}_{\hat{t}}\hat{u}\|_{l,\hat{t}} && (|.| \leq \|.\|) \\ &\leq c n^l \|B_t^{-1}\|^l |\mathrm{det} B_t|^{\frac{1}{2}} |\hat{u}|_{k,\hat{t}} && \text{(Bramble-Hilbert)} \\ &\leq c n^l \|B_t^{-1}\|^l |\mathrm{det} B_t|^{\frac{1}{2}} n^k \|B_t\|^k |\mathrm{det} B_t|^{-\frac{1}{2}} |u|_{k,t} && \text{(use (8.11a))} \\ &= c n^{k+l} (\|B_t\|\|B_t^{-1}\|)^l \|B_t\|^{k-l} |u|_{k,t} \\ &\leq c n^{k+l} \kappa_2^l \left(\frac{h(\hat{t})}{\rho(\hat{t})}\right)^l \left(\frac{h(t)}{\rho(\hat{S}_n)}\right)^{k-l} |u|_{k,t} \\ &= C(\Omega, k, l, n, \hat{S}_n, \kappa_2)\, h(t)^{k-l}\, |u|_{k,t}\end{aligned}$$

□

Remark 8.11. Shape regularity of the mesh is required for the constant to remain bounded as $h_\nu \to 0$. There is a refined analysis that shows that:

- bound on *smallest* angle from below is a sufficient condition.
- bound on largest angle away from π is a necessary condition.

See Babuska and Aziz [1976]. □

Proposition 8.12 (Inverse Estimate)**.** Let $\{\mathcal{T}_\nu\}$ be a family of affine and shape regular triangulations with corresponding finite element spaces $P_k(\mathcal{T}_\nu)$. There exists a constant c (depending on $\kappa_2, n, m, \ldots$), such that

$$\|v_h\|_1 \leq c h_\nu^{-1} \|v_h\|_0 \qquad \forall v_h \in P_k(\mathcal{T}_\nu).$$

Proof. [Braess, 2003, Thm. 6.8]. □

In the standard interpolation error estimates the m-norm is estimated by the k-norm with $m < k$. In the inverse estimate the 1-norm is estimated by the 0-norm, but this is only possible for finite element functions!

8.3. Error Estimates

Regularity The interpolation error estimate in Theorem 8.10 requires the solution of the variational problem to be in $H^k(\Omega)$ with the polynomial degree $k - 1 \geq 1$, i.e. $k \geq 2$. This cannot be deduced from the existence result (the Lax-Milgram Theorem) alone.

Definition 8.13. For $H_0^1(\Omega) \subseteq V \subseteq H^1(\Omega)$ let $a : V \times V \to \mathbb{R}$ be a coercive bilinear form on V. The variational problem

$$u \in V : a(u, v) = (f, v)_{0,\Omega} \quad \forall v \in V$$

is called H^s-regular for $s \geq 1$ if for every $f \in H^{s-2}$ there exists a (unique) solution $u \in H^s(\Omega)$ and a constant $c = c(\Omega, a, s)$ such that

$$\|u\|_{s,\Omega} \leq c \|f\|_{s-2m,\Omega} \ .$$

□

For $s = 1$ this is our existence result. If $s > 1$ one speaks of *higher* regularity. The existence of solutions with higher regularity depends on the form of Ω and the coefficients of the PDE.

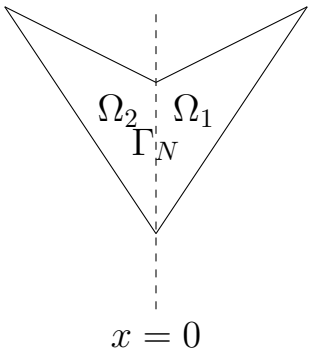


Figure 8.1.: Example of a low regularity of problem with mixed-type boundary conditions in a convex domain.

Theorem 8.14. Let $V = H_0^1(\Omega)$ and $a(u,v) = \int_\Omega (A\nabla u) \cdot \nabla v \, dx$ a coercive bilinear form on V with sufficiently smooth coefficients (e.g. $(A)_{\alpha\beta}$ Lipschitz continuous). Then the following holds true:

a) If Ω is convex the Dirichlet problem is H^2-regular.

b) Assume $s \geq 2$. If Ω has a C^s-boundary then the Dirichlet problem is H^s-regular.

See [Braess, 2003, Theorem 7.2]. □

Remark 8.15 ([Braess, 2003, §7c])**.** The following example illustrates that Neumann (mixed) boundary conditions may lead to low regularity even for convex domains. Consider the domain shown in Figure 8.1. Ω_1 is a convex domain with $\partial\Omega \cap \{(x,y) : x = 0\} = \Gamma_N$ and $\Omega_2 = \{(x,y) : (-x,y) \in \Omega_1\}$. The combined domain $\Omega = \Omega_1 \cup \Gamma_N \cup \Omega_2$ is a non-convex domain and the solution is in general *not* H^2-regular. Let now u be a solution of $-\Delta u = 0$ in Ω, $u = g$ on $\partial\Omega$ with correspondingly low regularity. Then $u_1 = u|_{\Omega_1}$ is the solution of the problem

$$\begin{aligned} -\Delta u &= 0 && \text{in } \Omega_1 \ , \\ u &= g && \text{on } \partial\Omega_1 \setminus \Gamma_N \ , \\ \frac{\partial u}{\partial n} &= 0 && \text{on } \Gamma_N \ . \end{aligned}$$

So u_1 on Ω_1 is in general not H^2-regular although Ω_1 is convex. □

A Priori Estimates We are now in a position to state the error estimates.

Theorem 8.16. Let Ω be a polyhedral domain and $\{\mathcal{T}_\nu\}$ a family of affine, shape regular triangulations. Assume that the solution $u \in V \subseteq H^1(\Omega)$ of the

variational problem is H^k-regular ($k > \frac{n}{2}$). Then we have the following a priori error estimate for the solution $u_h \in V_h = P_{k-1}(\mathcal{T}_h)$ of the discrete variational problem:

$$\|u - u_h\|_{1,\Omega} \leq Ch^{k-1}\|f\|_{k-2} \ .$$

Proof.

$$\begin{aligned}
\|u - u_h\|_{1,\Omega} &\leq \frac{C}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega} && \text{(Céa)} \\
&\leq \frac{C}{\alpha}\|u - \mathcal{I}_h u\|_{1,\Omega} && \text{(Use Lagrange interpolant)} \\
&\leq Ch^{k-1}|u|_{k,\Omega} && \text{(Interpolation error estimate)} \\
&\leq Ch^{k-1}\|f\|_{k-2,\Omega} && \text{(Regularity) .}
\end{aligned}$$

Here C is a generic constant, i.e. it may have different values at different occurences. □

Remark 8.17. In particular

i) The use of polynomial degree $k - 1 > 1$ requires regularity $s = k > 2$.

ii) Away from singularities (reentrant corners, discontinuous coefficients) the solution of elliptic problems is typically very regular and high polynomial degree is efficient

iii) Especially for H^2-regularity ($k = 2$) and polynomial degree $k - 1 = 1$, we get

$$\|u - u_h\|_{1,\Omega} \leq Ch\|f\|_{0,\Omega} \ ,$$

i.e. convergence is $\mathcal{O}(h)$.

iv) Polynomial degree $k > s - 1$ (i.e. without sufficient regularity) does not *hurt*.

v) The constant C depends on continuity and coercivity of the bilinear form as well as shape regularity κ_2 among others. □

Since $\|u - u_h\|_{1,\Omega}^2 = \|u - u_h\|_{0,\Omega}^2 + \|\nabla(u - u_h)\|_{0,\Omega}^2$ the 1-norm measures also the gradient of the error. Can we estimate also $\|u - u_h\|_{0,\Omega}$ alone?

Theorem 8.18. Under the same assumptions as in Theorem 8.16 with $k = 2$ the following estimate holds:

$$\|u - u_h\|_{0,\Omega} \leq Ch^2\|f\|_{0,\Omega} \ .$$

Proof. [Braess, 2003, §7]. The situation is as follows:

$$V_h \subset V = H_0^1(\Omega) \subset L^2(\Omega) = H \ .$$

For any $g \in H$ define the so-called *dual problem*:

$$\varphi_g \in V : a(w, \varphi_g) = (g, w)_{0,\Omega} \qquad \forall w \in V$$

(only for unsymmetric a this is a different problem). Using the error $u - u_h$ as a test function in the dual problem we obtain:

$$\begin{aligned}
(g, u-u_h)_{0,\Omega} &= a(u-u_h, \varphi_g) && (\varphi_g \text{ solution of dual problem}) \\
&= a(u-u_h, \varphi_g) - \underbrace{a(u-u_h, v_h)}_{=0 \text{ for } v_h \in V_h} && (\text{Galerkin orthogonality}) \\
&= a(u-u_h, \varphi_g - v_h) && (\text{linearity}) \\
&\le C\|u-u_h\|_{1,\Omega}\|\varphi_g - v_h\|_{1,\Omega} && (\text{continuity}) \\
&\le C\|u-u_h\|_{1,\Omega} \inf_{v_h \in V_h} \|\varphi_g - v_h\|_{1,\Omega} && (v_h \text{ was arbitrary}) \ .
\end{aligned}$$

Note: On the right hand side $\|u-u_h\|_{1,\Omega} = \mathcal{O}(h)$, $\inf_{v_h \in V_h} \|\varphi_g - v_h\|_{1,\Omega} = \mathcal{O}(h)$.

The norm $\|w\|_{0,\Omega}$ of any $w \in L^2(\Omega)$ can be characterized as

$$\|w\|_{0,\Omega} = \sup_{0 \neq g \in L^2(\Omega)} \frac{(g,w)_{0,\Omega}}{\|g\|_{0,\Omega}}$$

since due to Cauchy-Schwarz

$$(g,w)_{0,\Omega} \le \|g\|_{0,\Omega}\|w\|_{0,\Omega} \Leftrightarrow \frac{(g,w)_{0,\Omega}}{\|g\|_{0,\Omega}} \le \|w\|_{0,\Omega}$$

and

$$\frac{(w,w)_{0,\Omega}}{\|w\|_{0,\Omega}} = \|w\|_{0,\Omega} \ .$$

Using this we obtain

$$\begin{aligned}
\|u-u_h\|_{0,\Omega} &= \sup_{0 \neq g \in L^2(\Omega)} \frac{(g, u-u_h)_{0,\Omega}}{\|g\|_{0,\Omega}} \\
&\le C\|u-u_h\|_{1,\Omega} \sup_{0 \neq g \in L^2(\Omega)} \left\{ \frac{\inf_{v_h \in V_h} \|\varphi_g - v_h\|_{1,\Omega}}{\|g\|_{0,\Omega}} \right\} && (\text{ins. from above}) \\
&\le C\|u-u_h\|_{1,\Omega} \sup_{0 \neq g \in L^2(\Omega)} \left\{ \frac{ch\|g\|_{0,\Omega}}{\|g\|_{0,\Omega}} \right\} && (H^2\text{-reg. of dual pr.}) \\
&\le C\|u-u_h\|_{1,\Omega} h \\
&\le ch^2\|f\|_{0,\Omega} && (\text{a priori est.})
\end{aligned}$$

This way of proof is called a *duality argument* as it involves the solution of the dual problem. □

Remark 8.19. The proof also shows

$$\|u - u_h\|_{0,\Omega} \leq Ch\|u - u_h\|_{1,\Omega}$$

(just omit the last step). □

8.4. Loss of Coercivity

The Lemma of Céa states that

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

where $u \in V \subseteq H^1(\Omega)$ is the solution of the variational problem, $u_h \in V_h \subset V$ is the solution of the discrete variational problem, C is the constant from continuity and α is the constant from coercivity.

If the ratio C/α is large then the error in the finite element solution u_h may be large even if the discrete space V_h has good approximation properties. This situation is called *loss of coercivity*. We now explore two situations where this is the case.

Four Corner Problem In the first case we consider the equation

$$-\nabla \cdot (A(x)\nabla u) = f \qquad \text{in } \Omega$$

with isotropic, heterogeneous diffusion coefficient $A(x) = k(x)I$. Assuming that

$$\sup_{x\in\Omega} k(x) = 1 \qquad \text{and} \qquad \inf_{x\in\Omega} k(x) = \epsilon$$

the analysis of continuity and coercivity gives

$$C = 1, \qquad \alpha = \frac{\epsilon}{1+s^2} \qquad \Rightarrow \qquad \frac{C}{\alpha} = \frac{1+s^2}{\epsilon} \tag{8.12}$$

where s is the constant from Friedrichs inequality which essentially is $s = L$ with L the diameter of the domain. This results states that a heterogeneous diffusion coefficient leads to loss of coercivity.

Example 8.20. In particular consider now the problem

$$-\nabla \cdot (k(x)\nabla u) = 0 \qquad \text{in } \mathbb{R}^2$$

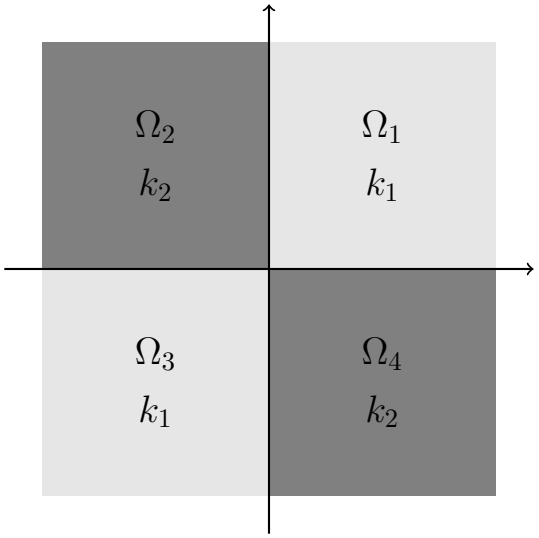


Figure 8.2.: Domain and diffusion coefficient distribution.

where the diffusion coefficient $k(x)$ contains a so-called cross point which is illustrated in Figure 8.2, i.e.

$$k(x) = \begin{cases} k_1 & x_1 x_2 > 0 \\ k_2 & \text{else} \end{cases} .$$

a) Exploiting the symmetry of the solution in polar coordinates

$$u(r, \theta) = -u(r, \theta - \pi)$$

the solution needs only to be determined in $\Omega_{1,2}$ where it is given by

$$u_i(r, \theta) = r^\alpha (a_i \sin(\alpha\theta) + b_i \cos(\alpha\theta))$$

with (assuming $0 < k_1 \leq k_2$)

$$\begin{aligned}
\alpha &= \frac{2}{\pi} \arctan\left(\frac{2\sqrt{k_1 k_2}}{k_2 - k_1}\right), \\
a_1 &= -\frac{k_2}{k_1}(a_2 \cos(\alpha\pi) - b_2 \sin(\alpha\pi)), \\
b_1 &= -(a_2 \sin(\alpha\pi) + b_2 \cos(\alpha\pi)), \\
a_2 &= -b_2 \frac{k_2 \sin(\alpha\pi)\cos(\alpha\pi/2) + k_1 \cos(\alpha\pi)\sin(\alpha\pi/2) + k_2 \sin(\alpha\pi/2)}{k_1 \sin(\alpha\pi)\sin(\alpha\pi/2) - k_2 \cos(\alpha\pi)\cos(\alpha\pi/2) - k_2 \cos(\alpha\pi/2)}, \\
b_2 &= 1.
\end{aligned}$$

For the case $k_1 \ll k_2$ we observe that $\alpha \approx (4/\pi)\sqrt{k_1/k_2}$. While for the reentrant corner problem the singularity always has $\alpha > 1/2$, here the exponent

α can be become arbitrarily close to zero and the solution has an extremely low regularity of $H^{1+\sqrt{k_1/k_2}}$.

Taking the exact solution as Dirichlet boundary data on the domain $\Omega = (-1,1)^2$ we obtain the results shown in Table 8.1. The table shows the L^2-norm as well as the energy norm $\|v\|_E = \sqrt{a(v,v)}$ and the corresponding rates computed on an adaptively refined mesh with up to 20 levels of refinement resulting in more than 2.7 million elements. Moreover, the left half of the table shows results for the lowest order conforming finite elements while the results in the right half are for a more elaborate method which is called *symmetric weighted interior penalty discontinuous Galerkin finite element method* (SWIPDG). The rates are close to the predicted rates. Looking at absolute errors one can see that both methods give relatively good results for a coefficient ratio of 1/10 but converge very slowly for the ratio 1/1000. Note however, that in this case the error in the energy norm with the SWIPDG method on the level 0 mesh with 8 elements is more smaller than the error of the standard conforming method on level 20!

b) Let us now consider the so-called flow cell setup with boundary conditions

$$u = 1 \text{ at } x_1 = -1, \quad u = -1 \text{ at } x_1 = 1, \quad \text{and } \frac{\partial u}{\partial n} = 0 \text{ else.}$$

Figure 8.3 shows the solution for two different values of $k_1/k_2 = 1/20$ and $k_1/k_2 = 1/10000$ for the standard P_1 Galerkin finite element method and the SWIPDG finite element method.

For these boundary condition the solution is not known analytically. However, one can show that the average flux

$$j(\xi) = \frac{1}{2} \int_{-1}^{1} k(\xi,\eta) \partial_{x_1} u(\xi,\eta)\, d\eta$$

along any line $\xi = \text{const}$ is given by

$$j(\xi) = \sqrt{k_1 k_2},$$

see [Keller, 1964; Mendelson, 1975]. Observe that due to Gauß' theorem the flux through any vertical line is the same.

Table 8.2 shows the average flux for two different values of k_1/k_2, two different lowest-order finite element methods and different mesh refinements indicated by the number of degrees of freedom N. For the larger ratio $k_1/k_2 = 1/10$ the flux converges reasonably well with a rate of 0.8 and relative errors of 10^{-3} and below can be obtained with about 10^6 degrees of freedom.

Table 8.1.: Convergence of two different finite element methods for the model problem with analytic solution.

$k_1/k_2 = 1/10$								
	conforming FEM				SWIPDG			
l	$\|u-u_h\|_0$	rate	$\|u-u_h\|_E$	rate	$\|u-u_h\|_0$	rate	$\|u-u_h\|_E$	rate
0	5.19E-1	-	1.47	-	2.05E-1	-	8.66E-1	-
5	2.71E-2	8.11E-1	3.28E-1	3.99E-1	1.08E-2	8.07E-1	2.42E-1	3.88E-1
10	1.82E-3	7.65E-1	8.65E-2	3.81E-1	7.75E-4	7.34E-1	6.53E-2	3.75E-1
15	1.32E-4	7.51E-1	2.35E-2	3.73E-1	6.45E-5	7.07E-1	1.81E-2	3.65E-1
20	1.10E-5	6.83E-1	6.67E-3	3.55E-1	6.66E-6	6.16E-1	5.33E-3	3.43E-1
$k_1/k_2 = 1/50$								
	conforming FEM				SWIPDG			
l	$\|u-u_h\|_0$	rate	$\|u-u_h\|_E$	rate	$\|u-u_h\|_0$	rate	$\|u-u_h\|_E$	rate
0	5.21E-01	-	1.25	-	1.21E-1	-	4.01E-1	-
5	8.29E-02	4.41E-01	4.47E-1	2.29E-1	3.04E-3	1.14	2.28E-1	1.78E-01
10	2.11E-02	3.77E-01	2.22E-1	1.92E-1	3.86E-4	3.62E-01	1.23E-1	1.78E-01
15	5.92E-03	3.63E-01	1.17E-1	1.83E-1	1.17E-4	4.34E-01	6.63E-2	1.80E-01
20	1.70E-03	3.60E-01	6.23E-2	1.80E-1	2.72E-5	4.08E-01	3.56E-2	1.80E-01
$k_1/k_2 = 1/100$								
	conforming FEM				SWIPDG			
l	$\|u-u_h\|_0$	rate	$\|u-u_h\|_E$	rate	$\|u-u_h\|_0$	rate	$\|u-u_h\|_E$	rate
0	5.45E-1	-	1.26	-	1.27E-1	-	3.11E-1	-
5	1.10E-1	3.65E-1	5.01E-1	1.96E-1	1.32E-2	4.16E-01	2.12E-1	1.26E-1
10	3.74E-2	2.89E-1	2.82E-1	1.52E-1	5.26E-3	2.51E-01	1.38E-1	1.27E-1
15	1.45E-2	2.67E-1	1.73E-1	1.37E-1	2.20E-3	2.54E-01	8.84E-2	1.28E-1
20	5.83E-3	2.59E-1	1.09E-1	1.31E-1	9.15E-4	2.54E-01	5.68E-2	1.27E-1
$k_1/k_2 = 1/1000$								
	conforming FEM				SWIPDG			
l	$\|u-u_h\|_0$	rate	$\|u-u_h\|_E$	rate	$\|u-u_h\|_0$	rate	$\|u-u_h\|_E$	rate
0	6.04E-1	-	1.34	-	1.41E-1	-	1.30E-1	-
5	1.74E-1	2.54E-1	6.52E-1	1.40E-1	2.74E-2	1.85E-1	1.40E-1	-1.56E-3
10	8.91E-2	1.66E-1	4.48E-1	9.43E-2	2.15E-2	6.13E-2	1.37E-1	1.19E-2
15	5.42E-2	1.31E-1	3.37E-1	7.53E-2	1.72E-2	6.55E-2	1.29E-1	2.31E-2
20	3.58E-2	1.14E-1	2.66E-1	6.46E-2	1.36E-2	6.87E-2	1.18E-1	2.74E-2

Table 8.2.: Evaluation of the average flux for two finite element methods and two different k_1/k_2 ratios.

$k_1/k_2 = 1/10$							
conforming FEM				SWIPDG			
N	j_h	$\lvert j - j_h\rvert/j$	rate	N	j_h	$\lvert j - j_h\rvert/j$	rate
179	3.31035154E-01	4.68E-02	-	972	3.20180997E-01	1.24E-02	-
681	3.24314950E-01	2.56E-02	0.873	3888	3.18136766E-01	6.04E-03	1.050
2657	3.20781044E-01	1.44E-02	0.829	15552	3.17242561E-01	3.19E-03	0.912
10497	3.18834185E-01	8.25E-03	0.805	62208	3.16791943E-01	1.78E-03	0.847
41729	3.17731742E-01	4.74-03	0.793	248832	3.16548101E-01	1.01E-03	0.817
166401	3.17098936E-01	2.75E-03	0.788	995328	3.16411466E-01	5.82E-04	0.802
664577	3.16733359E-01	1.60E-03	0.785	3981312	3.16333634E-01	3.35E-04	0.795
$k_1/k_2 = 1/1000$							
conforming FEM				SWIPDG			
N	j_h	$\lvert j - j_h\rvert/j$	rate	N	j_h	$\lvert j - j_h\rvert/j$	rate
179	1.63862045E-01	4.17	-	972	1.04923502E-02	0.667	-
681	1.42964281E-01	3.51	2.48E-01	3888	1.11906625E-02	0.645	4.86E-02
2657	1.27369212E-01	3.03	2.18E-01	15552	1.19371940E-02	0.623	5.37E-02
10497	1.15161562E-01	2.64	1.97E-01	62208	1.26778950E-02	0.598	5.53E-02
41729	1.05289827E-01	2.33	1.81E-01	248832	1.34063023E-02	0.575	5.66E-02
166401	9.71270673E-02	2.07	1.69E-01	995328	1.41194863E-02	0.553	5.76E-02
664577	9.02652109E-02	1.85	1.60E-01	3981312	1.48160358E-02	0.531	5.86E-02

For the smaller ratio $k_1/k_2 = 1/1000$ the convergence is down to 0.06 and the relative error obtained with the standard finite element method on a mesh with 664577 unknowns is still 185% (!). The discontinuous Galerkin method starts with an error of 66% on the coarsest mesh with 972 unknowns but the error decreases only to 53% using 3981312 degrees of freedom. □

The particular difficulty with the example presented is that the loss of coercivity goes hand in hand with a loss of regularity. However, in general these two properties are not related: The reentrant corner problem from example 7.14 has reduced regularity without loss of coercivity while the convection example 8.21 below has high regularity with loss of coercivity.

Convection-Diffusion Problem As second example we consider the stationary convection-diffusion equation with homogeneous isotropic diffusion coefficient ϵ and constant (divergence free) velocity field b:

$$-\epsilon\Delta u + b \cdot \nabla u = f \qquad \text{in } \Omega.$$

An analysis of continuity and coercivity gives in this case

$$C = \epsilon + s\|b\|, \qquad \alpha = \frac{\epsilon}{1+s^2} \qquad \Rightarrow \qquad \frac{C}{\alpha} = \frac{(\epsilon + s\|b\|)(1+s^2)}{\epsilon}$$

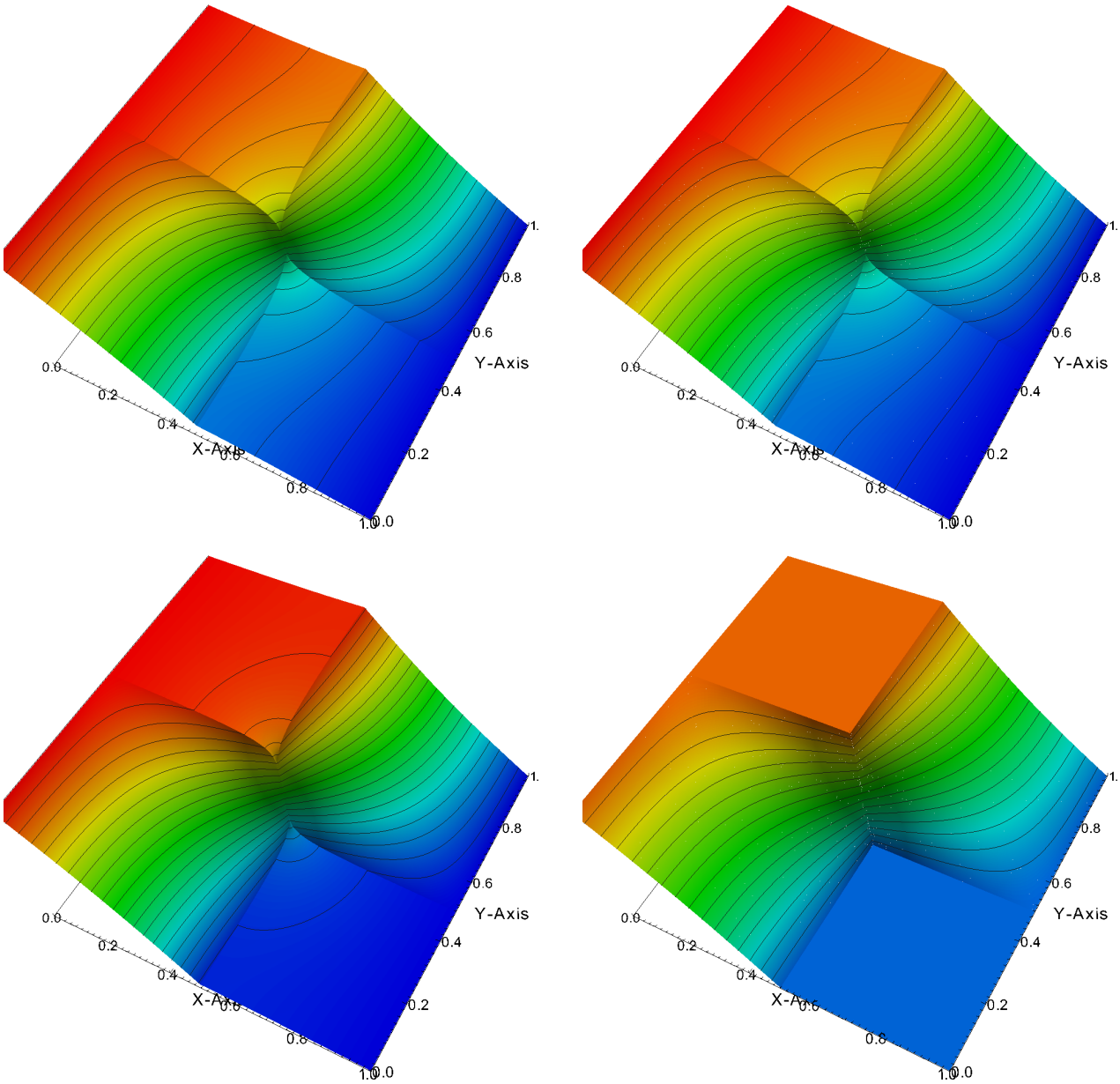


Figure 8.3.: Solution of the four corner problem: $k_1/k_2 = 5 \cdot 10^{-2}$ (top row), Standard Galerkin method with P_1 elements (top left), symmetric interior penalty Galerkin method with P_1 (top right), $k_1/k_2 = 1 \cdot 10^{-4}$ (bottom row), standard Galerkin method with P_1 elements (bottom left), symmetric interior penalty Galerkin method with P_1 (bottom right)

We rewrite this as

$$\frac{C}{\alpha} = (1 + \mathrm{Pe})(1 + s^2)$$

with the Peclet number

$$\mathrm{Pe} = \frac{s\|b\|}{\epsilon}. \tag{8.13}$$

The Peclet number measures the relative strength of diffusion and convection. This relative strength depends on the length scale which is measured by the constant s from Friedrichs inequality (The factor $1 + s^2$ comes in for different reasons, it is also present in the pure diffusion equation, compare (8.12). The standard argument would be a dimension analysis, see [Elman et al., 2005, §3.1].) A Peclet number $\mathrm{Pe} > 1$ indicates that the problem is convection dominated.

Example 8.21. We consider the two-point boundary value problem

$$-\epsilon u'' + u' = 0 \qquad \text{in } (0, 1)$$

with boundary conditions

$$u = 1 \text{ at } x_1 = 0, \qquad u = 0 \text{ at } x_1 = 1.$$

It has the exact solution

$$u(x) = \frac{1 - \exp((x-1)/\epsilon)}{1 - \exp(-1/\epsilon)}$$

shown in Figure 8.4 for various values of ϵ. The limit problem $u' = 0$ for $\epsilon = 0$ is first order hyperbolic and has the solution $u(x) = 1$. According to the method of characteristics a boundary condition can be given at $x = 0$ but no boundary condition can be given at $x = 1$. The limit problem is incompatible with the boundary condition at $x = 1$ for the problem with $\epsilon > 0$ which results in a so-called *exponential boundary layer*. This boundary layer has a width of $O(\epsilon)$. According to the definition given above, the Peclet number of this problem is $\mathrm{Pe} = 1/\epsilon$ and there is loss of coercivity for small ϵ. However, in comparison to the elliptic example with piecewise constant permeabilities the solution is perfectly smooth, i.e. it has a high regularity. Note also, that the exact solution satisfies a maximum principle.

Figure 8.5 shows numerical results for this problem using the standard P_1 Galerkin finite element method (shown in blue) and the discontinuous Galerkin finite element method (DG). On coarse meshes the results for the standard finite element method are very bad in the whole domain and small variations in the boundary condition result in large changes in the numerical solution (loss of stability). As the mesh is refined the oscillations are reduced and they are concentrated near the exponential boundary layer. When the mesh size reaches

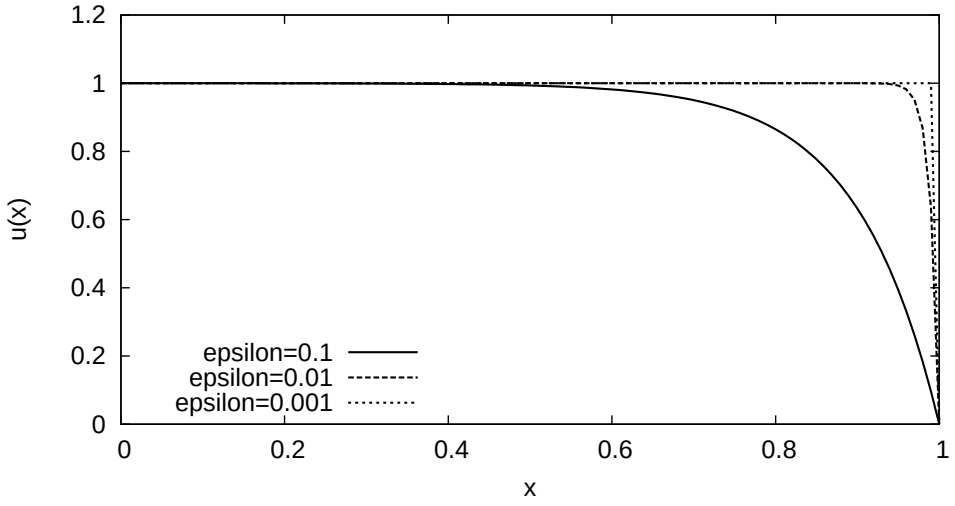


Figure 8.4.: Exact solution of the one-dimensional convection-diffusion problem

the value 2ϵ the oscillations vanish and the numerical solution obeys a maximum principle.

The discontinuous Galerkin scheme on the other hand shows only comparatively small oscillation which are always located close to the boundary layer. Moreover, the DG scheme obeys the Dirichlet boundary conditions only weakly, depending on the mesh size. As the mesh is refined the scheme "senses" the boundary condition and when $h \approx \epsilon$ the boundary conditions and the maximum principle are satisfied. □

Example 8.22. We consider the two-dimensional convection-diffusion problem

$$-\epsilon\Delta u + b \cdot \nabla u = 0 \qquad \text{in } (0,1)^2$$

with boundary conditions

$$u = 0 \text{ at } x_1 = 1, \qquad u = 1 \text{ if } x_1 < 1/2 \text{ and } x_2 < 1/4,$$
$$u = 0 \text{ if } x_1 = 0 \text{ and } x_2 > 1/4, \qquad \frac{\partial u}{\partial n} = 0 \text{ at } x_2 = 1.$$

Figure 8.6 shows numerical results for $\epsilon = 10^{-4}$ and $b = (1, 3/2)^T$ using Q_1 standard Galerkin finite elements (left column) and the P_1 DG method (right column) on a uniform quadrilateral mesh. The discontinuities in the Dirichlet

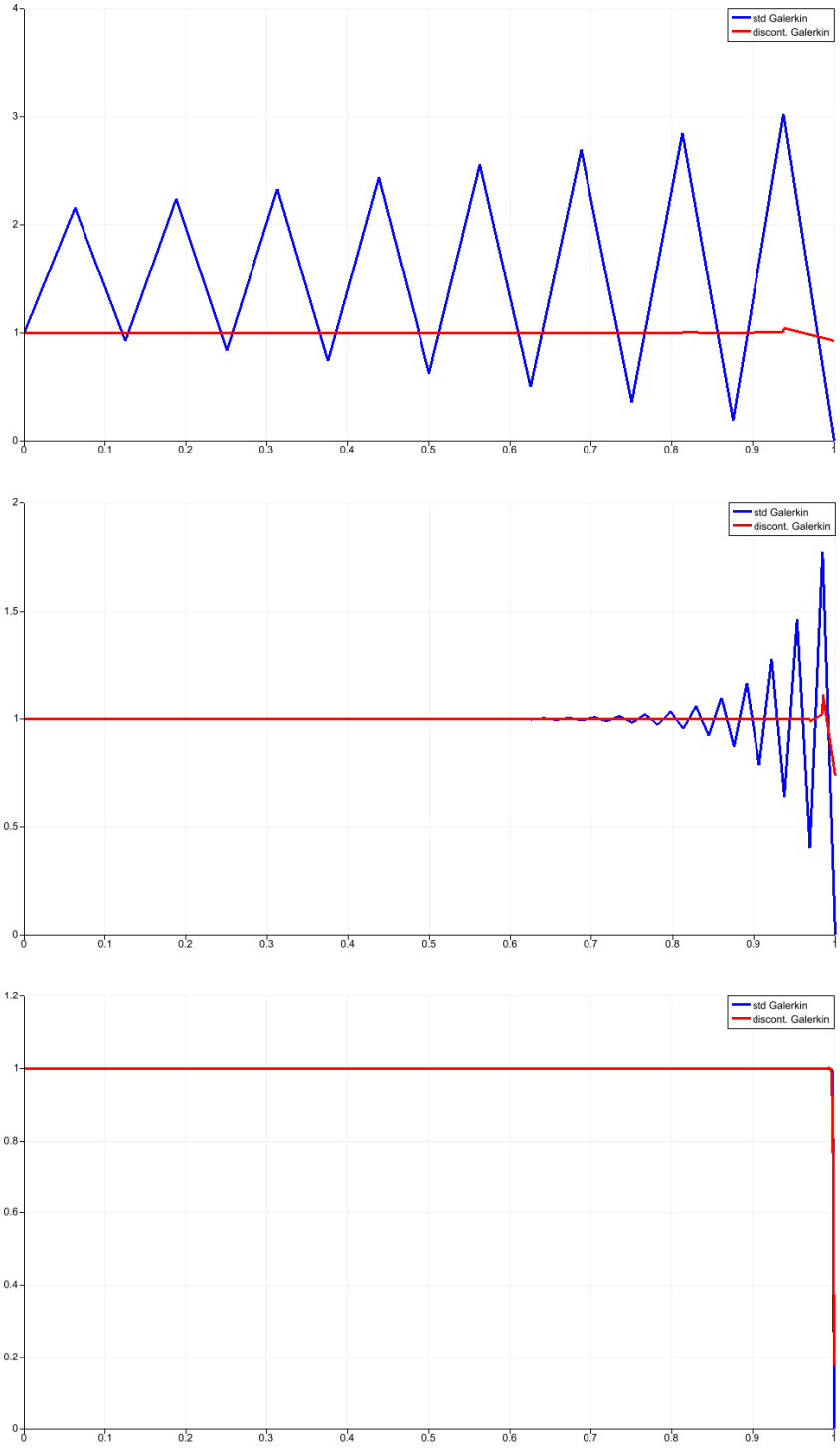


Figure 8.5.: One-dimensional convection-diffusion problem with $\epsilon = 10^{-3}$ and $b = 1$ solved with standard P_1 Galerkin and symmetric interior penalty discontinuous Galerkin method using P_1. From top to bottom $h = 1/16$, $1/64$ and $1/512$.

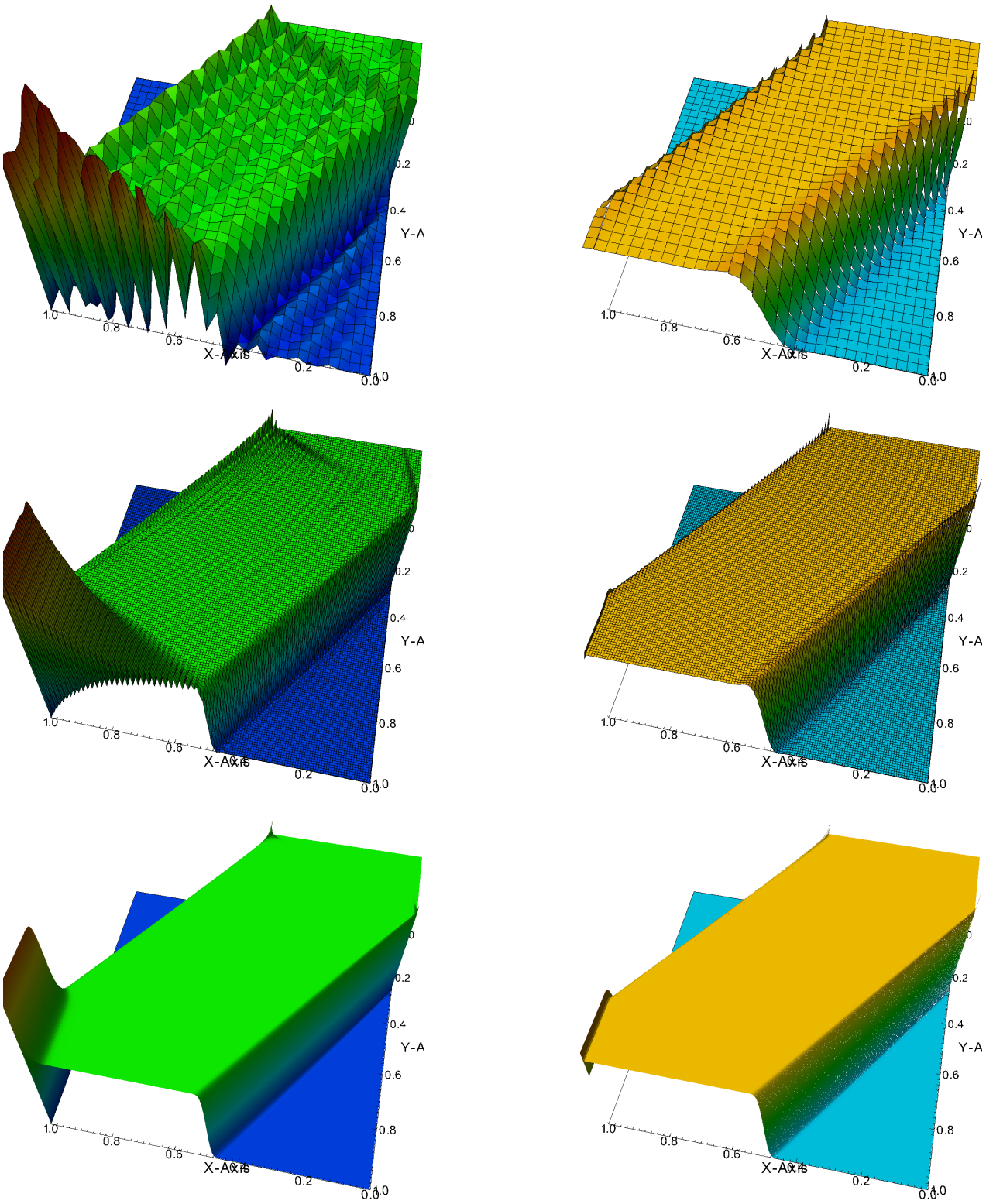


Figure 8.6.: Solution of the two-dimensional convection-diffusion problem in with $\epsilon = 10^{-4}$ and $b = 1$ solved with standard Q_1 Galerkin (left column) and symmetric interior penalty discontinuous Galerkin method using P_1 on structured quadrilateral mesh (right column). Mesh size (top to bottom): $h = 1/32$, $1/128$ and $1/512$.

boundary condition at $(0, 1/4)$ and $(1/2, 0)$ result in so-called *characteristic interior layers*. In the limit problem $\epsilon = 0$ the discontinuity would be transported along the characteristic. If $\epsilon > 0$ the solution is continuous and the discontinuity is smeared over a layer of width $O(\sqrt{\epsilon})$. Near the boundary $\{1\} \times (3/4, 1)$ an exponential boundary layer is formed since the solution of the limit problem is incompatible to the Dirichlet boundary condition. Moreover, the true solution obeys a maximum principle.

The standard Galerkin scheme shows qualitatively the same behaviour as in the one-dimensional case. For coarse meshes the solution quality is poor in the whole domain. As the mesh is refined the oscillations concentrate near the exponential boundary layer. The internal characteristic layer poses no problem as soon as $h \approx O(\sqrt{\epsilon})$. The DG scheme on the other hand provides a reasonable accurate solution also on coarse meshes. In both schemes the exponential boundary layer is not resolved and the maximum principle is not satisfied as the mesh is still too coarse. □

Chapter 9.

Adaptive Finite Element Methods

9.1. Introduction

Accepting a finite element solution without controlling the discretization error may have disastrous consequences as is e.g. illustrated in the failure of the Sleipner A[1] offshore platform. The collapse of this platform during final construction phase has been attributed to an underestimation of the stresses in numerical simulations.

The goal of this chapter therefore is to devise a practical method to ensure that

$$\|u - u_h\| \leq \text{TOL} \tag{9.1}$$

with $u_h \in V_h$ the computed finite element solution, $\dim V_h$ as small as possible, $\|.\|$ an appropriate norm and TOL a user given tolerance.

The *a-priori* estimates proven in the last chapter are not suitable to ensure (9.1) for practical computations since the constant C involved and the regularity of the exact solution are not known. In the following we will derive so-called *a-posteriori* error estimates of the form

$$\|u - u_h\| \leq \eta(u_h) = \left(\sum_{t \in \mathcal{T}_h} \eta_t^2(u_h) \right)^{\frac{1}{2}} \tag{9.2}$$

where the global error estimate η is computed from local error contributions $\eta_t(u_h)$ that depend on the computed solution u_h on t (or a small patch around t). Based on (9.2) the global error estimate η is used for *error control* by ensuring $\eta \leq \text{TOL}$ and the local estimators η_t are used within a heuristic algorithm for *local mesh refinement*.

There are many ways to derive a-posteriori error estimates. An overview of different techniques is given in Ainsworth and Oden [2000] (and this presentation follows that text). So-called goal-oriented refinement allowing to estimate functionals $J(u - u_h)$ of the error is treated in Bangerth and Rannacher [2003], a-posteriori error estimates for many different applications are discussed in Eriksson et al. [1996] and latest results for elliptic problems with discontinuous diffusion coefficients are given by Vohralik [2011].

[1] `http://en.wikipedia.org/wiki/Sleipner_A`

As a prerequisite we state the following approximation result.

Theorem 9.1. Let $t \in \mathcal{T}_h$ be a mesh element, V_h a conforming finite element space of polynomial order $k-1$ and let $\tilde{t} = \{\cup t', t' \in \mathcal{T}_h \,:\, \bar{t}' \cap \bar{t} \neq \emptyset\}$ denote the patch of elements having at least a common vertex with t. Then for $0 \leq m \leq k$ there exists a linear operator $\mathcal{I} : H^m(\Omega) \to V_h$ and a constant C depending only on the shape regularity of the mesh such that

$$\|u - \mathcal{I}u\|_{0,t} \leq C h_t^m |u|_{m,\tilde{t}}, \tag{9.3a}$$

$$\|u - \mathcal{I}u\|_{0,\gamma} \leq C h_t^{m-1/2} |u|_{m,\tilde{t}}, \tag{9.3b}$$

where γ is any face of the element t (of dimension $n-1$).
Proof. See [Ainsworth and Oden, 2000, THM 1.7] and Bernardi and Girault [1998]. □

Remark 9.2. The proof cited in the theorem above is only for the two-dimensional case. Various other constructions of this type exist like that of Clément [1975] or Scott and Zhang [1990]. Note the interpolation operator $\mathcal{I}$ can be applied in the case $m \leq n/2$ where Lagrange interpolation is not defined. The price to pay for this is that the norm on the right hand side is taken over the patch $\tilde{t}$. □

9.2. Residual-based a-posteriori Error Estimator

We consider the second order elliptic boundary value problem

$$\begin{aligned} -\nabla \cdot (K \nabla u) + cu &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \Gamma_D \subseteq \partial\Omega, \\ -(K\nabla u) \cdot n &= j && \text{on } \Gamma_N = \partial\Omega \setminus \Gamma_D, \end{aligned}$$

with the corresponding weak formulation

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

with $V = \{v \in H^1(\Omega) \,:\, v = 0 \text{ on } \Gamma_D\}$ and

$$a(u,v) = \int_\Omega (K\nabla u) \cdot \nabla v + cuv \, dx, \qquad l(v) = \int_\Omega fv \, dx - \int_{\Gamma_N} jv \, ds.$$

For the *error* $e = u - u_h$ we have due to linearity

$$a(e,v) = a(u,v) - a(u_h,v) = l(v) - a(u_h,v).$$

Thus, for a given computed solution $u_h \in V_h \subset V$ the error $e \in V$ satisfies a problem with the same bilinear form but a different right hand side.

Now for any $v \in V$ we obtain using integration by parts:

$$\begin{aligned} a(e,v) &= \sum_{t\in\mathcal{T}_h} \left\{ \int_t fv\,dx - \int_{\partial t\cap\Gamma_N} jv\,ds - \int_t (K\nabla u_h)\cdot\nabla v + cu_h v\,dx \right\} \\ &= \sum_{t\in\mathcal{T}_h} \left\{ \int_t rv\,dx + \int_{\partial t\cap\Gamma_N} R_b v\,ds - \int_{\partial t\setminus\Gamma_N} (K\nabla u_h)\cdot n v\,ds \right\} \end{aligned} \tag{9.4}$$

with

$$r = f + \nabla\cdot(K\nabla u_h) - cu_h, \qquad \text{(interior residual)} \tag{9.5a}$$

$$R_b = -(K\nabla u_h)\cdot n - j. \qquad \text{(boundary residual)} \tag{9.5b}$$

The element boundary $\partial t \setminus \Gamma_N$ can be split further into a part $\partial t \cap \Gamma_D$ covering the Dirichlet boundary and the remaining interior boundary. We introduce the notation

$$\mathcal{F}_h = \mathcal{F}_h^i \cup \mathcal{F}_h^D \cup \mathcal{F}_h^N$$

with

$$\begin{aligned} \mathcal{F}_h^i &= \{\gamma = \bar{t}\cap\bar{t}' \,:\, t,t'\in\mathcal{T}_h, t\neq t', \gamma \text{ has dimension } n-1\}, \\ \mathcal{F}_h^D &= \{\gamma = \bar{t}\cap\Gamma_D \,:\, t\in\mathcal{T}_h\}, \\ \mathcal{F}_h^N &= \{\gamma = \bar{t}\cap\Gamma_N \,:\, t\in\mathcal{T}_h\}. \end{aligned}$$

For each $\gamma \in \mathcal{F}_h$ a unit normal n_γ is selected which coincides with the exterior unit normal for the boundary faces. For interior faces $\gamma \in \mathcal{F}_h^i$ an arbitrary orientation can be selected and we denote by $T^+(\gamma)$ the element $t \in \mathcal{T}_h$ in direction of the normal n_γ and by $T^-(\gamma)$ the element in the opposite direction. For boundary faces $\gamma = \partial t \cap \partial\Omega$ we set $T^-(\gamma) = t$. Finally we denote the *jump* of a function that is discontinuous at element boundaries by

$$[w](x) = \lim_{s\to 0+} (w(x - sn_\gamma) - w(x + sn_\gamma)) \qquad \forall x \in \gamma \in \mathcal{F}_h^i. \tag{9.6}$$

We now observe that in the error representation (9.4) we can further split $\partial t \setminus \Gamma_N = (\partial t\cap\Gamma_D)\cup(\partial t\cap\Omega)$. Since $v = 0$ on Γ_D there is no contribution from the Dirichlet boundary and only the interior faces remain. Every interior face $\gamma \in \mathcal{F}_h^i$ is visited twice, once from $T^-(\gamma)$ and once from $T^+(\gamma)$ which gives

$$a(e,v) = \sum_{t\in\mathcal{T}_h} \int_t rv\,dx + \sum_{\gamma\in\mathcal{F}_h^N} \int_\gamma R_b v\,ds + \sum_{\gamma\in\mathcal{F}_h^i} \int_\gamma [-(K\nabla u_h)\cdot n_\gamma] v\,ds.$$

Introducing

$$R(x) = \begin{cases} [-(K\nabla u_h)\cdot n] & x \in \gamma \in \mathcal{F}_h^i \\ -(K\nabla u_h)\cdot n - j & x \in \gamma \in \mathcal{F}_h^N \end{cases} \qquad \text{(face residual)} \qquad (9.7)$$

we can combine the surface integrals to obtain the error representation formula

$$a(e,v) = \sum_{t\in\mathcal{T}_h} \int_t rv\,dx + \sum_{\gamma\in\mathcal{F}_h^i\cup\mathcal{F}_h^N} \int_\gamma Rv\,ds\ . \qquad (9.8)$$

The further steps exploit (9.8) to yield an estimate of the finite element error. Employing the interpolation operator $\mathcal{I}$ from Theorem 9.1 and Galerkin orthogonality we observe for any $v \in V$:

$$0 = a(e,\mathcal{I}v) = \sum_{t\in\mathcal{T}_h} \int_t r\mathcal{I}v\,dx + \sum_{\gamma\in\mathcal{F}_h^i\cup\mathcal{F}_h^N} \int_\gamma R\mathcal{I}v\,ds\ .$$

Subtracting this from the error representation we obtain the estimate:

$$\begin{aligned} a(e,v) &= \sum_{t\in\mathcal{T}_h} \int_t r(v-\mathcal{I}v)\,dx + \sum_{\gamma\in\mathcal{F}_h^i\cup\mathcal{F}_h^N} \int_\gamma R(v-\mathcal{I}v)\,ds \\ &\le \sum_{t\in\mathcal{T}_h} \|r\|_{0,t}\|v-\mathcal{I}v\|_{0,t} + \sum_{\gamma\in\mathcal{F}_h^i\cup\mathcal{F}_h^N} \|R\|_{0,\gamma}\|v-\mathcal{I}v\|_{0,\gamma} \qquad \text{(C. S.)} \\ &\le \sum_{t\in\mathcal{T}_h} \|r\|_{0,t} C h_t \|v\|_{1,\tilde{t}} + \sum_{\gamma\in\mathcal{F}_h^i\cup\mathcal{F}_h^N} \|R\|_{0,\gamma} C h_{T^-(\gamma)}^{\frac{1}{2}} \|v\|_{1,\tilde{T}^-(\gamma)} \qquad \text{(Thm. 9.1)} \\ &\le C\|v\|_{1,\Omega} \left\{ \sum_{t\in\mathcal{T}_h} h_t^2 \|r\|_{0,t}^2 + \sum_{\gamma\in\mathcal{F}_h^i\cup\mathcal{F}_h^N} h_{T^-(\gamma)} \|R\|_{0,\gamma}^2 \right\}^{\frac{1}{2}} . \qquad \text{(C. S., s. r.)} \end{aligned}$$

In the final step we have used the fact that due to the shape regularity there is a bound on the maximum number of neighbors of every triangle and thus every triangle is contained in a finite number of patches.

Now use the error itself as a test function and exploit coercivity $\alpha\|e\|_{1,\Omega}^2 \le a(e,e)$ to obtain

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

Finally, we distribute the contributions of the interior face residuals equally to the neighboring elements in order to obtain an element-wise form of the error

estimator:

$$\|e\|_{1,\Omega} \leq C\eta, \qquad \eta = \left\{ \sum_{t \in \mathcal{T}_h} \eta_t^2 \right\}^{\frac{1}{2}} \tag{9.9}$$

with

$$\eta_t^2 = h_t^2 \|r\|_{0,t}^2 + \sum_{\gamma \in \mathcal{F}_h^N \cap \partial t} h_t \|R\|_{0,\gamma}^2 + \sum_{\gamma \in \mathcal{F}_h^i \cap \partial t} \frac{h_t}{2} \|R\|_{0,\gamma}^2 \, . \tag{9.10}$$

This is the final form of the residual-based error estimator that will subsequently be used for error control and local mesh adaptation.

Remark 9.3. a) Note that the derivation of the error estimator did not require any additional regularity beyond $u \in H^1(\Omega)$. This is very important as error control and local mesh adaption is especially required in problems with low regularity.

b) The estimate above is not robust with respect to coefficient variations (loss of coercivity). However, in the derivation only coercivity but not continuity was employed. Thus one could always scale the problem in such a way that $\alpha = 1$ without altering the constant in the right hand side.

c) The constant C in (9.9) is usually not known exactly and is absorbed into the given user tolerance TOL to stop the calculations.

d) The estimate (9.9) does not imply that the error estimator is efficient. It could still happen that e.g. $\|e\|_{1,\Omega} = O(h)$ and $\eta = O(h^\beta)$ with $\beta < 1$, i.e. asymptotically the error decays faster than the error estimator and the stopping criterion would be very pessimistic. Therefore one would like to have also an estimate of the form $\eta \leq C\|e\|_{1,\Omega}$. If this is the case the error estimator is called *efficient* and the quantity $\eta / \|e\|_{1,\Omega}$ is called *efficiency index*. The numerical experiments below show that for the reentrant corner problem the efficiency index is about 3 to 6 depending only slightly on the type of mesh refinement and the polynomial degree. □

9.3. Local Mesh Adaptation

The a-posteriori error estimate is now used within an adaptive algorithm with the goal to construct a mesh that achieves the error tolerance with as few elements as possible.

Algorithm 9.4. The basic local mesh adaption algorithm reads:

1) Choose an initial mesh $\mathcal{T}_0$ that is sufficiently fine.

2) Compute u_h on the current mesh $\mathcal{T}_h$.

3) Compute the error estimate $\eta(u_h)$. If $\eta(u_h) \leq \text{TOL}$ then STOP.

4) If the tolerance is not reached refine the mesh according to the local quantities η_t (see below).

5) Interpolate the current solution u_h to the new mesh.

6) Go to 2). □

In step 4 of the algorithm two questions arise: (i) which elements should be refined and (ii) how to locally refine a mesh. In order to answer the first question we recall from Propostion 8.6 the estimate of the interpolation error (which bounds the true error):

$$|u - \mathcal{I}_h u|^2_{1,\Omega} = \sum_{t \in \mathcal{T}_h} |u - \mathcal{I}_t u|^2_{1,t} \leq \sum_{t \in \mathcal{T}_h} C h_t^2 |u|^2_{2,t} = \sum_{t \in \mathcal{T}_h} s_t^2.$$

Here, $\mathcal{I}$ is the Lagrange interpolation operator and we assume H^2-regularity. Each element makes a contribution $s_t^2 = C h_t^2 |u|^2_{2,t}$ to the squared error. If we refine element t into 2^n children $t' \in c(t)$ we can estimate their error contribution by

$$\sum_{t' \in c(t)} |u - \mathcal{I}_{t'} u|^2_{1,t'} \leq \sum_{t' \in c(t)} C \left(\frac{h_t}{2}\right)^2 |u|^2_{2,t'} \approx \sum_{t' \in c(t)} \left(\frac{1}{2}\right)^2 C h_t^2 \frac{|u|^2_{2,t}}{2^n} = \left(\frac{1}{2}\right)^2 s_t^2.$$

In the case of reduced regularity $u \in H^{1+\alpha}$ a reduction by a factor $(1/2)^{2\alpha}$ can be expected. Since our error estimator is efficient we assume it behaves the same way, i.e. if we refine element t contributing η_t^2 to the squared error the total error is *reduced* from η^2 to $\eta^2 - (1 - (1/2)^{2\alpha})\eta_t^2$ and the number of elements in the mesh is *increased* by $2^n - 1$. The optimal mesh achieves a given tolerance with fewest elements. So, if the tolerance is not yet met we seek the largest error reduction for a given increase in number of elements. Since the local regularity of the solution is in general not known we simply refine the elements with the largest error contribution. Ultimately this leads to an equilibration of the error contribution of each element. Since the mesh refinement algorithm and subsequent finite element solution has computational complexity at least proportional to the number of elements it is not efficient to refine just very few elements. All these consideration lead to the so-called *bulk fraction strategy*:

(i) Order all elements according to increasing error contribution:

$$\eta^2_{t_{i_1}} \leq \eta^2_{t_{i_2}} \leq \ldots \leq \eta^2_{t_{i_{m_h}}}.$$

(ii) For a given parameter $\rho \in (0, 1]$ determine

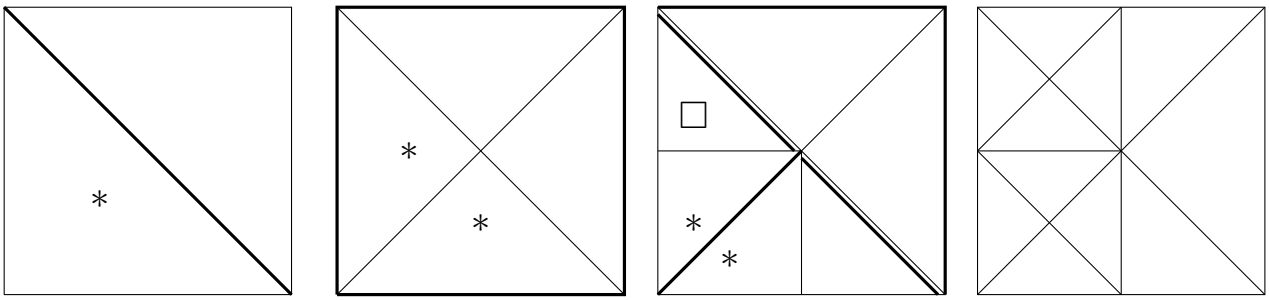
$$J = \max \left\{ j \, : \, \sum_{k=j}^{m_h} \eta_{t_{i_k}}^2 \geq \rho \sum_{t \in \mathcal{T}_h} \eta_t^2 \right\}.$$

(iii) Refine all elements $t_{i_J}, \ldots, t_{i_{m_h}}$.

This strategy assumes that the action of refinement is local, i.e. a refinement of an element leads to a reduction of the error in the refined elements without affecting the error in the elements that are not refined. This is true for the interpolation error (as shown above) but not necessarily for the finite element error. As an example where refinement acts less local consider a first-order hyperbolic problem (which is solved by the method of characteristics). There a large error upstream will produce a large error in all downstream elements. Consequently, problems can be expected for a convection-dominated convection-diffusion problem as well. The strategy outlined above is, however, very effective in pure diffusion problems.

Local Mesh Refinement In section 7.3 several algorithms for mesh construction through mesh refinement were discussed. These algorithms are now extended to the case of local refinement. The discussion is restricted to the two-dimensional case for simplicity (but all algorithms can be extended to three, or even arbitrary, dimensions).

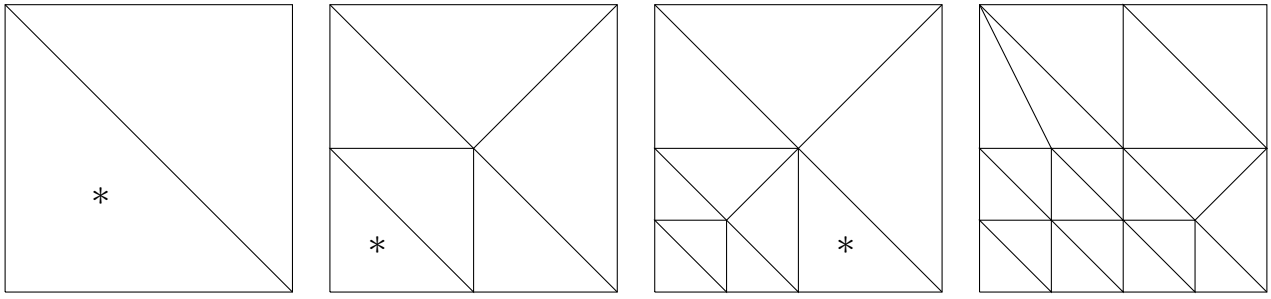
We begin with the discussion of newest vertex bisection refinement:



The edges outlined in thick are the only edges of a triangle that may be refined in a given step. Suppose that the triangle in $\mathcal{T}_0$ marked with an asterisk "$*$" is to be refined. Accordingly the thick edge is refined by introducing a new vertex and four new triangles are created replacing the old ones. In every new triangle the edge *opposite the newest vertex* is marked as the bisection edge. In the step $\mathcal{T}_1 \to \mathcal{T}_2$ two triangles are refined accordingly with the subsequent assignment of refinement edges. Now consider the step $\mathcal{T}_2 \to \mathcal{T}_3$. The two triangles marked with an asterisk are easy to do. However, refining the triangle marked by "$\square$"

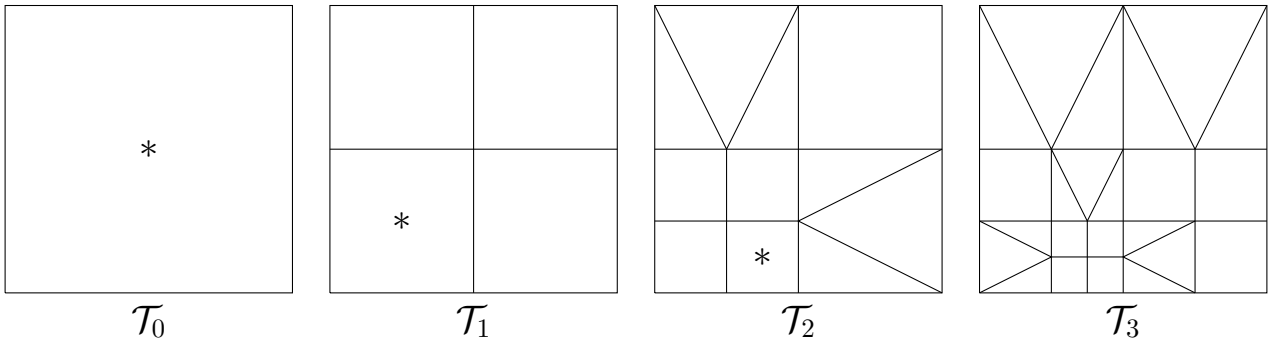
first requires to subdivide its neighbor to the right twice as only the splitting of a triangle along the refinement edge is allowed. The example shows that local refinement of an element may require substantial, non-local changes in the mesh.

The regular refinement algorithm can be extended to local refinement as well:



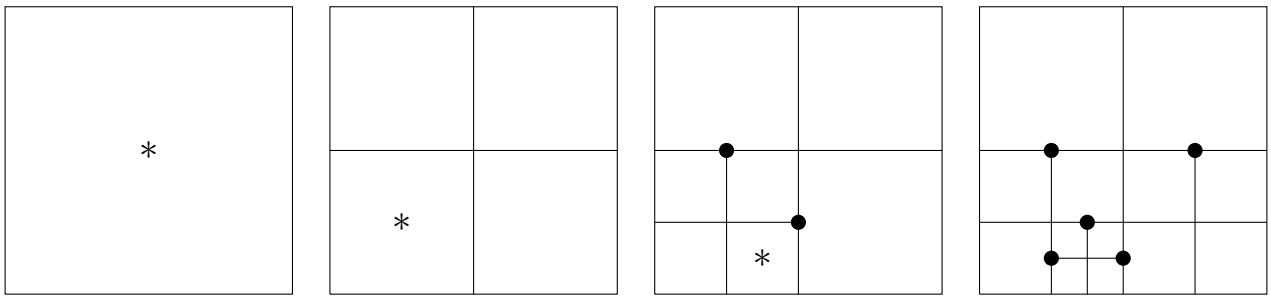
In order keep the refinement local one combines regular refinement with bisection (also called irregular refinement in this context). A problem arises when the refinement of triangle created through bisection is required as in step $\mathcal{T}_2 \to \mathcal{T}_3$ in the figure. Since arbitrary bisection may result in arbitrarily small angles the idea is to remove the bisection refinement and replace it by a regular refinement with subsequent bisection to make the mesh conforming again. As in the case of pure bisection refinement of a single element may require substantial, non-local changes in the mesh.

The combination of regular and irregular refinement can also be extended to quadrilateral meshes but requires then the combination of quadrilateral and triangular elements:



Note that this may become exceedingly complicated in three space dimensions as hexahedra, tetrahedra and (at least) four-sided pyramids are required.

A lot of the complexity of the algorithms above is introduced by the fact that the mesh is required to be conforming (in particular the replacement of previous refinements in the regular/irregular refinement scheme). This can be overcome by allowing the mesh to be non-conforming. In particular this also avoids mixed element type meshes in the case of quadrilaterals:



Omitting the irregular refinements results in so-called *hanging nodes* shown as filled circles in the figure. The rule that irregular elements are not allowed to be refined is now replaced by rule that only one hanging node is allowed on an edge or in other words, that neighboring elements differ at most in one level of refinement (in two space dimensions). However, some of the algorithmic complexity is now shifted into the finite element procedure. Since the finite element functions are required to be continuous over edges the value in a hanging node needs to coincide with the interpolated value from the coarse side. In P_1 using a Lagrange basis this means that the nodal value in the hanging node is not a degree of freedom but is interpolated from the degrees of freedom at the two end points of the edge.

The figures 9.1 and 9.2 below illustrate the different local mesh refinement procedures for the L-shaped domain. All these mesh refinement algorithms are available in the DUNE software framework.

9.4. Numerical Results

In order to evaluate the adaptive algorithm we consider again the reentrant corner problem given in example 7.14.

In a first experiment the choice of the bulk fraction parameter ρ is investigated for P_1 finite elements. Table 9.1 shows some properties of the meshes generated for a fixed tolerance value of 0.05. For small values of $\rho < 0.5$ the tolerance is reached with about 5000 degrees of freedom. Large values of $\rho > 0.6$ results in about 10000 degrees of freedom. On the other hand the number of iterations of the adaptive algorithm (each time requiring the solution of a finite element problem) decreases from 39 to 6. So in terms of computation time a value of $\rho \approx 0.5 \ldots 0.8$ is most effective. Also note that the meshes resulting in about the same finite element error may be quite different. For $\rho = 0.1$ the smallest mesh size is $h_t \approx 2^{-15} = 1/32768$ while for $\rho = 0.9$ the smallest mesh size is $h_t \approx 2^{-8} = 1/256$.

Next we compare different types of mesh refinement for P_1 and Q_1 finite elements in figure 9.3. The figure shows the estimated and the true error for non-conforming refinement (hanging nodes) with triangles, conforming triangu-

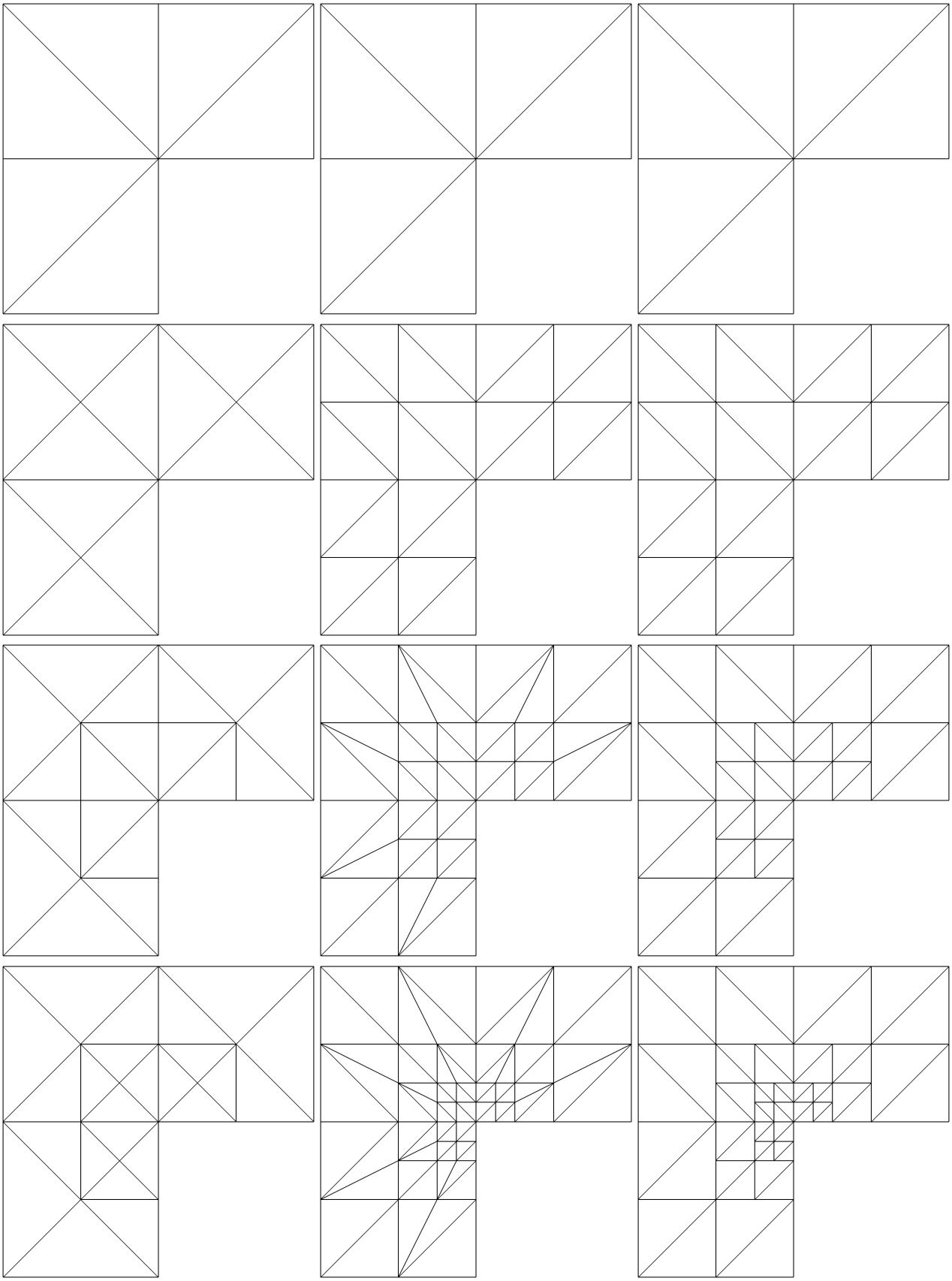


Figure 9.1.: Illustration of different local mesh refinement techniques for triangular elements. From left to right: Bisection refinement, regular refinement with conforming closure and regular refinement with hanging nodes.

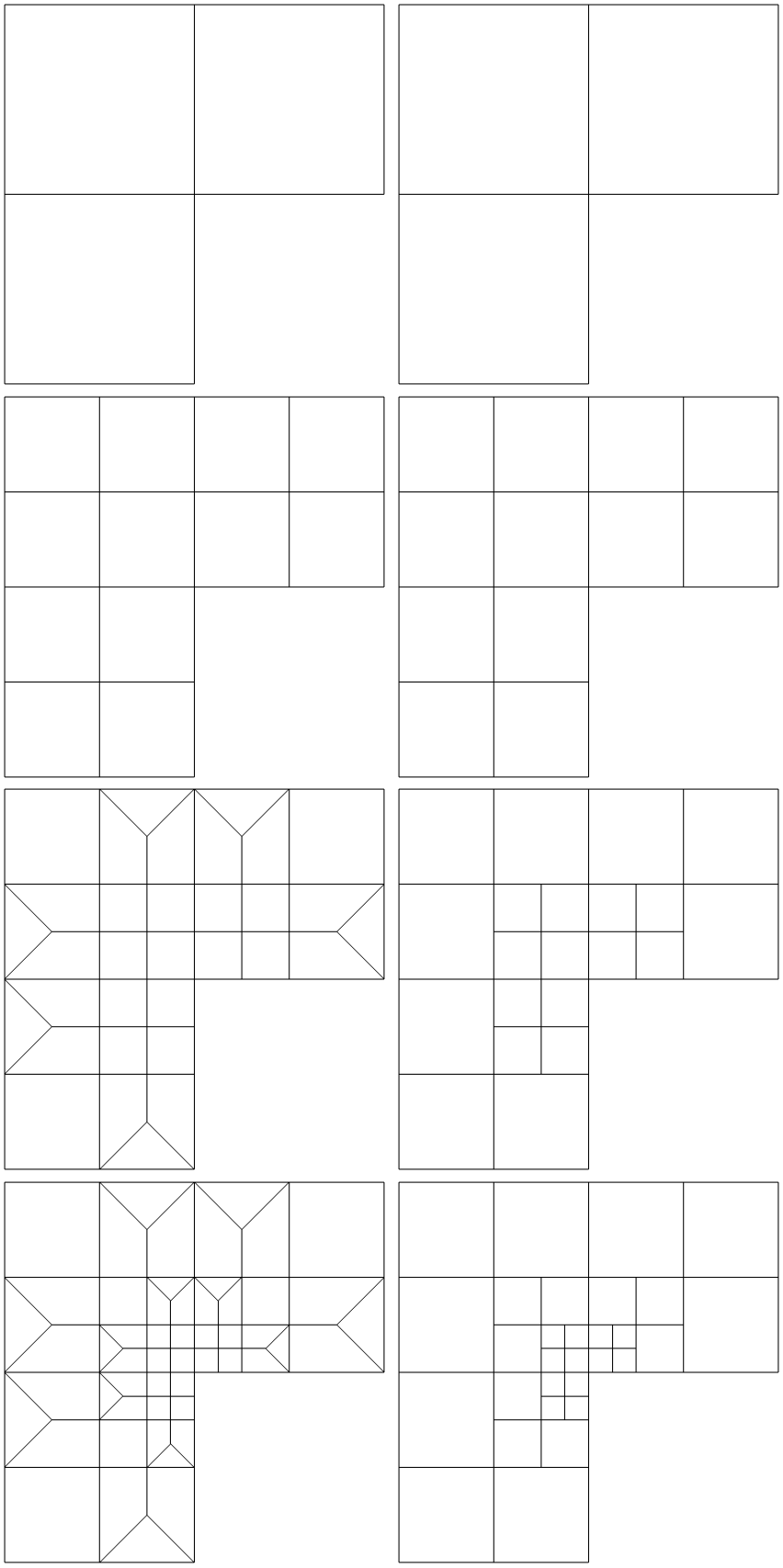


Figure 9.2.: Illustration of different local mesh refinement techniques for quadrilateral elements: Regular refinement with conforming closure and regular refinement with hanging nodes.

Table 9.1.: Efficiency in terms of accuracy per degrees of freedom of meshes generated with different values for the bulk fraction parameter ρ and a tolerance of 0.05.

ρ	iterations	depth	N	$\lvert u - u_h \rvert_{1,\Omega}$	η
0.1	39	15	5395	1.30e-02	4.94e-02
0.2	22	14	5498	1.29e-02	4.89e-02
0.3	15	14	5323	1.30e-02	4.99e-02
0.4	12	14	5933	1.22e-02	4.71e-02
0.5	10	12	6996	1.12e-02	4.30e-02
0.6	9	11	8312	1.04e-02	3.99e-02
0.7	8	10	10129	1.01e-02	3.77e-02
0.8	7	9	10501	1.15e-02	4.05e-02
0.9	6	8	10714	1.47e-02	4.83e-02

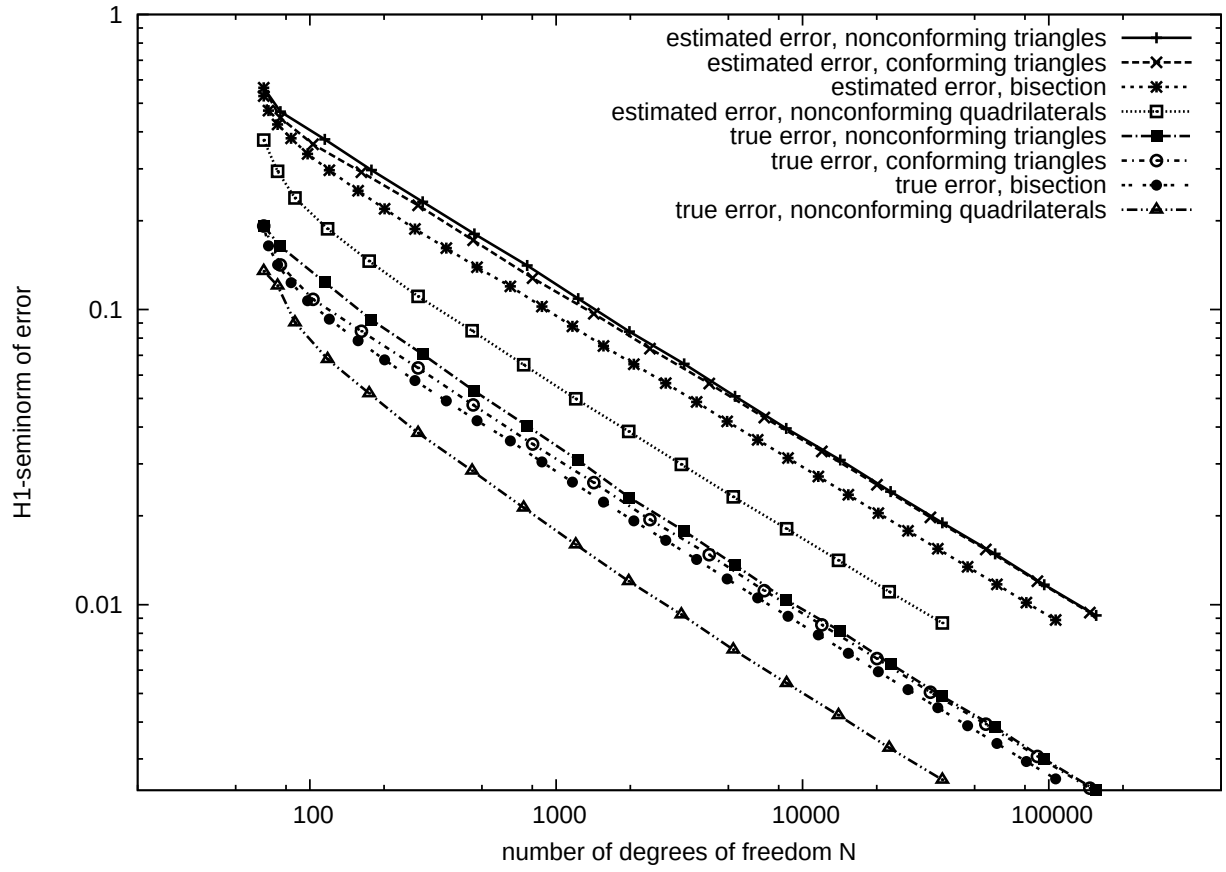


Figure 9.3.: Estimated and true error versus number of degrees of freedom for different local mesh refinement types. A bulk fraction parameter $\rho = 1/2$ was used.

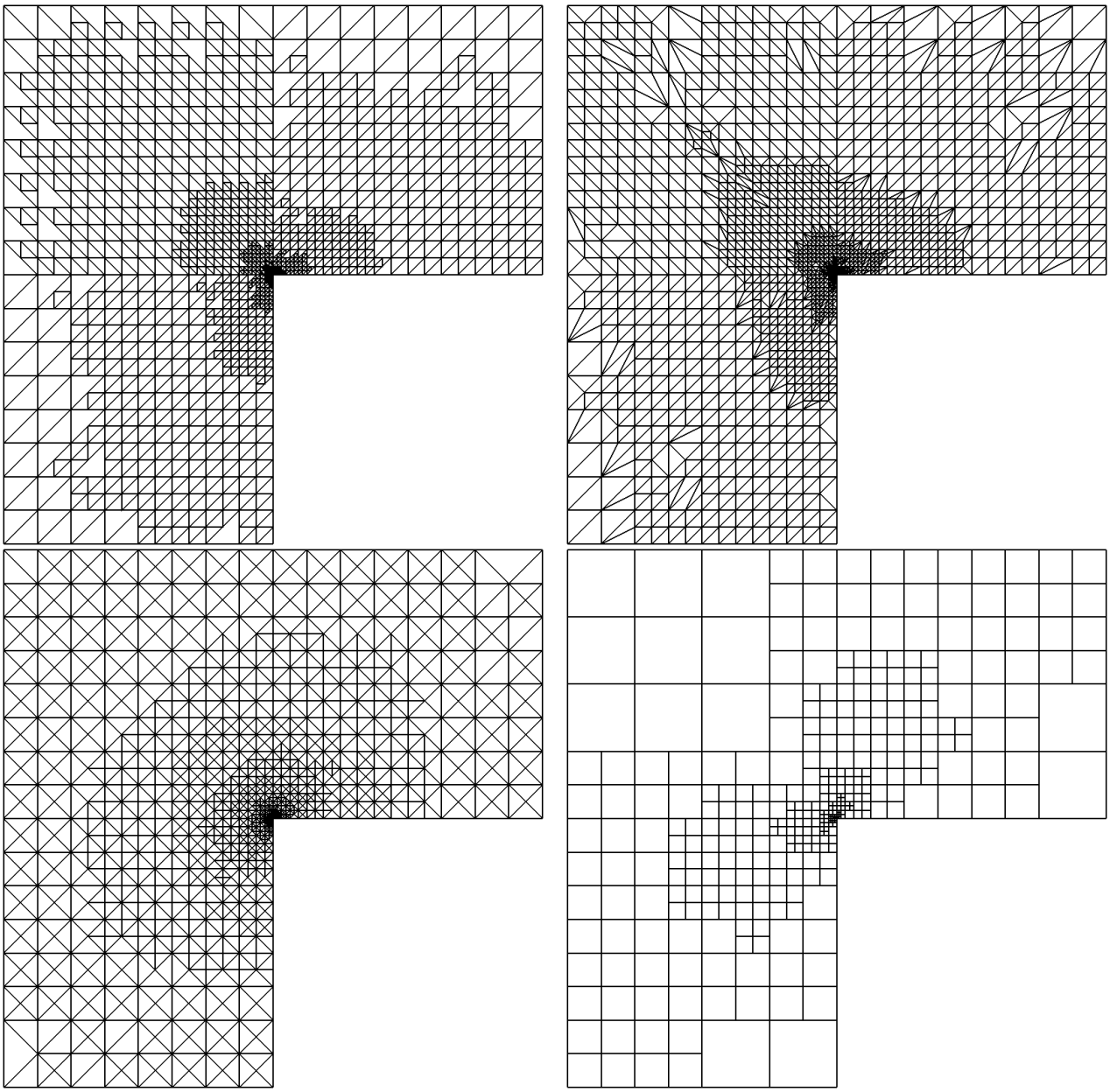


Figure 9.4.: Comparison of different local mesh refinement techniques at about the same absolute error of $|u - u_h|_{1,\Omega} = 0.03$: Nonconforming and conforming refinement with triangles, bisection refinement and nonconforming refinement with quadrilaterals (from top to bottom, left to right).

lar meshes, bisection type refinement and non-conforming quadrilateral mesh refinement. With respect to degrees of freedom conforming and non-conforming regular refinement on triangles is asymptotically identical (with small advantages for conforming refinement on coarse meshes). Bisection refinement is a bit more efficient and quadrilateral meshes are substantially more efficient. The efficiency index is about 3 for all types of meshes. The corresponding meshes for the four different refinement types are shown in figure 9.4. In the top row nonconforming and conforming regular refinement with triangles is shown. The refinement regions have very similar shapes (the conforming mesh is more refined). The bottom row shows bisection and nonconforming quadrilateral refinement. Clearly, these two meshes look different. What is very interesting that there is less refinement along the diagonal $y = -x$ in the quadrilateral case.

Finally, we turn to the combination of adaptive refinement and higher (but fixed) polynomial degree. Figure 9.5 shows the true error versus number of degrees of freedom (as a measure of computational complexity) for polynomial degree 1 and uniform refinement as well as polynomial degrees $1 \ldots 4$ using bisection type refinement. This figure can be compared directly to figure 7.10 in chapter 7. Clearly, in comparison to the case of uniform refinement shown in figure 7.10 the asymptotic convergence rate now improves with increasing polynomial degree. A quantitative comparison is shown in table 9.2. In the case of full regularity the convergence rate in the H^1-norm is $O(h^k)$ for polynomial degree k. For uniform refinement we have $h = N^{-1/2}$, i.e. the optimal convergence rate with respect to N is $O(N^{-k/2})$. The third column in table 9.2 shows that *we can recover the convergence rate expected for a fully regular solution* with respect to number of degrees of freedom! This is only possible through the combination of adaptive mesh refinement and increase of polynomial degree. The situation can be improved further by varying also the polynomial degree from element to element, choosing a high polynomial degree away from the reentrant corner and a low polynomial degree close to the corner. This leads to exponential convergence with respect to N. Table 9.2 also illustrates that the efficiency index depends slightly on the polynomial degree.

The meshes generated by the adaptive algorithm using P_1 and P_2 elements are shown in figure 9.6. The P_2 mesh is much more locally refined as is expected from the equilibration strategy. Away from the corner the error is reduced by $(1/2)^2$ for each refinement and near the corner it is only reduced by $(1/2)^{2/3}$ due to the low regularity.

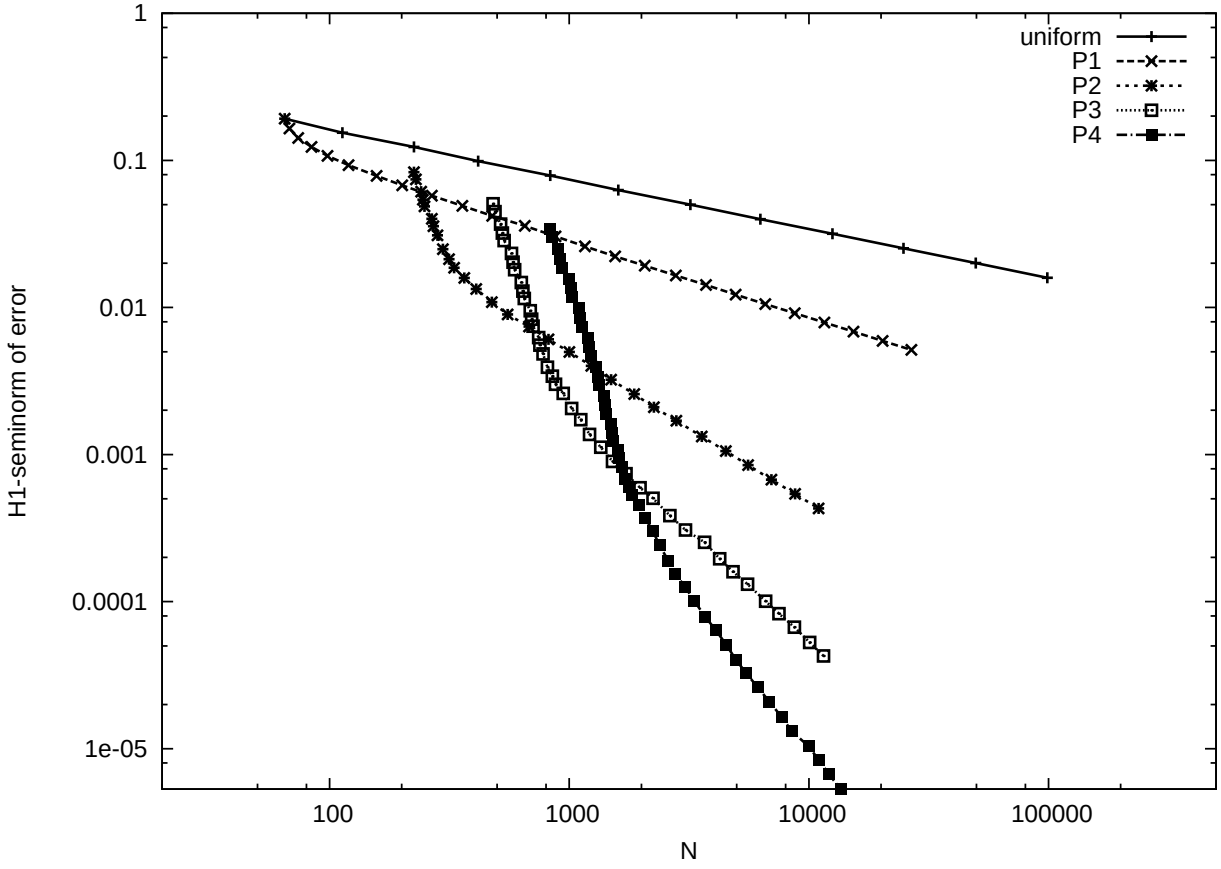


Figure 9.5.: Error versus number of degrees of freedom when combining adaptive mesh refinement with increasing polynomial degree.

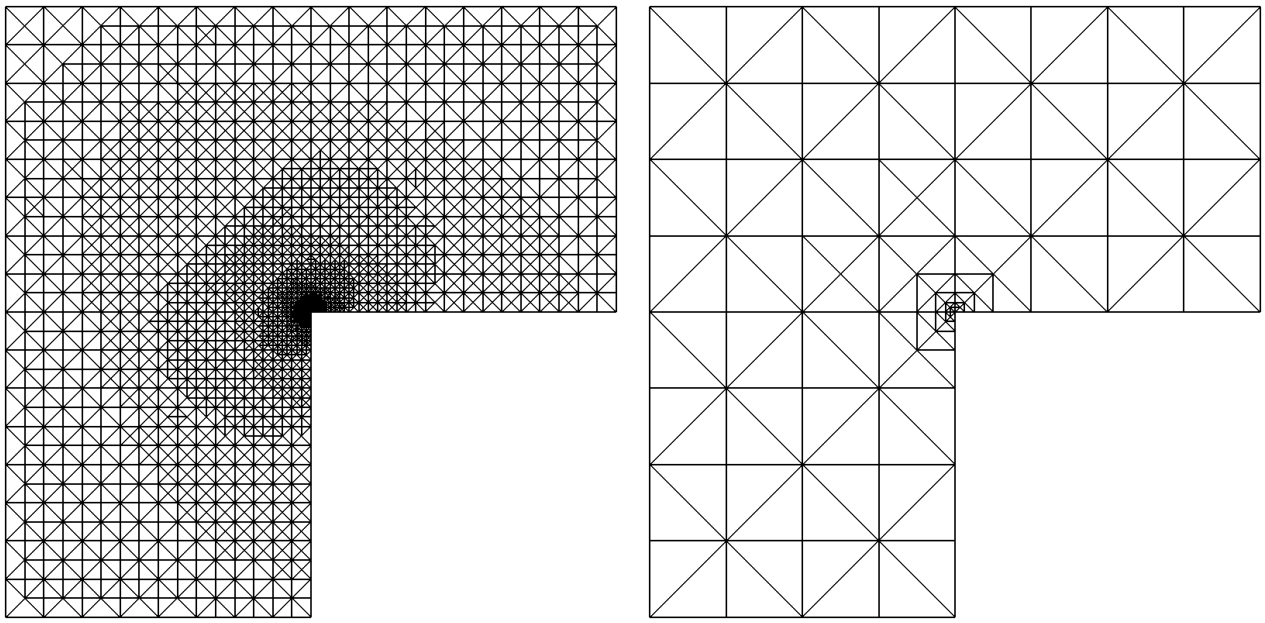


Figure 9.6.: Comparison of adaptive meshes using P_1 (left) and P_2 conforming finite elements at the same absolute error of $|u - u_h|_{1,\Omega} \approx 0.02$.

Table 9.2.: Convergence order and efficiency index for adaptive mesh refinement with varying polynomial degree. Bisection type refinement was used as refinement technique.

N	$\|u-u_h\|_{1,\Omega}$	|rate|	η	$\eta/\|u-u_h\|_{1,\Omega}$
$k=1$				
15355	6.84e-03	0.51	2.36e-02	3.44
20312	5.92e-03	0.52	2.04e-02	3.44
26779	5.15e-03	0.50	1.78e-02	3.44
$k=2$				
6979	6.76e-04	1.01	2.61e-03	3.86
8765	5.41e-04	0.98	2.08e-03	3.84
10985	4.30e-04	1.02	1.66e-03	3.86
$k=3$				
8701	6.70e-05	1.42	2.88e-04	4.30
10072	5.28e-05	1.63	2.27e-04	4.31
11515	4.28e-05	1.57	1.83e-04	4.27
$k=4$				
10993	8.39e-06	2.19	4.76e-05	5.68
12109	6.74e-06	2.26	3.81e-05	5.65
13541	5.32e-06	2.12	3.00e-05	5.64

Chapter 10.

Multigrid Methods

The linear systems arising in the finite element method are large, sparse and symmetric positive definite (assuming the bilinear form is symmetric and coercive). We now turn to the question how to solve them efficiently.

There are two basic approaches: Direct methods such as LU-decomposition or Cholesky decomposition produce a solution after a finite number of steps depending only on the size of the matrix A. However, this cost may be prohibitively large, the methods may need an excessive amount of memory and roundoff error might be a problem for ill-conditioned matrices. On the other hand iterative methods typically require little memory but the number of iterations needed to reduce the error to a acceptable amount depends strongly on the type of problem to be solved. Fortunately, symmetric positive definite linear systems are among the systems most amenable to iterative solution.

In particular the number of iterations needed in iterative methods often depends on the spectral condition number of the matrix $\kappa_2(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$. For a Lagrange finite element basis one can show:

$$\kappa_2(A) = O(h^{-2})$$

where h is the mesh size (uniform mesh).

The following table gives the number of arithmetic operations needed to solve a symmetric and positive definite linear system arising from the finite element discretization (operations are up to a constant factor depending on the coefficients and the required accuracy):

Scheme	$d = 2$	$d = 3$
Gaussian elimination	N^3	N^3
Banded-Gauß	N^2	$N^{7/3}$
Nested dissection	$N^{3/2}$	N^2
Gauß-Seidel, Jacobi	N^2	$N^{5/3}$
CG, SOR(ω_{opt})	$N^{3/2}$	$N^{4/3}$
Multigrid	N	N

The multigrid method is one of the few methods that is able to solve the systems in question with a computational complexity that is linear in the number of unknowns. The main purpose of the rest of this chapter is then to introduce the multigrid method and to prove its optimal convergence.

10.1. Some Examples

As a motivation to develop fast methods we now illustrate the performance of several iterative methods for different model problems

All tables in this section give the number of iterations needed to reduce the defect norm $\|b - Ax^0\|$ with respect to the initial guess x^0 by the factor 10^{-8}. Run-times are given in seconds (Intel 2.5 GHz Core 2 Duo Processor, gcc-4.2 with -O2 optimization). The maximum number of iterations allowed was 20000. Empty entries indicate that the required reduction was not reached within the maximum number of iterations.

Example 10.1. Model problem A reads as follows:

$$\begin{aligned} -\Delta u &= (2d - 4\|x\|^2)e^{-\|x\|^2} && \text{in } \Omega = (0,1)^d, \\ u &= e^{-\|x\|^2} && \text{on } \partial\Omega \end{aligned}$$

with the exact solution

$$u(x) = e^{-\|x\|^2}.$$

This model problem illustrates the most basic case with constant coefficients, Dirichlet boundary conditions and full regularity. The exact solution is illustrated in figure 10.1.

Table 10.1 shows results for seven different methods. Jacobi, Gauß-Seidel, gradient method (steepest descent) and gradient method preconditioned by the symmetric Gauß-Seidel (SGS) method all have a number of iterations that is $O(h^{-2})$. Consequently, the number of iterations increases by a factor of four with each mesh refinement and the computation time increases by a factor of 16 in 2d and 32(!) in 3d for each refinement.

The conjugate gradient method without preconditioning (CG) and preconditioned by SGS and incomplete LU-decomposition (ILU0) exhibit a number of iterations which is $O(h^{-1})$ and therefore the iteration number doubles with each refinement and the computation increases by factors 8 and 16 in 2d and 3d, respectively. It is clear that asymptotically as $N \to \infty$ the overall computation time is dominated by the time it takes to solve the system of linear equations as all other parts of the finite element procedure scale linearly in N. □

Example 10.2. Model problem B reads

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega = (0,1)^d, \\ u &= g && \text{on } \Gamma_D, \\ -\nabla u \cdot \nu &= j && \text{on } \Gamma_N, \end{aligned}$$

.png

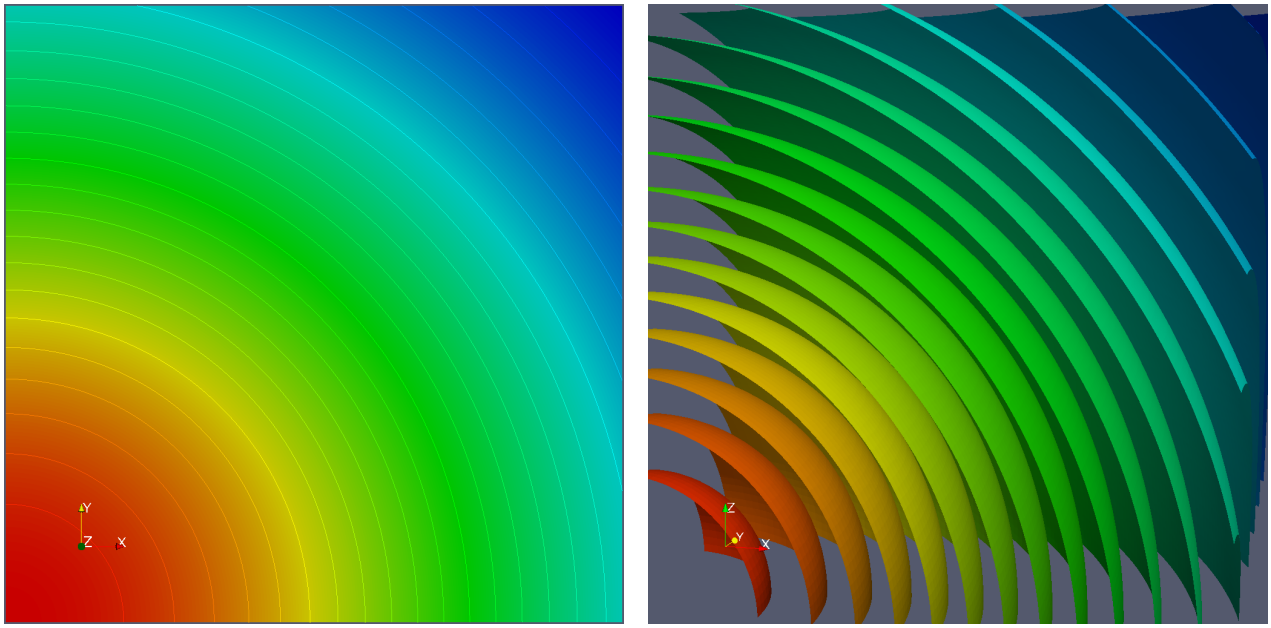


Figure 10.1.: Solution of model problem A in two and three space dimensions.

Table 10.1.: Results for model problem A.

Model problem A, Q_1, 2d														
h	Jacobi		Gauß-Seidel		Gradient		Grad+SGS		CG		CG+SGS		CG+ILU0	
	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time
1/8	147		75		113		24		16		10		8	
1/16	562	0.01	282		431		79		35		18		14	
1/32	2113	0.15	1056	0.08	1621	0.06	275	0.03	69		34		25	
1/64	7886	2.18	3939	1.10	6059	0.94	1011	0.43	136	0.03	64	0.03	46	0.01
1/128			14615	16.1			3741	6.42	266	0.18	120	0.22	87	0.10
1/256							13823	115	521	1.94	217	1.89	162	1.23
Model problem A, P_1, 2d														
h	Jacobi		Gauß-Seidel		Gradient		Grad+SGS		CG		CG+SGS		CG+ILU0	
	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time
1/8	218		112		220		51		22		13		13	
1/16	840	0.02	427		854		177		48		26		24	
1/32	3165	0.21	1607	0.11	3230	0.12	645	0.07	98		49		45	
1/64	11820	3.04	6004	1.57	12096	1.74	2403	0.95	193	0.03	95	0.04	88	0.02
1/128							8955	13.9	378	0.24	184	0.30	172	0.20
1/256									739	2.25	359	2.58	336	2.18
Model problem A, Q_1, 3d														
h	Jacobi		Gauß-Seidel		Gradient		Grad+SGS		CG		CG+SGS		CG+ILU0	
	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time
1/8	98	0.01	51		77		18		16		9		8	
1/16	376	0.24	189	0.12	290	0.10	55	0.05	34	0.01	17	0.02	15	0.01
1/32	1416	10.1	708	4.87	1087	4.10	187	1.95	67	0.26	32	0.34	27	0.25
1/64	5287	304.	2641	152.	4063	129.	681	65.6	132	4.43	59	5.86	51	4.18

with

$$f(x) = \begin{cases} 50 & 0.25 \leq x_0, x_1 \leq 0.375 \\ 0 & \text{else} \end{cases},$$

and

$$\Gamma_N = \{x \,|\, x_1 = 0 \vee x_1 = 1 \vee (x_0 = 1 \wedge x_1 > 1/2)\} \qquad \Gamma_D = \partial\Omega \setminus \Gamma_N,$$

and

$$g(x) = e^{-\|x - x_0\|^2}, \qquad x_0 = (1/2, \ldots, 1/2)^T,$$

as well as

$$j(x) = \begin{cases} -5 & x_0 = 1 \wedge x_1 > 1/2 \\ 0 & \text{else} \end{cases}.$$

This problem illustrates a case with mixed boundary conditions inducing a singularity. The solution is illustrated in figure 10.2 and the corresponding results are given in table 10.2. The behavior is pretty much similar to model problem A. In particular there is no influence of the regularity of the solution on the number of iterations needed. This can be proven in general: The asymptotic convergence rate for a linear iterative method only depends on the iteration matrix but not on the right hand side. It may depend on the initial guess but this would be very lucky (one would have to choose a suitable initial guess in a subspace spanned by eigenvectors corresponding to small eigenvalues of the iteration matrix). □

Example 10.3. Model problem C reads

$$\begin{aligned} -\nabla \cdot \{k(x)\nabla u\} &= 1 && \text{in } \Omega = (0,1)^d, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

with

$$k(x) = \begin{cases} 20.0 & \lfloor x_0/H \rfloor \text{ even}, \lfloor x_1/H \rfloor \text{ even}, \lfloor x_2/H \rfloor \text{ even} \\ 0.002 & \lfloor x_0/H \rfloor \text{ odd}, \lfloor x_1/H \rfloor \text{ even}, \lfloor x_2/H \rfloor \text{ even} \\ 0.2 & \lfloor x_0/H \rfloor \text{ even}, \lfloor x_1/H \rfloor \text{ odd}, \lfloor x_2/H \rfloor \text{ even} \\ 2000.0 & \lfloor x_0/H \rfloor \text{ odd}, \lfloor x_1/H \rfloor \text{ odd}, \lfloor x_2/H \rfloor \text{ even} \\ 1000.0 & \lfloor x_0/H \rfloor \text{ even}, \lfloor x_1/H \rfloor \text{ even}, \lfloor x_2/H \rfloor \text{ odd} \\ 0.001 & \lfloor x_0/H \rfloor \text{ odd}, \lfloor x_1/H \rfloor \text{ even}, \lfloor x_2/H \rfloor \text{ odd} \\ 0.1 & \lfloor x_0/H \rfloor \text{ even}, \lfloor x_1/H \rfloor \text{ odd}, \lfloor x_2/H \rfloor \text{ odd} \\ 10.0 & \lfloor x_0/H \rfloor \text{ odd}, \lfloor x_1/H \rfloor \text{ odd}, \lfloor x_2/H \rfloor \text{ odd} \end{cases}.$$

This model problem illustrates a case with highly variable diffusion coefficient (in fact a coefficient field with cross-points leading to a solution with very low regularity).

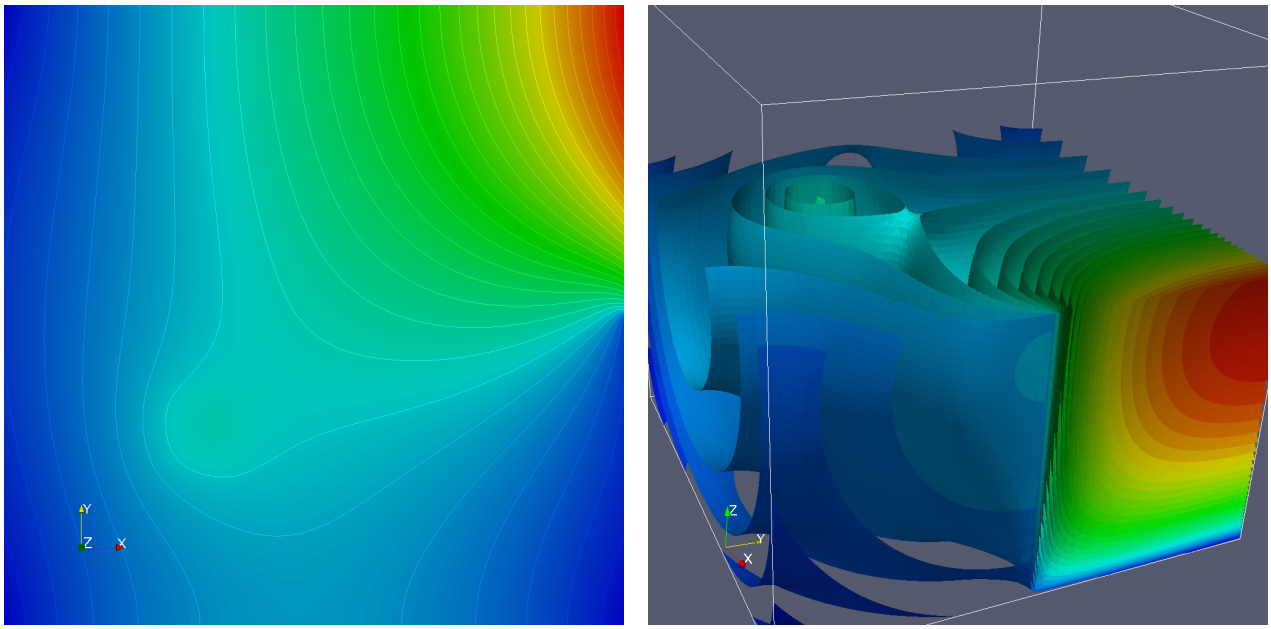


Figure 10.2.: Solution of model problem B in two and three space dimensions.

Table 10.2.: Results for model problem B.

Model problem B, Q_1, 2d														
h	Jacobi		Gauß-Seidel		Gradient		Grad+SGS		CG		CG+SGS		CG+ILU0	
	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time
1/8	456		230		424		65		32		14		11	
1/16	1770	0.07	888	0.02	1504	0.01	237		59		24		18	
1/32	6720	0.43	3364	0.21	5436	0.22	877	0.09	112		45		32	
1/64			12614	3.20	19895	3.11	3249	1.28	215	0.04	87	0.04	61	0.02
1/128							12055	18.8	415	0.28	168	0.27	118	0.13
1/256									806	2.88	328	2.63	231	1.71
Model problem B, P_1, 2d														
h	Jacobi		Gauß-Seidel		Gradient		Grad+SGS		CG		CG+SGS		CG+ILU0	
	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time
1/8	667		338		830		138		41		18		16	
1/16	2619	0.04	1327	0.02	2969	0.03	525	0.01	82		35		32	
1/32	10009	0.60	5075	0.32	10778	0.40	2017	0.20	159		68		62	
1/64			19131	4.57			7637	2.81	306	0.05	133	0.05	124	0.04
1/128									590	0.36	259	0.39	244	0.28
1/256									1143	3.45	505	3.47	478	3.08
Model problem B, Q_1, 3d														
h	Jacobi		Gauß-Seidel		Gradient		Grad+SGS		CG		CG+SGS		CG+ILU0	
	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time
1/8	180	0.01	92		176		29		29		12		10	
1/16	694	0.42	349	0.21	596	0.22	95	0.09	54	0.02	22	0.02	19	0.01
1/32	2622	17.6	1313	8.74	2126	7.86	343	3.54	102	0.39	42	0.44	35	0.32
1/64	9813	531.	4908	263.	7747	240.	1269	119.	197	6.42	80	7.70	67	5.40

The solution for this model problem is illustrated in figure 10.3 and results are shown in table 10.3. In comparison to model problems A and B the iteration numbers are much higher. In fact, only the preconditioned conjugate gradient method (last two columns) is able to solve this model problem for reasonable mesh sizes. This illustrates that a robust and efficient linear solver is crucial for the finite element method! □

10.2. Smoothing Property of Richardson Iteration

We want to solve

$$Ax = b$$

with symmetric and positive definite A. One of the simplest iterative methods is the so-called Richardson iteration given by

$$x^{k+1} = x^k + \omega(b - Ax^k)$$

which converges for $\omega < 2/\lambda_{\max}(A)$. Since A is s.p.d. we have the spectrum

$$\sigma(A) = \{\lambda_{\min}(A) = \lambda_1, \ldots, \lambda_N = \lambda_{\max}(A)\}$$

with the ordered eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$.

For the iteration error $x - x^k$ we have the recursion

$$e^{k+1} = x - x^{k+1} = x - x^k - \omega(b - Ax^k) = (I - \omega A)(x - x^k) = (I - \omega A)e^k.$$

$M = I - \omega A$ is the *iteration matrix* of Richardson's method. Now let (λ_i, z_i) be an eigenpair of A. Then

$$Mz_i = (I - \omega A)z_i = (1 - \omega\lambda_i)z_i \ .$$

Since the z_i form a basis of $\mathbb{R}^N$, any error e^k can be written in this basis and we get

$$Me = M\sum_{i=1}^{N} c_i z_i = \sum_{i=1}^{N} c_i(1 - \omega\lambda_i)z_i \ .$$

Setting $\omega = 1/\lambda_N$ we observe for the reduction factor

$$1 - \omega\lambda_i = 1 - \frac{\lambda_i}{\lambda_N} = \begin{cases} \text{small} & (\to 0) & i \text{ large} \\ \text{large} & (\to 1) & i \text{ small} \end{cases} .$$

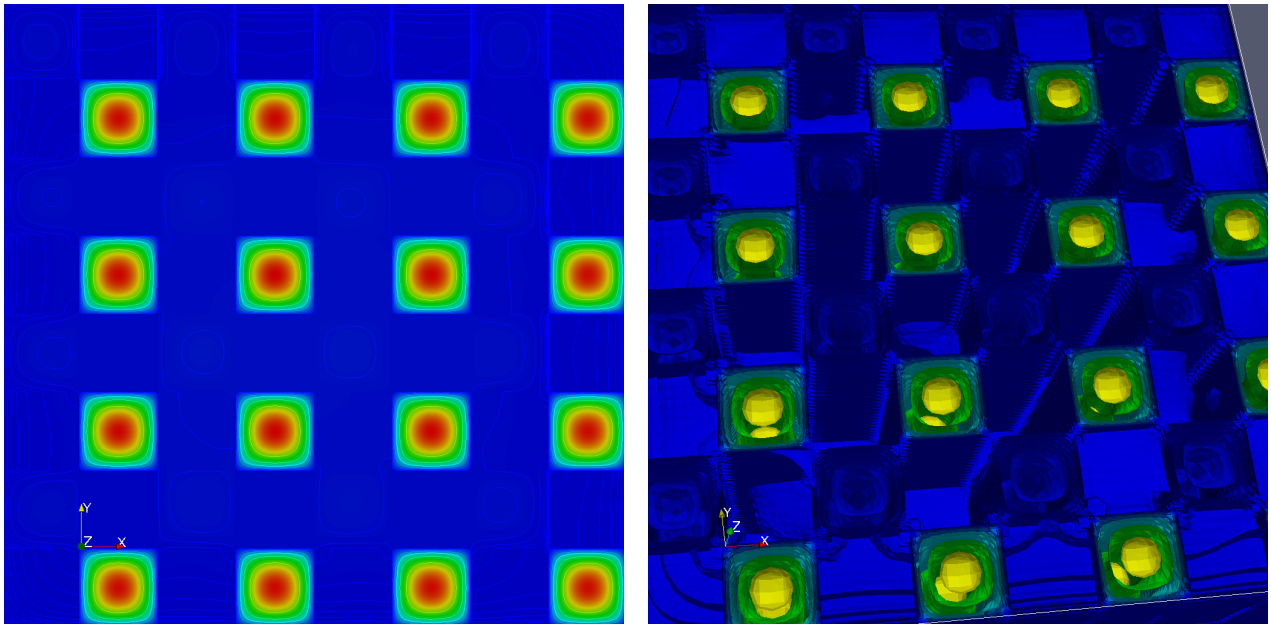


Figure 10.3.: Solution of model problem C in two and three space dimensions.

Table 10.3.: Results for model problem C.

Model problem C, Q_1, 2d														
h	Jacobi		Gauß-Seidel		Gradient		Grad+SGS		CG		CG+SGS		CG+ILU0	
	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time
1/8	4665	0.06	2354	0.01	3334	0.01	724		27		17		8	
1/16			13573	0.26			4335	0.12	281		38		27	
1/32							17512	1.91	1761	0.08	73		52	
1/64									8644	1.48	142	0.06	99	0.03
1/128											282	0.49	196	0.22
1/256											577	4.82	405	2.96
Model problem C, Q_1, 3d														
h	Jacobi		Gauß-Seidel		Gradient		Grad+SGS		CG		CG+SGS		CG+ILU0	
	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time	IT	Time
1/8	127	0.01	65		96		22		21		10		8	
1/16	1326	0.83	667	0.42			208	0.20	1179	0.45	32	0.03	23	0.02
1/32	9966	68.2	4996	34.8			1425	14.8	8594	32.9	71	0.76	56	0.51
1/64							8382	792.			151	14.6	124	9.96

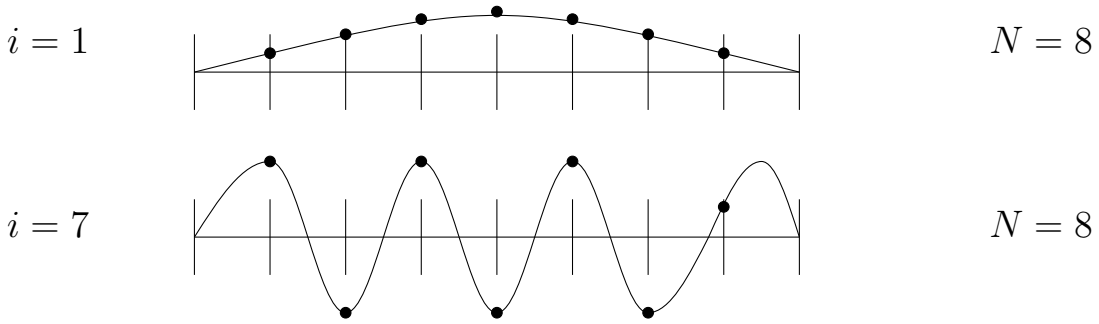
Example 10.4. For a discretization of $-u'' = f$ with $u(0) = u(1) = 0$ using P_1 finite elements on an equidistant mesh we obtain the tridiagonal matrix (scaled by h):

$$A = \text{tridiag}(-1, 2, -1)$$

which has eigenvalues $\lambda_i = 4\sin^2(i\pi h/2)$, $1 \le i < N$, and corresponding eigenvectors

$$(z_i)_k = \sin\left(\frac{k}{N} i\pi\right) \qquad 1 \le i, k < N.$$

These eigenvectors are illustrated for a low frequency $i = 1$ and a high frequency $i = 7$ (relative to $N = 8$) in the following figure:

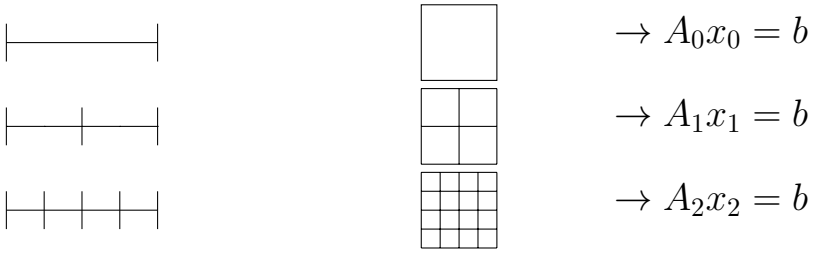


Whether $\sin(i\pi x)$ is a high or low frequency function depends on the mesh size $h = 1/N$. □

A possible remedy of the problem is the following:

- High frequency errors $i > N/2$ are damped efficiently by Richardson iteration.
- Low frequency errors $i \le N/2$ are damped slowly by Richardson iteration.
- The transfer of low frequency errors to a coarser grid would make them appear more high frequent there and Richardson iteration would remove them there.

The realization of this idea requires the representation of errors on a hierarchy of coarser grids as they are produced naturally by the refinement algorithms:



10.3. Variational Multigrid

Let $V_h = \text{span}\{\varphi_1, \ldots, \varphi_{N_h}\}$ be the discrete finite element space. The finite element ismorphism

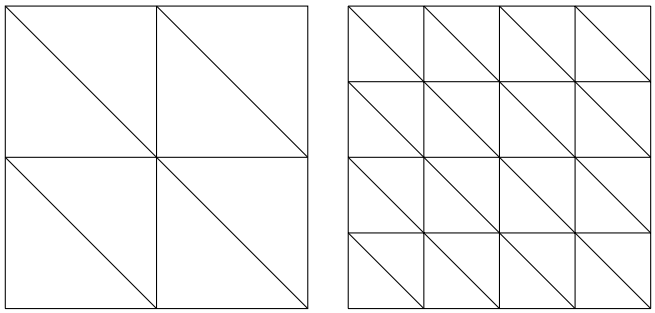
$$v = \text{FE}(x) = \sum_{i=1}^{N_h} x_i \varphi_i$$

establishes a one-to-one correspondence between

$$x \in \mathbb{R}^{N_h} \longleftrightarrow v \in V_h.$$

Below we will make use of the fact that multigrid components can be either interpreted in terms of matrices and vectors or function spaces and weak formulations. The derivation of the two- and multi-grid method in terms of the variational formulation is called variational multigrid.

Consider two levels of mesh hierarchy obtained from hierarchic refinement.



$\mathcal{T}_{2h}$ $\mathcal{T}_h$

Then the corresponding finite element spaces are nested, i.e.

$$V_{2h} = P_1(\mathcal{T}_{2h}) \subset V_h = P_1(\mathcal{T}_h).$$

The two-grid method is then defined as follows:

1) Given $u_h^k \in V_h$ define $u_h^{k,1}$ to be the finite element function obtained after ν-fold application of Richardson's method. I.e. given that $u_h^k = \text{FE}(x_h^k)$, set

$$x_h^{k,0} = x_h^k, \qquad x_h^{k,\frac{i}{\nu}} = x_h^{k,\frac{i-1}{\nu}} + \omega \left(b - A_h x_h^{k,\frac{i-1}{\nu}} \right), \quad 1 \le i \le \nu,$$

and $u_h^{k,1} = \text{FE}(x_h^{k,1})$.

2) Apply "coarse grid correction" which means to solve the finite element problem

$$\text{Find } w \in V_{2h}: \qquad a(u_h^{k,1} + w, v) = l(v) \quad \forall v \in V_{2h} \tag{10.1}$$

and set $u_h^{k+1} = u_h^{k,1} + w$.

The algebraic interpretation of the coarse grid correction step (10.1) is obtained by inserting a basis representation of the function spaces involved. Let us define the basis by $V_h = \text{span}\{\varphi_1^h, \ldots, \varphi_{N_h}^h\}$, $V_{2h} = \text{span}\{\varphi_1^{2h}, \ldots, \varphi_{N_{2h}}^{2h}\}$. Since $V_{2h} \subset V_h$, there exist coefficients r_{ij} such that:

$$\varphi_i^{2h} = \sum_{j=1}^{N_h} r_{ij} \varphi_j^h, \qquad \forall 1 \leq i \leq N_{2h} \; . \tag{10.2}$$

Now insert the basis representation into (10.1):

$$\begin{aligned}
& a(u_h^{k,1} + w, v) = l(v) \qquad v \in V_{2h} \\
\Leftrightarrow \quad & a(w, v) = l(v) - a(u_h^{k,1}, v) \qquad v \in V_{2h} \\
\Leftrightarrow \quad & a\left(\sum_{j=1}^{N_{2h}} y_j \varphi_j^{2h}, \varphi_i^{2h}\right) = l(\varphi_i^{2h}) - a\left(\sum_{n=1}^{N_h} x_n^{k,1} \varphi_n^h, \varphi_i^{2h}\right) \qquad 1 \leq i \leq N_{2h} \\
\Leftrightarrow \quad & \sum_{j=1}^{N_{2h}} y_j a(\varphi_j^{2h}, \varphi_i^{2h}) = l\left(\sum_{m=1}^{N_h} r_{im} \varphi_m^h\right) - a\left(\sum_{n=1}^{N_h} x_n^{k,1} \varphi_n^h, \sum_{m=1}^{N_h} r_{im} \varphi_m^h\right) \\
& \qquad\qquad = \sum_{m=1}^{N_h} r_{im} \left[l(\varphi_m^h) - \sum_{n=1}^{N_h} x_n^{k,1} a(\varphi_n^h, \varphi_m^h) \right] \\
\Leftrightarrow \quad & A_{2h} y = R_{2h}^h (b_h - A_h x_h^{k,1}) \; .
\end{aligned}$$

where the rectangular restriction matrix R_{2h}^h is given by $(R_{2h}^h)_{ij} = r_{ij}$.

This allows us now to formulate the two-grid method in an algebraic way.

Algorithm 10.5 (Two-grid method)**.** Denote the systems on the fine grid by $A_h x_h = b_h$ and on the coarse grid by $A_{2h} x_{2h} = b_{2h}$. Let the current iterate x_h^k on the fine grid be given. The following function computes the next iterate x_h^{k+1}.

```
TGM(x_h^k)
{
  x_h^{k,1} = x_h^k;
  for(κ = 1,...,ν) x_h^{k,1} = x_h^{k,1} + ω(b_h − A_h x_h^{k,1}); //Pre-smoothing
  d_h = b_h − A_h x_h^{k,1}; // Calculate defect
  d_2h = R_2h^h d_h; // Restriction
  y_2h = A_2h^{-1} d_2h; // Coarse grid solve
  y_h = (R_2h^h)^T y_2h; // Prolongation
  x_h^{k,2} = x_h^{k,1} + y_h; // Coarse grid correction
  return x_h^{k,2};
}
```

□

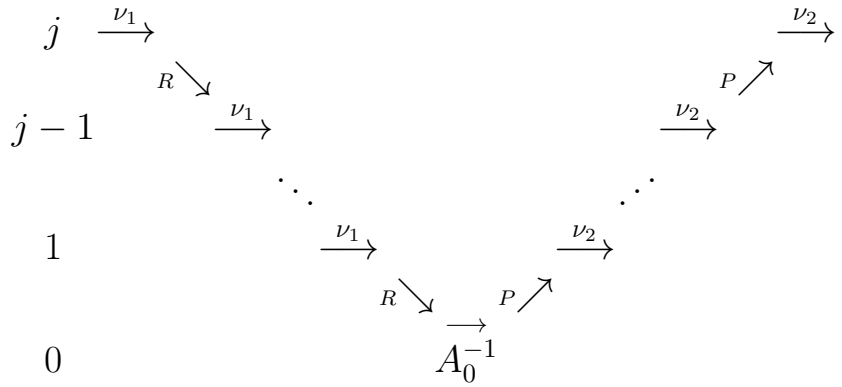
In the multigrid method the exact coarse grid solve is replaced by a recursive application of the method. The systems on the levels 0 (coarsest) to J (finest) are now denoted by $A_j x_j = b_j$.

Algorithm 10.6 (Multigrid method)**.** Let $\nu_1, \nu_2, \gamma \in \mathbb{N}$ be given parameters. Then the following function computes the new iterate x_j^{k+1} on mesh level j:

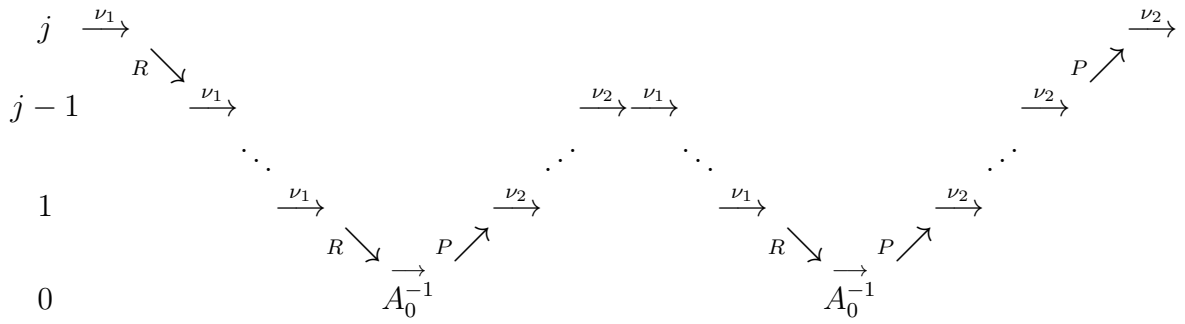
```
MGM(j, x_j^k, b_j)
{
  if (j == 0) { x_j^k = A_j^{-1} b_j; return x_j^k; }
  x_j^{k,1} = x_j^k;
  for(κ = 1,...,ν_1) x_j^{k,1} = x_j^{k,1} + ω(b_j − A_j x_j^{k,1}); //Pre-smoothing
  d_j = b_j − A_j x_j^{k,1}; // Calculate defect
  d_{j−1} = R_{j−1}^j d_j; // Restriction
  x_{j−1} = 0; // Initial value for coarse grid
  if (j == 1) γ̄ = 1; else γ̄ = γ;
  for (i = 1,...,γ̄)
    x_{j−1} = MGM(j − 1, x_{j−1}, d_{j−1}); // Approximate coarse grid solve
  y_j = (R_{j−1}^j)^T x_{j−1}; // Prolongation
  x_j^{k,2} = x_j^{k,1} + y_j; // Coarse grid correction
  x_j^{k,3} = x_j^{k,2};
  for (κ = 1,...,ν_2) x_j^{k,3} = x_j^{k,3} + ω(b_j − A_j x_j^{k,3}); //Post-smoothing
  return x_h^{k,3};
}
```

□

The parameter γ is called the cycle form parameter. For $\gamma = 1$ the call of MGM on level j carries out the following steps:



Because of this form $\gamma = 1$ is called the V-cycle. Setting $\gamma = 2$ results in the following sequence of steps:



This cycle is therefore called the W-cycle.

10.4. Convergence Analysis

There are many different proofs of convergence for the multigrid method. We concentrate here on the first rigorous multigrid proof due to W. Hackbusch, see e.g. the seminal book Hackbusch [1985]. The essence of this proof is to establish the optimal convergence of the two-grid method. From two-grid convergence the multigrid convergence for the W-cycle can be obtained. The advantage of this proof is that it is rather simple and it shows an improvement of the convergence factor with the number of smoothing steps. It has also several disadvantages: it does not show V-cycle convergence, it requires H^2-regularity and it requires that the number of smoothing steps is large enough. In fact non of these requirements is actually necessary for the method to work optimally.

Given x_h^k we obtain the following error recursion for the smoothing step

$$e^{k,1} = x_h - x_h^{k,1} = (I_h - \omega A_h)^{\nu_1}(x_h - x_h^k) = S_h^{\nu_1} e^k$$

and for the coarse grid correction step:

$$\begin{aligned} e^{k+1} &= x_h - x_h^{k+1} \\ &= x_h - \left(x_h^{k,1} + (R_{2h}^h)^T A_{2h}^{-1} R_{2h}^h (b_h - A_h x_h^{k,1}) \right) \\ &= e^{k,1} - (R_{2h}^h)^T A_{2h}^{-1} R_{2h}^h A_h e^{k,1} \\ &= \left(I_h - (R_{2h}^h)^T A_{2h}^{-1} R_{2h}^h A_h \right) e^{k,1}. \end{aligned}$$

The complete iteration matrix of one step of the two-grid method is:

$$\begin{aligned} e^{k+1} &= \left(I_h - (R_{2h}^h)^T A_{2h}^{-1} R_{2h}^h A_h \right) S_h^{\nu_1} e^k \\ &= \left(A_h^{-1} - (R_{2h}^h)^T A_{2h}^{-1} R_{2h}^h \right) A_h S_h^{\nu_1} e^k \end{aligned}$$

Taking appropriate norms and splitting up the operator into two parts we get

$$\|e^{k+1}\| \leq \underbrace{\|A_h^{-1} - (R_{2h}^h)^T A_{2h}^{-1} R_{2h}^h\|}_{\text{“approximation property”}} \underbrace{\|A_h S_h^{\nu_1}\|}_{\text{“smoothing property”}} \|e^k\|.$$

It turns out that the appropriate norm is the Euclidean norm.

A scale of norms Since A is s.p.d. we can define the following norms:

$$|||x|||_s := (x, A^s x)^{\frac{1}{2}} \qquad s = 0, 1, 2$$

(for $s = 0, 2$ s.p.d. is not necessary). In detail:

$s = 0$	$\|\|\|x\|\|\|_0 = (x, x)^{\frac{1}{2}}$	Euclidean norm
$s = 1$	$\|\|\|x\|\|\|_1 = (x, Ax)^{\frac{1}{2}}$	Energy norm
$s = 2$	$\|\|\|x\|\|\|_2 = (Ax, Ax)^{\frac{1}{2}}$	Defect norm

The last name stems from the fact

$$|||x - x^k|||_2^2 = (A(x - x^k), A(x - x^k)) = \underbrace{\|b - Ax^k\|^2}_{\text{Euclidean norm!}}.$$

The norm $|||.|||_s$ can be extended to the case $s \in \mathbb{R}$ as follows: Since A is s.p.d. there exists unitary Q (i.e. $Q^T = Q^{-1}$) such that $A = Q^T D Q$, $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$, $\lambda_i \in \sigma(A)$. Then set $A^s = Q^T D^s Q$ with $D^s := \operatorname{diag}(\lambda_1^s, \ldots, \lambda_N^s)$. So we get:

$$\begin{aligned} |||x|||_s &= (x, A^s x)^{\frac{1}{2}} = (x, Q^T D^s Q x)^{\frac{1}{2}} \\ &= (x, Q^T D^{\frac{s}{2}} Q Q^T D^{\frac{s}{2}} Q x)^{\frac{1}{2}} = (Q^T D^{\frac{s}{2}} Q x, Q^T D^{\frac{s}{2}} Q x)^{\frac{1}{2}} \\ &= \|A^{\frac{s}{2}} x\| \ . \end{aligned}$$

We now relate the Sobolev norms $\|v_h\|_{0,\Omega}$ and $\|v_h\|_{1,\Omega}$ of a finite element function with the corresponding norms of its coefficient vector $|||x|||_0$ and $|||x|||_1$.

From $|||x|||_1^2 = (x, Ax) = a(v_h, v_h)$, coercivity and continuity we conclude:

$$\alpha \|v_h\|_{1,\Omega} \leq |||x|||_1^2 = a(v_h, v_h) \leq C \|v_h\|_{1,\Omega}^2 \tag{10.3}$$

with α, C independent of h. Note that (10.3) is independent of the basis of V_h.

Lemma 10.7. Let $\Phi_h = \{\varphi_1, \ldots, \varphi_N\}$ be the P_1 Lagrange-basis for V_h on a family of *uniform* and shape regular triangulations and assume $v_h = \mathrm{FE}(x)$. Then there exist constants c_1, c_2 independent of h, but dependent on the mesh $\mathcal{T}_h$ such that

$$c_1 h^{\frac{n}{2}} |||x|||_0 \leq \|v\|_{0,\Omega} \leq c_2 h^{\frac{n}{2}} |||x|||_0 \tag{10.4}$$

with n the space dimension.

Proof:

a) Let $\hat{S}_n$ be the n-dimensional reference simplex. Set $\hat{v} = \sum_{i=0}^{n} x_i \hat{\varphi}_i$ with $\hat{\varphi}_i$ the basis function on $\hat{S}_n$. Then

$$\begin{aligned}\|\hat{u}\|^2_{0,\hat{S}_n} = (\hat{u},\hat{u})_{0,\hat{S}_n} &= \left(\sum_{i=0}^{n} x_i\hat{\varphi}_i, \sum_{j=0}^{n} x_j\hat{\varphi}_j\right)_{0,\hat{S}_n} = \sum_{i=0}^{n}\sum_{j=0}^{n} x_i x_j \underbrace{(\hat{\varphi}_i,\hat{\varphi}_j)_{0,\hat{S}_n}}_{=:\hat{M}_{ij}} \\ &= (x,\hat{M}x).\end{aligned}$$

$\hat{M}$ is the s.p.d. *mass matrix on the reference simplex.* Since $\hat{M}$ is s.p.d. we have

$$\lambda_{\min}(\hat{M})(x,x) \le (x,\hat{M}x) \le \lambda_{\max}(\hat{M})(x,x)$$

and therefore

$$\lambda_{\min}(\hat{M})\|x\|^2 \le \|\hat{u}\|^2_{0,\hat{S}_n} = (x,\hat{M}x) \le \lambda_{\max}(\hat{M})\|x\|^2 \ .$$

b) On the transformed element we get

$$\begin{aligned}\|u\|^2_{0,t} &= \int_t u^2(x)\,dx = \int_{\hat{S}_n} u^2(\mu_t(\xi))|\det B_t|\,d\xi \\ &= |\det B_t| \int_{\hat{S}_n} \hat{u}^2(\xi)\,d\xi = |\det B_t|\|\hat{u}\|^2_{0,\hat{S}_n} \ .\end{aligned}$$

c) Observing $c_1 h^n \le |\det B_t| \le c_2 h^n$ due to uniformity and shape regularity we obtain

$$\begin{aligned}\|u\|^2_{0,\Omega} &= \sum_{t\in\mathcal{T}_h} \|u\|^2_{0,t} = \sum_{t\in\mathcal{T}_h} |\det B_t|\|\hat{u}\|^2_{0,\hat{S}_n} \le c_2 h^n \lambda_{\max}(\hat{M}) \sum_{t\in\mathcal{T}_h} \|x_t\|^2 \\ &\le \bar{C} h^n |||x|||_0^2 \ .\end{aligned}$$

Here we used that x_t^2 contributes to finitely many $t \in \mathcal{T}_h$. Similarly we get for the lower bound

$$\|u\|^2_{0,\Omega} = \ldots \ge c_1 h^n \lambda_{\min}(\hat{M}) \sum_{t\in\mathcal{T}_h} \|x_t\|^2 \ge \underline{C} |||x|||_0^2 \ .$$

□

Approximation and Smoothing Property

Lemma 10.8 (Approximation property)**.** Let $A_h x_h = b_h$ be the discretized variational problem on the fine grid. $x_h^{k,1}$ denotes the iterate after smoothing and $x_h^{k,2}$ denotes the iterate after coarse grid correction (cf. Algorithm 10.5). Provided the mesh is uniform and shape-regular and the variational problem is H^2-regular there exists a constant c such that

$$|||x_h - x_h^{k,2}|||_0 \leq ch^{2-n}|||x_h - x_h^{k,1}|||_2 \ .$$

Proof:

a) From the proof of Theorem 8.18 (L^2 error estimate) we conclude (requires H^2-regularity)

$$\|u - u_h\|_{0,\Omega} \leq Ch\|u - u_h\|_{1,\Omega} \ .$$

(Just omit the last line in the last derivation).

b) We have the Galerkin orthogonality property for the error on the coarse grid:

$$\begin{aligned} a(u_h^{k,1} + w_{2h}, v) &= l(v) \qquad \forall v \in V_{2h} \qquad (u_h^{k,1} = \mathrm{FE}(x_h^{k,1})) \\ a(u_h, v) &= l(v) \qquad \forall v \in V_h \qquad (u_h = \mathrm{FE}(x_h)) \\ \Rightarrow a(u_h - u_h^{k,1} - w_{2h}, v) &= 0 \qquad \forall v \in V_{2h} \end{aligned}$$

With that we prove:

$$\begin{aligned} \alpha\|u_h - u_h^{k,2}\|_{1,\Omega}^2 &= \alpha\|u_h - (u_h^{k,1} + w_{2h})\|_{1,\Omega}^2 \\ &\leq a(u_h - u_h^{k,1} - w_{2h}, u_h - u_h^{k,1} - w_{2h}) && \text{(coercivity)} \\ &= a(u_h - u_h^{k,1} - w_{2h}, u_h - u_h^{k,1}) && \text{(orth.)} \\ &= (x_h - x_h^{k,1} - y_{2h}, A_h(x_h - x_h^{k,1})) && \text{(goto coeff.)} \\ &\leq |||x_h - x_h^{k,1} - y_{2h}|||_0 \ |||x_h - x_h^{k,1}|||_2 && \text{(C.S.)} \\ &\leq ch^{-\frac{n}{2}}\|u_h - u_h^{k,1} - w_{2h}\|_{0,\Omega}|||x_h - x_h^{k,1}|||_2 && \text{(go back)} \\ &\leq ch^{-\frac{n}{2}}h\|u_h - u_h^{k,1} - w_{2h}\|_{1,\Omega}|||x_h - x_h^{k,1}|||_2 && (L^2\text{-est.}) \\ &\leq ch^{-\frac{n}{2}}h\|u_h - u_h^{k,2}\|_{1,\Omega}|||x_h - x_h^{k,1}|||_2 \ . \end{aligned}$$

In the second to last step the L^2-estimate has been used for the problem $a(w_{2h}, v) = a(u_h - u_h^{k,1}, v)\ \forall v \in V_{2h}$ which has the exact solution $u_h - u_h^{k,1}$. Dividing by $\|u_h - u_h^{k,2}\|_{1,\Omega}$ and α results in

$$\|u_h - u_h^{k,2}\|_{1,\Omega} \leq Ch^{1-\frac{n}{2}}|||x_h - x_h^{k,1}|||_2 \ .$$

Note that this estimate is not robust with respect to the coercivity constant α which has been absorbed into C.

c) Finally, we obtain

$$\begin{aligned}
|||x_h - x_h^{k,2}|||_0 &\le ch^{-\frac{n}{2}}\|u_h - u_h^{k,2}\|_{0,\Omega} && \text{(use 10.7)}\\
&= ch^{-\frac{n}{2}}\|u_h - u_h^{k,1} - w_{2h}\|_{0,\Omega} && \text{(definition of } u_h^{k,2}\text{)}\\
&\le ch^{-\frac{n}{2}}h\|u_h - u_h^{k,2}\|_{1,\Omega} && \text{(use } L^2 \text{ estimate once)}\\
&\le ch^{2-n}|||x_h - x_h^{k,1}|||_2 \ . && \text{use b)}
\end{aligned}$$

□

Lemma 10.9 (Smoothing property)**.** Let A_h be symmetric positive definite. The Richardson iteration $x_h^{k+1} = x_h^k + \omega(b_h - A_h x_h^k)$ with $\omega = 1/\lambda_{\max}(A_h)$ satisfies:

$$|||x_h - x_h^\nu|||_2 \le \frac{\lambda_{\max}(A_h)}{\nu}|||x_h - x_h^0|||_0 \ .$$

Proof: With $e_h^\nu := x_h - x_h^\nu$

$$\begin{aligned}
|||e_h^\nu|||_2 &= \|A_h e_h^\nu\| = \|A_h(I - \omega A_h)^\nu e_h^0\| && \text{(Definitions)}\\
&= \|Q^T DQ(Q^T Q - \omega Q^T DQ)^\nu e_h^0\| && (A \text{ s.p.d.})\\
&= \|Q^T DQ(Q^T(I - \omega D)Q)^\nu e_h^0\|\\
&= \|Q^T DQQ^T(I - \omega D)Q \dots Q^T(I - \omega D)Qe_h^0\|\\
&= \|Q^T D(I - \omega D)^\nu Qe_h^0\|\\
&\le \|Q^T\|\|D(I - \omega D)^\nu\|\|Q\|\|e_h^0\| && (\|Q\| = 1\ )\\
&= \|D(I - \omega D)^\nu\|\|e_h^0\|
\end{aligned}$$

$\|Q\| = 1$ follows from $\|Qz\|^2 = (Qz, Qz) = (z, Q^T Qz) = (z, z) = \|z\|$ and $\sup_{x\neq 0}\frac{\|Qz\|}{\|z\|} = \sup_{x\neq 0}\frac{\|z\|}{\|z\|} = 1$.

Now since $D = \operatorname{diag}(\lambda_i)$ with $\lambda_i > 0$ we have

$$D(I - \omega D)^\nu = \operatorname{diag}(\lambda_i(1 - \omega\lambda_i)^\nu)$$

and

$$\begin{aligned}
\|D(I - \omega D)^\nu\| &= \max_{i=1,\dots,N}\lambda_i(1 - \omega\lambda_i)^\nu\\
&= \max_{i=1,\dots,N}\lambda_{\max}\frac{\lambda_i}{\lambda_{\max}}(1 - \frac{\lambda_i}{\lambda_{\max}})^\nu && (\omega = \frac{1}{\lambda_{\max}})\\
&\le \lambda_{\max}\max_{\xi\in[0,1]}\xi(1 - \xi)^\nu\\
&= \lambda_{\max}\frac{1}{1+\nu}(\frac{\nu}{\nu+1})^\nu\\
&\le \frac{\lambda_{\max}}{\nu}
\end{aligned}$$

□

Two- and Multigrid Convergence

Theorem 10.10. Assume that the variational problem has H^2-regularity. Then the two-grid method with ν steps of Richardson iteration as smoother satisfies the estimate

$$|||x_h - x_h^{k+1}|||_0 \leq \frac{c}{\nu}|||x_h - x_h^0|||_0 \ .$$

This says, that for ν large enough, the two-grid method converges independent of of the mesh size h.
Proof: Combine Lemma 10.8 and Lemma 10.9 to get

$$|||x_h - x_h^{k+1}|||_0 \leq ch^{2-n}\frac{\lambda_{\max}(A_h)}{\nu}|||x_h - x_h^k|||_0 \ .$$

It remains to prove an estimate for the maximum eigenvalue $\lambda_{\max}(A_h)$. The entries of the stiffness matrix can be estimated by

$$\begin{aligned}(A_h)_{ij} = a(\varphi_j, \varphi_i) &= \int_\Omega (K\nabla\varphi_j)\cdot\nabla\varphi_i \, dx \\ &\leq c \int_{\operatorname{supp}\varphi_j \cap \operatorname{supp}\varphi_i} h^{-1}h^{-1} \leq ch^{n-2} \qquad (\Omega \subset \mathbb{R}^n) \ .\end{aligned}$$

Here the essential property $|\operatorname{supp}\varphi_i| = O(h_t^n)$ of the Lagrange basis functions enters. Using the Gerschgorin circle theorem we obtain $\lambda_{\max}(A_h) \leq Ch^{n-2}$. Finally

$$ch^{2-n}\frac{\lambda_{\max}(A_h)}{\nu} \leq \frac{C}{\nu} \ .$$

□

Now W-cycle multigrid convergence follows from two-grid convergence by an induction argument.

Lemma 10.11. Let ρ_1 be the convergence rate of the two-grid method and ρ_l the convergence rate of a multigrid method with with cycle parameter γ and $l \geq 2$ levels. Then we have the recursive relation

$$\rho_l \leq \rho_1 + (1+\rho_1)\rho_{l-1}^\gamma \ .$$

Proof: Braess [2003]. Denote by $\hat{u}_l^{k,2} = u_l^{k,1} + \hat{w}_{l-1}$ the solution after *exact* coarse grid correction (i.e. the two-grid method on level l) and by $u_l^{k,2} = u_l^{k,1} + w_{l-1}^\gamma$ the solution after γ steps of multigrid on level $l-1$.

Then $\hat{w}_{l-1}$ is given by

$$a(\hat{w}_{l-1}, v) = l(v) - a(u_h^{k,1}, v) \qquad \forall v \in V_{l-1}$$

and w_{l-1}^γ is an *approximate* solution of *this* system using γ steps of the MG method on level $l-1$ with initial value $w_{l-1}^0 = 0$. So

$$\begin{aligned} \|\hat{w}_{l-1} - w_{l-1}^\gamma\|_{0,\Omega} &\le \rho_{l-1}^\gamma \|\hat{w}_{l-1} - \underbrace{w_{l-1}^0}_{=0}\|_{0,\Omega} = \rho_{l-1}^\gamma \|\hat{w}_{l-1}\|_{0,\Omega} \qquad (*) \\ &= \rho_{l-1}^\gamma \|\hat{u}_l^{k,2} - u_l^{k,1}\|_{0,\Omega} \ . \end{aligned}$$

Then we have:

$$\begin{aligned} \|u_l - u_l^{k+1}\|_{0,\Omega} &\le \|u_l - u_l^{k,2}\|_{0,\Omega} && \text{(conv. smoother)} \\ &= \|u_l - \hat{u}_l^{k,2} + \hat{u}_l^{k,2} - u_l^{k,2}\|_{0,\Omega} \\ &\le \|u_l - \hat{u}_l^{k,2}\|_{0,\Omega} + \|\hat{u}_l^{k,2} - u_l^{k,2}\|_{0,\Omega} && \text{(tri. ineq.)} \\ &\le \rho_1 \|u_l - u_l^k\|_{0,\Omega} + \|u_l^{k,1} + \hat{w}_{l-1} - u_l^{k,1} - w_{l-1}^\gamma\|_{0,\Omega} \\ &= \rho_1 \|u_l - u_l^k\|_{0,\Omega} + \|\hat{w}_{l-1} - w_{l-1}^\gamma\|_{0,\Omega} \\ &\le \rho_1 \|u_l - u_l^k\|_{0,\Omega} + \rho_{l-1}^\gamma \|\hat{u}_l^{k,2} - u_l^{k,1}\|_{0,\Omega} && \text{(use } (*)\text{)} \\ &= \rho_1 \|u_l - u_l^k\|_{0,\Omega} + \rho_{l-1}^\gamma \| \underbrace{\hat{u}_l^{k,2} - u_l}_{\text{two-grid}} + \underbrace{u_l - u_l^{k,1}}_{\text{smoothing}} \|_{0,\Omega} \\ &\le \rho_1 \|u_l - u_l^k\|_{0,\Omega} + \rho_{l-1}^\gamma \rho_1 \|u - u_l^k\|_{0,\Omega} + \rho_{l-1}^\gamma \|u_l - u_l^k\|_{0,\Omega} && \text{(conv. smoother)} \\ &= [\rho_1 + \rho_{l-1}^\gamma (1 + \rho_1)] \|u_l - u_l^k\|_{0,\Omega} \end{aligned}$$

□

This Lemma proves the following Theorem.

Theorem 10.12. If $\rho_1 \le \frac{1}{5}$ and $\gamma = 2$ then $\rho_l \le \frac{1}{3}$ for $l \ge 1$.
Proof: By induction.

$$\begin{aligned} l = 1: &\qquad \rho_l \le \frac{1}{5} \le \frac{1}{3} \qquad \checkmark \\ l - 1 \to l: &\qquad \rho_l \le \frac{1}{5} + (1 + \frac{1}{5})(\frac{1}{3})^2 = \frac{1}{5} + \frac{6}{5}\frac{1}{9} = \frac{9+6}{45} = \frac{1}{3} \ . \end{aligned}$$

□

Remark 10.13. Another interpretation of Lemma 10.11 is

$$\rho_l = f(\rho_{l-1}) = \rho_1 + (1 + \rho_1)\rho_{l-1}^\gamma \ .$$

This fixed-point iteration converges when f is a contraction. For $\gamma = 1$ we get

$$|f(x) - f(y)| \le |\rho_1 + (1 + \rho_1)x - \rho_1 - (1 + \rho_1)y| \le |1 + \rho_1||x - y| \ .$$

But since $1 + \rho_1 > 1$ f is no contraction for $\gamma = 1$. □

Finally, the optimal computational complexity requires that the work per iteration scales only linearly with the number of degrees of freedom.

Lemma 10.14. Let N_l be the number of unknowns on level l using P_1 finite elements. Then the amount of arithmetic operations A_l on level l is

$$A_l = \mathcal{O}(N_l) \ .$$

Proof: For uniform refinement we have $\frac{N_l}{N_{l-1}} = \omega = 2^n$ (for $l \to \infty$). Then

$$\begin{aligned} A_l &\leq CN_l + \gamma CN_{l-1} + \gamma^2 CN_{l-2} + \ldots + \gamma^l CN_0 \\ &= CN_l + \gamma C\frac{N_l}{\omega} + \gamma^2 C\frac{N_l}{\omega^2} + \ldots + \gamma^l C\frac{N_l}{\omega^l} \\ &\leq CN_l(1 + \frac{\gamma}{\omega} + (\frac{\gamma}{\omega})^2 + \ldots) \ . \end{aligned}$$

The geometric series converges for $\gamma < \omega \Leftrightarrow \frac{\gamma}{2^n} < 1 \Leftrightarrow \gamma < 2^n$. □

Chapter 11.

Finite Element Methods for Parabolic Problems

We consider the parabolic PDE

$$\partial_t u - \nabla \cdot (A \nabla u) = f \qquad \text{in } \Omega \times \Sigma \tag{11.1a}$$
$$u = 0 \qquad \text{on } \partial\Omega \times \Sigma \tag{11.1b}$$
$$u(x, t_0) = u_0 \qquad x \in \Omega \tag{11.1c}$$

Extension to non-homogeneous Dirichlet and flux boundary conditions follows as usual.

11.1. Method of Lines

We first derive a weak formulation of problem (11.1). For any $t \in \Sigma$ we set $u(t) = u(\cdot, t) \in V = H_0^1(\Omega)$.

Then require

$$\partial_t (u, v)_{0,\Omega} + (A \nabla u, \nabla v)_{0,\Omega} = (f, v)_{0,\Omega} \qquad \forall t \in \Sigma, \qquad \underbrace{\forall v \in V}_{\text{Note } v \text{ does not depend on } t} . \tag{11.2}$$

In the method of lines (MoL), discretization is done *in space first*, i.e. $V_h \subset V$ is a FE-space and $u_h : \Sigma \to V_h$ with

$$\frac{d}{dt}(u_h, v)_{0,\Omega} + (A \nabla u_h, \nabla v)_{0,\Omega} = (f, v)_{0,\Omega} \qquad \underbrace{\forall t \in \Sigma}_{\text{continuous}}, \ \underbrace{\forall v \in V_h}_{\text{discrete}} .$$

Inserting the basis representations yields:

$$\frac{d}{dt}\left(\sum_{j=1}^{N_h} x_j(t)\psi_j(x), \psi_i(x)\right)_{0,\Omega} + \left(A\sum_{j=1}^{N_h} x_j(t)\nabla\psi_j, \nabla\psi_i\right)_{0,\Omega} = (f,\psi_i)_{0,\Omega}\ \forall t\in\Sigma,\ i=1,..$$

$$\Leftrightarrow \sum_{j=1}^{N_j}\frac{dx_j}{dt}(t)\underbrace{(\psi_j(x),\psi_i(x))}_{(M_h)_{ij}}{}_{0,\Omega} + \sum_{j=1}^{N_j} x_j(t)\underbrace{(A\nabla\psi_j,\nabla\psi_i)}_{(A_h)_{ij}}{}_{0,\Omega} = (f(t,\cdot),\psi_i)_{0,\Omega}$$

$$\Leftrightarrow M_h\frac{dx}{dt}(t) + A_h x(t) = b(t)$$

$$\Leftrightarrow \frac{dx}{dt}(t) = -M_h^{-1}A_h x(t) + b(t)\ . \tag{11.3}$$

This is a linear system of ordinary differential equations.

What are the properties of $-M_h^{-1}A_h$?

- A_h, M_h are symmetric positive definite.
- $\sigma(-M_h^{-1}A_h) = \sigma(M_h^{\frac{1}{2}}M_h^{-1}A_hM_h^{-\frac{1}{2}}) = \sigma(\underbrace{M_h^{-\frac{1}{2}}A_hM_h^{-\frac{1}{2}}}_{\text{s.p.d.}})$

Therefore: All eigenvalues of $-M_h^{-1}A_h$ are real and negative.

A linear system of ODEs $\frac{dx}{dt}(t) = Bx(t) + b(t)$ with negative definite B is called stiff if $\lambda_{\min}(B) \ll \lambda_{\max}(B)$.

Lemma 11.1. Let M_h and A_h be the mass matrix and diffusion matrix obtained with the Lagrange basis functions on a uniform and shape regular mesh $\mathcal{T}_h$ of size h and $\Omega\subset\mathbb{R}^d$.

Then

$$\lambda_{\min}(A_h) \geq c_1 h^d\ , \qquad \lambda_{\max}(A_h) \leq c_2 h^{d-2}\ ,$$
$$\lambda_{\min}(M_h) \geq c_3 h^d\ , \qquad \lambda_{\max}(M_h) \leq c_4 h^d\ .$$

Proof:

a) Rayleigh quotient: For any s.p.d. matrix B the extreme eigenvalues are characterized by

$$\lambda_{\min}(B) = \inf_{x\neq 0}\frac{(Bx,x)}{(x,x)} \qquad \lambda_{\max}(B) = \sup_{x\neq 0}\frac{(Bx,x)}{(x,x)}$$

b) Mass matrix.

$$x\neq 0:\ (M_hx,x) \overset{u_h=\mathrm{FE}(x)}{=} (u_h,u_h)_{0,\Omega} \leq ch^d(x,x)\ \ \text{(Lemma 10.7, } (x,x) = |||x|||_0)$$

and

$$(M_h x, x) = (u_h, u_h)_{0,\Omega} \geq c h^d (x, x) \ .$$

Therefore

$$\lambda_{\min}(M_h) = \inf_{x \neq 0} \frac{(M_h x, x)}{(x, x)} \geq c h^d$$
$$\lambda_{\max}(M_h) = \sup_{x \neq 0} \frac{(M_h x, x)}{(x, x)} \leq c h^d$$

c) Diffusion matrix.

$$\begin{aligned}
(A_h x, x) &= \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} x_i x_j a(\psi_j, \psi_i) \\
&= \sum_i \sum_j \int\limits_{\operatorname{supp}\psi_i \cap \operatorname{supp}\psi_j} (A \nabla \psi_i) \cdot \nabla \psi_j \, dx \\
&= \sum_i \sum_j (x_i x_j \sum_{t \in \operatorname{supp}\psi_i \cap \operatorname{supp}\psi_j} \int\limits_{\hat{t}} (A B_t^{-T} \nabla \hat{\psi}_i) \cdot \underbrace{B^{-T}}_{=\frac{1}{h}} \nabla \hat{\psi}_j \underbrace{|\det B_t|}_{=h^d} \, d\hat{x}) \\
&\leq \sum_i x_i \sum_j x_j \sum_{t \in \operatorname{supp}\psi_i \cap \operatorname{supp}\psi_j} C h^{d-2} \qquad (\text{ since } (Az, z) \leq \bar{C}(z, z)) \\
&= C h^{d-2} (Ex, x) \qquad \text{where} (E)_{ij} = \begin{cases} 1 & \operatorname{supp}\psi_i \cap \operatorname{supp}\psi_j \neq 0 \\ 0 & \text{else} \end{cases} \\
&\leq C h^{d-2} \|Ex\| \|x\| \quad \text{(Cauchy-Schwarz)} \\
&\leq C h^{d-2} \|E\| (x, x)
\end{aligned}$$

Spectral norm $\|E\| = \lambda_{\max}(E)$ since E is s.p.d.

$$\lambda_{\max} \leq 1 + \max_i \sum_{j \neq i} (E)_{ij} \leq K \ . \qquad \text{(Gerschgorin)}$$

Therefore $\sup_{x \neq 0} \frac{(A_h x, x)}{(x, x)} \leq C h^{d-2}$.

$$\begin{aligned}
(A_h x, x) &= a(u_h, u_h) \geq \alpha \|u_h\|_{1,\Omega}^2 = \alpha (\|u\|_{0,\Omega}^2 + |u_h|_{1,\Omega}^2) \\
&\geq \alpha (\|u_h\|_{0,\Omega}^2 + \frac{1}{s^2} \|u_h\|_{0,\Omega}^2) \qquad \text{(Friedrich inequality)} \\
&= \alpha \frac{1 + s^2}{s^2} \|u_h\|_{0,\Omega}^2 \\
&\geq \alpha \frac{1 + s^2}{s^2} h^d (x, x) \ . \qquad \text{(Lemma 10.7)}
\end{aligned}$$

Therefore

$$\inf_{x\neq 0}\frac{(A_h x,x)}{(x,x)} \geq \alpha\frac{1+s^2}{s^2}h^d \ .$$

□

Corollary:

$$\kappa_2(A_h)=\frac{\lambda_{\max}(A_h)}{\lambda_{\min}(A_h)} \leq Ch^{-2} \ .$$

Lemma 11.2. The problem of equation (11.2) is stiff. There exist constants c_1, c_2:

$$\lambda_{\min}(M_h^{-1}A_h) \geq c_1 \ , \qquad \lambda_{\max}(M_h^{-1}A_h) \leq c_2 h^{-2} \ .$$

Proof:

$$\begin{aligned}
\sup_{x\neq 0}\frac{(M_h^{-1}A_h x,x)}{(x,x)} &\overset{\text{similarity transformation}}{=} \sup_{x\neq 0}\frac{(M_h^{\frac{1}{2}}M_h^{-1}A_h M_h^{-\frac{1}{2}}x,x)}{(x,x)} \\
&= \sup_{y=M_h^{\frac{1}{2}}x\neq 0}\frac{(M_h^{-\frac{1}{2}}A_h M_h^{-\frac{1}{2}}M_h^{\frac{1}{2}}x, M_h^{\frac{1}{2}}x)}{(M_h^{\frac{1}{2}}x, M_h^{\frac{1}{2}}x)} \\
&= \sup_{x\neq 0}\frac{(A_h x,x)}{(M_h x,x)} \\
&\leq C\frac{h^{d-2}(x,x)}{h^d(x,x)} = Ch^{-2} \ . \qquad \text{(Lemma 11.1)}
\end{aligned}$$

$$\inf_{x\neq 0}\frac{(M_h^{-1}A_h x,x)}{(x,x)} = \inf_{x\neq 0}\frac{(A_h x,x)}{(M_h x,x)} \geq C\frac{h^d(x,x)}{h^d(x,x)} \geq C \ .$$

□

Conclusion: The ODE system (11.3) needs to be solved with A-stable methods. E.g. implicit Euler:

$$t_0 = t^0 \leq t^1 \leq \ldots \leq t^M = t_0 + T \ , \qquad \Delta t^i = t^i - t^{i-1} \ .$$

$$x^0 = x(t^0)$$

$$M_h\frac{1}{\Delta t}(x^m - x^{m-1}) + A_h x^m = b(t^m)$$

$$\Leftrightarrow (M_h + \Delta t A_h)x^m = M_h x^{m-1} + \Delta t b(t^m)$$

Solve one linear system per time step.
$\frac{\Delta t}{h^2}$ small $\Rightarrow$ easy to solve
$\frac{\Delta t}{h^2}$ large $\Rightarrow$ same as elliptic equation

11.2. Rothe Method

Starting from (11.1) we first *discretize in time*, e.g. with implicit Euler:

$$\partial_t u(x,t^m) \approx \frac{1}{\Delta t^m}(u(x,t^m)) - u(x,t^{m-1}) \ .$$

Introduce U^m , $m = 0,\ldots,M$ and set

$$U^m - U^{m-1} - \Delta t^m \nabla \cdot (A\nabla U^m) = \Delta t^m f^m \ .$$

Discretization in space with the FEM yields the fully discrete version:

$$U_h^m \in V_h^m : \qquad (U_h^m - U_h^{m-1}, v)_{0,\Omega} + \Delta t^m (A\nabla U_h^m, v) = \Delta t^m (f^m, v) \ \ \forall v \in V_h^m \ .$$

Note: Here it is conceptually easy to have a *different* finite element space V_h^m, i.e. a different mesh/polynomial degree, in *each time step*. This has advantages for the theoretical treatment in a priori and a posteriori error analysis.

Theorem 11.3 (Rannacher Satz 5.5, p.200)**.** Rothe method with implicit Euler/piecewise linear FEM. Special case $A = I$, spatial problem $a(u,v) = (\nabla u, \nabla v)_{0,\Omega}$ H^2-regular, and $V_h^m = V_h$ fixed for all time steps. Then the following error estimate holds:

$$\max_{i\le m\le M} \|u(\cdot,t^m) - U_h^m\|_{0,\Omega} \le c_1 T^{\frac{1}{2}} \max_{0\le m\le M}\{h^2\|u\|_{2,\Omega}\} + c_2 \left(\sum_{m=1}^{M} (\Delta t^m)^2 \int_{t^{m-1}}^{t^m} \|\nabla \partial_t u\|^2 \, dt\right)^{\frac{1}{2}}$$

Proof: See Rannacher.

□

Note: The convergence rate is $\mathcal{O}(\Delta t + h^2)$.
Smoothing effect: Consider

$$\begin{aligned} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} &= 0 && \text{in } (0,1)\times(0,\infty) \\ u(x,0) &= f(x) && \text{at } t_0 = 0 \\ u(0,t) &= u(1,t) = 0 \ . \end{aligned}$$

For $f(x) = \sum_{n=1}^{\infty} A_n \sin n\pi x$ the solution is

$$u(x,t) = \sum_{n=1}^{\infty} A_n e^{-n^2\pi^2 t} \sin n\pi x \ .$$

(Proof: Separation of variables.)

This shows that $\partial_t u$ decreases exponentially, a fact called *smoothing property* of parabolic equations. A consequence is that first-order convergence in time is often acceptable.

11.3. Space-time Method

Simultaneous discretization in space and time is possible as well.

Set

$$u_h(x,t) = \sum_{i=1}^{N} \sum_{j=1}^{m} x_{ij} \psi_i(x) \varphi_j(t) \ ,$$

integrate in space and time...

Problem: Full coupling of degrees of freedom in space and time.

Solution: Discontinuous polynomials in time + upwind.

Chapter 12.

Numerical Methods for First-order Hyperbolic Equations

In this section we consider the linear model problem in one space dimension

$$\partial_t u + \partial_x(au) = 0 \tag{12.1a}$$
$$u(0,t) = g(t) \tag{12.1b}$$
$$u(x,0) = u_0(x) \tag{12.1c}$$

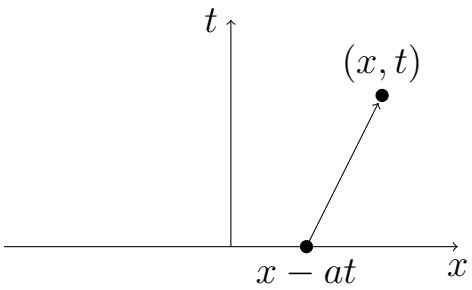
with $a > 0$.

This is a special case of $\partial_t u + \partial_x f(u) = 0$ with the "flux function" $f : \mathbb{R} \to \mathbb{R}$.

The method of characteristics provides the exact solution of (12.1):

$$u(x,t) = u_0(x - at) \ .$$

Graphically



12.1. Finite Difference Methods

Approach:

- Method of Lines
- Use Finite Difference Method instead of Finite Elements

Space Discretization

Recall the *centered difference*

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} + \mathcal{O}(h^2)$$

Space grid $x_i = ih$, $0 \le i \le N$, $h = \frac{1}{N}$.
Set $U_i(t) \approx u(x_i, t)$ given by

$$\partial_t U_i(t) + a\frac{U_{i+1}(t) - U_{i-1}(t)}{2h} = 0 \ , \qquad 0 < i < N \ .$$

ODE-system

$$\begin{aligned} U_0(t) &= g(t) \\ U_N(t) &= ?? \\ U_i(0) &= u_0(x_i, 0) \end{aligned}$$

Time Discretization

θ-method:

$$\frac{du}{dt} = f(t, u(t)) : \qquad \frac{u(t+\Delta t) - u(t)}{\Delta t} = \theta f(t, u(t+\Delta t)) + (1-\theta) f(t, u(t))$$

$\theta \in [0,1]$: $\theta = 1$ implicit Euler, $\theta = \frac{1}{2}$ trapezoidal rule, $\theta = 0$ explicit Euler.
Time grid: $t^k := \Delta t k$; $0 \le k$
Set $U_i^k \approx U_i(t^k)$ which is then given by

$$\frac{1}{\Delta t}(U_i^{k+1} - U_i^k) + \frac{a(1-\theta)}{2h}(U_{i+1}^k - U_{i-1}^k) + \frac{a\theta}{2h}(U_{i+1}^{k+1} - U_{i-1}^{k+1}) = 0$$

$$\overset{\cdot \Delta t}{\Leftrightarrow} -\frac{a\theta\Delta t}{2h}U_{i-1}^{k+1} + U_i^{k+1} + \frac{a\theta\Delta t}{2h}U_{i+1}^{k+1} = \frac{a(1-\theta)\Delta t}{2h}U_{i-1}^k + U_i^k - \frac{a(1-\theta)\Delta t}{2h}U_{i+1}^k \ ,$$

$k > 0, 0 < i < N$.

$$\begin{aligned} U_0^{k+1} &= g(t^k) \\ U_N^{k+1} &= ?? \\ U_i^0 &= u_0(x_i, t^0) \end{aligned}$$

$\Leftrightarrow L_h U^{k+1} = M_h U^k$. (Note: M_h is *not* a mass matrix.)

Analysis of this system:

- L_h is not symmetric for $\theta \neq 0$
- L_h is not an M-Matrix (positive off-diagonal entry) for $\theta \neq 0$
- L_h is diagonally dominant provided
  - $\theta = 0$ (explicit Euler): $L_h = I$
  - $\theta \neq 0$: $2\frac{a\theta\Delta t}{2h} < 1 \Leftrightarrow \Delta t < \frac{h}{a\theta}$, i.e. $\Delta t = \mathcal{O}(h)$.

- Consider $\theta = 0$: $U^{k+1} = M_h U^k$, $M_h = \text{tridiag}(\frac{a\Delta t}{2h}, 1, -\frac{a\Delta t}{2h})$

$$\|U^{k+1}\|_\infty \leq \|M_h\|_\infty \|U^k\|_\infty$$
$$\|M_h\|_\infty = 1 + \frac{|a|\Delta t}{h} > 1$$

  $\Rightarrow$ method is uncoditionally unstable!

- Consider $\theta = 1$: $L_h U^{k+1} U_k$, $L_h = \text{tridiag}(-\frac{a\Delta t}{2h}, 1, \frac{a\Delta t}{2h})$ Practical experience: stable for Δt *large enough* (see below), unphysical oscillations for Δt small (for fixed h).

Upwind

Idea: Use first order finite difference in space. Two possibilities:

(i) $f'(x) = \frac{f(x+h)-f(x)}{h} + \mathcal{O}(h)$

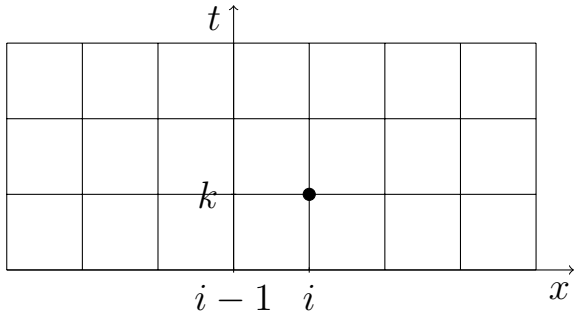
(ii) $f'(x) = \frac{f(x)-f(x-h)}{h} + \mathcal{O}(h)$

Which one to use?

(i) if $a > 0$

(ii) if $a < 0$

in order to reflect dependence in method of characteristic.



For $a > 0$ (left to right) this leads to

$$\frac{1}{\Delta t}(U_i^{k+1} - U_i^k) + \frac{a(1-\theta}{h}(U_i^k - U_{i-1}^k) + \frac{a\theta}{h}(U_i^{k+1} - U_{i-1}^{k+1}) = 0$$

$$\Leftrightarrow -\frac{a\theta\Delta t}{h}U_{i-1}^{k+1} + (1 + \frac{a\theta\Delta t}{h})U_i^{k+1} = \frac{a(1-\theta)\Delta t}{h}U_{i-1}^k + (1 - \frac{a(1-\theta)\Delta t}{h}U_i^k \ , \qquad 0 < i \leq$$

$$U_0^{k+1} = g(t^{k+1}) \qquad k \geq 0$$
$$U_i^0 = u_0(x_i, t^0)$$

Note: No boundary condition at $x = 1$ is necessary!

- $\theta = 0$, explicit case $U^{k+1} = M_h U^k$, $M_h = \text{tridiag}(\frac{a\Delta t}{h}, 1 - \frac{a\Delta t}{h}, 0)$

$$\|U^{k+1}\|_\infty \leq \|M_h\|_\infty \|U^k\|_\infty$$

with

$$\|M_h\|_\infty = \frac{a\Delta t}{h} + 1 - \frac{a\Delta t}{h} = 1$$

if

$$1 - \frac{a\Delta t}{h} \geq 0 \Leftrightarrow \boxed{\frac{a\Delta t}{h} \leq 1}$$

Famous Courant-Friedrich-Levy (CFL) condition.

Physical interpretation:

a	$\cdot$	Δt	$\leq$	h	
velocity $\left[\frac{m}{s}\right]$		timestep$[s]$		meshsize	Particle does not move more than one cell.

- $\theta = 1$, implicit case $L_h U^{k+1} = U^k$, $L_h = \text{tridiag}(-\frac{a\Delta t}{h}, 1 + \frac{a\Delta t}{h}, 0)$ One can show: L_h is M-Matrix and $\|L_h^{-1}\|_\infty \leq 1$ for all $\Delta t, h > 0$! Method is unconditionally stable! $\Rightarrow$ Numerical results. $a = 1, h = \frac{1}{200}$.

Numerical Diffusion

Question: Why does the Upwind difference work? Taylor expansion in more detail. For the *exact, smooth* solution $U(x,t)$:

$$\frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} = \frac{\partial u}{\partial t}\underbrace{(x,t+\Delta t)}_{\text{expansion point}} - \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}(x,t+\Delta t) + \mathcal{O}(\Delta t^2)$$

$$\frac{u(x,t+\Delta t) - u(x-h,t+\Delta t)}{h} = \frac{\partial u}{\partial x}(x,t+\Delta t) - \frac{h}{2}\frac{\partial^2 u}{\partial x^2}(x,t+\Delta t) + \mathcal{O}(h^2)$$

For u smooth enough we have:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial t} = 0 \left\{\begin{array}{ll} \Rightarrow \frac{\partial^2 u}{\partial t^2} + a\frac{\partial^2 u}{\partial x \partial t} & = 0 \\ \Rightarrow \frac{\partial^2 u}{\partial t \partial x} + a\frac{\partial^2 u}{\partial x^2} & = 0 \end{array}\right\} \Rightarrow \frac{\partial^2 u}{\partial t^2} - a^2\frac{\partial^2 u}{\partial x^2} = 0$$

$$\Leftrightarrow \frac{\partial^2 u}{\partial t^2} = a^2\frac{\partial^2 u}{\partial x^2}\ .$$

Combining this gives for the exact (smooth) solution u:

$$\underbrace{\frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} + a\frac{u(x,t+\Delta t) - u(x-h),t+\Delta t)}{h}}_{\text{implicit upwind difference scheme at } (x,t+\Delta t)}$$

$$= \left(\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x}\right)\Big|_{(x,t+\Delta t)} - \left(\frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2} + \frac{ah}{2}\frac{\partial^2 u}{\partial x^2}\right)\Big|_{(x,t+\Delta t)} + \mathcal{O}(\Delta t^2 + h^2)$$

$$= \left(\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x}\right)\Big|_{(x,t+\Delta t)} - \left(\frac{a^2\Delta t + ah}{2}\frac{\partial^2 u}{\partial x^2}\right)\Big|_{(x,t+\Delta t)} + \mathcal{O}(\Delta t^2 + h^2)$$

- Leading order term of consistency error is a diffusion term (note minus sign!)
- Upwind difference and *implicit* Euler *in time* add diffusion in space!
- For *large* Δt this stabilizes the implicit Euler / central scheme.
- We effectively solve an advection-diffusion equation.

Appendix A.

Nabla and Friends

A.1. Notation for Derivatives

The partial derivative

$$\frac{\partial u}{\partial x_i}(x) = \lim_{h\to\infty} \frac{u(x+he_i)-u(x)}{h}$$

of a scalar function $u : \mathbb{R}^n \to \mathbb{R}$ is written in short notation as

$$\partial_{x_i} u(x) = \frac{\partial u}{\partial x_i}(x).$$

Similarly we have for the higher derivatives

$$\partial^2_{x_i} u(x) = \frac{\partial^2 u}{\partial x_i^2}(x), \qquad \partial_{x_i}\partial_{x_j} u(x) = \frac{\partial^2 u}{\partial x_i \partial x_j}(x), \qquad \dots$$

A vector $\alpha = (\alpha_1, \dots, \alpha_n)^T$ of nonnegative integers α_i is called a *multiindex* of order

$$|\alpha| = \sum_{i=1}^{n} \alpha_i.$$

Sometimes

$$|\alpha|_\infty = \max_{i=1,\dots,n} \alpha_i.$$

is referred to as the maximum order.

For a given multiindex α we set

$$\partial^\alpha u(x) = \partial^{\alpha_1}_{x_1} \dots \partial^{\alpha_n}_{x_n} u(x) = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}}(x)$$

For a given nonnegative integer k

$$D^k u(x) = \{\partial^\alpha u(x) \,:\, |\alpha| = k\}$$

denotes the ordered set of all partial derivatives of order k at the point x. Note that $D^k u(x)$ has n^k elements, i.e. $\partial_{x_i}\partial_{x_j} u(x)$ and $\partial_{x_j}\partial_{x_i} u(x)$ are different elements although they have the same value.

For the special cases $k = 1$ and $k = 2$ we identify $D^1u(x)$ with the gradient $\nabla u(x)$ and $D^2u(x)$ with the Hessian matrix $\nabla^2 u(x)$ (see below for the definition of gradient and Hessian).

In the case of a function $u(x, y)$, $u : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, we write $D^1_x u$ or $D^2_y u$ to indicate the variable with respect to which differentiation is to be applied.

A.2. Vector Differential Calculus

The whole presentation treats the differential operators only in cartesian coordinates.

A.2.1. Nabla Operator

The nabla operator formally is a row or column vector of partial derivatives with respect to all variables of its argument:

$$\nabla = (\partial_1, \ldots, \partial_n)^T \tag{A.1}$$

(when we assume that the argument has n variables).

A.2.2. Gradient

Gradient of a Scalar Nabla applied to a scalar function $u(x_1, \ldots, x_n)$ in n variables gives a vector called "gradient" of the function:

$$\nabla u = (\partial_1 u, \ldots, \partial_n u)^T . \tag{A.2}$$

We can imagine ∇ to be a column vector in this case applied to a scalar which gives a vector.

The gradient of a scalar function in point x is a vector which is perpendicular to the level set $l(c) = \{y : u(y) = c\}$ for $c = u(x)$ pointing in the direction of the steepest increase of the function u.

Gradient of a Vector-valued Function Nabla applied to a vector-valued function

$$u(x) = (u_1(x_1, \ldots, x_n), \ldots, u_m(x_1, \ldots, x_n))^T$$

with m components in n variables gives a matrix called the "Jacobian" of the function:

$$\nabla u = \begin{pmatrix} (\nabla u_1)^T \\ \vdots \\ (\nabla u_m)^T \end{pmatrix} = \begin{pmatrix} \partial_1 u_1 & \ldots & \partial_n u_1 \\ \vdots & & \vdots \\ \partial_1 u_m & \ldots & \partial_n u_m \end{pmatrix} \qquad \text{or} \qquad (\nabla u)_{i,j} = \partial_j u_i. \tag{A.3}$$

If we wish to view the gradient as a column vector and the function u also as a column vector (of possibly different size) then we formally have:

$$\text{“}\nabla u\text{”} := \left(\nabla u^T\right)^T. \tag{A.4}$$

Here ∇u^T acts as an outer product producing a matrix.

In the case of a scalar function u the matrix $\nabla\nabla u = \nabla^2 u$ is called the Hessian matrix.

A.2.3. Divergence

Divergence of a Vector Field The scalar product of nabla with a vector-valued function gives a scalar called the “divergence” of the function:

$$\nabla \cdot u = \sum_{i=1}^{n} \partial_i u_i.$$

Divergence of a Matrix-valued Function The divergence operator applied to a matrix-valued function

$$\sigma(x_1, \ldots, x_n) = \begin{pmatrix} \sigma_1 \\ \vdots \\ \sigma_m \end{pmatrix} = \begin{pmatrix} \sigma_{1,1}(x) & \ldots & \sigma_{1,n}(x) \\ \vdots & & \vdots \\ \sigma_{m,1}(x) & \ldots & \sigma_{m,n}(x) \end{pmatrix}$$

in n variables is defined to yield the divergence for each row of the matrix. Note that σ needs to have as many columns as there are variables. It produces a vector-valued function:

$$\nabla \cdot \sigma = \begin{pmatrix} \nabla \cdot \sigma_1 \\ \vdots \\ \nabla \cdot \sigma_m \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} \partial_j \sigma_{1,j} \\ \vdots \\ \sum_{j=1}^{n} \partial_j \sigma_{m,j} \end{pmatrix} \qquad \text{or} \qquad (\nabla \cdot \sigma)_i = \sum_{j=1}^{n} \partial_j \sigma_{i,j}. \tag{A.5}$$

If we regard the divergence as a row vector and σ an $m \times n$ matrix with n also the number of variables, then we can formally write

$$\text{“}\nabla \cdot \sigma\text{”} := \left(\nabla \cdot (\sigma^T)\right)^T. \tag{A.6}$$

Here the inner product $\nabla \cdot (\sigma^T)$ produces a row vector. Note the similarity to the formula (A.4).

A.2.4. Curl

The "curl" (also called "rot", which is exactly the same thing) of a vector field is defined as

$$\nabla \times u = \begin{pmatrix} \partial_2 u_3 - \partial_3 u_2 \\ \partial_3 u_1 - \partial_1 u_3 \\ \partial_1 u_2 - \partial_2 u_1 \end{pmatrix} \tag{A.7}$$

which corresponds to the vector (cross) product $a \times b = (a_2 b_3 - a_3 b_2, a_3 b_1 - a_1 b_3, a_1 b_2 - a_2 b_1)^T$. As stated, it makes only sense for $u : \mathbb{R} \to \mathbb{R}^3$ and there is no obvious extension of the curl operator to n dimensions. However, the related Stokes theorem (see below) can be extended to arbitrary dimensions.

A.2.5. Convection Term in Navier-Stokes Equations

For a vector-valued function u, the convection term in the Navier-Stokes equations is written as $u \cdot \nabla u$ which is formally defined as

$$u \cdot \nabla u = (\nabla u)u = \begin{pmatrix} \nabla u_1 \cdot u \\ \vdots \\ \nabla u_n \cdot u \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n u_i \partial_i u_1 \\ \vdots \\ \sum_{i=1}^n u_i \partial_i u_n \end{pmatrix}. \tag{A.8}$$

Note that the scalar product of a vector with a matrix (∇u is a matrix!) is defined as a vector where each component is the scalar multiplication of the vector with a row of the matrix.

A.2.6. Laplacian

Laplacian of a scalar function The Laplacian takes second order derivatives of a scalar function and is defined as

$$\Delta u = \nabla \cdot \nabla u = \sum_{i=1}^n \partial_i^2 u. \tag{A.9}$$

Laplace of Vector-valued function The definition of the Laplacian is extended to vector-valued functions by applying it to each component, i.e. the Laplacian of a vector-valued function is again a vector-valued function. In agreement with the conventions above we have:

$$\Delta u = \nabla \cdot \nabla u = \begin{pmatrix} \nabla \cdot \nabla u_1 \\ \vdots \\ \nabla \cdot \nabla u_n \end{pmatrix} = \begin{pmatrix} \Delta u_1 \\ \vdots \\ \Delta u_n \end{pmatrix}. \tag{A.10}$$

A.3. Vector Integral Calculus

A.3.1. Matrix Product

Let T, S be two $m \times n$ matrices, then we define

$$T : S = \sum_{i=1}^{m} \sum_{j=1}^{n} T_{i,j} S_{i,j}. \tag{A.11}$$

Applied to two vector-valued functions u, v with m components in n variables we have with the definitions from above:

$$\nabla u : \nabla v = \sum_{i=1}^{m} \nabla u_i \cdot \nabla v_i. \tag{A.12}$$

Now let T, S, Q be $n \times n$ matrices. Then the following holds:

$$T : (QSQ^T) = (Q^T T Q) : S. \tag{A.13}$$

This can be shown as follows:

$$\begin{aligned}
T : (QSQ^T) &= \sum_{i=1}^{n} \sum_{j=1}^{n} T_{i,j} (e_i^T Q S Q^T e_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} T_{i,j} \left(\sum_{k=1}^{n} Q_{i,k} \left(\sum_{l=1}^{n} S_{kl} Q_{l,j}^T \right) \right) \\
&= \sum_{k=1}^{n} \sum_{l=1}^{n} S_{kl} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} T_{i,j} Q_{i,k} Q_{l,j}^T \right) \\
&= \sum_{k=1}^{n} \sum_{l=1}^{n} S_{kl} \sum_{i=1}^{n} Q_{k,i}^T \left(\sum_{j=1}^{n} T_{i,j} Q_{j,l} \right) \\
&= (Q^T T Q).
\end{aligned}$$

A.3.2. Integration by Parts

Green's formula for sufficiently smooth scalar functions u, v and a suitable bounded domain Ω is

$$\int_{\Omega} (\partial_i u) v = -\int_{\Omega} u \partial_i v + \int_{\partial\Omega} u v n_i \tag{A.14}$$

where n_i is the i-th component of the outer unit normal vector n.

For a vector-valued function u and a scalar function v we then have

$$\int_\Omega (\nabla \cdot u)v = -\int_\Omega u \cdot \nabla v + \int_{\partial\Omega} u \cdot n\, v \quad . \tag{A.15}$$

For a matrix-valued function T and a vector valued function v one shows the corresponding formula

$$\int_\Omega (\nabla \cdot T) \cdot v = -\int_\Omega T : \nabla v + \int_{\partial\Omega} (T \cdot n) \cdot v \tag{A.16}$$

which is needed in the variational formulation of the Navier-Stokes equations. Indeed using the definitions above one obtains:

$$\begin{aligned}
\int_\Omega (\nabla \cdot T) \cdot v &= \int_\Omega \sum_{i=1}^n \left(\sum_{j=1}^n \partial_j T_{i,j} \right) v_i \\
&= \sum_{i=1}^n \sum_{j=1}^n \left\{ -\int_\Omega T_{i,j} \partial_j v_i + \int_{\partial\Omega} T_{i,j} v_i n_j \right\} \\
&= -\int_\Omega \sum_{i=1}^n \sum_{j=1}^n T_{i,j} (\nabla v)_{i,j} + \int_{\partial\Omega} \sum_{i=1}^n \left(\sum_{j=1}^n T_{i,j} n_j \right) v_i \quad .
\end{aligned}$$

Bibliography

R. A. Adams. *Sobolev Spaces*. Academic Press, 1978.

M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis*. Wiley, 2000.

I. Babuska and A.K. Aziz. On the angle condition in the finite element method. *SIAM J. Numer. Anal.*, 13(2):214–226, 1976.

W. Bangerth and R. Rannacher. *Adaptive Finite Element methods for differential equations*. Birkhäuser, 2003.

E. Bänsch. Local mesh refinement in 2 and 3 dimensions. *IMPACT of Computing in Science and Engineering*, 3(3):181–191, 1991. ISSN 0899-8248.

C. Bernardi and V. Girault. A local regularization operator for triangular and quadrilateral finite elements. *SIAM J. Numer. Anal.*, 35(5):1893?–1916, 1998.

J. Bey. Simplicial grid refinement: on Freudenthal's algorithm and the optimal number of congruence classes. *Numerische Mathematik*, 85(1):1–29, 2000. ISSN 0029-599X.

D. Braess. *Finite Elemente*. Springer, 3rd edition, 2003.

S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*. Springer, 1994.

T. J. Chung. *Applied Continuum Mechanics*. Cambridge University Press, 1996.

P. G. Ciarlet. *The finite element method for elliptic problems*. Classics in Applied Mathematics. SIAM, 2002.

P. Clément. Approximation by finite element functions using local regularization. *RAIRO Modél. Math. Anal. Numér*, 9:77?–84, 1975.

H. Elman, D. Silvester, and A. Wathen. *Finite Elements and Fast Iterative Solvers*. Oxford University Press, 2005.

K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational Differential Equations*. Cambridge University Press, 1996.

A. Ern and J.-L. Guermond. *Theory and practice of finite element methods*. Springer, 2004.

L. C. Evans. *Partial Differential Equations*. American Mathematical Society, 2nd edition, 2010.

C. Geuzaine and J.-F. Remacle. Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79(11):1309–1331, 2009.

C. Großmann and H.-G. Roos. *Numerische Behandlung partieller Differentialgleichungen*. Teubner, 2006.

W. Hackbusch. *Multi–Grid Methods and Applications*. Springer–Verlag, 1985.

W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner, 1986. `http://www.mis.mpg.de/preprints/ln/lecturenote-2805.pdf`.

R. Hiptmair. Numerical methods for partial differential equations. Lecture slides, ETH Zürich, `http://www.sam.math.ethz.ch/~hiptmair/tmp/NPDE10.pdf`, 2010.

J. B. Keller. A theorem on the conductivity of a composite medium. *Journal of Mathematical Physics*, 5(4):548–549, 1964.

S. M. Kozlov, O. A. Oleinik, and V. V. Zhikhov. *Homogenization of differential operators and integral functionals*. Springer-Verlag, 1994.

R. J. Leveque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

K. S. Mendelson. Effective conductivity of two-phase material with cylindrical phase boundaries. *Journal of Applied Physics*, 46(2):917–918, 1975.

William F. Mitchell and Marjorie A. McClain. A survey of *hp*-adaptive strategies for elliptic partial differential equations. In Theodore E. Simos, editor, *Recent Advances in Computational and Applied Mathematics*, pages 227–258. Springer Netherlands, 2011. ISBN 978-90-481-9981-5. doi: 10.1007/978-90-481-9981-5_10.

R. Rannacher. Einführung in die Numerische Mathematik II (Numerik partieller differentialgleichungen). `http://numerik.iwr.uni-heidelberg.de/~lehre/notes`, 2006.

M. Renardy and R. C. Rogers. *An Introduction to Partial Differential Equations*, volume 13 of *Texts in Applied Mathematics*. Springer, 1993.

B. Rivière. *Discontinuous Galerkin methods for solving elliptic and parabolic equations.* Frontiers in Applied Mathematics. SIAM, 2008.

L. R. Scott and S. Zhang. Finite element interpolation of non-smooth functions satisfying boundary conditions. *Math. Comput.*, 54:483?–493, 1990.

W. I. Smirnow. *Lehrgang der höheren Mathematik - Teil II.* VEB Verlag der deutschen Wissenschaften, 15. edition, 1981.

Andrea Toselli and Olof B. Widlund. *Domain Decomposition Methods – Algorithms and Theory.* Springer-Verlag Berlin Heidelberg, 2005.

M. Vohralik. Guaranteed and fully robust a posterior error estimates for conforming discretizations of diffusion problems with discontinuous coefficients. *Journal of Scientific Computing*, 46(3):397–438, 2011.