Freie Universität Berlin

FB Mathematik und Informatik

Numerische Mathematik/Scientific Computing

Numerics of Partial Differential Equations

Ralf Kornhuber and Christof Schütte

3rd Edition: Summer Semester 2007
– English Version from September 2012 –

Translation: Alexander Sikorski and Jonas Krone

Contents

1 Where do Partial Differential Equations Come from? 1

1.1 Variational Principle 1

1.1.1 Deflection of a Membrane 1

1.1.2 Potential Equation 5

1.2 Conservation Principle 6

1.2.1 Mass Conservation, Diffusion and Convection 6

1.2.2 Conservation of Energy (Heat Equation) 11

1.2.3 Groundwater Flow in Saturated Soil 12

1.2.4 Conservation of Momentum (Wave Equation) 14

1.2.5 Traffic Flow (Car Conservation) 16

2 Elliptic, Parabolic, Hyperbolic 17

2.1 The Cauchy Problem 17

2.2 Classification 20

2.2.1 Quasi-linear differential equations of second order in two variables . . 20

2.2.2 Systems of first order in two variables 22

2.3 Well-Posed or Ill-Posed Problems 25

2.3.1 Initial Value Problems 25

2.3.2 Boundary and Initial Value Problems 28

3 Classical Solutions to Elliptic Problems 29

3.1 Representation Formulas and Green's Function 29

3.2 Existence 34

3.3 Uniqueness and Continuous Dependence on the Data 35

4 Difference Methods 37

4.1 Grids and Difference Quotients 37

4.2 The Shortley–Weller Method 40

4.3 Consistency, Stability and Convergence 44

5 Weak Solutions 47

5.1 Hilbert Spaces 48

5.2 Sobolev Spaces 57

5.2.1 Completion 57

5.2.2 Fundamental properties of Sobolev spaces 58

5.3 Weak Solutions of Elliptic Boundary Value Problems 66

6 Finite Elements 71

6.1 Construction of FE Spaces 71

6.2 Error Estimates . . . 75
6.2.1 Error Estimates in the H^1-Norm . . . 75
6.2.2 Error Estimates in the L^2-Norm . . . 81
6.2.3 Adaptive Step-Size . . . 82
6.3 Condition Number of the Stiffness Matrix and Fourier Method . . . 84
6.4 Multi-Grid Methods . . . 92

7 Parabolic Differential Equations 101
7.1 Classical Solutions . . . 101
7.1.1 The Cauchy-Problem for the Heat Equation . . . 101
7.1.2 Initial–Boundary–Value Problems . . . 104
7.2 Weak Solutions . . . 109

8 Numerical Methods for Parabolic Problems 117
8.1 Time Integration . . . 117
8.2 Semidiscrete Methods . . . 119
8.2.1 Method of Lines . . . 119
8.2.2 Rothe's Method . . . 124

Bibliography 127

1 Where do Partial Differential Equations Come from?

1.1 Variational Principle

1.1.1 Deflection of a Membrane

We consider a membrane (thin plate without bending-stiffness) under strain by a vertically acting force. At the boundary, the membrane is fixed by a planar frame. The for the mathematical formulation essential quantities are

deflection	:	u	(m)	
force density	:	f	$(Newton/m^2)$	(acting in direction u).

We represent the membrane by a set $\Omega \subset \mathbb{R}^2$. In the following Ω shall always be a *domain*, that is

$$\Omega \text{ is open and connected}$$

with sufficiently smooth boundary (i.e. Ω Riemann-measurable). Now we check how big the deflection $u(x)$ in every point $x = (x_1, x_2) \in \Omega \subset \mathbb{R}^2$ with a given force density f will be. First, the fixation forces the following **boundary condition**:

$$\boxed{u(x) = 0 \quad \forall x \in \partial\Omega \quad .}$$

In order to characterize u, we use the **Principle of Minimal Energy**: The deflection u has the property

$$\boxed{u \in H : \quad J(u) \leq J(v) \quad \forall v \in H \quad .} \tag{1.1}$$

Where H is the set of all allowed deflections, and $J(v)$ the *energy* of a deflection $v \in H$. We gain the *energy functional* $J : H \to \mathbb{R}$ from the following *energy balancing*. For now, v shall be an arbitrary, smooth enough function on Ω.

- The *strain energy* $J_1(v)$ is proportional to the change of area of the surface.

  Surface area of Ω : $\int_\Omega 1 \, dx$

  Surface area after deflection v : $\int_\Omega \sqrt{1 + v_{x_1}^2 + v_{x_2}^2} \, dx$

  So altogether we have

$$J_1(v) = \alpha \int\limits_\Omega \left(\sqrt{1 + v_{x_1}^2 + v_{x_2}^2} - 1 \right) dx \quad .$$

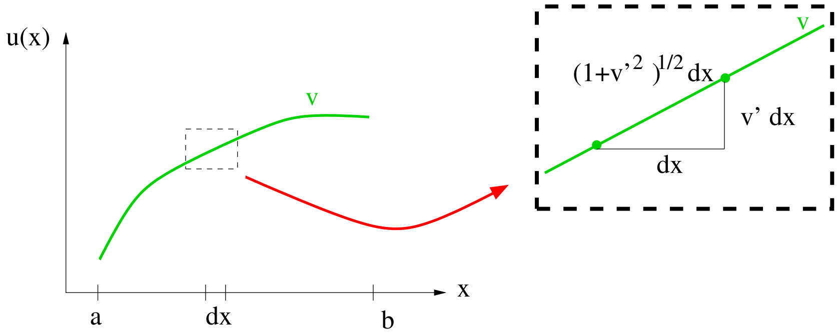


Figure 1.1: Calculation of the arc length $l(a,b)$ of the graph of a function $v = v(x)$ as $l(a,b) = \int_a^b \sqrt{1+v'(x)^2}\,dx$. Here $v' = v_x$ is the derivative of v. The two-dimensional generalization yields the above formula for surfaces after deflection.

The material constant $\alpha > 0$ is called *elasticity* of Ω. Under certain conditions one can simplify this expression: For *small deflections* ∇v is

$$\sqrt{1+v_{x_1}^2+v_{x_2}^2} - 1 \doteq 1 + \frac{1}{2}(v_{x_1}^2+v_{x_2}^2) - 1 = \frac{1}{2}(v_{x_1}^2+v_{x_2}^2) \quad .$$

This yields

$$J_1(v) \doteq \frac{1}{2}\int\limits_\Omega \alpha\,|\nabla v|^2\,dx \quad .$$

- As *potential energy* (= force × path) $J_2(v)$ one gains

$$J_2(v) = -\int\limits_\Omega f v\,dx \quad .$$

- The resulting total energy $J(v) = J_1(v) + J_2(v)$ is

$$\boxed{J(v) = \frac{1}{2}\int\limits_\Omega \alpha\,|\nabla v|^2\,dx - \int\limits_\Omega f v\,dx \quad .}$$

We will come across the following function spaces (linear IRvector spaces) frequently in this lecture.

Definition 1.1 *Suppose $\overline{\Omega}$ the closure of Ω. Then we set:*

$$\begin{aligned}
C(\Omega) &:= \{v : \Omega \to \mathbb{R} \,|\, v \text{ continuous on } \Omega\}\\
C(\overline{\Omega}) &:= \{v \in C(\Omega) \,|\, v \text{ continuously extendable on } \overline{\Omega}\}\\
C^m(\Omega) &:= \{v \in C(\Omega) \,|\, \partial^\gamma v \in C(\Omega),\ |\gamma| \le m\}\\
C^m(\overline{\Omega}) &:= \{v \in C(\Omega) \,|\, \partial^\gamma v \in C(\overline{\Omega}),\ |\gamma| \le m\}
\end{aligned}$$

For convenience reasons we set the higher derivatives ∂^γ to be

$$\partial^\gamma = \frac{\partial^k}{\partial x_{\gamma_1} \dots \partial x_{\gamma_k}}, \; \gamma_i = 1, 2, \; |\gamma| = k$$

using the multi index γ. *In the case* $\Omega \subset \mathrm{I\!R}^d$, $\gamma_i \in 1, \dots, d$.

From the *data* f we require

$$f \in C(\overline{\Omega}) \quad .$$

A glance at the energy functional J suggests to choose (1.1) the linear space

$$\boxed{H_C := \{ v \in C^1(\overline{\Omega}) \, | \, v|_{\partial\Omega} = 0 \} \quad .}$$

as solution space H for our minimization problem. Then, on the one hand, the fixation condition is fulfilled, and on the other hand

$$J(v) < \infty \quad \forall v \in H_C \quad .$$

We will later see, that, in the sense of a complete existence theory, it is an advantage to allow larger spaces H.

Theorem 1.2 (variational formulation) *A function $u \in H_C$ is a solution of* (1.1) *with $H = H_C$ if and only if it fulfills the variational equation*

$$\boxed{\int_\Omega \alpha \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \forall v \in H_C \quad .} \tag{1.2}$$

Proof:

First consider $u \in H_C$ to be a solution of the minimization problem (1.1) and choose an arbitrary $v \in H_C$ and $t > 0$. Now, $J(u) \le J(u + tv)$, because u minimizes the functional J, so

$$0 \le J(u+tv) - J(u) = t\Big(\int_\Omega \alpha \nabla u \cdot \nabla v \, dx - \int_\Omega f v \, dx \Big) + t^2 \frac{1}{2} \int_\Omega \alpha |\nabla v|^2 \, dx \quad .$$

Dividing by t and letting $t \to 0$ we get

$$\int_\Omega \alpha \nabla u \cdot \nabla v \, dx \ge \int_\Omega f v \, dx \quad .$$

The inverse estimate follows analogously by choosing $t < 0$.
Now consider $u \in H_C$ to fulfill the variational equation (1.2), then for arbitrary $v \in H_C$

$$J(u+v) - J(u) = \int_\Omega \alpha \nabla u \cdot \nabla v \, dx - \int_\Omega f v \, dx + \frac{1}{2} \int_\Omega \alpha |\nabla v|^2 \, dx \ge 0 \quad .$$

Note:

In Theorem 1.2 there is nothing said about the existence of u. ◁

Note:
For every set $u \in H_C$ by

$$J'(u)(v) = \int\limits_{\Omega} (\alpha \nabla u \nabla v - fv)\, dx, \quad v \in H_C$$

defines a linear mapping from H_C to $\mathrm{I\!R}$. If one now sets

$$\|v\|_1 = \left(\int\limits_{\Omega} (v^2 + |\nabla v|^2)\, dx \right)^{1/2}$$

then it follows naturally, that

$$|J(u+v) - J(v) - J'(u)(v)| = \frac{1}{2} \int\limits_{\Omega} \alpha |\nabla v|^2\, dx \le \frac{1}{2} \alpha \|v\|_1^2 \quad .$$

Thereby $J'(u)$ is precisely the *Fréchet derivative* of J in u. The variational formulation (1.2) suggests, that the minimum of J is to be found in u just as $J'(u)$ vanishes. ◁

So we have cleared the relation between the minimization problem (1.1) and the variational formulation (1.2). Next, we want to derive a relation between these formulations and a partial differential equation (PDE). For this we need

Theorem 1.3 (Green's formula, partial integration) *Let $v, w \in C^1(\overline{\Omega})$ and $\partial\Omega$ be smooth (sufficient: $\partial\Omega$ has a continuous differentiable parametrization). Then*

$$\boxed{\int\limits_{\Omega} v_{x_i} w\, dx = -\int\limits_{\Omega} v w_{x_i}\, dx + \int\limits_{\partial\Omega} vw\, n_i\, d\sigma \qquad i = 1, 2 \quad ,} \tag{1.3}$$

where $n = (n_1, n_2)$ is the outward-oriented normal on $\partial\Omega$.

Proof:
Exercise. □

For simplicity we will call a domain where Green's formulas (1.3) hold *Green domain*.

Note:
Compare this to the product rule from Calculus

$$\int\limits_a^b vw'\, dx = -\int\limits_a^b v'w\, dx + vw|_a^b$$

for $v, w \in C^1[a, b]$. ◁

Theorem 1.4 *Let Ω be a Green domain. A function $u \in C^2(\overline{\Omega})$ is a solution to the minimization problem* (1.1) *if and only if it is a solution to the boundary value problem*

$$\begin{array}{rcll} -\alpha\Delta u(x) & = & f(x) & \forall x \in \Omega \\ u(x) & = & 0 & \forall x \in \partial\Omega \quad . \end{array} \tag{1.4}$$

Proof:
Suppose $u \in C^2(\overline{\Omega})$ to be a solution of the minimization problem. Using Theorem 1.2 it suffices that u fulfills the variational equation (1.2). Green's formula yields

$$\int\limits_{\Omega} (-\alpha\Delta u - f)v\,dx = 0 \quad \forall v \in H_C \quad . \tag{1.5}$$

In contradiction to the claim, we assume that there is a $x_0 \in \Omega$, s.t.

$$-\alpha\Delta u(x_0) - f(x_0) > 0 \quad .$$

Since Ω is open and $-\alpha\Delta u - f$ is continuous in Ω, we can find an open neighbourhood $U(x_0)$ of x_0, s.t.

$$-\alpha\Delta u(x) - f(x) > 0 \qquad \forall x \in U(x_0) \quad .$$

But now we can construct a negative function $v \in H_C$ with the property

$$v(x) > 0 \quad \forall x \in \bar{K}_\varepsilon(x_0) \subset U(x_0), \quad v(x) = 0 \quad \forall x \in \Omega \setminus U(x_0) \quad .$$

Here, $\bar{K}_\varepsilon(x_0)$ is the closed ball around x_0 with radius $\varepsilon > 0$. Inserting such a function v in (1.5) leads to a contradiction. Analogously one can argument in the case $-\alpha\Delta u(x_0) - f(x_0) < 0$.
If conversely $u \in C^2(\overline{\Omega})$ is a solution of the BVP, then by multiplying of the differential equation with an arbitrary $v \in H_C$, integrating over Ω and applying Green's formula, u fulfills the variational equation (1.2)□

Note:
The differential equation $-\Delta u = f$ is called *Poisson's equation.* In variational calculus, one calls (1.4) *Euler's differential equation* of the minimization problem (1.1). ◁

Note:
Observe that the equivalence of minimization and boundary value problem are only true under an additional *regularity assumption.* Particularly, it can happen that (1.1) has a (physically reasonable!) solution, whereas (1.4) has no solution. ◁

1.1.2 Potential Equation

On a bounded subset $\Omega \subset \mathbb{R}^3$ with a given *charge density* f we want to determine the *potential* u.

electric potential	:	u	$(Volt)$
charge density	:	f	$(A\,s/m^3)$
applied voltage	:	$u\|_{\partial\Omega} = g$	$(Volt)$
dielectric constant	:	ε	$(A\,s/Volt\,m)$

Following the Principle of Minimal Energy and a correspondent energy balance the solution is characterized by the **minimization problem**

$$\boxed{u \in H: \quad J(u) = \frac{1}{2} \int\limits_{\Omega} \varepsilon \, |\nabla u|^2 \, dx - \int\limits_{\Omega} f u \, dx \leq J(v) \quad \forall v \in H \quad .}$$

Under appropriate regularity assumptions on the applied voltage g one could consider as the solution set $H = H_C$,

$$H_C^g := \{ v \in C^1(\overline{\Omega}) \, | \, v|_{\partial\Omega} = g \} \quad .$$

Notice that this solution set is no linear space. Similar as above, one can derive a **variational formulation** (How?). From this, one gains under an additional **regularity condition** the corresponding **Euler differential equation**

$$\boxed{- \operatorname{div}(\varepsilon \nabla u) = f \quad \text{in } \Omega}$$

together with boundary conditions (Which?). In the case $f = 0$ this is called *potential equation.*

Note:
Again, we derived the potential equation only under certain smoothness conditions *a posteriori.* The original problem was a minimization problem. ◁

1.2 Conservation Principle

1.2.1 Mass Conservation, Diffusion and Convection

We consider the movement of a fluid in a given domain $\Omega \subset \mathrm{I\!R}^3$.

density	:	ϱ	(kg/m^3)
velocity	:	v	(m/s)
source density	:	f	$(kg/(s\,m^3))$

As a foundation for the mathematical description of this situation, we use the **Principle of Mass Conservation**: In every fixed *control volume* $\Omega' \subset \Omega$ the following *mass equilibrium* holds

mass change in time = – mass outflow + mass inflow.

If mass is flowing into Ω', then we have a negative mass outflow, so outflow = – inflow. Similarly, a sink is to be interpreted as a negative source.
So let $\Omega' \subset \Omega$ be such a control volume with smooth enough boundary $\partial\Omega'$. We want to express the change in mass, inflow and outflow in the period from t to $t + \Delta t$ through the function ϱ, v and f.

- Mass change in time:

$$\int\limits_{\Omega'} \varrho(x, t + \Delta t) \, dx - \int\limits_{\Omega'} \varrho(x, t) \, dx \quad .$$

- Mass outflow: Mass escapes through a small piece of $\partial\Omega'$ with area $b = \Delta\sigma$ in normal direction n (directed outwards!) with the velocity $(n\cdot v)n$. The velocity v is considered to be constant on that piece. Then, the escaping mass covers in the time interval Δt the distance
$$a = n \cdot v \; \Delta t \quad .$$
Thereby, at this time, a cuboid of the volume
$$ab = n \cdot v \; \Delta\sigma \; \Delta t$$
escapes the volume Ω' (see figure 1.2).

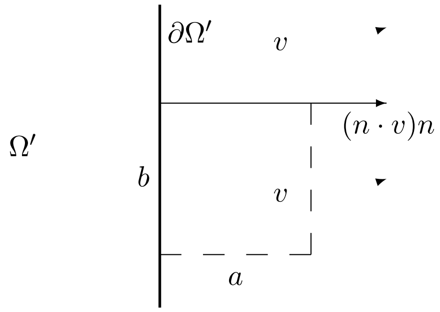


Figure 1.2: flow through the boundary of Ω'

So through this surface patch the following mass flows out
$$\varrho n \cdot v \; \Delta\sigma \; \Delta t \quad .$$
Summing up and going to the limit, the mass outflow through all $\partial\Omega'$ becomes the surface integral
$$\Delta t \int\limits_{\partial\Omega'} \varrho n \cdot v \; d\sigma \quad .$$

- Source:

If the source density f is constant in the time interval Δt, then the mass change in Ω' in this period is given by
$$\int\limits_{\Omega'} f(x,t)\Delta t \; dx \quad .$$

Then the overall mass change in the period Δt becomes
$$\frac{1}{\Delta t}\left(\int\limits_{\Omega'} \varrho(x, t+\Delta t)\; dx - \int\limits_{\Omega'} \varrho(x,t)\; dx\right) = -\int\limits_{\partial\Omega'} \varrho v \cdot n \; d\sigma + \int\limits_{\Omega'} f \; dx \quad .$$
Going to the limit $\Delta t \to 0$, this leads to the **mass conservation in integral form**:
$$\boxed{\frac{d}{dt}\int\limits_{\Omega'} \varrho \; dx + \int\limits_{\partial\Omega'} \varrho v \cdot n \; d\sigma = \int\limits_{\Omega'} f \; dx \quad \forall \Omega' \subset \Omega \quad .} \tag{1.6}$$

In the stationary case, this becomes

$$\int\limits_{\partial\Omega'} \varrho v \cdot n \, d\sigma = \int\limits_{\Omega'} f \, dx \quad \forall \Omega' \subset \Omega \quad . \tag{1.7}$$

Now we want to express the mass conservation equation in differential form, which in this case is as a partial differential equation. The essential tool is again Green's formula (Thm. 1.3). A direct consequence of this is

Theorem 1.5 (Divergence Theorem) *Suppose $\Omega' \subset \mathbb{R}^3$ a Green domain and $v = (v_1, v_2, v_3)^T$ a vector field in Ω' with $v_i \in C^1(\overline{\Omega'})$, $i = 1, 2, 3$, abbreviated $v \in \left(C^1(\overline{\Omega'})\right)^3$. Then:*

$$\boxed{\int\limits_{\Omega'} \operatorname{div} v \, dx = \int\limits_{\partial\Omega'} (v \cdot n) \, d\sigma \quad .}$$

Theorem 1.6 (Continuity Equation) *Choose an arbitrary $t \in \mathbb{R}_+$. Under the* regularity assumptions

$$(\varrho v)_i(\cdot, t) \in C^1(\overline{\Omega}), \ i = 1, 2, 3, \quad \varrho_t(\cdot, t) \in C(\overline{\Omega}), \quad f(\cdot, t) \in C(\overline{\Omega})$$

the functions ϱ, v and f fulfill the integral conservation equation (1.6) *for all $\Omega' \subset \Omega$ if and only if they fulfill the continuity equation*

$$\boxed{\varrho_t + \operatorname{div}(\varrho v) = f \quad \text{in } \Omega \quad .} \tag{1.8}$$

Proof:

First, consider the integral equation (1.6) to be true. Contradictory to the claim, suppose for a certain $x_0 \in \Omega$,

$$\varrho_t(x_0, t) + \operatorname{div}(\varrho v)(x_0, t) - f(x_0, t) > 0 \quad .$$

Since Ω open and $\varrho_t + \operatorname{div}(\varrho v) - f$ continuous in Ω, there is an open ball $K_\varepsilon(x_0)$, s.t $\overline{K_\varepsilon(x_0)} \subset \Omega$ and

$$\varrho_t(x, t) + \operatorname{div}(\varrho v)(x, t) - f(x, t) > 0 \qquad \forall x \in K_\varepsilon(x_0) \quad .$$

We choose a control volume $\Omega' = K_\varepsilon(x_0)$. Applying the divergence theorem (1.6), this yields

$$\frac{d}{dt} \int\limits_{\Omega'} \varrho \, dx + \int\limits_{\Omega'} (\operatorname{div}(\varrho v) - f) \ dx = 0 \quad .$$

By the premises, ϱ_t is uniformly continuous in Ω' and thereby we are allowed to interchange differentiation and integration. Thus

$$\int\limits_{\Omega'} (\varrho_t + \operatorname{div}(\varrho v) - f) \ dx = 0 \quad .$$

But from the construction, the integrand is larger than 0 on $\Omega' = K_\varepsilon(x_0)$. That is a contradiction.
The converse follows directly from the divergence theorem. □

Note:

Observe, that again integral and differential form are equivalent only under *additional regularity assumptions*. So the use of (1.8) in place of (1.6) means a restriction. Particularly, there can be (physically sensible!) solutions to (1.6), whereas the differential equation (1.8) has no solution. ◁

In many problem cases the source density f is not known. Unfortunately, the unknown functions v and ϱ are left over, that cannot be uniquely determined from just one equation. To the (indisputable) Principle of Mass Conservation, there are additional material equations or **equation of state** added, which describe the present mass flux closer. The form of these equations of state is a question of physical *modelling*, that aims for a viable mathematical description of the whole problem, but that, from a mathematical point of view, constitutes a (at first sight) completely arbitrary additional condition. This condition usually is only under certain physical premisses true, and therefore in general to be handled with caution!
We introduce a dimensionless *concentration* u and set

$$\varrho = \varrho_0 u\ , \quad \varrho_0 = \text{const.} > 0 \quad .$$

The fundamental transportation processes are *diffusion* and *convection*.

Diffusion: Ficks' Law (1855)

$$\varrho v = -\alpha \nabla u, \quad \alpha > 0$$

In words: The mass flux is proportional to the steepest decent of the concentration.

Convection:

$$\varrho v = \vec{\beta} u, \quad \vec{\beta} = \text{const.}$$

In words: The expansion velocity $v = \vec{\beta}/\varrho_0$ is fixed.

Convection–Diffusion:

$$\varrho v = -\alpha \nabla u + \vec{\beta} u$$

In words: Both phenomena coincide.
Applying this to the conservation equation (1.8), we gain

$$\boxed{\varrho_0 u_t = \operatorname{div}(\alpha \nabla u - \vec{\beta} u) + f \quad .}$$

In general, α, $\vec{\beta}$ are functions depending on x and u! In the case $\alpha, \vec{\beta} = \text{const.}$, this becomes the **convection–diffusion equation**:

$$\varrho_0 u_t = \alpha \Delta u - \vec{\beta} \nabla u + f \quad .$$

Note:
From the conservation of mass and momentum together with the equation of state for so called incompressible fluids with constant density, this yields the *Navier–Stokes Equations*

$$\begin{array}{rcl} \vec{u}_t + (\vec{u} \cdot \nabla)\vec{u} + \nabla p &=& f + \varepsilon \Delta \vec{u} \\ \nabla \cdot \vec{u} &=& 0 \end{array} \qquad \text{in } \Omega \quad .$$

There ε stands for the viscosity of the fluid, $\vec{u} = (u_1, u_2, u_3)^T$ for the velocity and p for the pressure. The Laplace operator $\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2}$ and $(\vec{u} \cdot \nabla) = u_1 \frac{\partial}{\partial x_1} + u_2 \frac{\partial}{\partial x_2} + u_3 \frac{\partial}{\partial x_3}$ are applied component-wise. Under the (severe!) simplification:

- known flux direction $\vec{\beta} \cdot \nabla = \vec{u} \cdot \nabla$ and
- known pressure p

one gains the convection–diffusion equation for every component u_i. ◁

Apparently, one can expect a unique solution only if one describes concentration and mass flux at the boundary appropriately. The most common boundary conditions (BC) are

Dirichlet boundary condition:

$$\boxed{\varrho = \varrho_0 u = g \qquad \text{on } \partial\Omega \quad .}$$

In words: prescription of the density resp. the concentration.

Von Neumann boundary condition:

$$\boxed{\varrho v \cdot n = g \qquad \text{on } \partial\Omega}$$

with outward pointing normal n on $\partial\Omega$.
In words: Prescription of the mass outflow.

Cauchy boundary condition:

$$\boxed{\varrho v \cdot n = \alpha u \qquad \text{on } \partial\Omega \quad .}$$

In words: The mass outflow is proportional to the concentration.
Instead of Cauchy BC one sometimes also speaks of *Robin boundary condition.* Dirichlet-, von Neumann- and Cauchy BC are also called *boundary conditions of* 1^{st}*,* 2^{nd}*, and* 3^{rd} *kind*, respectively.

Note:
Of course, one can also prescribe different boundary conditions on different parts of $\partial\Omega$.
In addition to these standard types of there might be other boundary conditions reasonable. As an example for a *non-local boundary condition* consider the (global) *mass conservation*

$$\int\limits_{\partial\Omega} \varrho v \cdot n \, d\sigma = 0 \quad .$$

Note:
Similarly to the equations of state, the boundary equations are supposed to describe *special physical situations. Mathematically* reasonable BC's together with present (differential) equations provide existence and uniqueness of a solution. The choice of physically and mathematically reasonable boundary conditions in the practical application is anything but easy. ◁

Naturally the concentration $u(x,t)$ depends on the situation at the initial time $t=0$. So we can expect a unique solution u, we have to choose an **initial condition (IC)**

$$\boxed{u(x,0) = u_0(x) \quad x \in \Omega \quad .}$$

1.2.2 Conservation of Energy (Heat Equation)

Additionally to mass there are other quantities that are conserved, for example energy. We consider the example of the heat flow in a domain $\Omega \subset \mathbb{R}^3$, described by the following quantities.

energy density	:	$\mathcal{E}$	$Joule/(m^3)$
heat flow	:	q	$Joule/(m^2\, s)$
heat source	:	f	$Joule/(m^3\, s)$

Exactly as in part 1.2.1 one gets the **energy conservation in integral form**:

$$\boxed{\frac{d}{dt}\int\limits_{\Omega'} \mathcal{E}\, dx + \int\limits_{\partial\Omega'} q \cdot n\, d\sigma = \int\limits_{\Omega'} f\, dx \quad \forall \Omega' \subset \Omega \quad .}$$

Under certain regularity assumptions (cf thm. 1.6) this is equivalent to the **continuity equation**:

$$\boxed{\mathcal{E}_t + \operatorname{div} q = f \quad \text{in } \Omega \quad .}$$

As a first **equation of state**, we postulate

$$\mathcal{E} = \mathcal{E}(\theta) = c\theta,$$

where θ and $c > 0$ stand for the temperature and the specific heat (capacity) respectively. A (simple, but for example near absolute zero completely wrong!) connection between heat flow and temperature provides the following equation of state.

Fourier's Law (1822):

$$q = -\kappa \nabla\theta \quad .$$

The proportionality constant $\kappa > 0$ is called *thermal conductivity*. Fourier's Law describes heat conduction as **diffusion**. In its differential form, this becomes the **heat equation**:

$$\boxed{\frac{\partial}{\partial t}(c\theta) = \operatorname{div}(\kappa\nabla\theta) + f \quad .}$$

In general, c, κ are functions depending on x and u. If these functions are constant, the heat equation simplifies to

$$\theta_t = \gamma\Delta\theta + g \quad \text{with} \quad \gamma = \frac{\kappa}{c},\ g = \frac{f}{c} \quad .$$

Note:
We will briefly discuss the heat conduct in a time variable medium, that is

$$x = x(t), \quad x \in \Omega \quad .$$

The velocity at a point $x = x(t)$ then is $\vec{\beta} = \dot{x}(t)$. In this case, following the chain rule one gets a **convection–diffusion Equation**

$$\frac{\partial}{\partial t}\theta(x(t), t) = \vec{\beta} \cdot \nabla\theta + \theta_t = \gamma\Delta\theta + g \quad .$$

Observe that the *convection* $\vec{\beta} \cdot \nabla\Theta$ is caused by the movement of the medium.
Again, one has to choose *boundary* and *initial conditions.* What physical meaning do Dirichlet, Neumann and Cauchy BC have in this case? ◁

1.2.3 Groundwater Flow in Saturated Soil

The groundwater flow through a domain $\Omega \subset \mathbb{R}^3$ can be described by the following quantities.

density	:	ϱ	kg/m^3
porosity	:	n	dimensionless
saturation	:	S	dimensionless
distance velocity	:	v	m/s
filtration velocity	:	v_F	m/s
pressure	:	p	$Newton/m^2$
source density	:	f	$(kg/m^3\,s)$

The *porosity* $n(x)$, $x \in \Omega$ describes the content of *pore space* in an (infinitesimal) reference volume. The *saturation* $S(x,t)$ describes the content of fluid in the pore space. Then the **mass conservation in differential form** is:

$$(Sn\varrho)_t + \operatorname{div}(Sn\varrho v) = f \quad . \tag{1.9}$$

Again, a number of equations of state are necessary. We wanted to consider a *saturated* flow, so

$$S(x,t) \equiv 1 \quad \forall x \in \Omega,\ t > 0 \quad .$$

Furthermore, temperature differences shall be ignored. Then

$$\varrho(x,t) \equiv \varrho_0 = \text{const.} \quad \forall x \in \Omega,\ t > 0,$$

because water is *incompressible.* The porosity is a function of the pressure, so

$$\varrho_0 \frac{\partial(n(p))}{\partial t} = \varrho_0 \frac{dn}{dp} p_t$$

Now let us assume, that

$$\varrho_0 \frac{dn}{dp} = \frac{S_0}{g} = \text{const.}, \tag{1.10}$$

where g denotes the *gravitational acceleration* (unit m/s^2) and S_0 the *specific storage* (unit $1/m$). In an extensive study of the wells of Dijon, Henry Darcy found the equation of state for the filtration velocity.

Darcy's Law (1856):

$$v_F = -K\nabla h$$

The symmetric 3×3-matrix K with columns K_i, $i = 1, \ldots, 3$, is called *hydraulic conductivity* and describes the soil's permeability. The function h is called *piezometric* or *(total) hydraulic head* and can be observed as the level of groundwater in a borehole. This reads

$$h = \frac{p}{\varrho_0 g} + x_3 \quad .$$

Finally we suggest, that the (microscopic) distance velocity v correlates to the (macroscopic) filtration velocity v_F in the following way.

$$v_F = nv$$

Applying all these relations to (1.9), we finally obtain **Darcy's Equation**:

$$\boxed{S_0 p_t = \operatorname{div}(K\nabla p) + \varrho_0 g \operatorname{div} K_3 + gf \quad .} \tag{1.11}$$

Note:
Instead of the situation described by the equation of state (1.10), we might also have a *pressure-stable grain structure* given, that is

$$\frac{dn}{dp} = 0 \; .$$

In this case, Darcy's Equation (1.11) reduces to

$$-\operatorname{div}(K\nabla p) = \varrho_0 g \operatorname{div} K_3 + gf \; .$$

Note:
We had assumed $S(x,t) \equiv 1$ and $K = K(x)$. In the case of *unsaturated currents* everything becomes more complicated, namely $S(x,t) \neq 1$ and $K = K(p,x)$. To eliminate the unknown saturation S from the differential equation, one uses in empirically acquired pressure–saturation relations to obtain a *non-linear* differential equation, the so called *Richards Equation.*

Prescribed pressure–saturation relations are only valid under certain conditions. If those are not fulfilled, one has to pass on to even more complex models, consisting of to separate non-linear differential equations for pressure and saturation. With such *multiphase flows*, we are quite close to current research. ◁

Now some appropriate **boundary conditions** are missing. What physical meaning do Dirichlet, Neumann and Cauchy BC have in this case?

1.2.4 Conservation of Momentum (Wave Equation)

Up until now the we focussed on the continuity equation. Finally the **Principle of Conservation of Momentum**:

The increase of momentum in a given material control
volume is equal to an outside force acting on it.

By the way: The conservation of momentum is also reflected in its measuring units (reminder: 1 Newton = 1 kg m/s^2). For every control volume $\Omega' \subset \Omega \subset \mathbb{R}^d$ (with smooth enough boundary), this physical law can be written directly as a formula,

$$\boxed{\frac{d}{dt}\int\limits_{\Omega'} \varrho v \, dx = \int\limits_{\Omega'} f \, dx + \int\limits_{\partial\Omega'} \sigma \cdot n ds \quad .}$$

Here we have used the following quantities.

density	:	ϱ	kg/m^d
velocity	:	v	m/s
volume force density	:	f	$Newton/m^d$
stress tensor	:	σ	$Newton/m^{d-1}$

Again one can reformulate this balance equation (under additional regularity conditions) into a differential equation (which one?).
Now, we want to use the conservation of momentum to describe a vibrating string (without gravitation and bending-stiffness). The string shall be fixed at a and b and we are interested in its **deflection** at every time $t > 0$:

$$\boxed{u : [a, b] \to \mathbb{R} \ .}$$

In this (one dimensional) case the momentum balancing takes the form (cf. figure 1.3)

$$\frac{d}{dt}\int\limits_{x_0}^{x_1} \varrho v \, dx = \int\limits_{x_0}^{x_1} f \, dx + \sigma(x_1) - \sigma(x_0) \quad \forall x_0, x_1, \quad a \le x_0 < x_1 \le b \quad .$$

We now have to determine ϱ, v, f and σ. Gravitational forces shall be omitted, so

$$f \equiv 0 \ .$$

For vertical oscillations, it holds

$$v = u_t \ .$$

For the stresses $\sigma(x_0)$ and $\sigma(x_1)$ we again need an *equation of state.* To this end we define a function $L : [a, b] \to \mathbb{R}$ as

$$L(x) = \text{Length of the curve}\,\Gamma$$

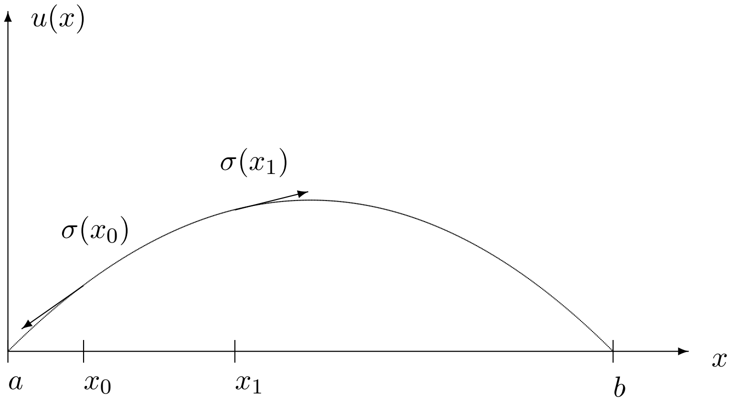


Figure 1.3: oscillating string

with $\Gamma(x) = \{(s, u(s)) \in \mathrm{I\!R}^2 \mid s \in [a, x]\}$.
The change in length $\frac{dL}{dx}$ is, as you might know,

$$\frac{dL}{dx}(x) = \frac{d}{dx}\int\limits_0^x \sqrt{1 + u_x^2(s)}\, ds = \sqrt{1 + u_x^2(x)} - 1 \ .$$

For an *elastic string* **Hooke's Law** comes into effect:

$$\sigma = \alpha\, \frac{dL}{dx} \quad .$$

The proportionality constant $\alpha > 0$ is called elasticity and depends on the material and thereby in general on the position. We want to limit ourselves to *small distortions* $u_x \approx 0$. Then

$$\frac{dL}{dx} \doteq \tfrac{1}{2} u_x \ .$$

Thereby Hooke's Law implies the following *linear equation of state*

$$\sigma = \frac{1}{2}\alpha u_x \quad .$$

Applying all these relations, this provides

$$\frac{d}{dt}\int\limits_{x_0}^{x_1} \varrho u_t \, dx = \tfrac{1}{2}\big(\alpha(x_1)u_x(x_1) - \alpha(x_0)u_x(x_0)\big) \qquad \forall x_0, x_1 \quad a < x_0 < x_1 < b \quad .$$

Under appropriate *regularity conditions* on the solution u and the data ϱ, α, we obtain in the usual way the **wave equation**:

$$\boxed{(\varrho u_t)_t = \frac{1}{2}(\alpha u_x)_x \quad \text{in } (a, b) \ .}$$

In the case $\varrho, \alpha = \text{const.}$ the wave equation simplifies with $\gamma = \frac{\alpha}{2\varrho}$ to

$$\boxed{u_{tt} - \gamma u_{xx} = 0 \quad \text{in } (a,b)\ .}$$

To this, there are for example **Dirichlet boundary conditions** added:

$$u(a,t) = u_a(t) \qquad\qquad u(b,t) = u_b(t) \quad t > 0 \quad .$$

By the way, what physical meaning do von Neumann BCs have? Are Cauchy BCs physically sensible?
Naturally the string's motion depends crucially on how the deflection looks at time $t = 0$ and what velocity it has at that time. To expect a unique solution, we presumably have to choose the **initial conditions**

$$u(x,0) = u_0(x), \qquad u_t(x,0) = u_1(x), \qquad x \in (a,b) \quad .$$

1.2.5 Traffic Flow (Car Conservation)

traffic density	:	ϱ	cars/m
velocity	:	v	m/s

right at the beginning we state the *continuum hypothesis*

$$\varrho : [a,b] \to \mathbb{R} \quad .$$

Despite a growing traffic volume, this is in the present case at least questionable. That alone is a reason not to take this model to serious. Similarly to the conservation of mass one gains the **conservation of cars**:

$$\boxed{\varrho_t + (\varrho v)_x = 0}$$

Apparently

$$0 \le \varrho(x) \le \varrho_{max} = \text{const.} \quad .$$

In the case $\varrho = \varrho_{max}$, the cars are driving bumper-to-bumper. Furthermore, we assume a somewhat sensible driving manner, that is

$$0 \le v(x) \le v_{max} = \text{const.}$$

Interpolating both extreme cases $v = 0$ (in the case $\varrho = \varrho_{max}$) and $v = v_{max}$ (in the case $\varrho = 0$) just linearly, one obtains the **equation of state**

$$v(\varrho) = v_{max}(1 - \varrho/\varrho_{max}) \quad .$$

Applying this yields a **nonlinear conservation law**:

$$\boxed{\varrho_t + v_{max}(\varrho(1 - \varrho/\varrho_{max}))_x = 0 \quad .}$$

2 Elliptic, Parabolic, Hyperbolic

2.1 The Cauchy Problem

We consider the following differential equation of *second order* in two variables x and y

$$\boxed{a u_{xx} + 2b u_{xy} + c u_{yy} = d.} \tag{2.1}$$

The differential equation (2.1) is called

$$\begin{array}{lll} \textit{quasi-linear}, & \text{if} & a,b,c,d = a,b,c,d(x,y,u,u_x,u_y) \\ \textit{semi-linear}, & \text{if} & a,b,c = a,b,c(x,y),\ \ d = d(x,y,u,u_x,u_y) \\ \textit{linear}, & \text{if} & a,b,c,d = a,b,c,d(x,y). \end{array}$$

In any case, a, b, c should not vanish all at the same time (2^{nd} order!). Note that linear combination are solutions again only in the linear case.
We now want to examine the existence of solutions to (2.1).We have already seen, that we can expect uniqueness only under some additional conditions. Since these differential equations are of second order, conditions on the function and its first derivative seem reasonable, that is

$$\begin{aligned} u|_\gamma &= u_0 \\ u_x|_\gamma &= g_1 \\ u_y|_\gamma &= g_2 \quad . \end{aligned}$$

Here $\gamma = (\gamma_1, \gamma_2)$ shall be a smooth curve in $\mathbb{R}^2$ (more specific: $\gamma_1,\ \gamma_2 \in C^\infty(I)$, $I \subset \mathbb{R}$, interval) with a tangent $\gamma' = (\gamma_1', \gamma_2')$ of length $\sqrt{(\gamma_1')^2 + (\gamma_2')^2} > 0$. A closer look shows, that the solution is overdetermined by these conditions. This is due to the *compatibility condition*

$$\frac{d}{ds} u_0(\gamma_1(s), \gamma_2(s)) = g_1 \gamma_1' + g_2 \gamma_2' \quad .$$

So there are only **two conditions free to choose**![1] So we should not, for example, appoint conditions for $u_x|_\gamma$ and $u_y|_\gamma$, but instead for the normal derivative w.r.t. γ:

$$\frac{\partial u}{\partial n} = \nabla u \cdot (\gamma')^\perp = \frac{-u_x \gamma_2' + u_y \gamma_1'}{\sqrt{{\gamma_1'}^2 + {\gamma_2'}^2}}$$

We thus consider the *Cauchy problem*

$$\boxed{\begin{array}{rcll} a u_{xx} + 2\, b u_{xy} + c u_{yy} &=& d & \text{in } \mathbb{R}^2 \\ u &=& u_0 & \text{on } \gamma \\ \dfrac{\partial u}{\partial n} &=& u_1 & \text{on } \gamma \quad . \end{array}} \tag{2.2}$$

[1] One can see this by differentiation following the chain rule and insertion of $u_x|_\gamma = g_1$ and $u_y|_\gamma = g_2$.

The conditions on γ shall be assumed to be continuous, so

$$\lim_{(x,y)\to\gamma(s_0)} u(x,y) = u_0(\gamma(s_0)) \quad .$$

Note that, with the help of the compatibility conditions, the partial derivatives $u_x|_\gamma$ and $u_y|_\gamma$ can be calculated from the function's value and its normal derivative. A (good!) idea for the proof of existence would be a Taylor expansion of the solution u in a neighbourhood of γ, i.e.

$$u(x,y) = u(\gamma(s_0)) + \sum_{k=1}^{\infty} \sum_{|\beta|=k} \frac{1}{\beta!} \partial^\beta u(\gamma(s_0)) h^\beta \qquad \forall x,y \in K_\varepsilon(\gamma(s_0)) \ .$$

Here, $\beta = (\beta_1, \ldots, \beta_k)$, $\beta_i \in \{x,y\}$, $i = 1, \ldots, k$, denotes a multi-index with $|\beta| = k$,

$$\partial^\beta = \partial_{\beta_1} \cdots \partial_{\beta_k}, \qquad h^\beta = h_{\beta_1} \cdots h_{\beta_k}, \ \ h_x = x - \gamma_1(s_0), h_y = y - \gamma_2(s_0)$$

and $\beta! = \beta_x!\beta_y!$. Finally, β_x and β_y shall be the number of indices $\beta_i = x$ and $\beta_i = y$ respectively.

So, in order to determine u, we need the higher derivatives $\partial^\beta u$ of u on γ. The first-order derivatives u_x and u_y, we already have. Now we have to take care of the second-order derivatives. The chain rule provides

$$\begin{aligned} \frac{d}{ds} u_x &= \gamma_1' u_{xx} + \gamma_2' u_{xy} \\ \frac{d}{ds} u_y &= \gamma_1' u_{xy} + \gamma_2' u_{yy} \quad , \end{aligned}$$

and we already have

$$a u_{xx} + 2b u_{xy} + c u_{yy} = d \quad .$$

Under the condition

$$\boxed{\begin{vmatrix} \gamma_1' & \gamma_2' & 0 \\ 0 & \gamma_1' & \gamma_2' \\ a & 2b & c \end{vmatrix} = a{\gamma_2'}^2 - 2b\gamma_1'\gamma_2' + c{\gamma_1'}^2 \neq 0} \tag{2.3}$$

this linear equation system has a uniquely determined solution u_{xx}, u_{xy}, u_{yy} for every right-hand side. If the second-order derivatives are known, we can calculate all higher partial derivatives $\partial^\beta u(\gamma(s_0))$, $|\beta| = 3, 4, \ldots$, by additional differentiation at each point $\gamma(s_0) \in \gamma$ (exercise).

The property (2.3) of γ is apparently of some importance, so, for the sake of completeness, we give it a name.

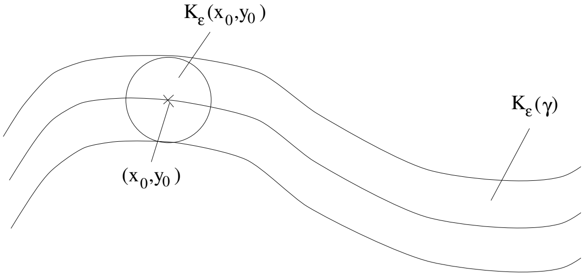
Definition 2.1 *A curve $\gamma(s)$ is said to be characteristic in $s \in I$ if*

$$\boxed{\det(s) := a\gamma_2'(s)^2 - 2b\gamma_1'(s)\gamma_2'(s) + c\gamma_1'(s)^2 = 0 \quad .}$$

A curve $\gamma(s)$ is said to be non-characteristic in $s \in I$ if $\det(s) \neq 0$. *Accordingly, if a curve γ is (non-)characteristic everywhere, i.e. in every $s \in I$, it is called a (non-)characteristic.*

To complete the existence proof foreshadowed above, we need to secure the local convergence of the Taylor expansion. This is the tough part.

Theorem 2.2 (Cauchy–Kovalevskaya) *For the Cauchy problem* (2.2)*, given on the curve* $\gamma = \gamma(s)$*, the following statement holds: Suppose* $\gamma(s)$ *to be non-characteristic in* s_0 *(and thus in a small neighbourhood* s_0*). Furthermore, suppose that in a neighbourhood of* $\gamma(s_0)$ [2]*,* u_0*,* u_1 *or equivalently the functions* a, b, c, d *are real analytic (i.e. representable by a multivariant power series). Then in a small enough neighbourhood* $K_\varepsilon(\gamma(s_0))$ *there is a uniquely determined analytic solution to the Cauchy problem* (2.2)*.*



Proof:
See for example John [13] chapter 3.3 or Renardy and Rogers [19] chapter 2.2 □

Note:
Though this fundamental existence theorem can be generalized a bit, there remain some significant flaws:

- very strong regularity assumptions on the data
- only local existence ◁

To illustrate the fundamental significance of characteristics we will quickly look at the *proliferation of discontinuities.*

Theorem 2.3 *Let* $\gamma = (\gamma_1, \gamma_2)$ *be a smooth curve and* $\Omega = \Omega^{(1)} \cup \gamma \cup \Omega^{(2)}$*. Suppose the functions* a, b, c, d *to be continuous in* Ω*. Furthermore, we define* $u \in C^1(\Omega)$ *by* $u|_{\Omega^{(i)}} = u^{(i)}$*, where*

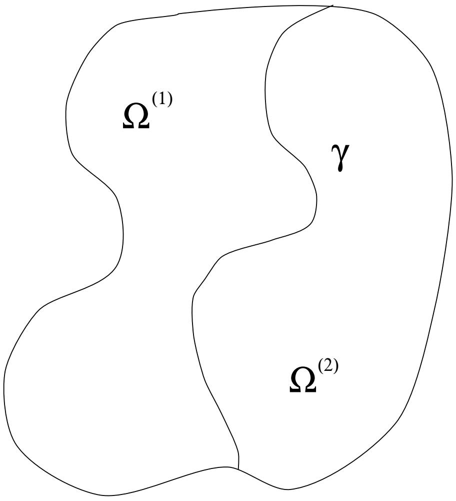
$$u^{(i)} \in C^2(\Omega^{(i)} \cup \gamma),\ i = 1, 2,$$

both shall be solutions to (2.1) *in* $\Omega^{(i)} \cup \gamma$*. If in every point on* γ *at least one of the second derivatives* u_{xx}*,* u_{xy} *or* u_{yy} *is discontinuous, then* γ *is a characteristic.*

Proof:
We define a jump in a function $v : \Omega \to \mathbb{R}$ as above along γ as

$$[v] := v^{(1)}(\gamma_1(s), \gamma_2(s)) - v^{(2)}(\gamma_1(s), \gamma_2(s)) \quad .$$

[2] More precisely: in a neighbourhood of $(\gamma(s_0), u_0(\gamma(s_0)), u_x(\gamma(s_0)),\ u_y(\gamma(s_0))$



Since $u \in C^1(\Omega)$, the chain rule provides

$$\begin{aligned} 0 &= \frac{d}{ds}[u_x] = u_{xx}^{(1)}\gamma_1' + u_{xy}^{(1)}\gamma_2' - u_{xx}^{(2)}\gamma_1' - u_{xy}^{(2)}\gamma_2' = [u_{xx}]\gamma_1' + [u_{xy}]\gamma_2' \\ 0 &= \frac{d}{ds}[u_y] = [u_{xy}]\gamma_1' + [u_{yy}]\gamma_2' \quad . \end{aligned}$$

The continuity of u, u_x, u_y and a, b, c, d yields

$$a[u_{xx}] + 2b[u_{xy}] + c[u_{yy}] = 0 \quad .$$

Altogether the following linear system of equations is obtained

$$\begin{pmatrix} \gamma_1' & \gamma_2' & 0 \\ 0 & \gamma_1' & \gamma_2' \\ a & 2b & c \end{pmatrix} \begin{pmatrix} [u_{xx}] \\ [u_{xy}] \\ [u_{yy}] \end{pmatrix} = 0 \quad . \tag{2.4}$$

Since by our assumption $[u_{xx}]^2 + [u_{xy}]^2 + [u_{yy}]^2 \neq 0$, γ has to be a characteristic. □

2.2 Classification

2.2.1 Quasi-linear differential equations of second order in two variables

The basis for classifying quasi-linear differential equations of second order (cf. (2.1)) is the existence or non-existence of characteristics. In order to calculate the *characteristic directions* $(\gamma_1', \gamma_2')^\top$ we are given the equation

$$a{\gamma_2'}^2 - 2b\gamma_1'\gamma_2' + c{\gamma_1'}^2 = 0 \quad .$$

Consider the arguments x, y, u, u_x, u_y fixed. Whether real solutions γ_1', γ_2' and thus characteristics exist is determined by the discriminant

$$b^2 - ac \begin{cases} < 0 : & \text{no real solution} \\ = 0 : & \text{one real solution} \\ > 0 : & \text{two real solutions} \quad . \end{cases} \tag{2.5}$$

This motivates the following

Definition 2.4 *The differential equation* (2.1) *is said to be in* x, y, u, u_x, u_y

$$\begin{array}{rcccl} \textit{elliptic} & \Longleftrightarrow & b^2 - ac < 0 & \Longleftrightarrow & \textit{no real characteristic} \\ \textit{parabolic} & \Longleftrightarrow & b^2 - ac = 0 & \Longleftrightarrow & \textit{exactly one real characteristic} \\ \textit{hyperbolic} & \Longleftrightarrow & b^2 - ac > 0 & \Longleftrightarrow & \textit{two real characteristics} \quad . \end{array}$$

Without the following result the whole definition would be futile!

Theorem 2.5 *The type of differential equation is invariant under coordinate transformations.*

Proof:
Exercise. □

Note:
The type of differential equation might depend on x, y and even on the solution u (determined by additional conditions). ◁

Note:
Only the main part

$$au_{xx} + 2bu_{xy} + cu_{yy}$$

and not $d(x, y, u, u_x, u_y)$ decides about the type of a differential equation. ◁

Note:
In the case $v \in C_0^\infty(\mathbb{R}^2)$, applying the Fourier transformation

$$\hat{v}(\xi, \eta) = \mathcal{F}v(\xi, \eta) = \frac{1}{2\pi} \int\limits_{\mathbb{R}^2} e^{-i(x\xi + y\eta)}\, v(x, y)\, d(x, y)$$

to the partial derivatives $v = u_x$ one obtains the relation $\mathcal{F}u_x = i\xi\hat{u}$ (partial integration). In the case of constant $a, b, c \in \mathbb{R}$ not depending on the solution and $u \in C_0^\infty(\mathbb{R}^2)$, applying the Fourier transformation to (2.1) accordingly provides

$$(a\xi^2 + 2b\xi\eta + c\eta^2)\, \hat{u}(\xi, \eta) = -\hat{d}(\xi, \eta).$$

The polynomial $a\xi^2 + 2b\xi\eta + c\eta^2$ is called the differential equation's *symbol.* In the elliptic, parabolic or hyperbolic case, the symbol describes an ellipse, a parabola or a hyperbola respectively. Hence these names. ◁

Example:

a) **Constant coefficients**

- Laplace equation:

$$u_{xx} + u_{yy} = 0 \qquad\qquad (a = c = 1,\ b = d = 0)$$

(elliptic, no characteristics)

- Heat equation:

$$u_y - u_{xx} = 0 \qquad\qquad (a = 1,\ b = c = 0,\ d = u_y)$$

(parabolic, characteristic: $\gamma = (s, 0)$)

- Wave equation:

$$u_{xx} - u_{yy} = 0 \qquad\qquad (a = -c = 1,\ b = d = 0)$$

(hyperbolic, characteristics: $\gamma_\pm = (s, \pm s + \text{const.})$) ◁

b) **Coefficients depending on solution**

- potential flow:

$$(c_0^2 - u_x^2)u_{xx} - 2u_x u_y u_{xy} + (c_0^2 - u_y^2)u_{yy} = 0$$

$$\begin{array}{ll} M = \sqrt{c_0^{-2}(u_x^2 + u_y^2)} & \text{Mach number} \\ M < 1 & \text{ellipic} \\ M = 1 & \text{parabolic} \\ M > 1 & \text{hyperbolic} \end{array}$$

2.2.2 Systems of first order in two variables

We now come to classifying systems of 1^{st} order. This shall not contradict our prior agreement. The examination of systems of first order will help us to better understand the meaning of characteristics especially for the hyperbolic case.

The above considered (scalar) equations of second order can be rewritten into a system of 1^{st} order. To this end we introduce new variables v, w

$$v := u_x \quad \text{and} \quad w := u_y \quad . \tag{2.6}$$

With this, we remodel the system of second order

$$au_{xx} + 2bu_{xy} + cu_{yy} = d$$

into the shape

$$\boxed{\begin{aligned} av_x + 2bv_y + cw_y &= d \\ w_x - v_y &= 0 \quad , \end{aligned}} \tag{2.7}$$

where the second equation simply results from

$$v_y = u_{xy} = u_{yx} = w_x \quad .$$

To make things more easy, we assume

$$a \neq 0 \quad .$$

The system (2.7) can be written in matrix notation as

$$\boxed{U_x + AU_y + D = 0 \qquad \text{with} \quad U := \begin{pmatrix} v \\ w \end{pmatrix},}$$

where
$$A = \begin{pmatrix} 2b/a & c/a \\ -1 & 0 \end{pmatrix} \quad \in \mathbb{R}^{2,2}, \qquad \text{and} \quad D = \begin{pmatrix} -d/a \\ 0 \end{pmatrix}.$$

The matrix A has eigenvalues

$$\boxed{\lambda_\pm = \frac{1}{a}\left(b \pm \sqrt{b^2 - ac}\right).}$$

So we have analogue to (2.5) the following situation

$$b^2 - ac \begin{cases} < 0 : & \text{no real eigenvalue} \\ = 0 : & \text{one (double) real eigenvalue} \\ > 0 : & \text{two (different) real eigenvalues} \end{cases}$$

More over, in the case $b^2 - ac > 0$, it happens that A is diagonalizable over $\mathbb{R}$! We use this observation to classify systems of 1^{st} order.

Definition 2.6 *The system*

$$U_x + AU_y + BU + D = 0 \tag{2.8}$$

with $A, B(x,y,v,w) \in \mathbb{R}^{n,n}$, $D(x,y,v,w) \in \mathbb{R}^n$ and an unknown function $U(x,y) \in \mathbb{R}^n$ is called elliptic if all eigenvalues of A are complex and hyperbolic if A is diagonalizable over $\mathbb{R}$.

Example:

- **Laplace equation.** Applying transformation (2.6) to the (elliptic!) Laplace equation $u_{xx} + u_{yy} = 0$, yields the *Cauchy–Riemann differential equation*, that is a system of the form (2.8) with
$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad a = c = 1,\ b = 0.$$
Apparently, A has imaginary eigenvalues $\pm i$. Attention! Connection to complex analysis: The complex-valued function $W(z) = w(z) + iv(z)$ with complex argument $z = x + iy$ is holomorphic (i.e. complex differentiable) if and only if v, w are solutions to the above elliptic problem.

- **Wave equation.** Applying the transformation (2.6) to the (hyperbolic!) wave equation $u_{xx} - u_{yy} = 0$, one obtains a system of the form (2.8) with
$$A = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \qquad a = 1,\ b = 0,\ c = -1.$$

Apparently, A has reel eigenvalues $\lambda_1 = -1$, $\lambda_2 = 1$ and orthogonal eigenvectors

$$e_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad e_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad .$$

Accordingly, we can diagonalize system (2.8) in this case through the transformation

$$T = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \, , \quad T^{-1} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad .$$

We obtain

$$T^{-1}U_x + T^{-1}ATT^{-1}U_y = 0 \tag{2.9}$$

and thus a *decoupled* system $\tilde{U}_x + T^{-1}AT\tilde{U}_y = 0$ with new unknowns

$$\tilde{U} := \begin{pmatrix} \tilde{v} \\ \tilde{w} \end{pmatrix} = T^{-1}U = \frac{1}{2} \begin{pmatrix} u_x + u_y \\ u_x - u_y \end{pmatrix}.$$

Note:
As in the above cases, we can transform the heat equation into a parabolic system of first order, which is actually neither hyperbolic nor elliptic. For lack of meaning though, one does not speak of a parabolic systems of first order. ◁

Characteristic for scalar differential equations of first order. In order to understand the meaning of these ideas closer, let us for a moment consider a scalar differential equation of the form

$$\boxed{a_1 u_x + a_2 u_y + bu + d = 0.} \tag{2.10}$$

with x, y-dependent coefficients from $C^1(\mathbb{R}^2)$. In this case one defines:

Definition 2.7 *A continuously differentiable curve $\gamma = (\gamma_1, \gamma_2) : I = [0, T] \to \mathbb{R}^2$, that fulfills on I the ordinary differential equation*

$$\boxed{\gamma_1' = a_1(\gamma), \qquad \gamma_2' = a_2(\gamma)} \tag{2.11}$$

is called characteristic of (2.10).

The following assertions hold:

Theorem 2.8 *Let $u \in C^1(\mathbb{R}^2)$ be a solution of* (2.10) *and γ a characteristic. Then $U(t) = u(\gamma(t))$ fulfills the ordinary differential equation*

$$\boxed{U'(t) + b(\gamma(t))\, U(t) + d(\gamma(t)) = 0. \quad \forall t \in I} \tag{2.12}$$

Proof:
Exercise. □

Theorem 2.9 *For every point $(x, y) \in \mathrm{I\!R}^2$ there is a characteristic γ with associated parameter interval $I = [0, T]$, such that $\gamma(t) = (x, y)$ for a $t \in [0, T]$ holds. Let $u \in C^1(\mathrm{I\!R}^2)$ be a function with the property, that the restriction $U(t) = u(\gamma(t))$ for any characteristic γ fulfills the ordinary differential equation* (2.12). *Then u is a solution to the partial differential equation* (2.10).

Proof:
Exercise. □

Both these theorems show, that the characteristics of (2.10) are precisely the curves, on which the values $u(\gamma(t))$ can be calculated from the *ordinary* differential equation (2.12). So along these characteristics, the partial differential equation becomes an ordinary differential equation! Since no interaction with values from outside the characteristic takes place, one also says, that the solution's value is being *transported* along the characteristics.

Note:
The meaning of these insights for our above examinations is seen best in the example of the wave equation $u_{xx} - u_{yy} = 0$: Since the two-dimensional system (2.9) can be decoupled, the system of first order belonging to the wave equation can be decomposed into two partial differential equations of the form (2.10). Their characteristics $\gamma^{(\pm)}$ are given as solutions to

$$(\gamma_1')^{(\pm)} = 1, \quad (\gamma_2')^{(\pm)} = \pm 1,$$

are thus two straight lines $\gamma^{(\pm)}(t) = \gamma^{(\pm)}(0) + (t, \pm t)$ and indeed identical to the characteristics that we determined when we examined the wave equation as an equation of second order. Because of $b = d = 0$, (2.12) here simply reads $\dot{U} = 0$, i.e. the initial conditions are being transported variation-less along the characteristic. As a solution of the so called *Cauchy initial value problem*

$$\begin{aligned} u_{xx} - u_{yy} &= 0 && (x, y) \in \mathrm{I\!R} \times \mathrm{I\!R}_+ \\ u(x, 0) &= u_0(x) && x \in \mathrm{I\!R} \\ u_y(x, 0) &= u_1(x) && x \in \mathrm{I\!R} \end{aligned}$$

one obtains in this way

$$\begin{aligned} u_x(x, y) &= \frac{1}{2}\Big(- u_1(x - y) + u_1(x + y) + u_0'(x - y) + u_0'(x + y)\Big) \\ u_y(x, y) &= \frac{1}{2}\Big(u_1(x - y) + u_1(x + y) - u_0'(x - y) + u_0'(x + y)\Big) \end{aligned}$$

where we used $u_x(x, 0) = u_0'(x)$. This yields the so called *d'Alembert solution* u of the Cauchy initial value problem. What does the solution look like? (exercise) ◁

2.3 Well-Posed or Ill-Posed Problems

2.3.1 Initial Value Problems

As an example, we consider the Cauchy problem (2.2) for the **Laplace equation**

$$\boxed{\begin{array}{rcll} \Delta u & = & 0 & \text{in } \Omega \\ u & = & g_0 & \text{on } \gamma \\ \dfrac{\partial}{\partial n} u & = & g_1 & \text{on } \gamma \end{array}}$$

on the domain Ω

$$\Omega = \mathbb{R} \times \mathbb{R}_+, \quad \mathbb{R}_+ = \{y \in \mathbb{R} \mid y > 0\},$$

with $\gamma = \partial\Omega = \{(x, y) \in \mathbb{R}^2 : \ y = 0\}$. Since the second variable y in most cases can be interpreted as time, one calls in this case the corresponding Cauchy condition

$$u(x, 0) = g_0, \qquad u_y(x, 0) = g_1 \qquad\qquad \forall x \in \mathbb{R}$$

for $y = 0$ *initial value problem* and the whole problem *Cauchy initial value problem.* Suppose some $j \in \mathbb{N}$ fixed. In the case

$$g_0(x) = \cos(j\pi x), \qquad g_1(x) = j\pi \cos(j\pi x) \quad , \tag{2.13}$$

the solution of the resulting initial value problem for the Laplace equation can easily be determined using the *separation of variables* ansatz

$$u_j(x, y) = w(y) \cos(j\pi x) \quad ,$$

because it follows that

$$u_j(x, y) = e^{j\pi y} \cos(j\pi x) \quad .$$

The solution

$$u^n(x, y) = \sum_{j=1}^{n} \frac{1}{(j\pi)^4} e^{j\pi y} \cos(j\pi x)$$

to the initial conditions

$$g_0^n(x) = \sum_{j=1}^{n} \frac{1}{(j\pi)^4} \cos(j\pi x), \qquad g_1^n(x) = \sum_{j=1}^{n} \frac{1}{(j\pi)^3} \cos(j\pi x) \quad ,$$

one obtains by superposition. Apparently, the function sequences g_0^n, g_1^n converge uniformly to the Lipschitz-continuous functions

$$\boxed{g_0(x) = \sum_{j=1}^{\infty} \frac{1}{(j\pi)^4} \cos(j\pi x), \qquad g_1(x) = \sum_{j=1}^{\infty} \frac{1}{(j\pi)^3} \cos(j\pi x).} \tag{2.14}$$

For any $y > 0$ however,

$$\boxed{u^n(0, y) = \sum_{j=1}^{n} \frac{1}{(j\pi)^4} e^{j\pi y} \to \infty.}$$

Thus, arbitrarily small changes in the initial condition have an (infinitely) large effect on the associated solution. This motivates the following definition.

Definition 2.10 *The Cauchy problem* (2.2) *is called* <u>*well-posed*</u> *(in X and Y wrt. the norms $\|\cdot\|$) if the following conditions hold:*

- *existence and uniqueness:* $\forall g_0,\, g_1 \in X$ *there exists a uniquely determined solution* $u \in Y$
- *continuous dependency on the data: there are* $c_0,\, c_1 \in \mathrm{I\!R}$*, such that*

$$\|u - \tilde{u}\| \le c_0\|g_0 - \tilde{g}_0\| + c_1\|g_1 - \tilde{g}_1\|.$$

Otherwise the problem is called <u>*ill-posed*</u> *(in X and Y wrt. the norms $\|\cdot\|$).*

Through our counterexample we have shown, that the Cauchy initial value problem for the Laplace equation with respect to the maximum norm $\|\cdot\|_\infty$ is ill-posed. Nevertheless, elliptic problems on unbounded domains do occur in practice (e.g. the potential of a point charge). In this case, one has to define appropriate fading conditions in order to obtain a well-posed problem.

Wave equation. Analogue to the above method (separation of variables, superposition, going to the limit), one obtains as a solution to the initial value problem for the (hyperbolic!) wave equation $u_{xx} - u_{yy} = 0$ for the initial conditions (2.14) the bounded in-limit-solution

$$u^n(x,y) \to u(x,y) = \sum_{j=1}^{\infty} \frac{1}{(j\pi)^4}(\sin(j\pi y) + \cos(j\pi y))\cos(j\pi x) \quad .$$

This problem is apparently **well-posed** wrt. the maximum norm.

Heat equation. In the case of the (parabolic!) heat equation $u_y = u_{xx}$ the solution is *over-determined* by the two initial conditions (2.13), because at least for smooth solutions

$$g_1(x) = u_y(x,0) = \lim_{y\to 0} u_y(x,y) = \lim_{y\to 0} u_{xx}(x,y) = g_0''(x) \quad .$$

This difficulty is not surprising, because $\gamma = \{(x,0) \mid x \in \mathrm{I\!R}\}$ is exactly the heat equation's characteristic! So, although we are dealing with a differential equation of second order, we can demand only one initial condition for the wave equation's Cauchy initial value problem, that is

$$u(x,0) = g_0(x) \qquad \forall x \in \mathrm{I\!R} \quad .$$

This circumstance can also be interpreted physically ("Infinite propagation velocity").

Note:
In our example

$$g_0(x) = \sum_{j=1}^{\infty} \frac{1}{(j\pi)^4}\cos(j\pi x)$$

by using separation of variables, superposition and going to the limit, one gets a solution

$$u(x,y) = \sum_{j=1}^{\infty} \frac{1}{(j\pi)^4} e^{-(j\pi)^2 y}\cos(j\pi x) \quad .$$

What difficulties arise if, instead of $u_y - u_{xx} = 0$, one considers the backward heat equation $u_y + u_{xx} = 0$? How can one interpret these difficulties physically? ◁

2.3.2 Boundary and Initial Value Problems

Boundary value problems. For partial differential equations with a bounded domain Ω one usually can only fix one condition at the boundary. We already got to know Dirichlet, Neumann and Cauchy boundary conditions in chapter 1. In this case, one speaks of a *boundary value problem.*

With elliptic boundary value problems we will deal in detail in the next chapter. Through counterexamples, one can show that boundary value problems for hyperbolic and parabolic differential equations are generally ill-posed (exercise).

Initial value problems. If the given domain is unbounded in one coordinate direction (usually in time direction) and bounded in all other directions, then additional to the initial conditions there are some boundary conditions necessary. For example, we can consider the heat equation on

$$\Omega = (0,1) \times \mathbb{R}_+$$

with the initial condition

$$u(x,0) = u_0(x), \qquad \forall x \in \mathbb{R}$$

and the Dirichlet boundary conditions

$$u(0,y) = g_0(y), \;\; u(1,y) = g_1(y) \qquad \forall y \in \mathbb{R}_+$$

(physical interpretation?) This is a typical *initial value problem.* Initial value problems for elliptic differential equations are generally ill-posed, too (counterexample?).

3 Classical Solutions to Elliptic Problems

3.1 Representation Formulas and Green's Function

As a prototype of elliptic differential equations we investigate *Poisson's equation*

$$-\Delta u = f \quad \text{in } \Omega \quad . \tag{3.1}$$

Here we assume $f \in C(\Omega)$.

$$\Delta u = \sum_{j=1}^{n} \frac{\partial^2 u}{\partial x_j^2}$$

is called the Laplace Operator or *Laplacian.* In the case $f \equiv 0$, (3.1) is called *Laplace's equation.* Every function u, that solves Laplace's equation, is called *harmonic.* We consider (3.1) on a bounded domain $\Omega \subset \mathbb{R}^n$ (i.e. open and connected) with smooth boundary $\partial\Omega$ (Green domain). On $\partial\Omega$ we assess Dirichlet boundary conditions, i.e.

$$u(x) = g(x) \qquad \text{for} \quad x \in \partial\Omega \ , \tag{3.2}$$

and suppose $g \in C(\partial\Omega)$.

Definition 3.1 *If $u \in C^2(\Omega) \cap C(\overline{\Omega})$ solves Poisson's equation* (3.1) *in Ω and fulfills the Dirichlet boundary conditions* (3.2) *on $\partial\Omega$, we call u classical solution to the boundary problem* (3.1)*,* (3.2)*.*

The solution space $C^2(\Omega)\cap C(\overline{\Omega})$ secures the existence of the necessary derivatives of u in the interior of Ω and the continuous taking of the boundary conditions. The following example shows, that it is useless to restrict the solution space any further.

Example:
Suppose $\Omega = \{(r,\varphi) \,|\, r \in (0,1) \ , \ \varphi \in (0, \frac{3}{2}\pi)\}$ and

$$\begin{array}{rcll} f(r,\varphi) & \equiv & 0 & (r,\varphi) \in \Omega \\ g(r,\varphi) & = & \sin\left(\frac{2}{3}\varphi\right) & (r,\varphi) \in \partial\Omega \quad . \end{array}$$

Observe that $g \in C(\partial\Omega)$! We transform the Laplacian into polar coordinates

$$\begin{array}{rcl} x & = & r\cos\varphi \\ y & = & r\sin\varphi \end{array}$$

and together with

$$u(r,\varphi) = r^{\frac{2}{3}} \sin(\tfrac{2}{3}\varphi)$$

we gain a solution to the corresponding boundary value problem. Now we will show, that $u \notin C^1(\overline{\Omega})$.

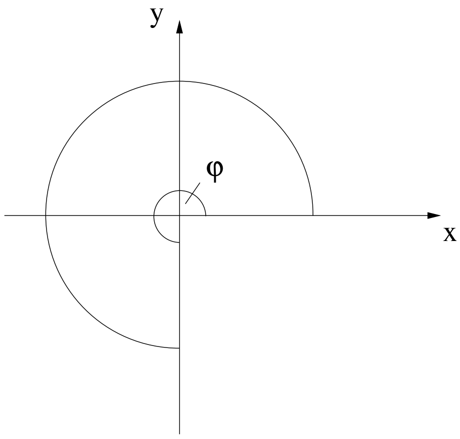


Figure 3.1: reentrant angle

Along with u_x and u_y, u_r should be bounded as well, because

$$u_r = u_x x_r + u_y y_r = u_x \cos\varphi + u_y \sin\varphi \quad ,$$

but

$$u_r = \tfrac{2}{3} r^{-\frac{1}{3}} \sin(\tfrac{2}{3}\varphi) \to \infty \quad \text{for} \quad r \to 0 \quad .$$

The reason for these problems is the *reentrant angle* in the domain Ω. ◁

We now want to investigate, whether a classical solution u is determined uniquely through g and f and whether it depends continuously on these data. To this end we will first derive an *integral representation* of (smooth enough) solutions. Fundamental to this are (once again) suitable product rules.

Lemma 3.2 *Suppose $u, v \in C^2(\overline{\Omega})$. Then Green's first identity*

$$\boxed{\int\limits_{\Omega} u\Delta v \, dx = -\int\limits_{\Omega} \nabla u \cdot \nabla v \, dx + \int\limits_{\partial\Omega} u \frac{\partial v}{\partial n} \, d\sigma} \tag{3.3}$$

and Green's second identity

$$\boxed{\int\limits_{\Omega} (u\Delta v - v\Delta u) \, dx = \int\limits_{\partial\Omega} (u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n}) \, d\sigma} \tag{3.4}$$

hold.

Proof:
(3.3) follows, by choosing $V = u\nabla v$, directly from the divergence theorem. (3.4) follows from interchanging u and v in (3.3) and subtracting. □

Conclusion 3.3 *The Dirichlet problem for Poisson's equation has* at most *one solution* $u \in C^2(\overline{\Omega})$ *(exercise). In particular,* u *is overdetermined by fixing* u and $\frac{\partial u}{\partial n}$ *at the boundary.*

Conclusion 3.4 *In the case of Neumann boundary conditions*

$$\frac{\partial}{\partial n} u = h \quad \text{on } \partial\Omega \quad ,$$

the compatibility condition

$$\int\limits_{\Omega} f \, dx = - \int\limits_{\partial\Omega} h \, d\sigma$$

is necessary *for the existence of a solution.*[1]

One can check that the Laplacian is invariant under rotation (physical interpretation?). This motivates the examination of rotationally invariant solution, i.e. solutions of the form

$$\boxed{u(x) = s(r) \qquad r = |x-a| \ , \quad a \in \mathbb{R}^n .}$$

Lemma 3.5 *Rotationally invariant, in* $\mathbb{R}^n \setminus \{a\}$ *harmonic functions have the form*

$$s(r) = \begin{cases} c \log r + C & n = 2 \\ c\, r^{2-n} + C & n > 2 \end{cases} \tag{3.5}$$

with arbitrary constants $c, C \in \mathbb{R}$.

Proof:
By transforming the Laplacian into polar coordinates. (Exercise). □

We now choose special values for the constants c, C.

Definition 3.6 *Suppose* $a \in \mathbb{R}^n$. *The in* $\mathbb{R}^n \setminus \{a\}$ *harmonic function*

$$\boxed{s(x,a) = \begin{cases} -\frac{1}{2\pi} \log |x-a| & n = 2 \\ -\frac{1}{n(2-n)\omega_n} \, |x-a|^{2-n} & n > 2 \end{cases}}$$

is called singularity function*. Here* $\omega_n = \int_{|x|=1} d\sigma$ *shall denote the surface area of the unit sphere in* $\mathbb{R}^n$.

By means of the singularity function, one defines the fundamental solutions:

Definition 3.7 *Suppose* $\Phi(\cdot, a) \in C^2(\overline{\Omega})$ *harmonic for all* $a \in \Omega$. *Then*

$$\gamma(x,a) := s(x,a) + \Phi(x,a)$$

is called fundamental solution *of Laplace's equation in* Ω.

[1] Directly by applying $\Delta v = -f$ and $\frac{\partial v}{\partial n} = g$ with $u = 1$ to (3.3).

Now we can formulate the representation theorem.

Theorem 3.8 (Representation Theorem) *Suppose $u \in C^2(\overline{\Omega})$ and γ being a fundamental solution. Then for every point $a \in \Omega$,*

$$\boxed{u(a) = \int\limits_{\partial\Omega} \left(\gamma(x,a)\,\frac{\partial}{\partial n}u(x) - u(x)\,\frac{\partial}{\partial n}\gamma(x,a)\right)\,d\sigma - \int\limits_{\Omega} \gamma(x,a)\Delta u(x)\,dx \quad .} \tag{3.6}$$

Proof:
One can find the proof for example in Jost [14, p. 11]. There, the following property of the singularity function is used: Suppose $K_\varrho(a) := \{x \in \mathbb{R}^n \mid |x-a| < \varrho\} \subset \Omega$ and $v \in C(\Omega)$ arbitrary. Then:

$$\lim_{\varrho\to 0} \int\limits_{K_\varrho(a)} v(x)\,s(x,a)\,dx = 0 \quad \text{and} \quad \lim_{\varrho\to 0} \int\limits_{\partial K_\varrho(a)} v(x)\,s(x,a)\,d\sigma = 0,$$

as well as

$$-\lim_{\varrho\to 0} \int\limits_{\partial K_\varrho(a)} v(x)\,\frac{\partial}{\partial n_K}s(x,a)\,d\sigma = v(a),$$

where n_K denotes the outward normal on $K_\varrho(a)$. These equations then imply the claim. □

Observe, that the formula (3.6) represents the value of the solution $u(a)$ through the right-hand side $f = -\Delta u$ and the Cauchy data u and $\frac{\partial u}{\partial n}$ on $\partial\Omega$. We have already seen, that in every $x \in \partial\Omega$, we can either prescribe only u or only $\frac{\partial u}{\partial n}$. To get to a representation of $u(a)$ depending only on known data, we now choose the formerly unknown function Φ properly. As in our case, with given Dirichlet boundary conditions, we want to **choose** Φ in a way, that for all $a \in \Omega$ the following condition

$$\boxed{\gamma(x,a) = 0 \quad \forall x \in \partial\Omega} \tag{3.7}$$

or equivalently

$$\Phi(x,a) = -s(x,a) \quad \forall x \in \partial\Omega$$

holds.

Definition 3.9 *A fundamental solution $G(x,a) = \gamma(x,a)$ with the property $G(x,a) = 0$ for all $x \in \partial\Omega$ is called Green's function of first kind.*

The representation formula (3.6) of a solution u for the Dirichlet problem the transforms into

$$\boxed{u(a) = \int\limits_{\Omega} G(x,a)\,f(x)\,dx - \int\limits_{\partial\Omega} \frac{\partial}{\partial n}G(x,a)\,g(x)\,d\sigma \quad .} \tag{3.8}$$

Example:
Suppose $g = 0$ and $f_n \in C(\Omega)$, $n \in \mathbb{N}$ a sequence of functions with the property

$$\int\limits_{\Omega} f_n(x)\,dx = 1, \qquad \operatorname{supp} f_n \subset K_{\frac{1}{n}}(x_0) \quad .$$

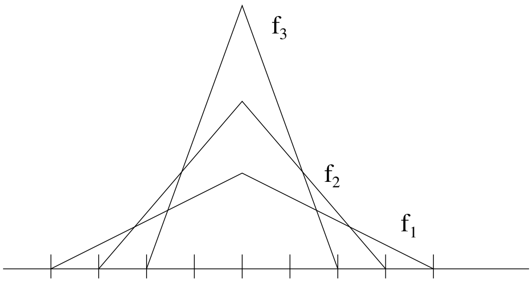


Figure 3.2: example for a sequence of functions f_n

The set

$$\operatorname{supp} f_n = \overline{\{x \in \Omega \,|\, f_n(x) \neq 0\}}$$

is called support of f_n. $x_0 \in \Omega$ is chosen arbitrarily, but fixed for all n. We consider the corresponding solution u_n to exist and to lie in $C^2(\overline{\Omega})$. Then for all $a \in \Omega$, $a \neq x_0$

$$u_n(a) \to u(a) = G(x_0, a) \quad \text{for} \quad n \to \infty \quad .$$

With this, one can imagine $u(x) = G(x_0, x)$ as a solution of

$$\boxed{-\Delta G(x_0, x) = \delta_{x_0}(x)}$$

with a right-hand side

$$\delta_{x_0} = \lim_{n\to\infty} f_n \tag{3.9}$$

with the following property

$$\int\limits_{\Omega} \delta_{x_0}(x)\, dx = 1\,, \qquad \delta_{x_0}(x) = 0 \quad \text{for} \quad x \neq x_0 \quad .$$

Clearly, such a function cannot exist (although the physicist Dirac used it successful and it is nowadays an indispensable tool in physics). This is a *generalized function*, called Dirac's δ-distribution. We will talk about distributions again later on. Then it will become clear, how to interpret the limit in (3.9).

For the time being, we can imagine δ_{x_0} as a point source. The decay the corresponding Green's function $G(x, x_0)$ especially describes the effect of small perturbations on the right-hand side f in x_0. The decay of $G(x, x_0)$ is dominated by the singularity function and therefore behaves like

$$\begin{aligned} &\mathcal{O}\left(-\log|x - x_0|\right) \quad \text{for} \quad n = 2 \\ &\mathcal{O}\left(|x - x_0|^{2-n}\right) \quad \text{for} \quad n > 2 \quad . \end{aligned}$$

$G(x_0, x) > 0$ holds (exercise). With this, though perturbations of the right-hand side f (at a point $x_0 \in \Omega$) affect the solution u globally (in every $x \in \Omega$), the effect will decay with growing distance to x_0 very rapidly. Local perturbation thereby have an (almost) local effect. ◁

In general, the existence of a Green's function is not sure. The following example illustrates the difficulties.

Theorem 3.10 *Suppose for all $a \in \Omega$ the Dirichlet problem*

$$\begin{aligned} \Delta\Phi(\cdot, a) &= 0 && \text{in } \Omega \\ \Phi(\cdot, a) &= -s(\cdot, a) && \text{on } \partial\Omega \end{aligned}$$

to be solvable and $\Phi(\cdot, a) \in C^2(\overline{\Omega})$. Then there is a Green's function of the first kind to the corresponding Dirichlet problem.

We have already seen, that especially in domains with a reentrant angle, the regularity condition $\Phi(\cdot, a) \in C^2(\overline{\Omega})$ can become problematic. For certain domains however, one can write down Green's function explicitly. As you might expect, the most simple case is a ball, cf. Jost [14, p. 14].

3.2 Existence

Up until low, we always presumed the existence of a solution $u \in C^2(\overline{\Omega})$ and derived its representation through the data by applying Green's function. Now one can hope that *conversely* by formula (3.8), there is always a solution to the Dirichlet problem for Poisson's equation *defined*. In general however, this is *wrong*. There are particularly right-hand sides $f \in C(\overline{\Omega})$, such that (3.8) is no solution to the Dirichlet problem. One needs in general more regularity.

Theorem 3.11 *Suppose $g = 0$ and f Hölder continuous with exponent $\lambda \in (0, 1)$, that is*

$$|f(x) - f(y)| \leq c\, |x - y|^\lambda \qquad \forall x, y \in \Omega$$

with a constant $c \in \mathbb{R}$ independent of x, y. In addition, suppose there is an associated Green's function of first kind $G(\xi, x)$. Then

$$\boxed{u(x) = \int\limits_\Omega G(\xi, x) f(\xi)\, d\xi}$$

is a solution to the Dirichlet problem.

Proof:
See for example Hackbusch [12] p. 38. □

Note:
Bear in mind the regularity conditions on f and above all on Ω (existence of a Green's function $G(\xi, x)$!). ◁

To guarantee the existence of a solution to the Dirichlet problem (3.1), (3.2), we just need an existence result for the homogeneous case $f = 0$. Here we will confine ourselves to a ball.

Theorem 3.12 (Poisson's Integral Formula) *Suppose* $\Omega = \{x \in \mathbb{R}^n \mid |x| < \varrho\}$, $f = 0$ *and* $g \in C(\partial\Omega)$. *Then*

$$u(x) = \frac{\varrho^2 - |x|^2}{\varrho\omega_n} \int\limits_{\partial\Omega} \frac{g(\xi)}{|\xi - x|^n}\, d\sigma \quad \text{for} \quad x \in \Omega \tag{3.10}$$

is a classical solution to the Dirichlet problem (3.1), (3.2). *In addition,* $u \in C^\infty(\Omega) \cap C(\overline{\Omega})$ *holds.*

Proof:
Cf. John [13] or Hackbusch [12]. The proof relies on a special form of Green's function for the ball and applying it to (3.8). □

More general existence results are to be found in John [13]. There, the regularity conditions at the boundary $\partial\Omega$ play a major role.

3.3 Uniqueness and Continuous Dependence on the Data

An important consequence from the representation formula (3.8) is the *Mean Value Property* (MVP).

Theorem 3.13 *Suppose* u *harmonic in* Ω. *If* $x \in \Omega$ *and* $\varrho > 0$ *small enough, such that* $\overline{K_\varrho(x)} \subset \Omega$, *then*

$$u(x) = \frac{1}{\omega_n \varrho^{n-1}} \int\limits_{\partial K_\varrho(x)} u(\xi)\, d\sigma \quad . \tag{3.11}$$

Proof:
For $x = 0$ the representation formula (3.8) has the form (3.10). Applying $u = g$ on $\partial\Omega$ and $|\xi| = \varrho$ yields (3.11). The case $x \neq 0$, can be reduced to $x = 0$ by a translation $\tilde{\Omega} := \Omega - x$. □

Theorem 3.14 (Strong Maximum Principle) *If* u *is harmonic in* $\Omega \subset \mathbb{R}^n$ *and has a maximum or minimum in* $x \in \Omega$, *then* u *is constant in* Ω.

Proof:
Suppose $M := \sup\{u(x) \mid x \in \Omega\} < \infty$. We divide Ω into

$$\Omega_1 = \{x \in \Omega \mid u(x) = M\} \cup \Omega_2 = \{x \in \Omega \mid u(x) < M\} \quad .$$

Since u is continuous, Ω_2 is open. We will show, that Ω_1 is open as well. Suppose $x \in \Omega_1$, that is $u(x) = M$. From the MVP it follows, that for small enough $\varrho > 0$

$$\begin{aligned} 0 &= \frac{1}{\omega_n \varrho^{n-1}} \int\limits_{\partial K_\varrho(x)} u(\xi)\, d\sigma - u(x) \\ &= \frac{1}{\omega_n \varrho^{n-1}} \int\limits_{\partial K_\varrho(x)} (u(\xi) - M)\, d\sigma \quad . \end{aligned}$$

Since $u(\xi) - M$ is continuous in ξ and $u(\xi) - M \leq 0$, it follows $u(\xi) = M \ \forall \xi \in \partial K_\varrho(x)$. By contraposition, we see that $K_\varrho(x) \subset \Omega_1$. Since Ω is (topologically) connected, we have $\Omega_1 = \emptyset$ or $\Omega_2 = \emptyset$. This implies the claim. The minimum property follows by applying the maximum property to the harmonic function $-u$. □

Note:
From the strong maximum principle follows

$$G(\xi, x) > 0 \qquad \forall \xi, x \in \Omega, \xi \neq x \quad .$$

Conclusion 3.15 (Weak Maximum Principle) *for every harmonic function* $u \in C(\overline{\Omega}) \cap C^2(\Omega)$

$$\min_{\xi \in \partial\Omega} u(\xi) \leq u(x) \leq \max_{\xi \in \partial\Omega} u(\xi) \quad \forall x \in \Omega.$$

Theorem 3.16 (Uniqueness) *The Dirichlet problem* (3.1)*,* (3.2) *has at most one solution* $u \in C^2(\Omega) \cap C(\overline{\Omega})$*.*

Proof:
Suppose $u_1, u_2 \in C(\overline{\Omega}) \cap C^2(\Omega)$ two solutions. Then $u = u_1 - u_2$ solves the homogeneous problem ($f = 0, g = 0$). If $u(x) \neq 0$ for a $x \in \Omega$, then there would be a (positive) maximum or a (negative) minimum in Ω. Because of thm. 3.14 this is not possible. □

One last important consequence from the weak maximum principle is the following theorem.

Theorem 3.17 (Continuous dependency on the boundary data) *Consider* u_1, u_2 *to be two classical solutions to the boundary value problems*

$$-\Delta u_i = f \text{ in } \Omega \;, \quad u_i = g_i \text{ on } \partial\Omega \;, i = 1, 2 \quad .$$

Then

$$\max_{x \in \Omega} |u_1(x) - u_2(x)| \leq \max_{y \in \partial\Omega} |g_1(y) - g_2(y)| \quad .$$

Proof:
From $g_1 = g_2$, from the uniqueness theorem 3.16 follows $u_1 = u_2$. So consider $g_1 \neq g_2$.
Obviously, $u = u_1 - u_2$ is harmonic and fulfills the boundary conditions $g = g_1 - g_2$. If

$$u(x) > \max_{y \in \partial\Omega} |g(y)| \geq \max_{y \in \partial\Omega} g(y) \quad ,$$

u would have a maximum in Ω without being constant. Similarly, one excludes the case

$$-u(x) > \max_{y \in \partial\Omega} |g(y)| \geq -\min_{y \in \partial\Omega} g(y) \quad .$$

4 Difference Methods

As in the previous chapter, we consider the Dirichlet problem

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega \\ u &= g \qquad \text{on } \partial\Omega \end{aligned} \tag{4.1}$$

with $f \in C(\Omega)$, $g \in C(\partial\Omega)$ and a smoothly bounded domain Ω (e.g. parametrization of $\partial\Omega$ continuously differentiable). We assume the existence of a classical solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$.

4.1 Grids and Difference Quotients

We want to develop a difference method, that approximatively solves (4.1). To this end, we first cover the domain Ω with a *equidistant grid* to the *step-size* $h > 0$,

$$\Omega_h = \{x \in \Omega \,|\, x = (ih, jh),\, i, j \in \mathbb{Z}\} \quad .$$

As *boundary discretization* we use

$$\partial\Omega_h = \{x \in \partial\Omega \,|\, x = (x_1, x_2)\ ,\ x_1 = ih \text{ or } x_2 = ih,\ \ i \in \mathbb{Z}\} \quad .$$

There might also be other discretizations possible. The above alternative will later lead to the so called *Shortley–Weller method*.

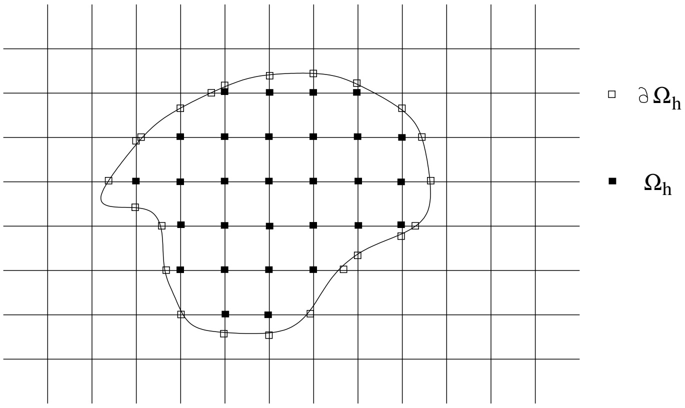


Figure 4.1: discretization of Ω

Observe the disturbance of the equidistant grid at the boundary. An important subset of the equidistant grid Ω_h is the set of *inner points*, denoted Ω_h°. This is defined as follows:

$$\Omega_h^\circ = \{x \in \Omega_h \mid Nb(x) \subset \Omega_h\}$$

There, we used the *set of neighbours* $Nb(x)$:

$$Nb(x) = \{y \in \Omega_h \cup \partial\Omega_h \,:\, |x_1 - y_1| + |x_2 - y_2| \leq h\}\ .$$

Neighbouring points are being labeled in accordance to figure 4.2. Similar to the continuous case, we finally set

$$\overline{\Omega}_h = \Omega_h \cup \partial\Omega_h \quad .$$

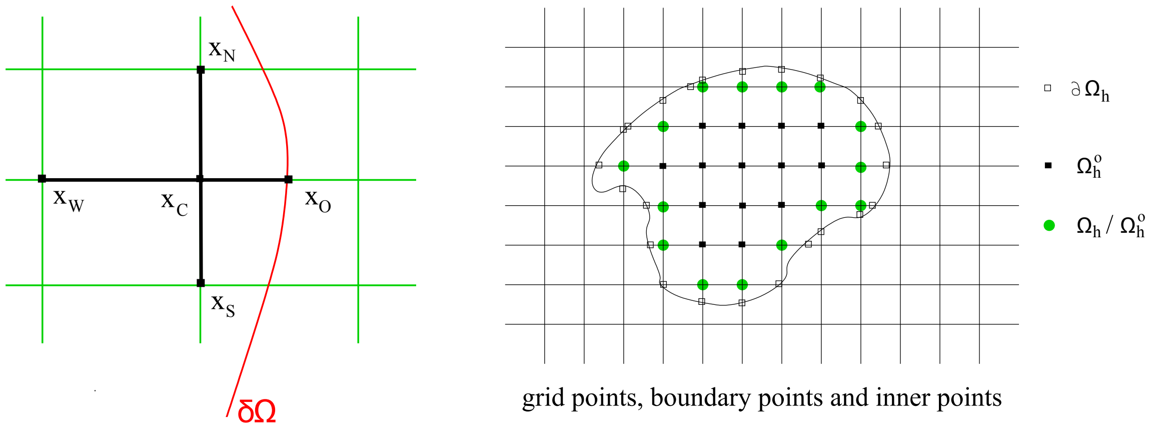


Figure 4.2: Left – grid points for approximation of the Laplacian. We see the labelling of the neighbours $Nb(x_C) = \{x_W, x_N, x_O, x_S\}$ of the central point $x_C \in \Omega_h$. In the present case apparently $x_C \notin \Omega_h^\circ$.
Right – illustration of inner points.

Definition 4.1 *$\overline{\Omega}_h$ is called discretely connected if every pair of grid points $x, y \in \overline{\Omega}_h$ can be connected by a sequence of neighbouring points.*

Lemma 4.2 *Let Ω be connected. Then there is a $h_0 > 0$, such that for all $h < h_0$, $\overline{\Omega}_h$ is discretely connected.*

Proof:
Exercise. □

From now on, $\overline{\Omega}_h$ shall be discretely connected, that is h chosen small enough. On $\overline{\Omega}_h$ we now search a *grid function*

$$U : \overline{\Omega}_h \to \mathbb{R} \quad ,$$

that approximates the continuous solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$ of (4.1) as well as possible. In order to obtain a calculation rule for U we have to approximate the differential equation on the one side, and the boundary condition on the other side.
To approximate the derivatives appearing in the Laplacian we want to use difference quotients (with the notation from fig. 4.2):

$$\boxed{\begin{array}{rcll} D_1^+U(x_C) & = & \dfrac{U(x_O)-U(x_C)}{|x_O-x_C|} & \text{(forward)} \\[2ex] D_1^-U(x_C) & = & \dfrac{U(x_C)-U(x_W)}{|x_C-x_W|} & \text{(backward)} \\[2ex] D_1U(x_C) & = & \frac{1}{2}\left(D_1^+U(x_C)+D_1^-U(x_C)\right) & \text{(central)} \quad . \end{array}}$$

Lemma 4.3 *Let $u \in C^2(\overline{\Omega})$ and $x \in \Omega_h$. Then for $D = D_1^+, D_1^-, D_1$,*

$$u_{x_1}(x) - Du(x) = \mathcal{O}\left(h\right) \tag{4.2}$$

holds. If even $u \in C^3(\overline{\Omega})$ and $x \in \Omega_h^\circ$, then for D_1 follows

$$\boxed{u_{x_1}(x) - D_1u(x) = \mathcal{O}\left(h^2\right) \ .} \tag{4.3}$$

Proof:
We will only show (4.3). Taylor expansion for $x = x_{ij} = (ih, jh) \in \Omega_h^\circ$ provides

$$u(x_{i\pm1,j}) = u(x_{ij}) \pm hu_{x_1}(x_{ij}) + \frac{h^2}{2}u_{x_1x_1}(x_{ij}) \pm \frac{h^3}{6}u_{x_1x_1x_1}(x_{ij} \pm \omega_\pm h)$$

with $\omega_\pm \in (0,1)$. Insertion yields

$$\begin{aligned} D_1u(x_{ij}) &= \frac{1}{2h}\left(u(x_{i+1,j}) - u(x_{i-1,j})\right) \\ &= u_{x_1}(x_{ij}) + \mathcal{O}\left(h^2\right) \quad . \end{aligned}$$

Central difference quotients thus have (under suitable smoothness requirements!) a higher approximation quality. To approximate the second order derivatives we thus use the *central divided differences of second order*

$$\boxed{\begin{aligned} D_{11}U(x_C) &= \frac{2}{|x_O - x_W|}\left(D_1^+U(x_C) - D_1^-U(x_C)\right) \\ D_{22}U(x_C) &= \frac{2}{|x_N - x_S|}\left(D_2^+U(x_C) - D_2^-U(x_C)\right) \quad . \end{aligned}}$$

As discretization Δ_h of the Laplacian Δ we finally choose

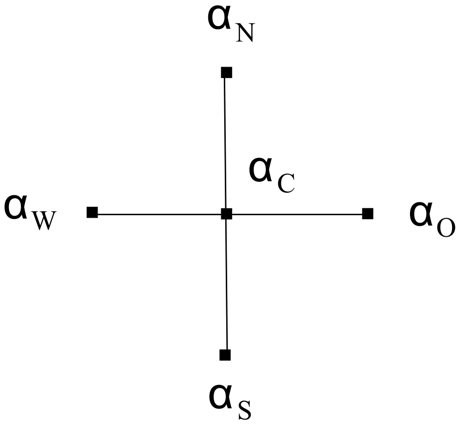
$$\boxed{\Delta_h = D_{11} + D_{22} \ .}$$

Expanded this becomes

$$\Delta_hU(x_C) = \alpha_C\,U(x_C) + \alpha_W\,U(x_W) + \alpha_O\,U(x_O) + \alpha_N\,U(x_N) + \alpha_S\,U(x_S) \quad .$$

with the corresponding weights α_C, α_W,α_O, α_S and α_N. These can also be interpreted as values of a grid function

$$\alpha(x, x_D) = \alpha_D, \qquad D = C, W, O, N, S$$



in $x = x_C \in \Omega_h$. More graphic is the following *difference star.*
In the case $x_C \in \Omega_h^\circ$ we have the following *standard 5-point star*

$$\boxed{\alpha_C = -\frac{4}{h^2}\ , \quad \alpha_W = \alpha_O = \alpha_N = \alpha_S = \frac{1}{h^2}.}$$

In general though, the weights depend on $x_C \in \Omega_h \setminus \Omega_h^\circ$. For all $x_C \in \Omega_h$ however, the following relations hold

$$\alpha_C < 0\ , \quad \alpha_W, \alpha_O, \alpha_N, \alpha_S > 0\ , \quad \alpha_C + \alpha_W + \alpha_O + \alpha_N + \alpha_S = 0 \quad .$$

Difference method with these properties are called *of positive type.*
We shortly want to investigate the accuracy of our difference approximation.

Lemma 4.4 *Consider $u \in C^3(\overline{\Omega})$ and $x \in \Omega_h$. Then*

$$\Delta u(x) - \Delta_h u(x) = \mathcal{O}\left(h\right).$$

If even $u \in C^4(\overline{\Omega})$ and $x \in \Omega_h^\circ$, then follows

$$\boxed{\Delta u(x) - \Delta_h u(x) = \mathcal{O}\left(h^2\right)\ .}$$

Proof:
Taylor expansion. □

4.2 The Shortley–Weller Method

As difference approximation of (4.1) we finally obtain the *Shortley–Weller method*:

$$\boxed{\begin{aligned} -\Delta_h U(x) &= f_h(x) \quad x \in \Omega_h \\ U(x) &= g_h(x) \quad x \in \partial\Omega_h\ . \end{aligned}} \tag{4.4}$$

Here, f_h and g_h are appropriate approximations of the data f and g. We simply choose

$$f_h(x) = f(x) \text{ for } x \in \Omega_h \quad , \quad g_h(x) = g(x) \text{ for } x \in \partial\Omega_h \quad . \tag{4.5}$$

Using the boundary condition in (4.4) we can eliminate the unknown $U(x)$, $x \in \partial\Omega_h$. In order to determine the remaining values we have to solve a linear equation system with $N = \#\Omega_h$ equations and unknowns. After the introduction of a convenient numeration of the grid points

$$\Omega_h = \{x_i \,|\, i = 1, \ldots, N\}$$

it becomes

$$\boxed{AU = F} \tag{4.6}$$

with a coefficient matrix A,

$$A = (a_{ij})_{i,j=1}^{N} \qquad a_{ij} = -\alpha(x_i, x_j) \quad ,$$

a vector of unknowns U,

$$U = (U_i)_{i=1}^{N} \quad ,$$

and a right-hand side F,

$$F = (F_i)_{i=1}^{N} \quad , \quad F_i = f_h(x_i) + \sum_{y \in \partial\Omega_h} \alpha(x_i, y) g_h(y) \quad .$$

Independent of the chosen numeration, in every row of A there are at most 5 non-zero entries. So A is *sparse.* In the case $\partial\Omega_h \subset \{x = (ih, jh) \,|\, i, j \in \mathbb{Z}\}$, A is *symmetric*, i.e. $a_{ij} = a_{ji}$. Unfortunately, in general this is wrong!

Definition 4.5 *The* graph *of a matrix $A = (a_{ij})_{i,j=1}^{N}$ has N knots and there is an oriented edge between the knots i and j if $a_{ij} \neq 0$. The matrix A is called irreducible if the graph of A is connected, i.e. if a pair of knots i, j can be connected by a sequence of knots.*

Note:
If Ω_h is discretely connected, then A is irreducible. $\lhd$

There are other properties of our discretization (4.4) that also can be translated into properties of the algebraic equation system (4.6) and vice-versa. We have two different points of view at our disposal. To stress the analogy with the continuous case, we prefer to talk about grid functions rather then vectors and will only once in a while come back to the algebraic point of view.

Definition 4.6 *A grid function $U : \overline{\Omega}_h \to \mathbb{R}$ with the property*

$$-\Delta_h U = 0 \qquad \text{on } \Omega_h$$

is called discretely harmonic. If only

$$-\Delta_h U \leq 0 \qquad \text{on } \Omega_h$$

holds, U is called discretely subharmonic.

Lemma 4.7 (Discrete mean value property) *If U is discretely subharmonic, U has the discrete mean value property*

$$U(x_C) \leq \frac{1}{\alpha_W + \alpha_O + \alpha_S + \alpha_N} \left(\alpha_W U(x_W) + \alpha_O U(x_O) + \alpha_S U(x_S) + \alpha_N U(x_N)\right) \tag{4.7}$$

for all $x_C \in \Omega_h$.

Proof:
This follows by

$$-\alpha_C = \alpha_W + \alpha_O + \alpha_S + \alpha_N > 0 \quad .$$

Since we assumed $\overline{\Omega}_h$ to be discretely connected, there is another analogy to the continuous case, following from the mean value property.

Theorem 4.8 (Discrete maximum principle) *Let U be discretely subharmonic. Then U reaches its maximal value on* $\partial\Omega_h$ *or is constant in* $\overline{\Omega}_h$ *.*

Proof:
Suppose $x_C \in \Omega_h$ and

$$U(x_C) = M = \max_{x \in \overline{\Omega}_h} U(x) \quad .$$

Then from the discrete mean value property follows

$$\begin{aligned} 0 \leq\ & \alpha_W \left(U(x_W) - U(x_C)\right) + \alpha_O \left(U(x_O) - U(x_C)\right) \\ & + \alpha_S \left(U(x_S) - U(x_C)\right) + \alpha_N \left(U(x_N) - U(x_C)\right) \quad . \end{aligned}$$

All summands are non-positive, since $U(x_C) = M$. It thus follows

$$U(x) = M, \quad \forall x \in Nb(x_C) \quad .$$

Since two grid points from $\overline{\Omega}_h$ can be connected by a sequence of finitely many neighbouring points, the assertion follows. □

We go on as in the continuous case.

Theorem 4.9 *The discrete boundary value problem* (4.4) *is uniquely solvable.*

Proof:
Given two solutions $U^{(1)}$, $U^{(2)}$ of (4.4), $U = U^{(1)} - U^{(2)}$ fulfills the equations

$$\begin{aligned} -\Delta_h U(x) &= 0 \qquad x \in \Omega_h \\ U(x) &= 0 \qquad x \in \partial\Omega_h \quad . \end{aligned}$$

In particular, U is discretely subharmonic. Setting $\max_{x\in\Omega_h} U(x) = M > 0$, from the discrete maximum principle follows $U(x) = M \ \forall x \in \overline{\Omega}_h$. This contradicts $U(x) = 0,\ x \in \partial\Omega_h$. So $U(x) \leq 0$ has to be true for all $x \in \Omega_h$. Assuming $\min_{x\in\Omega_h} U(x) = M < 0$ leads in the same way to a contradiction, since $-U$ is discretely subharmonic as well. We are dealing with a linear mapping from an *finite-dimensional space.* From injectivity the surjectivity follows. □

As a last consequence from the discrete maximum principle we deduce the *inverse monotony* of $-\Delta_h$.

Lemma 4.10 *The discrete Laplacian* $-\Delta_h$ *is inversly monotone, i.e. from the conditions*

$$\begin{aligned} -\Delta_h U(x) &\leq -\Delta_h V(x) \qquad x \in \Omega_h \\ U(x) &\leq V(x) \qquad\quad\ \ x \in \partial\Omega_h \end{aligned}$$

follows that

$$U(x) \leq V(x) \ , \quad x \in \overline{\Omega}_h \quad .$$

Proof:
Apparently, for $W = U - V$, we have

$$\begin{aligned} -\Delta_h W(x) &\le 0 \qquad x \in \Omega_h \\ W(x) &\le 0 \qquad x \in \partial\Omega_h \quad . \end{aligned}$$

If then, in contradiction to the assertion

$$\max_{x \in \Omega_h} W(x) = M > 0 \quad ,$$

following the discrete maximum principle $W(x) = M > 0$, $\forall x \in \overline{\Omega}_h$ would hold. This however contradicts $W(x) \le 0$, $x \in \partial\Omega_h$. □

We will need lemma 4.10 later on for the stability. For now there is another interesting property for the coefficient matrix A following.

Note:
Since (4.4) is solvable uniquely, A has to be regular. If one appoints in lemma 4.10 $V(x) = 0$, $\forall x \in \partial\Omega_h$, one in particular gets

$$AU = F \le 0 \quad \Rightarrow \quad U \le 0 \quad ,$$

where "$\le$" is meant component-wise. Choosing $F = -e_i$ (i-th unit vector), one obtains the solution

$$U = -(A^{-1})_i \le 0 \qquad \text{(i-th column } A^{-1}) \quad .$$

With this, we have shown that component-wise

$$A^{-1} \ge 0$$

holds. A more thorough analysis shows even $A^{-1} > 0$ (exercise). This is no mere coincidence, because A^{-1} plays the role of a discrete Green's function! The interpretation of the columns $(A^{-1})_i$ as grid functions corresponds to $G(x_i, \cdot)$, since

$$A(A^{-1})_i = e_i$$

is a discrete analogon to

$$-\Delta_x G(x_i, x) = \delta_{x_i}(x) \quad x \in \Omega \quad .$$

The Green's function $G(x_i, \cdot)$ is positive. Discretizations that preserve this property, are of particular importance. This motivates

Definition 4.11 (M-Matrix) *A matrix $A = (a_{ij})_{i,j=1}^N$ is called M-matrix if it fulfills*

$$\begin{gathered} a_{ii} > 0 \quad , \quad a_{ij} \le 0 \quad , \quad i, j = 1, \dots, N \\ A \textit{ is regular and } A^{-1} \ge 0 \quad . \end{gathered}$$

We have shown above that our discretization (4.6) yields an M-matrix. ◁

4.3 Consistency, Stability and Convergence

In preparation for the concluding convergence theorem we will first show a discrete analogon to the *a priori stability estimate*

$$\max_{x\in\overline{\Omega}} |u(x)| \leq \max_{x\in\partial\Omega} |u(x)| + c \sup_{x\in\Omega} |\Delta u(x)| .$$

Reminder: From this estimate follows that the continuous boundary value problem is represented properly.

Theorem 4.12 (Stability) *Let U be a grid function on $\overline{\Omega}_h$ and R the radius of a circle $K_R(0)$ with $\overline{\Omega}_h \subset K_R(0)$. Then the following estimate holds:*

$$\max_{x\in\overline{\Omega}_h} |U(x)| \leq \max_{x\in\partial\Omega_h} |U(x)| + \frac{R^2}{2} \max_{x\in\Omega_h} |\Delta_h U(x)| \quad .$$

Proof:
The proof is somewhat technical (tricky). We examine the effect the Laplacian Δ_h has on

$$W(x) = \frac{1}{2}x_1^2 \quad , \quad V(x) = 1 \quad , \quad x = (x_1, x_2) \in \overline{\Omega}_h \quad .$$

Inserting yields

$$\Delta_h W(x) = 1 \quad , \quad \Delta_h V(x) = 0 \quad , \quad x \in \Omega_h \quad .$$

We define

$$M = \max_{x\in\Omega_h} |\Delta_h U(x)| \quad , \quad N = \max_{x\in\partial\Omega_h} |U(x)| \quad .$$

With this, for both signs follows

$$\begin{aligned} -\Delta_h (\pm U + MW)(x) \leq 0 &= -\Delta_h \left((N + \frac{R^2}{2}M)V \right)(x) \ , \qquad && x \in \Omega_h \\ \pm U(x) + MW(x) &= (N + \frac{R^2}{2}M)V(x) \ , && x \in \partial\Omega_h \quad . \end{aligned}$$

The inverse monotonicity from lemma 4.10 provides

$$\pm U(x) + MW(x) \leq N + \frac{R^2}{2}MV(x) \quad , \quad x \in \overline{\Omega}_h \quad .$$

So, because $MW(x) \geq 0$,

$$\pm U(x) \leq N + \frac{R^2}{2}M \quad , \quad x \in \overline{\Omega}_h \quad .$$

This is exactly what we wanted to show. □

The above theorem ensures the stability of U against perturbations of $\Delta_h U$ and perturbations of the boundary data. We will exploit that now.

Definition 4.13 *Inserting the exact solution into the difference approximation one obtains the local truncation error*

$$\boxed{\begin{aligned} \tau_h(x) &= f_h(x) + \Delta_h u(x) && x \in \Omega_h \\ \tau_h(x) &= u(x) - g_h(x) && x \in \partial\Omega_h \ . \end{aligned}}$$

The difference method is called consistent if

$$\max_{x\in\overline{\Omega}_h} |\tau_h(x)| \to 0 \ , \quad h \to 0$$

and consistent of order p if

$$\max_{x\in\overline{\Omega}_h} |\tau_h(x)| = \mathcal{O}\left(h^p\right)$$

holds.

Theorem 4.14 (Consistency) *Suppose $u \in C^3(\overline{\Omega})$. Then we have*

$$\boxed{\max_{x\in\overline{\Omega}_h} |\tau_h(x)| = \mathcal{O}\left(h\right) \ .}$$

Proof:
Evidently $\tau_h(x) = 0$ for $x \in \partial\Omega_h$, and because of $f_h(x) = f(x)$, we have

$$\tau_h(x) = f_h(x) + \Delta_h u(x) = -\Delta u(x) + \Delta_h u(x) = \mathcal{O}\left(h\right) \qquad \text{for } x \in \Omega_h$$

by lemma 4.4. □

Note:
We even have $\tau_h(x) = \mathcal{O}\left(h^2\right)$ if $x \in \Omega_h^\circ$ (and u smooth enough). ◁

From consistency and stability now convergence follows.

Theorem 4.15 (Convergence) *Suppose $u \in C^3(\overline{\Omega})$. Then:*

$$\max_{x\in\overline{\Omega}_h} |U(x) - u(x)| = \mathcal{O}\left(h\right) \quad .$$

Proof:
We set $V(x) = U(x) - u(x)$ for $x \in \overline{\Omega}_h$. Then follows by definition of the truncation error

$$\begin{aligned} -\Delta_h V(x) &= f_h(x) + \Delta_h u(x) = \tau_h(x) && x \in \Omega_h \\ V(x) &= \tau_h(x) && x \in \partial\Omega_h \quad . \end{aligned}$$

From the stability we get

$$\begin{aligned} \max_{x\in\overline{\Omega}_h} |U(x) - u(x)| &= \max_{x\in\overline{\Omega}_h} |V(x)| \\ &\leq \max_{x\in\partial\Omega_h} |\tau_h(x)| + \frac{R^2}{2} \max_{x\in\Omega_h} |\tau_h(x)| = \mathcal{O}\left(h\right). \end{aligned}$$

So our difference method is consistent. Convergence and consistency order coincide.
A glance at the convergence proof shows that only a bad consistency near the boundary

$$\tau_h(x) = \mathcal{O}\left(h\right) \quad , \quad x \in \Omega_h \setminus \Omega_h^\circ$$

prevents the convergence order 2 (if $u \in C^4(\overline{\Omega})$).

Note:
For vanishing step-size h, the number of irregular points $\#\Omega_h \setminus \Omega_h^\circ$ is one order of magnitude smaller than the number of inner points Ω_h°. This makes us hope for a tightening of our convergence result. A closer analysis indeed gives the following theorem. ◁

Theorem 4.16 *Suppose $u \in C^4(\overline{\Omega})$. Then:*

$$\max_{x \in \overline{\Omega}_h} |U(x) - u(x)| = \mathcal{O}\left(h^2\right) \quad .$$

Proof:
See for example Hackbusch [12, p. 82]. □

5 Weak Solutions

From section 1.1.1 we know that the deflection u of a fixed membrane Ω caused by a force density f is the solution to the elliptic boundary value problem

$$\begin{aligned} -\operatorname{div}(\alpha\nabla u) &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{5.1}$$

Here $\alpha > 0$ is a material constant. We now consider a membrane made from to different materials. Then

$$\alpha(x) = \begin{cases} \alpha_1 & x \in \Omega_1 \\ \alpha_2 & x \in \Omega_2 \end{cases} ,$$

where α_1, $\alpha_2 \in \mathbb{R}$ and the partial domains Ω_1, Ω_2, with $\Omega_1 \cup \Omega_2 \cup \Gamma = \Omega$, $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$, represent the different materials.

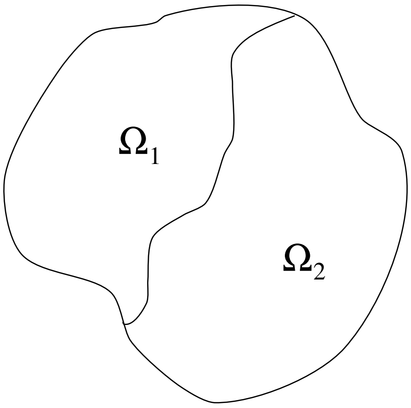


Figure 5.1: transition problem

From $\alpha_1 \neq \alpha_2$ the dilemma arises, that *either* ∇u *or* $\alpha\nabla u$ can be continuous in Ω. In both cases, (5.1) cannot be fulfilled (in the classical sense). A possible way out might be to replace (5.1) by a so called *transition problem.* Here we demand (5.1) to be valid only in the interior of Ω_1 and Ω_2. At the inner boundary $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$ with normal vector n (in whatever direction!) one sets the transition conditions

$$\begin{aligned} u_1 &= u_2 & \text{on } \Gamma \\ \alpha_1 \frac{\partial}{\partial n} u_1 &= \alpha_2 \frac{\partial}{\partial n} u_2 & \text{on } \Gamma \quad , \end{aligned} \tag{5.2}$$

where $u_i = u_{|\Omega_i}$ for $i = 1, 2$. Now we have to clear existence, uniqueness and stability of the solution to transition problems.

In order to apply difference schemes, the transition conditions (5.2) have to be discretized separately. For a curved transition Γ especially the discretization of the normals becomes tricky. The problems become even worse if $\Gamma = \Gamma(u)$ depends on an unknown function u (heat conduct, porous media). In the course of the iterative solution of the resulting non-linear system of equations one has to adapt the whole discretization of the transition condition to the at each time effective transition $\Gamma(U_h^\nu)$.
Such problems are typical for difference schemes. Their fundamental disadvantages are

- missing flexibility,
- high regularity requirements.

To construct more resistant solvers, we remember, that our mathematical modeling first lead to the minimization problem

$$u \in H : \;\; J(u) \leq J(v) \quad \forall v \in H \tag{5.3}$$

for the energy functional

$$J(v) = \frac{1}{2} \int\limits_\Omega \alpha \, |\nabla v|^2 \; dx - \int\limits_\Omega f v \, dx \quad . \tag{5.4}$$

Only afterwards and under additional regularity assumptions, we were able to derive a corresponding differential equation. These assumption are for non-smooth date, for example for discontinuous α, just plainly wrong! This thought in mind, we will from now on concentrate on solving the minimization problem (5.3) directly. Again we have to clarify existence, uniqueness and continuous dependency on the data. In particular, this calls for the right choice of a solution space H.

5.1 Hilbert Spaces

We will give a short introduction to the basic concepts of functional calculus. A more detailed and application-oriented introduction to this field is for example Alt [2].

Definition 5.1

(a) Suppose V is a linear space. Then a mapping

$$\|\cdot\| : \; V \to \mathbb{R}$$

is called norm on V if for all $v, w \in V$ and $\mu \in \mathbb{R}$, the following properties hold:

$$\begin{gathered} \|v\| \geq 0 \quad \text{and} \quad \|v\| = 0 \iff v = 0 \\ \|\mu v\| = |\mu| \, \|v\| \\ \|v + w\| \leq \|v\| + \|w\| \end{gathered}$$

If there is a norm $\| \cdot \|_V$ defined on V, then V is a normed, lineare space.

(b) *A normed, linear space V is called complete if every Cauchy sequence in V converges (in V, of course!).*

(c) *A complete, normed, linear space B is called Banach space.*

Example:
Suppose $\Omega \subset \mathbb{R}^n$ to be a bounded domain. Then $C(\overline{\Omega})$, together with the norm

$$\|v\|_\infty = \max_{x\in\overline{\Omega}} |v(x)| \quad ,$$

is a Banach space. ◁

Example:
The linear space $C^1(\overline{\Omega})$, together with the norm

$$\|v\|_{1,\infty} = \|v\|_\infty + \sum_{i=1}^{n} \|v_{x_i}\|_\infty \quad ,$$

is a Banach space. ◁

Definition 5.2

(a) *Suppose V and W are normed, linear spaces with norms $\|\cdot\|_V$ and $\|\cdot\|_W$ respectively. Then, a linear mapping $L : V \to W$ is called bounded if there is a constant $c > 0$, such that*

$$\|Lv\|_W \le c\,\|v\|_V \quad \forall v \in V \quad .$$

(b) *We denote the set of all bounded, linear mappings $L : V \to W$ with $\mathcal{L}(V,W)$.*

(c) *$V' = \mathcal{L}(V,\mathbb{R})$ is the dual space of V. The elements of V' are called functionals.*

Theorem 5.3 *A linear mapping $L : V \to W$ is continuous iff it is bounded.*

Theorem 5.4 *By*

$$\|L\| = \sup_{v\in V} \frac{\|Lv\|_W}{\|v\|_V} \quad , \quad L \in \mathcal{L}(V,W) \quad ,$$

a canonical norm is defined on $\mathcal{L}(V,W)$.

Definition 5.5

(a) *A mapping $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is called bilinear form (on V) if it is linear in both arguments.*

(b) *$a(\cdot,\cdot)$ is called symmetric if*

$$a(v,w) = a(w,v) \quad \forall v,w \in V \quad .$$

(c) *$a(\cdot,\cdot)$ is called positive definite if*

$$a(v,v) \ge 0 \quad \text{and} \quad a(v,v) = 0 \iff v = 0 \quad \forall v \in V \quad .$$

(d) $a(\cdot,\cdot)$ is called bounded if there is a constant Γ, such that

$$|a(v,w)| \le \Gamma \|v\|_V \|w\|_V \quad \forall v,w \in V \quad .$$

Theorem 5.6 *A bilinear form $a(\cdot,\cdot)$ is continuous iff it is bounded.*

Theorem 5.7 *A symmetric, positive definite bilinear form $a(\cdot,\cdot)$ fulfills the Cauchy–Schwarz inequality*

$$|a(v,w)| \le a(v,v)^{\frac{1}{2}} \cdot a(w,w)^{\frac{1}{2}} \quad \forall v,w \in V$$

and the triangle inequality

$$a(v+w,v+w)^{1/2} \le a(v,v)^{\frac{1}{2}} + a(w,w)^{\frac{1}{2}} \quad \forall v,w \in V$$

Particularly, by

$$\|v\| = a(v,v)^{\frac{1}{2}}, \quad v \in V,$$

a norm is defined, the so called energy norm.

Definition 5.8

(a) A symmetric, positive definite bilinear form is called dot product or scalar product.

(b) A linear space V together with a scalar product $(\cdot,\cdot)$ and the corresponding norm $\|\cdot\|_V$ is called pre-Hilbert space.

(c) A complete pre-Hilbert space is called Hilbert space.

Example:
$\mathbb{Q}^n$ together with the Euclidean scalar product $(v,w) = \sum_i v_i w_i$ is a pre-Hilbert space. The completion $\mathbb{R}^n$ is a Hilbert space. ◁

Example:
The linear space

$$X = \{v \in C(\Omega) \mid \int\limits_\Omega v^2(x)\,dx < \infty\} \quad ,$$

together with the scalar product

$$(v,w)_{L^2(\Omega)} = \int\limits_\Omega v(x)w(x)\,dx \quad ,$$

is a pre-Hilbert space. ◁

Example:
The linear space

$$X = \{v \in C^1(\Omega) \mid (v,v)_{H^1(\Omega)} < \infty\} \quad ,$$

together with the scalar product

$$(v,w)_{H^1(\Omega)} = (v,w)_{L^2(\Omega)} + \sum_{i=1}^n (v_{x_i}, w_{x_i})_{L^2(\Omega)} \quad ,$$

is a pre-Hilbert space. ◁

Example:
Consider $\alpha : \Omega \to \mathrm{I\!R}$ the piecewise continuous (i.e. continuous except for Riemann-measure-zero sets) and suppose

$$0 < \alpha_0 \leq \alpha(x) \leq \alpha_1 < \infty \quad \forall x \in \Omega$$

with $\alpha_0, \alpha_1 \in \mathrm{I\!R}$. Then

$$X = \{v \in C^1(\Omega) \cap C(\overline{\Omega}) \mid v|_{\partial\Omega} = 0, (v,v)_{H^1(\Omega)} < \infty\} \subset H_C$$

is a pre-Hilbert space with the energy scalar product

$$a(v,w) = \int\limits_{\Omega} \alpha \nabla v \cdot \nabla w \, dx \quad .$$

Theorem 5.9 *Every pre-Hilbert space V is in some way up to an isomorphism uniquely extendible to a Hilbert space H. V then is said to lie densely in H, and H is called completion of V.*

Theorem 5.10 *Suppose V to be a linear space with norm $\|\cdot\|_V$. On V there is exactly one scalar product $(\cdot,\cdot)$, that induces $\|\cdot\|_V$, i.e.*

$$\|v\|_V = (v,v)^{\frac{1}{2}} \quad \forall v \in V$$

if and only if $\|\cdot\|_V$ fulfills the parallelogram equation

$$\|v+w\|_V^2 + \|v-w\|_V^2 = 2\|v\|_V^2 + 2\|w\|_V^2 \quad \forall v,w \in V \quad .$$

Example:
The norm $\|\cdot\|_{1,\infty}$ does not fulfills the parallelogram equation, so $C^1(\overline{\Omega})$ in *no* Hilbert space!◁

Before we go on to special function spaces, we want to exploit this abstract framework for a bit to provide some fundamental statements . From now on, consider H to be a Hilbert space equipped with a scalar product $(\cdot,\cdot)$ and the corresponding norm $\|\cdot\|_H = (\cdot,\cdot)^{1/2}$.

Theorem 5.11 (Riesz representation theorem) *For every $l \in H'$, the variational problem*

$$u \in H : \quad (u,v) = l(v) \quad \forall v \in H \tag{5.5}$$

has a unique solution and it holds

$$\|u\|_H = \|l\|_{H'} \quad .$$

Proof:
The proof for existence is a case of *direct methods of variational calculus.* We set $J(v) = \frac{1}{2}(v,v) - l(v)$. We can copy the proof to theorem 1.2 (page 3) verbatim to see the equivalence of (5.5) and the minimization problem

$$u \in H : \quad J(u) \leq J(v) \quad \forall v \in H \quad . \tag{5.6}$$

Thus, it is sufficient to show the existence of a unique solution.

1. First, we will show, that $J(v)$ is bounded from below, meaning there is a $M \in \mathbb{R}$, s.t.
$$-\infty < M \le J(v) \quad \forall v \in H \quad .$$
Because of the boundedness of l,
$$|l(v)| \le \|l\|_{H'} \|v\|_H$$
and it follows, that
$$\begin{aligned} J(v) &= \frac{1}{2}(v,v) - l(v) = \frac{1}{2}\|v\|_H^2 - l(v) \\ &\ge \frac{1}{2}\|v\|_H^2 - |l(v)| \\ &\ge \frac{1}{2}\|v\|_H^2 - \|l\|_{H'}\|v\|_H \quad . \end{aligned}$$
The parabola $\frac{1}{2}\|v\|_H^2 - \|l\|_{H'}\|v\|_H$ is bounded from below for all $\|v\|_H$, and thereby $J(v)$ as well.
2. Since $J(v)$ is bounded from below, the infimum exists
$$\beta = \inf_{v \in H} J(v) \quad .$$
By definition of the infimum, there is a minimal sequence $(u_n)_{n\in\mathbb{N}} \subset H$, that is a sequence with the property
$$J(u_n) \to \beta \quad \text{for} \quad n \to \infty \quad .$$
We show, that $(u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. First, we notice, that because of the *convexity of* J, i.e.
$$\beta \le J\left(\tfrac{u_n+u_m}{2}\right) \le \frac{1}{2}\left(J(u_n) + J(u_m)\right) \quad \to \beta$$
$J(\frac{u_n+u_m}{2})$ converges to β as well. From the parallelogram equation, we get
$$\begin{aligned} \|u_n - u_m\|_H^2 &= 2\|u_n\|_H^2 + 2\|u_m\|_H^2 - \|u_n + u_m\|_H^2 \\ &= 4J(u_n) + 4J(u_m) + 4l(u_n+u_m) - 4\left\|\frac{u_n+u_m}{2}\right\|_H^2 \\ &= 4J(u_n) + 4J(u_m) - 8\left(\frac{1}{2}\left\|\frac{u_n+u_m}{2}\right\|_H^2 - l(\frac{u_n+u_m}{2})\right) \\ &= 4J(u_n) + 4J(u_m) - 8J(\frac{u_n+u_m}{2}) \to 0 \quad . \end{aligned}$$
Then $(u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence and because of the completeness of H there is a $u^* \in H$ with
$$\lim_{n\to\infty} u_n = u^* \qquad \text{in } H \quad .$$
3. The continuity of J yields
$$\beta = \lim_{n\to\infty} J(u_n) = J(u^*) \quad .$$
So u^* is a solution to the minimization problem (5.6).
4. The solution u^* is unique. If there were two solutions u_1 and u_2 of (5.5), then
$$(u_1, v) = l(v) = (u_2, v) \quad \forall v \in H$$
so
$$(u_1 - u_2, v) = 0 \quad \forall v \in H \quad .$$
Insertion of $v = u_1 - u_2$ provides the claim.
5. $\|u\|_H = \|l\|_{H'}$ follows from the inequalities
$$\|l\|_{H'} = \sup_{v\neq 0}\frac{|l(v)|}{\|v\|_H} = \sup_{v\neq 0}\frac{|(u,v)|}{\|v\|_H} \le \|u\|_H \qquad \text{Cauchy–Schwarz}$$
and
$$\|u\|_H^2 = (u,u) = l(u) \le \|l\|_{H'}\|u\|_H \quad .$$

Note:
Riesz' representation theorem tells us, that every $l \in H'$ can, via the scalar product, be represented by a $u \in H$ with equal norm. *The Hilbert space H and its dual space H' are isometrically isomorphic*, this implies, amongst other things, that they are topologically or algebraically indistinguishable. ◁

Corollary 5.12 *Suppose $l \in H'$ and $a(\cdot,\cdot)$ to be a symmetric bilinear form on H with the property*

$$\gamma \|v\|_H^2 \leq a(v,v) \qquad \forall v \in H \tag{5.7}$$

where γ does not depend on v. Then the minimization problem

$$u \in H: \quad J(u) \leq J(v) \quad \forall v \in H$$

for the energy functional

$$J(v) = \frac{1}{2}a(v,v) - l(v), \quad v \in H \quad ,$$

or respectively its variational formulation

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

has a uniquely determined solution. Furthermore

$$\|u\|_H \leq \frac{1}{\gamma}\|l\|_{H'} \quad .$$

Proof:
Because of (5.7), $a(\cdot,\cdot)$ is a scalar product on H and $l \in H'$ is bounded with respect to the energy norm as well. So we can apply 5.11 and get existence and uniqueness. The rest follows from

$$\gamma \|u\|_H^2 \leq a(u,u) = l(u) \leq \|l\|_{H'} \|u\|_H \quad .$$

Corollary 5.12 motivates the following

Definition 5.13 *A bilinear form $a(\cdot,\cdot)$ is called (H-) elliptic or coercive if the estimates*

$$\boxed{\gamma\|v\|_H^2 \leq a(v,v), \quad |a(v,w)| \leq \Gamma\|v\|_H\|w\|_H \qquad \forall v,w \in H} \tag{5.8}$$

with constants $\gamma, \Gamma > 0$ hold.

Corollary 5.14 (best approximation) *Suppose S is a closed subspace of H and $w_0 \notin S$. Then there exists a $w \in S$ with*

$$\|w - w_0\|_H = \min_{v \in S} \|v - w_0\|_H \quad ,$$

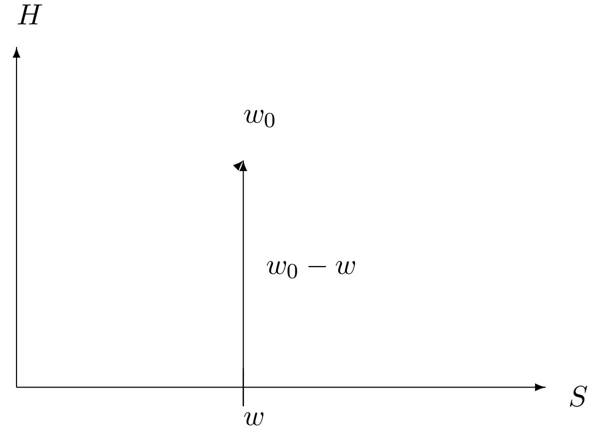


Figure 5.2: best approximation in Hilbert spaces

and the following orthogonality holds:

$$(w_0 - w, v) = 0 \quad \forall v \in S \quad .$$

Proof:
Fix $J(v) = \frac{1}{2}(v,v) - \ell_0(v)$ with $\ell_0(v) = (v, w_0)$, then evidently

$$\|v - w_0\|_H^2 = 2\,J(v) + (w_0, w_0) \quad .$$

S is as a closed subspace again a Hilbert space. Thus, by the Riesz representation theorem (Theorem 5.11) there exists exactly one $w \in S$ with the property

$$J(w) \leq J(v) \quad \forall v \in S \quad .$$

The orthogonality is precisely the equivalent variational formulation

$$w \in S : \quad (w, v) = (w_0, v) \quad \forall v \in S \quad .$$

We will now generalize the notion of corollary 5.12. The central idea is, that the bilinear form no longer has to be symmetric. *Note, that the variational problem is no longer based upon a minimization problem*!

Theorem 5.15 (Lax–Milgram lemma) *Suppose $a(\cdot,\cdot)$ is H-elliptic (not necessarily symmetric!). Then the variational equation*

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

has for every $l \in H'$ a uniquely defined solution, and it holds

$$\|u\|_H \leq \frac{1}{\gamma}\,\|l\|_{H'} \quad .$$

Proof:

1. We find ourselves an operator representation for the bilinear form. Suppose $w \in H$ arbitrary. Then

$$f_w := a(w, \cdot) \in H' \quad ,$$

because f_w is linear and continuous, since

$$|f_w(v)| = |a(w,v)| \le \Gamma \left\| w \right\|_H \left\| v \right\|_H \quad .$$

Then we also have

$$\left\| f_w \right\|_{H'} \le \Gamma \left\| w \right\|_H \quad .$$

Thus, by the Riesz representation theorem there is for every $w \in H$ a $Aw \in H$ with the property

$$(Aw, v) = f_w(v) = a(w,v) \quad \forall v \in H \quad .$$

In this sense, the so-defined mapping

$$A: H \to H$$

represents the bilinear form $a(\cdot,\cdot)$.

2. Properties of A:
   - A is linear, because for all $v \in H$

$$\begin{aligned}(A(\mu_1 w_1 + \mu_2 w_2), v) &= a(\mu_1 w_1 + \mu_2 w_2, v)\\ &= \mu_1 a(w_1, v) + \mu_2 a(w_2, v)\\ &= \mu_1 (Aw_1, v) + \mu_2 (Aw_2, v) \quad .\end{aligned}$$

   - A is bounded, i.e. $\left\| Av \right\|_H \le \Gamma \left\| v \right\|_H \quad \forall v \in H$, since

$$\left\| Av \right\|_H^2 = (Av, Av) = a(v, Av) \le \Gamma \left\| v \right\|_H \left\| Av \right\|_H \quad .$$

   - A is inversely bounded, i.e. $\gamma \left\| v \right\|_H \le \left\| Av \right\|_H \quad \forall v \in H$, since

$$\gamma \left\| v \right\|_H^2 \le |a(v,v)| = |(Av, v)| \le \left\| Av \right\|_H \left\| v \right\|_H$$

   - A is injective, since $Av = 0 \Rightarrow v = 0$.

3. Properties of $R(A)$ and $R(A)^\perp$:
   We define for A the image $R(A)$ and its orthogonal complement $R(A)^\perp$.

$$\begin{aligned}R(A) &:= \{ w \in H \,|\, \exists v \in H \text{ with } w = Av \}\\ R(A)^\perp &:= \{ v \in H \,|\, (w,v) = 0 \quad \forall w \in R(A) \}\end{aligned}$$

   - $R(A)$ is closed, since from

$$(w_n)_{n \in \mathbb{N}} \in R(A) \quad \text{and} \quad w_n \to w_0$$

   follows with $Av_n = w_n$ due to

$$\left\| w_n - w_m \right\|_H = \left\| A(v_n - v_m) \right\|_H \ge \gamma \left\| v_n - v_m \right\|_H \quad ,$$

   that $(v_n)_{n \in \mathbb{N}}$ is a Cauchy sequence. Then $v_n \to v_0$ converges, and from the continuity of A follows $Av_0 = w_0$, so $w_0 \in R(A)$.
   - We have $R(A)^\perp = \{0\}$, since for $w_0 \in R(A)^\perp$, that is

$$(w_0, w) = 0 \quad \forall w \in R(A) \quad ,$$

   follows

$$0 = (w_0, w) = (w_0, Av) \quad \forall v \in H \quad ,$$

   in particular

$$0 = (w_0, Aw_0) = a(w_0, w_0) \ge \gamma \left\| w_0 \right\|_H^2 \quad .$$

4. We will now show $R(A) = H$. For this, we use Corollary 5.14 with $S = R(A)$ (R(A) is closed!). Because if $w_0 \in H \setminus R(A)$, there is a $w \in R(A)$ with

   $$(w_0 - w, v) = 0 \quad \forall v \in R(A) \quad ,$$

   so $w_0 - w \in R(A)^\perp$ and thereby $w_0 - w = 0$, which contradicts $w_0 \notin R(A)$.
5. Existence and uniqueness:

   Since we now know that $R(A) = H$, for every $w \in H$ there has to be a $u \in H$ with $Au = w$. Choosing in accordance to the Riesz representation theorem a w, such that

   $$(w, v) = l(v) \quad \forall v \in H \quad ,$$

   it follows, that

   $$a(u, v) = (Au, v) = (w, v) = l(v) \quad \forall v \in H \quad .$$

   The uniqueness simply follows from the injectivity of A. The stability estimate works just as in corollary 5.12. □

Solution Approximation via Galerkin Method. So much about existence and uniqueness of solutions to variational equations. Now we go on to the approximation of u in closed subspaces, so called Galerkin approximations.

Theorem 5.16 (Céa's lemma) *Suppose $a(\cdot,\cdot)$ is H-elliptic, $l \in H'$ and $S \subset H$ a closed subspace of H. Then the variational problem*

$$u_s \in S : \quad a(u_s, v) = l(v) \quad \forall v \in S$$

has exactly one solution u_s, and the following a priori error estimate holds

$$\|u - u_s\|_H \le \frac{\Gamma}{\gamma} \inf_{v \in S} \|u - v\|_H \quad ,$$

where u is the solution to the variational problem in H.

Proof:

1. Existence and uniqueness:

   The uniqueness of the solution follows directly from the Lax–Milgram lemma, because S, as a closed subspace of a Hilbert space, is again a Hilbert space.
2. Galerkin orthogonality:

   Apparently

   $$a(u_s, v) = l(v) = a(u, v) \quad \forall v \in S \quad ,$$

   so

   $$a(u - u_s, v) = 0 \quad \forall v \in S \quad .$$

   Thereby the error is a-orthogonal on S.
3. Error estimate:

   From

   $$\begin{aligned} \gamma \|u - u_s\|^2 &\le a(u - u_s, u) - \underbrace{a(u - u_s, u_s)}_{=0} = a(u - u_s, u) \\ &= a(u - u_s, u) - \underbrace{a(u - u_s, v)}_{=0} = a(u - u_s, u - v) \\ &\le \Gamma \|u - u_s\| \, \|u - v\| \end{aligned}$$

for all $v \in S$, it directly follows, that

$$\|u - u_s\| \leq \frac{\Gamma}{\gamma} \|u - v\| \quad .$$

5.2 Sobolev Spaces

5.2.1 Completion

In order to be able to apply the abstract results from the previous section to our minimization problem (5.3), we need to choose an appropriate Hilbert space H as solution space. In doing so, we come across the following dilemma.

a) The space of all functions $v \in C^1(\overline{\Omega})$ with the property $v|_{\partial\Omega} = 0$, together with the norm $\|v\|_{1,\infty}$, is, though complete, no Hilbert space. Because

$$\|v + w\|_{1,\infty}^2 + \|v - w\|_{1,\infty}^2 \neq 2(\|v\|_{1,\infty}^2 + \|w\|_{1,\infty}^2)$$

(exercise!) there is no scalar product, inducing $\|\cdot\|_{1,\infty}$.

b) As an example for a pre-Hilbert space we introduced on page 50 the space

$$X = \{v \in C^1(\Omega) \cap C(\overline{\Omega}) \mid v|_{\partial\Omega} = 0, (v, v)_{H^1(\Omega)} < \infty\} \tag{5.9}$$

with the scalar product

$$(v, w)_{H^1(\Omega)} = (v, w)_{L^2(\Omega)} + \sum_{i=1}^{n} (v_{x_i}, w_{x_i})_{L^2(\Omega)} \quad .$$

This space might be larger than our next best solution space H_C for the deflected membrane (cf. page 3); but again, X is not complete.

Example:
To see this, consider the following counter-example. Suppose $\Omega = (-1, 1)$. The sequence $(v_k)_{k \in \mathbb{N}} \subset X$ defined by

$$v_k(x) = \begin{cases} 1 + x & -1 \leq x \leq -\frac{1}{n} \\ 1 - \frac{1}{2n} - \frac{n}{2}x^2 & -\frac{1}{n} \leq x \leq +\frac{1}{n} \\ 1 - x & +\frac{1}{n} \leq x \leq 1 \end{cases}$$

is a Cauchy sequence with respect to $\|\cdot\|_{H^1(\Omega)}$ and converges uniformly to

$$v^*(x) = \begin{cases} 1 + x & -1 \leq x \leq 0 \\ 1 - x & 0 \leq x \leq 1 \end{cases}$$

but $v^* \notin X$, since $v^* \notin C^1(\Omega)$ (exercise). ◁

To turn the pre-Hilbert space X from (5.9) into a Hilbert space, we need to complete X. We fix

$$H_0^1(\Omega) = \text{completion of } X \text{ with respect to } \|\cdot\|_{H^1(\Omega)} = (\cdot, \cdot)_{H^1(\Omega)}.$$

and now can choose $H = H_0^1(\Omega)$ as our solution space. Unfortunately, we do not yet know, how such a solution, or more general, a function $v \in H$, might look like.

5.2.2 Fundamental properties of Sobolev spaces

$H_0^1(\Omega)$ is a special *Sobolev space.* In the following, we will define additional Sobolev spaces by completion and give some basic properties. If you want to learn more, you could try Alt [2]. Even more is written in Adams [1].

1. Equivalence classes. The elements of Sobolev spaces are equivalence classes $[v]$ of functions. The corresponding equivalence relation is given by

$$v_1 \sim v_2 \iff \text{meas}\{x \in \Omega \,|\, v_1(x) \neq v_2(x)\} = 0 \quad .$$

Here, meas means the n-dimensional Lebesgue-measure.

Definition 5.17 $\Omega' \subset \mathbb{R}^n$ *is called (Lebesgue-)measure-zero or null set if there is for every* $\varepsilon > 0$ *a set of square boxes* $\{I_k \,|\, k \in \mathbb{N}\}$*, such that*

$$\Omega' \subset \bigcup_{k\in\mathbb{N}} I_k \quad \text{and} \quad 0 \leq \sum_{k\in\mathbb{N}} \text{meas}\{I_k\} < \varepsilon \quad .$$

If some assertion is to be true for all $x \in \Omega$ *except for a Lebesgue-null set, we say the assertion is true almost everywhere (a.e.) in* Ω*.*

Note:
In contrast to the Riemann-measure, *countably infinite* boxes are allowed! ◁

Example:
Every countable set $\Omega' = \{z_k \in \mathbb{R}^n \,|\, k \in \mathbb{N}\}$ is a Lebesgue-null set. ◁

First of all, we characterize the Sobolev space, which is generated by completion of the continuous, square-integrable function.

Theorem 5.18 *Through completion of*

$$X = \{v \in C(\Omega) \mid \int_\Omega v^2 \, dx < \infty\}$$

with the scalar product

$$(v, w)_{L^2(\Omega)} = \int_\Omega vw \, dx$$

one obtains the space of equivalence classes $[v]$ *of (Lebesgue-measurable) functions* v *with the property*

$$\|v\|_{L^2(\Omega)} = \left(\int_\Omega |v(x)|^2 \, dx \right)^{\frac{1}{2}} < \infty \quad .$$

Example:
Dirichlet's function

$$v = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

lies in $L^2(\mathbb{R})$ and it holds $v \in [0]$. ◁

All equivalent functions have the same norm! To simplify things, we will from now on no longer differentiate between $[v]$ and its representative $v \in [v]$. But be careful!

Finally, one last approximation result.

Theorem 5.19 *The subspace*

$$C_0^\infty(\Omega) = \{v \in C^\infty(\Omega) \mid \operatorname{supp} v \subset \Omega\}$$

of infinitely often differentiable functions with compact support in Ω

$$\operatorname{supp} \varphi = \overline{\{x \in \Omega \,|\, \varphi(x) \neq 0\}}$$

lies dense in $L^2(\Omega)$, *i.e for all* $v \in L^2(\Omega)$ *and every* $\varepsilon > 0$ *there is a* $\varphi_\varepsilon \in C_0^\infty$ *with the property*

$$\|v - \varphi_\varepsilon\|_{L^2(\Omega)} \leq \varepsilon \quad .$$

Proof:
see Alt [2], page 73. □

2. Weak Derivatives. Functions $v \in L^2(\Omega)$ are only determined up to (Lebesgue)-null sets and thus certainly no longer differentiable in the classical sense. So we generalize our notion of derivative.

Definition 5.20 (weak derivative) *Suppose* $u \in L^2(\Omega)$. *If there is a* $g \in L^2(\Omega)$ *with the property*

$$\int\limits_\Omega u\varphi_{x_k}\, dx = -\int\limits_\Omega g\varphi\, dx \qquad \forall \varphi \in C_0^\infty(\Omega) \quad ,$$

$u_{x_k} := g$ *is called* *weak derivative* *of* u *in direction* x_k.

Note:
The weak derivative fulfills *by definition* the product rule (Green's formula!). ◁

Example:
We consider the piecewise linear spline functions

$$S = \{v \in C[0,1] \mid v_{[x_{i-1},x_i]} \in \Pi_1,\ i = 1, \ldots, N\}$$

(Π_1 denotes the polynomials of 1^{st} order) with respect to a grid

$$0 = x_0 < x_1 < x_2 < \ldots < x_N = 1 \quad .$$

Apparently $v \in S$ is almost never classically differentiable. In the weak sense, however, it is:

With $\varphi \in C_0^\infty(0,1)$ we have

$$\begin{aligned}\int_0^1 v\varphi' \, dx &= \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} v\varphi' \, dx \\ &= \sum_{i=1}^{N} \left(- \int_{x_{i-1}}^{x_i} v'\varphi \, dx \right) + \sum_{i=1}^{N-1} \underbrace{\left(v(x_i)^{\text{links}} - v(x_i)^{\text{rechts}}\right)}_{=0} \varphi(x_i) \\ &= - \int_0^1 g\varphi \, dx \quad ,\end{aligned}$$

with $g(x) := v'(x) \; \forall x \neq x_i$. ◁

Example:
Consider $H : [-1,1] \to \mathbb{R}$ given by

$$H(x) = \begin{cases} 0 & -1 \leq x < 0 \\ 1 & 0 \leq x \leq 1 \end{cases} \qquad \text{(Heaviside function)} \quad .$$

To see, whether H is weakly differentiable, we choose $\varphi \in C_0^\infty(-1,1)$ and calculate

$$\int_{-1}^{1} H\varphi' \, dx = 0 + \int_0^1 1 \cdot \varphi' \, dx = \varphi(1) - \varphi(0) = -\varphi(0) =: -\delta_0\varphi \quad .$$

It appears, that by $\delta_0\varphi = \varphi(0)$ a linear mapping from $C_0^\infty(-1,1)$ to $\mathbb{R}$ is defined. This mapping is called *Dirac delta distribution.* δ_0 is not defined as a mapping from $L^2(-1,1)$ to $\mathbb{R}$, so there is no $g \in L^2(-1,1)$ with the property $\int_{-1}^{1} g\varphi \, dx = \delta_0\varphi$. The Heaviside side function H is thus *not* weakly differentiable. ◁

Recall the multi-index notation $\partial^\beta = \frac{\partial}{\partial_{x_{\beta_1}} \cdots \partial_{x_{\beta_k}}}$, $\beta_i = 1, \ldots, n, |\beta| = k$.

Definition 5.21 (Sobolev spaces) *The completion of*

$$X = \{v \in C^\infty(\Omega) \mid (v,v)_{H^m(\Omega)} < \infty\}$$

with respect to the norm, induced by the scalar product

$$(v,w)_{H^m(\Omega)} = \sum_{|\beta| \leq m} (\partial^\beta v, \partial^\beta w)$$

is the Sobolev space $H^m(\Omega)$.

Theorem 5.22 *The Sobolev space $H^m(\Omega)$ consists of all (equivalence classes of) functions $v \in L^2(\Omega)$ with weak derivatives*

$$\partial^\beta v \in L^2(\Omega), \quad |\beta| \leq m \quad .$$

$H^m(\Omega)$ is a Hilbert space with the scalar product

$$(v, w)_m = \sum_{|\beta| \leq m} (\partial^\beta v, \partial^\beta w) \quad .$$

More details are given in Alt [2] on pages 31 ff. and 78.

Note:
By construction, $C^\infty(\Omega) \cap H^m(\Omega)$ lies dense in $H^m(\Omega)$. ◁

3. Boundary Conditions in the Weak Sense. We have seen, that functions $u \in L^2(\Omega)$ are only determined up to (Lebesgue-)null sets. Unfortunately, the boundary $\partial\Omega$ is a null set. But then, what sense should for example Dirichlet boundary conditions $v|_{\partial\Omega} = 0$ make for functions $v \in H_0^1(\Omega)$, that do not lie in $C(\overline{\Omega})$? A first answer, for the case

$$\Omega = \mathbb{R}^n_+ = \{(x, y) \,|\, x \in \mathbb{R}^{n-1}, y > 0\}, \qquad n \in \mathbb{N} \geq 2$$

with boundary $\partial\Omega = \mathbb{R}^{n-1}$, gives the following *trace theorem.*

Theorem 5.23 (trace theorem in the half space) *Suppose $\Omega = \mathbb{R}^n_+$. Then, there is a bounded linear mapping*

$$\mathrm{tr} : H^1(\Omega) \to L^2(\partial\Omega)$$

with the property

$$\mathrm{tr}\, v = v|_{\partial\Omega} \qquad \forall v \in H^1(\Omega) \cap C(\overline{\Omega}) \quad .$$

Proof:
We will later see, that for smoothly bounded domains the linear space

$$X = \{u|_\Omega \mid u \in C_0^\infty(\mathbb{R}^n)\}$$

lies dense in $H^1(\Omega)$. The boundary $\partial\Omega = \mathbb{R}^{n-1}$ of $\Omega = \mathbb{R}^n_+$ is extremely smooth. We set

$$\mathrm{tr}\, v = v|_{\partial\Omega} \qquad \forall v \in X$$

and want to show now, that for a proper $c > 0$ the estimate

$$\|\mathrm{tr}\, v\|_{L^2(\partial\Omega)} \leq c\|v\|_{H^1(\Omega)} \qquad \forall v \in X$$

holds. To this, choose any $v \in X$.

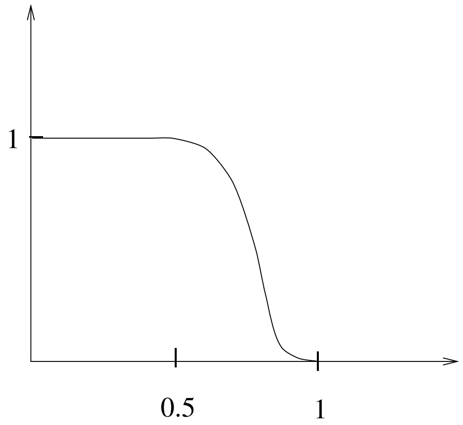


Figure 5.3: cut-off function φ

1. Localization:
   First, we construct a *cut-off function* $\varphi \in X$ with the property

$$\varphi(x,y) = \begin{cases} 1 & 0 \le y \le \frac{1}{2} \\ 0 & 1 \le y \end{cases} \qquad \forall (x,y) \in \overline{\mathbb{R}^n_+} \quad .$$

The for $w := \varphi v \in C^\infty(\overline{\mathbb{R}^n_+})$

$$w(x,0) = w(x,y) - \int\limits_0^y w_y(x,s)\, ds \quad .$$

Integration over y from 0 to 1 yields

$$w(x,0) = \int\limits_0^1 w(x,y)\, dy - \int\limits_0^1 1 \cdot \int\limits_0^y w_y(x,s)\, ds\, dy \quad .$$

From the product rule, we get

$$w(x,0) = \int\limits_0^1 w(x,y)\, dy - \int\limits_0^1 (1-y) w_y(x,y)\, dy \quad .$$

Young's inequality ($(a+b)^2 \le 2a^2 + 2b^2$) and the Cauchy–Schwarz inequality provide

$$\begin{aligned} w^2(x,0) &\le 2\int\limits_0^1 1^2\, dy \cdot \int\limits_0^1 w^2(x,y)\, dy + 2\int\limits_0^1 (1-y)^2\, dy \int\limits_0^1 w_y^2(x,y)\, dy \\ &= 2\int\limits_{\mathbb{R}_+} w^2(x,y)\, dy + \frac{2}{3}\int\limits_{\mathbb{R}_+} w_y^2(x,y)\, dy \quad . \end{aligned}$$

Integration over x yields

$$\int\limits_{\mathbb{R}^{n-1}} w^2(x,0)\, dx \le 2\int\limits_{\mathbb{R}^n_+} w^2(x,y)\, d(x,y) + 2\int\limits_{\mathbb{R}^n_+} w_y^2(x,y)\, d(x,y) \quad ,$$

so, because of $\operatorname{tr} w = w|_{\partial\mathbb{R}^n_+}$,

$$\|\operatorname{tr} w\|^2_{L^2(\partial\mathbb{R}^n_+)} \le 2(\|w\|^2_{L^2(\mathbb{R}^n_+)} + \|w_y\|^2_{L^2(\mathbb{R}^n_+)}) \quad .$$

2. Stability estimate for $v \in X$
   Inserting $w = v\varphi$ provides because of $\varphi_{|\partial \mathbb{R}^n_+} \equiv 1$
   $$\operatorname{tr} w = \operatorname{tr} v \quad .$$
   Furthermore
   $$\int\limits_{\mathbb{R}^n_+} w^2 \, d(x,y) = \int\limits_{\mathbb{R}^n_+} \varphi^2 v^2 \, d(x,y) \le \max_{(x,y)\in\mathbb{R}^n_+} \{\varphi(x,y)\} \int\limits_{\mathbb{R}^n_+} v^2 \, d(x,y) \le \|v\|^2_{L^2(\mathbb{R}^n_+)}$$
   and
   $$\begin{aligned}\int\limits_{\mathbb{R}^n_+} w_y^2 \, d(x,y) &= \int\limits_{\mathbb{R}^n_+} (\varphi_y v + v_y \varphi)^2 \, d(x,y) \\ &\le 2 \int\limits_{\mathbb{R}^n_+} \varphi_y^2 v^2 + v_y^2 \varphi^2 \, d(x,y) \\ &\le 2(\max_{(x,y)\in\mathbb{R}^n_+} \{\varphi_y^2(x,y)\} \, \|v\|^2_{L^2(\mathbb{R}^n_+)} + \max_{x\in\mathbb{R}^n_+} \{\varphi^2(x)\} \, \|v_y\|^2_{L^2(\mathbb{R}^n_+)}) \\ &\le c \, \|v\|^2_{H^1(\mathbb{R}^n_+)} \quad .\end{aligned}$$
   All in all, we get
   $$\|\operatorname{tr} v\|^2_{L^2(\partial\Omega)} \le c \, \|v\|^2_{H^1(\Omega)} \qquad \forall v \in X \quad . \tag{5.10}$$
3. Denseness approach:
   Suppose $v \in H^1(\Omega)$. Then there is a sequence $(v_k)_{k\in\mathbb{N}} \in X$ with
   $$\|v - v_k\|_{H^1(\Omega)} \to 0 \quad .$$
   So, $(v_k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $H^1(\Omega)$. Thanks to
   $$\|\operatorname{tr} v_k - \operatorname{tr} v_j\|_{L^2(\partial\Omega)} \le c \, \|\operatorname{tr} v_k - \operatorname{tr} v_j\|_{H^1(\Omega)}$$
   $\operatorname{tr} v_k = v_{k|\partial\Omega}$ is a Cauchy sequence in $L^2(\partial\Omega)$. $v^* \in L^2(\partial\Omega)$ shall be the limit. Then we set $\operatorname{tr} v := v^*$. This definition is independent of the choice of the sequence $(v_k)_{k\in\mathbb{N}}$ (check). All $v_k \in X$ meet the estimate (5.10). Passing to the limit, together with the continuity of the norm, provides
   $$\|\operatorname{tr} v\|_{L^2(\partial\Omega)} \le c \, \|v\|_{H^1(\Omega)} \qquad \forall v \in H^1(\Omega) \quad ,$$
   that is the continuity of $\operatorname{tr} : H^1(\Omega) \to L^2(\Omega)$. Finally, to show that
   $$\operatorname{tr} v = v|_{\partial\Omega} \qquad \forall v \in H^1(\Omega) \cap C(\overline{\Omega})$$
   holds, we need to construct for every $v \in H^1(\Omega) \cap C(\overline{\Omega})$ a sequence $(v_k)_{k\in\mathbb{N}} \in X$, converging to v, and fulfilling additionally the property
   $$v_n|_{\partial\Omega} \to v|_{\partial\Omega} \qquad \text{in } L^2(\Omega) \quad .$$
   The technical set of tools to this are to be found in Alt [2] in section 2.9 and 2.12. □

One can generalize the trace theorem in the half-space on domains with *Lipschitz boundary.*

Definition 5.24 (Lipschitz boundary) *A bounded domain $\Omega \subset \mathbb{R}^n$ has a Lipschitz boundary if there are finitely many open sets O_i and a number $\varepsilon > 0$, such that for all $x \in \partial\Omega$ the ball $K_\varepsilon(x)$ lies in some O_i and $O_i \cap \Omega = O_i \cap \Omega_i$ holds. Here*
$$\Omega_i = \{(x,y) \in \mathbb{R}^n \,|\, x \in \mathbb{R}^{n-1}, y \in \mathbb{R}, y < \varphi_i(x)\}$$
with Lipschitz-continuous φ_i, that is
$$|\varphi_i(x_1) - \varphi_i(x_2)| \le L|x_1 - x_2|$$
for some fixed (Lipschitz-)constant $L \in \mathbb{R}$.

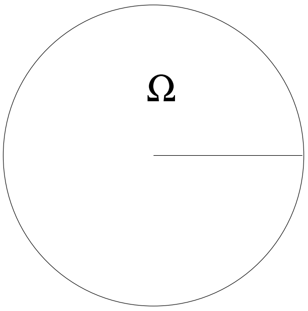


Figure 5.4: slit circular domain

Example:
The boundary $\partial\Omega$ of the slit circular domain in figure 5.4 is not Lipschitz. Note: In this case, for $v \in C(\overline{\Omega})$ the boundary values $v_{|\partial\Omega}$ are not defined just like that (double values for $y = 0, x \in (0,1)$)! ◁

Theorem 5.25 (Trace theorem for Lipschitz domains) *Let $\Omega \subset \mathbb{R}^n$ be a domain with Lipschitz boundary. Then for every $v \in H^1(\Omega)$ there exists a sequence $(v_k)_{k\in\mathbb{N}}$,*

$$(v_k)_{k\in\mathbb{N}} \subset \{w|_\Omega \mid w \in C_0^\infty(\mathbb{R}^n)\}$$

with the property

$$v = \lim_{k\to\infty} v_k \qquad \textit{in } H^1(\Omega) \quad .$$

The trace operator $\operatorname{tr}: \ H^1(\Omega) \to L^2(\partial\Omega)$

$$\operatorname{tr} v = \lim_{k\to\infty} v_k|_{\partial\Omega} \qquad \textit{in } L^2(\partial\Omega) \quad ,$$

is well-defined and bounded. Furthermore

$$\operatorname{tr} v = v|_{\partial\Omega} \qquad \forall v \in H^1(\Omega) \cap C(\overline{\Omega}) \quad .$$

Proof:
See Alt [2], page 190. □

Note:
This theorem can be tightened to

$$\|\operatorname{tr} v\|_{H^{\frac{1}{2}}(\partial\Omega)} \le c\,\|v\|_{H^1(\Omega)} \quad .$$

The Sobolev space (of rational order!) $H^{\frac{1}{2}}(\partial\Omega)$ is exactly the right trace space, since the trace operator $\operatorname{tr}: H^1(\Omega) \to H^{\frac{1}{2}}(\partial\Omega)$ is surjective, i.e. for every $g \in H^{\frac{1}{2}}(\partial\Omega)$ there is at least one $v \in H^1(\Omega)$ with the property $\operatorname{tr} v = g$. ◁

Let us go back to $H_0^1(\Omega)$ for awhile.

Theorem 5.26 *The subspace*

$$C_0^\infty(\Omega) \subset H_0^1(\Omega)$$

lies dense in $H_0^1(\Omega)$.

So we could also have defined $H_0^1(\Omega)$ directly as completion of $C_0^\infty(\Omega)$ with respect to $\|\cdot\|_{H^1(\Omega)}$. Usually it is done that way.

Note:
Apparently by Theorem 5.25 we have

$$\operatorname{tr} v = 0 \qquad \forall v \in H_0^1(\Omega)$$

in the trace operator sense. Equivalently,

$$H_0^1(\Omega) = \{ v \in H^1(\Omega) \mid \operatorname{tr} v = 0 \} \quad .$$

So that is how to understand zero-boundary conditions. ◁

We have seen that functions $v \in H^k(\Omega)$, at least for large k, show a similar behaviour as continuous functions (boundary values). This leads to the question, how big we need $k > m$, in order for

$$v \in H^k(\Omega) \Longrightarrow v \in C^m(\Omega)$$

to hold. The answer is the following *embedding theorem*

Theorem 5.27 (Sobolev embedding theorem) *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with Lipschitz boundary. If $v \in H^k(\Omega)$ with*

$$k > m + \frac{n}{2} \quad , \tag{5.11}$$

follows that

$$v \in C^m(\Omega) \qquad (\textit{there exists a } v \in [v] \cap C^m(\Omega))$$

and

$$\sup_{x\in\Omega} |\partial^\beta v| \leq c\,\|v\|_{H^k(\Omega)} \quad \forall |\beta| \leq m \quad .$$

Proof:
See Alt [2], page 244. □

Note:
So $H^1(\Omega) \subset C(\overline{\Omega})$ if $n = 1$. As we have only an equality in (5.11) for $n = 2$, this is no longer the case for $n \geq 2$. There are counterexamples (exercise). ◁

Theorem 5.28 (Rellich selection theorem) *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with Lipschitz boundary. If $(v_k)_{k\in\mathbb{N}} \subset H^{m+1}(\Omega)$ is bounded, that is*

$$\|v_k\|_{H^{m+1}} \le c \qquad \forall k \in \mathbb{N} \quad ,$$

then there exists a subsequence $\left(v_{k_j}\right)_{k_j\in\mathbb{N}}$ and a $v \in H^{m+1}(\Omega)$, such that

$$v_{k_j} \to v\,, \quad \text{for} \quad j \to \infty \quad \text{in} \quad H^m(\Omega) \quad .$$

Proof:
For those familiar with functional calculus: $H^{m+1}(\Omega)$ is reflexive, so a subsequence $\left(v_{k_j}\right)_{k_j\in\mathbb{N}}$ converges weakly in $H^{m+1}(\Omega)$ to a $v \in H^{m+1}(\Omega)$. The rest can be found in Alt [2], page 184. □

5.3 Weak Solutions of Elliptic Boundary Value Problems

As a solution space for our minimization problem (5.3), or more precisely for its variational formulation

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

we have now identified $H = H_0^1(\Omega)$. In order to apply the Riesz representation theorem, we must make sure that $a(\cdot,\cdot)$ and l fulfill their respective preconditions. Before that, we declare the following abbreviations

$$\|v\|_{m,\Omega} = \|v\|_{H^m(\Omega)} \qquad \|v\|_{0,\Omega} = \|v\|_{L^2(\Omega)} \quad .$$

If it is clear, that the norms relate to Ω, we sometimes omit the domain in the index. We also write

$$(v,w) = (v,w)_{L^2(\Omega)} \qquad (\nabla v, \nabla w) = \sum_{k=1}^{n} (v_{x_k}, w_{x_k}) \quad .$$

Lemma 5.29 *Let $f \in L^2(\Omega)$. Then*

$$l(v) = (f,v) \tag{5.12}$$

is a bounded functional on $H = H_0^1(\Omega)$ and $\|l\|_{H'} \le \|f\|_0$ holds.

Proof:
The Schwarz inequality provides $v \in H_0^1(\Omega)$

$$|l(v)| \le \|f\|_0\, \|v\|_0 \le \|f\|_0\, \|v\|_1$$

and thus the claim. □

Lemma 5.30 (Poincaré–Friedrich Inequality) *Let Ω be bounded and $0 < \alpha_0 \le \alpha(x) \le \alpha_1 < \infty$ almost everywhere in Ω. Then there is a $\gamma > 0$, such that the bilinear form*

$$a(u,v) = \int\limits_{\Omega} \alpha\, \nabla u \cdot \nabla v\; dx \tag{5.13}$$

fulfills the Poincaré–Friedrich inequality

$$\gamma\, \|v\|_0^2 \le a(v,v) \quad \forall v \in H_0^1(\Omega) \quad .$$

Proof:
Because of $\alpha(x) \geq \alpha_0 > 0$, almost everywhere, we have

$$a(v,v) \geq \alpha_0 \int_\Omega |\nabla v|^2 \, dx \quad .$$

So it suffices to find a $c > 0$, such that

$$\|v\|_0^2 \leq c \| \, |\nabla v| \, \|_0^2 \tag{5.14}$$

holds. We now choose an open square $Q = (a,b) \times (a,b)$ with $\overline{\Omega} \subset Q$ and extend an arbitrarily fixed $v \in C_0^\infty(\Omega)$ onto $v \in C_0^\infty(Q)$ by prescribing it to be zero everywhere else. Then

$$\begin{aligned} v(x_1,x_2) &= v(a,x_2) + \int_a^{x_1} v_{x_1}(\xi,x_2)\, d\xi = \int_a^{x_1} v_{x_1}(\xi,x_2)\, d\xi \\ &\overset{\text{Cauchy–Schwarz}}{\leq} \left(\int_a^b 1^2 \, dx \right)^{\frac{1}{2}} \cdot \left(\int_a^{x_1} v_{x_1}^2(\xi,x_2)\, d\xi \right)^{\frac{1}{2}} \quad . \end{aligned}$$

Integration provides

$$\begin{aligned} \|v\|_0^2 &= \int_a^b \int_a^b v^2(x_1,x_2)\, dx_1\, dx_2 \\ &\leq (b-a)^2 \int_a^b \int_a^b v_{x_1}^2(\xi,x_2)\, d\xi\, dx_2 \\ &\leq (b-a)^2 \, \| \mid \nabla v \mid \|_0^2 \quad . \end{aligned}$$

As a corollary of the Lax–Milgram lemma, we now obtain

Theorem 5.31 *Let $H = H_0^1(\Omega)$, $a(u,v) = \int_\Omega \alpha \nabla u \cdot \nabla v \, dx$ and $l(v) = (f,v)$ given as above in (5.13) and (5.12) respectively. Also let $0 < \alpha_0 \leq \alpha(x) \leq \alpha_1 < \infty$, almost everywhere, and $f \in L^2(\Omega)$. Then, the variational problem*

$$u \in H: \quad a(u,v) = l(v) \qquad \forall v \in H \tag{5.15}$$

is uniquely solvable and (with γ from lemma 5.30)

$$\|u\|_1 \leq \frac{1}{\gamma} \|f\|_0$$

holds.

Proof:
From lemma 5.29 $l \in H'$. Furthermore $a(\cdot,\cdot)$ is obviously symmetric and bilinear. Due to lemma 5.30 and

because of

$$\begin{aligned}
|a(v,w)| &\leq \alpha_1 \int\limits_\Omega |\nabla v \cdot \nabla w| \ dx \\
&\leq \alpha_1 \int\limits_\Omega |v_{x_1}|\,|w_{x_1}| + |v_{x_2}|\,|w_{x_2}| \ dx \\
&\overset{\text{Cauchy-Schwarz in } \mathbb{R}^2}{\leq} \alpha_1 \int\limits_\Omega (|v_{x_1}|^2 + |v_{x_2}|^2)^{\frac{1}{2}} \cdot (|w_{x_1}|^2 + |w_{x_2}|^2)^{\frac{1}{2}} \ dx \\
&\overset{\text{Cauchy-Schwarz in } L^2(\Omega)}{\leq} \alpha_1 \left(\int\limits_\Omega (|v_{x_1}|^2 + |v_{x_2}|^2) \ dx \right)^{\frac{1}{2}} \cdot \left(\int\limits_\Omega (|w_{x_1}|^2 + |w_{x_2}|^2) \ dx \right)^{\frac{1}{2}} \\
&\leq \alpha_1 \ \|v\|_1 \cdot \|w\|_1
\end{aligned}$$

$a(\cdot,\cdot)$ is H-elliptic. □

Note:
According to theorem 5.31 the variational problem (5.15) is posed correctly: The solution operator $f \to u$ as a mapping from $L^2(\Omega)$ to $H_0^1(\Omega)$ is continuous. ◁

Note:
If the solution of (5.15) is a smooth function, namely $u \in C^2(\overline{\Omega}) \cap H_0^1(\Omega)$, Green's formula provides that u is also a classical solution of the boundary value problem

$$\begin{aligned}
-\operatorname{div}(\alpha \nabla u) &= f \quad \text{in } \Omega \\
u &= 0 \quad \text{on } \partial\Omega \quad .
\end{aligned} \tag{5.16}$$

Definition 5.32 *The solution u to the variational problem* (5.15) *is called* <u>*weak solution*</u> *of the boundary value problem* (5.16)*.*

We now want to deduce the weak formulation of some other boundary value problems. The bilinear form $a(\cdot,\cdot)$ has to be chosen in each case in such a way, that for smooth enough solutions Green's formula leads back to the original boundary value problem.
If $a(\cdot,\cdot)$ is symmetric and positive definite on a (w.r.t. energy norm!) closed solution space $H \subset H^1(\Omega)$, then there is a uniquely defined weak solution (Riesz). If $a(\cdot,\cdot)$ is H-elliptic in the sense of (5.8), we do not need the symmetry and it suffices to know, that H is closed w.r.t. $\|\cdot\|_1$, and in addition, the $H^1(\Omega)$-stability follows (Lax–Milgram).

Example:
We consider the weak form of the boundary value problem (5.16) with

$$\alpha(x) = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} (x) \quad .$$

For every $x \in \Omega$ let $\alpha(x)$ be a symmetric matrix. For the eigenvalues $\lambda_1(x)$, $\lambda_2(x)$ suppose

$$0 < \alpha_0 \leq \lambda_1(x), \lambda_2(x) \leq \alpha_1 \text{ o.e.} \quad .$$

Then the bilinear form

$$a(v,w) = \int\limits_\Omega (\alpha \nabla v) \cdot \nabla w \ dx$$

is $H_0^1(\Omega)$-elliptic. ◁

Example:
We consider the boundary value problem

$$\begin{aligned} -\Delta u + \beta \cdot \nabla u &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \partial\Omega \quad . \end{aligned}$$

The corresponding weak formulation leads to the bilinear form

$$a(v,w) = \underbrace{(\nabla v, \nabla w)}_{\text{symmetric}} + \underbrace{(\beta \cdot \nabla v, w)}_{\text{not symmetric}} \quad .$$

This bilinear for in not symmetric, but $H_0^1(\Omega)$-elliptic (exercise). ◁

Example:
We consider the boundary value problem

$$\begin{aligned} -\Delta u + \lambda u &= f \qquad \text{in } \Omega \\ \frac{\partial}{\partial n} u &= 0 \qquad \text{on } \Omega \end{aligned} \tag{5.17}$$

with $\lambda > 0$. The corresponding weak formulation (5.15) leads to the bilinear form

$$a(v,w) = (\nabla v, \nabla w) + \lambda (v,w)$$

and to the solution space $H = H^1(\Omega)$. Apparently $a(v,w)$ $H^1(\Omega)$-elliptic for $\lambda > 0$.
In contrast to Dirichlet boundary conditions Neumann boundary conditions are not forced by the choice of a subspace of $H^1(\Omega)$. Of course they cannot be, because $\nabla u \in (L^2(\Omega))^2$ has no boundary conditions! Smooth enough solution of the variational problem due to Green's formula automatically fulfill homogeneous Neumann conditions (exercise)!
Let us briefly consider the case $\lambda = 0$. $a(v,w) = (\nabla v, \nabla w)$ is not positive definite on $H^1(\Omega)$, since

$$(\nabla v, \nabla w) = 0 \ \forall w \in H^1(\Omega) \iff v = \text{const.} \quad .$$

Thus, we identify functions, that only differ by a constant,

$$[v] = \{ w \mid w(x) - v(x) = \text{const., almost everywhere in } \Omega \} \quad .$$

We can define a scalar product on the quotient space H of the thus induced equivalence classes $[v]$ via

$$a([v],[w]) = (\nabla v, \nabla w), \qquad v \in [v], \ w \in [w]$$

If and only if the (already known!) compatibility condition

$$\int_\Omega f \, dx = 0$$

is fulfilled by $f \in L^2(\Omega)$,

$$l([v]) = \int_\Omega f v \, dx, \qquad v \in [v]$$

defines a linear functional.
In this case, the variational problem

$$[u] \in H : \qquad a([u],[v]) = l([v]) \qquad \forall v \in H$$

has a uniquely determined solution $[u]$. All functions $u \in [u]$ are solutions to the original problem with $\lambda = 0$. ◁

Example:
We will now consider inhomogeneous boundary conditions. To this end, let $\partial_D \cup \partial_N = \partial\Omega$ be a non-overlapping partition of the boundary into two parts of non-zero length each. The given boundary value problem is

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega \\ u &= g_D && \text{on } \partial_D \\ \frac{\partial u}{\partial n} &= g_N && \text{on } \partial_N \quad . \end{aligned} \tag{5.18}$$

One gets a weak formulation (5.15) by

$$a(v,w) = \int\limits_\Omega \nabla v \cdot \nabla w \, dx$$

and

$$l(v) = \int\limits_\Omega fv \, dx + \int\limits_{\partial_N} g_N v \, d\sigma \quad .$$

As reasonable solution space is

$$X = \{v \in H^1(\Omega) \mid \operatorname{tr}_{\partial_D} v = g_D\} \quad .$$

Problem: X is no linear space! We now assume, that at least $X \neq \emptyset$, i.e. that there is a $w_0 \in H^1(\Omega)$ with the property

$$\operatorname{tr}_{\partial_D} w_0 = g_D$$

(This is the case if and only if $g_D \in H^{\frac{1}{2}}(\partial_D)$). Then we calculate u_0 from

$$u_0 \in H : \qquad a(u_0, v) = l(v) - a(w_0, v) \qquad \forall v \in H$$

with the Hilbert space

$$H = \{v \in H^1(\Omega) \mid \operatorname{tr}_{\partial_D} v = 0\}$$

and get a (uniquely determined!) solution $u = u_0 + w_0$.
We have seen, that g_D has to fulfill certain conditions. Which conditions on g_N are necessary for solvability? ◁

6 Finite Elements

6.1 Construction of FE Spaces

Consider the variational equality

$$u \in H: \; a(u,v) = l(v) \quad \forall v \in H \tag{6.1}$$

a closed subspace $H \subset H^1(\Omega)$, an H-elliptic bilinear form $a(\cdot,\cdot)$, i.e.

$$\gamma \|v\|_1^2 \leq a(v,v) \;, \quad |a(v,w)| \leq \Gamma \|v\|_1 \|w\|_1, \quad \forall v,w \in H$$

and $l \in H'$. The Céa-Lemma 5.16 implies directly

Theorem 6.1 *Let $S \subset H$ be a closed subspace. Then u_S given by*

$$u_S \in S: \quad a(u_S,v) = l(v) \qquad \forall v \in S$$

is unique, and the following estimate holds

$$\|u - u_S\|_1 \leq \frac{\Gamma}{\gamma} \inf_{v \in S} \|u - v\|_1 \quad .$$

Note:
This class of approximation methods is called *Galerkin method.* If $a(\cdot,\cdot)$ is symmetric, they are called *Ritz–Galerkin method.* ◁

For S we want to consider finite elements.

Definition 6.2 *Let $\Omega \subset \mathbb{R}^2$ have a polygonal boundary. A set $\mathcal{T}$ of triangles t is called a triangulation of Ω if*

(i) $\overline{\Omega} = \bigcup_{t \in \mathcal{T}} t$

(ii) The intersection of two triangles from $\mathcal{T}$ is either a common edge, a common node or empty.

The set of all inner edges is called $\mathcal{E}_{\mathcal{T}}$. The set of all inner nodes $\mathcal{N}_{\mathcal{T}}$.

Definition 6.3 *Let Π_m denote the set of all polynomials of m-th order. Then*

$$\boxed{S^{(m)} := \{v \in C(\overline{\Omega}) \mid v|_t \in \Pi_m \quad \forall t \in \mathcal{T}\} \;.}$$

is the space of the (polynomial) m-th order finite elements (relating the triangulation $\mathcal{T}$).

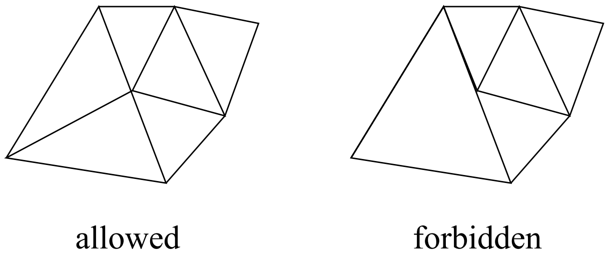


Figure 6.1: triangulation and forbidden triangle-decomposition

Note:
Aside from trivial cases $S^{(m)} \not\subset C^1(\Omega)$. In the case $m = 1, 2, 3, \ldots$, one speaks of (piecewise) linear, quadratic, cubic, ... finite elements. ◁

To construct a basis of $S^{(m)}$, we first want to clarify how to determine a polygon $P \in \Pi_m$.

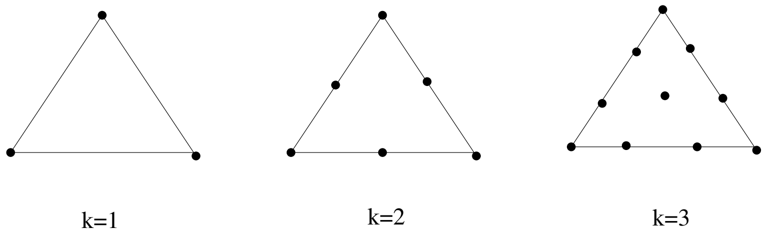


Figure 6.2: nodes

Lemma 6.4 *Let $m \geq 0$. Arrange*

$$s = 1 + 2 + \cdots + (m+1) = \frac{(m+1)(m+2)}{2}$$

nodes $\mathcal{N}^{(m)}(t) = \{z_1, z_2, \ldots, z_s\}$ on $m+1$ parallel lines on the triangle t, as shown in figure 6.2. Then the interpolation problem

$$P \in \Pi_m : \qquad P(z_i) = P_i \qquad \forall i = 1, \ldots, s$$

has a unique solution for all node values P_i.

Proof:
By induction:
The case $m = 0$ is clear. Let the claim be true for $m - 1 \geq 0$. Without loss of generality let the edge with the nodes $z_1, z_2, \ldots, z_{m+1}$ lie on the x-axis. Then the lines above are $\{(x, y) \mid y = y_i\}$, with positive $y_k \in \mathbb{R}$, $k = 1, \ldots, m$. As we know, the univariate interpolation problem

$$P_0 \in \Pi_m : \qquad P_0(z_i) = P_i \qquad \forall i = 1, \ldots, m+1$$

has a unique solution for all P_i, which we can extend, e.g. constant in y-direction, to all $\mathbb{R}^2$. By induction hypothesis there is exactly one $Q_0 \in \Pi_{m-1}$ with the property

$$Q_0(z_i) = \frac{1}{y_{k_i}}(P_i - P_0(z_i)) \qquad \forall i = m+2, \ldots, s \quad .$$

Thereby k_i denotes the line which contains the node z_i. Then $P(x, y) = P_0(x, y) + yQ_0(x, y)$ is the unique solution. □

We now consider the whole triangulation $\mathcal{T}$.

Corollary 6.5 *Let*

$$\mathcal{N}^{(m)} = \bigcup_{t \in \mathcal{T}} \mathcal{N}^{(m)}(t).$$

Then the interpolation problem

$$v \in S^{(m)} : \qquad v(p) = v_p \qquad \forall p \in \mathcal{N}^{(m)} \tag{6.2}$$

has a unique solution for all node values v_p.

Proof:
For each triangle $t \in \mathcal{T}$, the polynomial $P_t \in \Pi_m$ is given uniquely by

$$P_t \in \Pi_m : \qquad P_t(p) = v_p \qquad \forall p \in t \cap \mathcal{N}^{(m)}$$

As all corners are nodes, the adjacent triangles have the same value in these points. If two triangles t_1, t_2 share the same edge $e = t_1 \cap t_2$, $P_{t_1}(x) = P_{t_2}(x)$ $\forall x \in e = t_1 \cap t_2$ holds, as on each edge lie $m+1$ nodes. In this way

$$v(x) = P_t(x) \qquad \text{falls } x \in t$$

defines a function $v \in S^{(m)}$, which solves the interpolation problem, which thus is a solution to (6.2). The uniqueness follows by contradiction. □

In the case $m = 1$ we have that

$$\mathcal{N}^{(1)} = \{p \mid p \in \Omega \ \text{ is a corner of a triangle } \ t \in \mathcal{T}\} \quad .$$

Definition 6.6 *By solving the interpolation problem*

$$\boxed{\lambda_p \in S^{(m)} : \quad \lambda_p(q) = \delta_{pq} \qquad \forall q \in \mathcal{N}^{(m)} \qquad \textit{(Kronecker-}\delta\textit{)}}$$

for all $p \in \mathcal{N}^{(m)}$ one gets the nodal basis

$$\Lambda^{(m)} = \{\lambda_p \mid p \in \mathcal{N}^{(m)}\}.$$

of $S^{(m)}$. If p is on the boundary of t, $\lambda_p|_t \in \Pi_m$ is called formfunction.

Note:
To justify this notation we have to check whether $\Lambda^{(m)}$ is a basis at all. Obviously each $v \in S^{(m)}$ has the representation

$$v = \sum_{p \in \mathcal{N}^{(m)}} v(p)\lambda_p \quad .$$

The uniqueness follow from Corollary 6.5. In the linear case (m=1) the nodal base functions are illustrated in figure 6.3. ◁

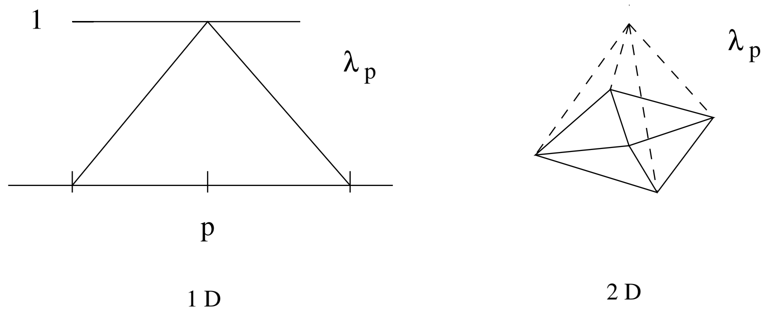


Figure 6.3: nodal basis functions for $m = 1$

We now consider the variational equality (6.1) in the case $H = H_0^1(\Omega)$, an $H_0^1(\Omega)$-elliptic bilinear form $a(\cdot,\cdot)$ an $l \in H'$. For example

$$a(v, w) = (\alpha\nabla v, \nabla w) = \int\limits_{\Omega} \alpha\nabla v \cdot \nabla w \; dx$$

with $0 < \alpha_0 \leq \alpha(x) \leq \alpha_1 < \infty$ a.e. on Ω, and

$$l(v) = (f, v) = \int\limits_{\Omega} f \; v \; dx$$

with $f \in L^2(\Omega)$ fulfill these assumptions. To approximate the solutions u of our model problem (6.1) we now choose

$$\boxed{S_h := \{v \in S^{(m)} \mid v|_{\partial\Omega} = 0\}} \tag{6.3}$$

The discretization parameter h,

$$\boxed{h = \max_{t \in \mathcal{T}_h} \operatorname{diam} t,}$$

describes the fineness of the mesh. Often h is called *stepsize.*

Theorem 6.7 *S_h is a closed subspace of $H_0^1(\Omega)$.*

Proof:
Exercise. □

Application of the Ritz–Galerkin method according to Theorem 6.1 provides the variational equality

$$u_h \in S_h : \qquad a(u_h, v) = l(v) \qquad \forall v \in S_h \quad . \tag{6.4}$$

Let

$$\mathcal{N}_h = \mathcal{N}^{(m)} \cap \Omega$$

denote the set of the *inner* nodes of $\mathcal{T}_h$, then apparently

$$\Lambda_h = \Lambda^{(m)} \cap H_0^1(\Omega) = \{\lambda_p \mid p \in \mathcal{N}_h\}$$

is the nodal basis of S_h, as the values on the boundary nodes are determined by the boundary conditions. Inserting of the nodal basis representation

$$\boxed{u_h = \sum_{p \in \mathcal{N}_h} u_p \lambda_p}$$

and of $v = \lambda_q$, $q \in \mathcal{N}_h$, provides a linear system of equations

$$\sum_{p \in \mathcal{N}_h} a(\lambda_p, \lambda_q) u_p = l(\lambda_q) \qquad \forall q \in \mathcal{N}_h$$

for the unknown coefficients $u_p = u(p)$. By defining

$$\begin{aligned} A &= (a_{p,q})_{p,q \in \mathcal{N}_h} & a_{p,q} &= a(\lambda_q, \lambda_p) \\ b &= (b_p)_{p \in \mathcal{N}_h} & b_p &= l(\lambda_p) \\ U &= (u_p)_{p \in \mathcal{N}_h} \end{aligned}$$

the variational problem (6.4) can be written as the equivalent system

$$\boxed{AU = b \quad .} \tag{6.5}$$

We now note some basic properties of A.

Theorem 6.8 *If the bilinear form $a(\cdot,\cdot)$ is elliptic, A positive definite. If the bilinear form $a(\cdot,\cdot)$ is symmetric, A is symmetric as well.*

Proof:
Exercise. □

We will come back to the solution of the (large!) system (6.5) after investigating the convergence of the method.

6.2 Error Estimates

6.2.1 Error Estimates in the H^1-Norm

We want to know, how the

$$\boxed{\textit{discretization error:} \ \ \|u - u_h\|_1}$$

behaves if we approximate $H_0^1(\Omega)$ by $S_h \subset S^{(m)}$ (see (6.3)). By theorem 6.1 it is enough to estimate the

$$\boxed{\textit{approximation error:} \ \inf_{v \in S_h} \|u - v\|_1}$$

We want to start with the easiest possible case with

$$m = 1, \qquad \Omega = (a, b) \subset \mathbb{R} \quad .$$

Then from $\mathcal{T}_h$ we get the *grid*

$$a = x_0 < x_1 < \ldots < x_{N-1} < x_N = b$$

and we set

$$t_i = [x_{i-1}, x_i], \qquad h_i = x_i - x_{i-1}, \qquad h = \max_{i=1,\ldots,N} h_i \quad .$$

S_h then denotes the space of the continuous, piecewise linear functions. The following methods are chosen such that they can be transferred to the two- and three-dimensional case.

Theorem 6.9 *For each $u \in H_0^1(\Omega)$ we have*

$$\inf_{v \in \mathcal{S}_h} \|u - v\|_1 \to 0$$

as $h \to 0$.

Proof:
Exercise. □

Note:
The theorem 6.9 together with the Céa Lemma 5.16 implies the convergence

$$\|u - u_h\|_1 \to 0$$

of the FE approximations u_h as $h \to 0$. Notice that there is no need for further regularity assumptions except those of the existence and uniqueness theorem 5.31! ◁

We now want to know, *how fast* the approximations converge. To acquire the appropriate sharper statements we need the

$$\boxed{\textit{regularity assumption:} \ u \in H^2(a, b).}$$

By the Sobolev embedding theorem 5.27 we have the continuous embedding

$$H^2(a, b) \subset C[a, b]$$

(This is also true in two or three dimensional spaces, but not in four!) Thus by corollary 6.5 the *interpolation operator* $I_h : H^2(\Omega) \to S_h$ given by

$$I_h v \in S_h : \qquad I_h v(x_i) = v(x_i) \qquad \forall i = 0, \ldots, N$$

is well defined. Apparently

$$\inf_{v \in S_h} \|u - v\|_1 \le \|u - I_h u\|_1$$

holds. Therefore it suffices to estimate the

$$\boxed{\textit{interpolation error:}\quad \|u - I_h u\|_1}$$

This is done in four steps.

- **Step 1: Localization.** By the linearity of integrals we have

$$\|u - I_h u\|_{1,\Omega}^2 = \sum_{i=1}^{N} \|u - I_h u\|_{1,t_i}^2$$

- **Step 2: Transformation to the unit interval.** For each fixed i we transform the unit interval $T = [0,1]$ affinely to t_i by

$$t_i = F_i(T), \qquad x = F_i(\xi) = x_{i-1} + h_i \xi, \quad \xi = F^{-1}(x) = h_i^{-1}(x - x_{i-1})$$

  and set

$$\hat{v}(\xi) = v(F_i(\xi)) = v(x) \qquad \forall v \in H^2(t_i) \quad .$$

  By the transformation rule for integrals

$$\|v\|_{0,t_i}^2 = \int_{x_{i-1}}^{x_i} v(x)^2\, dx = h_i \int_0^1 \hat{v}(\xi)^2\, d\xi = h_i \,\|\hat{v}\|_{0,T}^2 \quad .$$

  And the chain rule provides

$$\hat{v}'(\xi) = \frac{d\hat{v}}{d\xi}(\xi) = \frac{dv}{dx}(x)\frac{dx}{d\xi} = h_i v'(x)$$

  and thus

$$\|v'\|_{0,t_i}^2 = h_i^{-1}\,\|\hat{v}'\|_{0,T}^2 \quad .$$

  Therefore for $h_i \le 1$ it holds

$$\|v\|_{1,t_i}^2 \le h_i^{-1}\,\|\hat{v}\|_{1,T}^2 \qquad \forall v \in H^2(t_i) \quad . \tag{6.6}$$

  Strictly speaking, we showed these estimates only for functions differentiable in the classical sense (point-wise application of the chain rule).

  The validity of the above chain rule for all $v \in H^2(t_i)$ follows by a *density argument* (in detail in the next step).

- **Step 3: Local interpolation error.** Applying the transformation rule (6.6) to $v = u - I_h u$ provides

$$\|u - I_h u\|_{1,t_i}^2 \le h_i^{-1}\|\hat{u}_i - \hat{I}\hat{u}_i\|_{1,T}^2$$

where

$$\hat{I}v(x) = v(0) + (v(1) - v(0))x \qquad v \in H^2(T)$$

denotes the interpolation operator on the reference interval T. We now want to prove the estimate

$$\|v - \hat{I}v\|_{1,T}^2 \leq c\,\|v''\|_{0,T}^2 \qquad \forall v \in H^2(T) \tag{6.7}$$

with $c = \frac{1}{3}(8 + 2\sqrt{3})$. Proof method: We first show the estimates for functions in a dense subspace of smooth functions and expand their validity by a density argument. As we know $X = \{\varphi|_T \mid \varphi \in C_0^\infty(\mathrm{I\!R})\} \subset H^2(T)$ dense. Let $\varphi \in X$ arbitrary but fixed. Taylor expansion provides

$$\varphi(x) = \varphi(0) + \varphi'(0)x + \int_0^x (x - z)\varphi''(z)\,dz \quad .$$

Inserting this for $\varphi(x)$ and $\varphi(1)$ we get

$$\varphi(x) - \hat{I}\varphi(x) = \int_0^x (x - z)\varphi''(z)\,dz - x\int_0^1 (1 - z)\varphi''(z)\,dz \quad .$$

Using the Cauchy–Schwarz inequality, we get

$$\begin{aligned}
\|\varphi - \hat{I}\varphi\|_{0,T}^2 &= \int_0^1 \left(\int_0^x (x - z)\varphi''(z)\,dz - x\int_0^1 (1 - z)\varphi''(z)\,dz\right)^2 dx \\
&\leq \int_0^1 \left(\left(\int_0^x (x - z)^2\,dz\right)^{\frac{1}{2}} \left(\int_0^x \varphi''(z)^2\,dz\right)^{\frac{1}{2}}\right. \\
&\qquad \left. + x\left(\int_0^1 (1 - z)^2\,dz\right)^{\frac{1}{2}} \left(\int_0^1 \varphi''(z)^2\,dz\right)^{\frac{1}{2}}\right)^2 dx \\
&\leq 4\int_0^1 (1 - z)^2\,dz \int_0^1 \varphi''(z)^2\,dz \\
&= \frac{4}{3}\,\|\varphi''\|_{0,T}^2 \quad .
\end{aligned}$$

By known differentiation rules

$$\left(\varphi - \hat{I}\varphi\right)'(x) = (x - x)\varphi''(x) + \int_0^x \varphi''(z)\,dz - \int_0^1 (1 - z)\varphi''(z)\,dz.$$

From this we get similar to the above

$$\|\left(\varphi - \hat{I}\varphi\right)'\|_{0,T}^2 \leq (1 + \tfrac{1}{3}\sqrt{3})^2\|\varphi''\|_{0,T}^2.$$

Altogether

$$\|\varphi - \hat{I}\varphi\|_{1,T}^2 \leq \tfrac{1}{3}(8 + 2\sqrt{3})\|\varphi''\|_{0,T}^2$$

Now the density conclusion. Therefore let $v \in H^2(T)$ arbitrary but fixed. As $X \subset H^2(T)$, dense, there exists a sequence $(\varphi_k)_{k\in\mathbb{N}} \subset X$, which converges to v in $H^2(T)$, so

$$\|v - \varphi_k\|_{2,T} \to 0, \qquad k \to \infty \quad .$$

From the embedding result, theorem 5.27, we furthermore know that

$$\|\hat{I}v\|_{1,T} \leq \sqrt{5} \max_{x\in[0,1]} |v(x)| \leq C \, \|v\|_{2,T} \qquad \forall v \in H^2(T)$$

Together with the triangle inequality we see that

$$\begin{aligned} \|v - \hat{I}v\|_{1,T} &\leq \|\varphi_k - \hat{I}\varphi_k\|_{1,T} + \|v - \varphi_k\|_{1,T} + \|\hat{I}(\varphi_k - v)\|_{1,T} \\ &\leq \sqrt{c} \left\|\varphi_k''\right\|_{0,T} + (1 + C) \, \|v - \varphi_k\|_{2,T} \quad . \end{aligned}$$

holds. This, for $k \to \infty$, proofs the claim (6.7) true.

- **Step 4: Back-transformation.** As above we show by the chain rule that

$$\left\|\hat{u}_i''\right\|_{0,T}^2 = h_i^3 \left\|u''\right\|_{0,t_i}^2 \quad .$$

Now we can assemble everything to get the desired estimate

$$\begin{aligned} \|u - I_h u\|_{1,\Omega}^2 &= \sum_{i=1}^{N} \|u - I_h u\|_{1,t_i}^2 \\ &\leq \sum_{i=1}^{N} h_i^{-1} \|\hat{u}_i - \hat{I}\hat{u}_i\|_{1,T}^2 \\ &\leq c \sum_{i=1}^{N} h_i^{-1} \left\|\hat{u}_i''\right\|_{0,T}^2 \\ &= c \sum_{i=1}^{N} h_i^{-1} h_i^3 \left\|u''\right\|_{0,t_i}^2 \\ &\leq c h^2 \left\|u''\right\|_{0,\Omega}^2 \quad . \end{aligned}$$

Thereby supposing the *regularity assumption* $u \in H^2(\Omega)$, we obtain the *a priori error estimate*

$$\boxed{\|u - u_h\|_1 \leq \tilde{c}h\|u''\|_0 \leq \tilde{c}h\|u\|_2}$$

with $\tilde{c} = \sqrt{c}\Gamma/\gamma$. This result can be extended to the case $m \geq 1$ and $\Omega \subset \mathbb{R}^2$. Analogously to the transformation to the unit interval, the affine transformation F_t from $t \in \mathcal{T}_h$ onto the unit triangle with corners $(0, 0)$, $(1, 0)$ and $(0, 1)$ plays a central role in the proof. This transformation is singular if t degenerates to an interval. The degree of degeneration of a triangle $t \in \mathcal{T}_h$ is described by the ratio of the circumcircle r_t to the incircle ϱ_t. The worst possible ratio

$$\sigma_h = \max_{t \in \mathcal{T}_h} \frac{r_t}{\varrho_t} \tag{6.8}$$

is called *form regularity* or shortly *regularity* of $\mathcal{T}_h$.

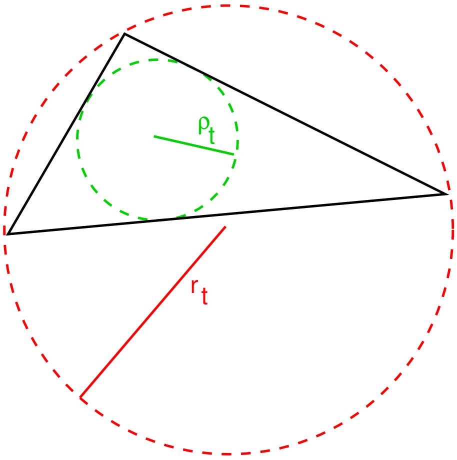


Figure 6.4: Incircle and circumcircle ϱ_t and r_t of a triangle.

Now we can formulate our central *approximation theorem.*

Theorem 6.10 *Let $m \geq 1$, h small enough and $u \in H_0^1(\Omega) \cap H^{m+1}(\Omega)$, $\Omega \subset \mathbb{R}^2$. Then the a priori estimate*

$$\|u - u_h\|_1 \leq c\, \sigma_h\, h^m\, |u|_{m+1}\,, \qquad |u|_{m+1} = \sum_{|\beta|=m+1} \|\partial^\beta u\|_0$$

holds with σ_h from (6.8)*.*

Proof:
The proof is analogous to the 1-D case. For the case m=1 the affine transformation is technically more demanding. To handle the general case $m > 1$, we have to estimate the interpolation error on the reference triangle more neatly (Keyword: Bramble–Hilbert-Lemma). A more elaborate description can be found in chapter II, paragraph 6 in the book of Braess [6]. □

If we now want to approximate the solution $u \in H_0^1(\Omega) \cap H^m(\Omega)$ by a *sequence of triangulations* $\mathcal{T}_h$, $h \in \mathcal{H} = \{h_1 > h_2 > \ldots\}$, we have to take care that σ_h remains bounded.

Definition 6.11 *Let $\mathcal{T}_h$, $h \in \mathcal{H} = \{h_1 > h_2 > \ldots\}$ be a family of triangulations with decreasing step–size. We call the sequence $(\mathcal{T}_h)_{h \in \mathcal{H}}$ regular if there exists a number $\sigma > 0$, such that*

$$\sup_{h \in \mathcal{H}} \sigma_h \leq \sigma$$

holds.

Apparently $(\mathcal{T}_h)_{h\in\mathcal{H}}$ is regular if and only if the interior angles of the triangulations $\mathcal{T}_h$ remain uniformly bounded from below. Then

$$\|u-u_h\|_1 = \mathcal{O}(h^m)$$

if $u \in H_0^1(\Omega)\cap H^{m+1}(\Omega)$. Caution: Our proof does *not* give

$$u \in H_0^1(\Omega) \quad \Longrightarrow \quad \|u-u_h\|_0 = \mathcal{O}\left(h\right) \quad .$$

6.2.2 Error Estimates in the L^2-Norm

Let $m=1$ from here on. From theorem 6.10 we can deduce for the L^2-error of the approximation, that it is of the order h. We now show with another technique that given certain conditions

$$u \in H^2(\Omega) \quad \Longrightarrow \quad \|u-u_h\|_0 = \mathcal{O}\left(h^2\right)$$

holds.

Theorem 6.12 *Let $\mathcal{T}_h$, $h\in\mathcal{H}$ be a regular family of triangulations. Let the so called dual problem*

$$w \in H_0^1(\Omega): \quad a(v,w) = (g,v) \qquad \forall v \in H_0^1(\Omega) \tag{6.9}$$

be H^2-regular, i.e. $\|w\|_2 \le c\,\|g\|_0$. Then we have

$$\boxed{\|u-u_h\|_0 \le ch\,\|u-u_h\|_1\,.}$$

If the original problem is H^2-regular, we even have

$$\boxed{\|u-u_h\|_0 \le ch^2\,\|f\|_0\,.}$$

Proof:
(Nitsche–Trick, Aubin(67)–Nitsche(68)-Lemma)

1) Dual Problem:
Solve the dual problem (6.9) for $g = u-u_h$. Then follows

$$a(v,w) = (u-u_h, v) \quad \forall v \in H_0^1(\Omega) \quad .$$

Finite element approximation of w

$$w_h \in S_h: \quad a(v,w_h) = (u-u_h,v) \qquad \forall v \in S_h \quad .$$

Error estimate ($w \in H^2(\Omega)\cap H_0^1(\Omega)$ holds by assumption)

$$\|w-w_h\|_1 \le ch|w|_2 \le ch\,\|g\|_0 = ch\,\|u-u_h\|_0 \quad .$$

2) Orthogonality:
As is well known

$$a(u-u_h,v) = 0 \qquad \forall v \in S_h \quad .$$

3) Inserting of $v = u - u_h$:

$$\begin{aligned} \|u-u_h\|_0^2 &= (u-u_h, u-u_h) = a(u-u_h, w) \\ &= a(u-u_h, w-w_h) && \text{(Orthogonality)} \\ &\le \Gamma \|u-u_h\|_1 \|w-w_h\|_1 \\ &\le \Gamma \|u-u_h\|_1 \, ch \, \|u-u_h\|_0 \quad , \end{aligned}$$

and the first part of the claim holds. The second follows directly from

$$\|u-u_h\|_1 \le ch \, \|f\|_0$$

for H^2-regular problems. □

6.2.3 Adaptive Step-Size

We shortly want to discuss the construction of a sequence of regular triangulations $(\mathcal{T}_h)_{h\in\mathcal{H}}$. Therefore we first need to construct a

$$\text{starting triangulation } \mathcal{T}_0$$

Given an easy geometry (unit square) we can do this by hand. For more complex cases there are standard techniques like for example the advancing front methods or Delaunay Triangulation. Details can be found in George [10, 11]. We don't want to conceal that these techniques give unsatisfying results (too many nodes) on special applications with multi-scale domains (semiconductors, humans with veins, porous media with gaps, ...). Especially in three space dimensions many questions are still open.

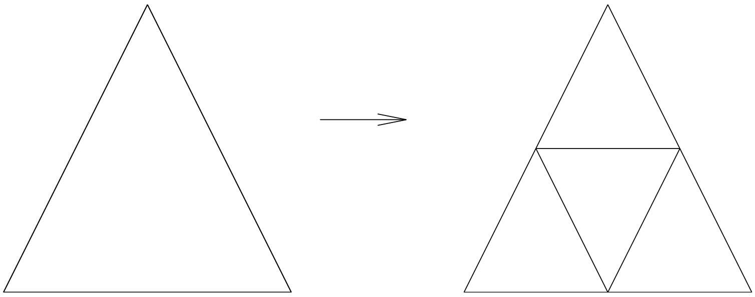


Figure 6.5: regular refinement of a triangle

We want to refine the starting triangulation. Therefore we want to divide a triangle into 4 similar triangles (see figure 6.5). Thereby the interior angles don't change. That's why this is called a regular refinement.

In the case that not all triangles should be divided one has to stop somewhere. To avoid "hanging nodes", one can introduce *irregular closures*, as shown in figure. 6.6. Apparently the interior angles are cut in half.

Continued bisection leads to degenerate triangles (figure 6.7). That's why R.E. Bank suggested already in the beginning of the 80's to reverse the irregular refinements before each new refinement step and avoid multiple irregular refinements of each triangle.

Altogether we thus get the following

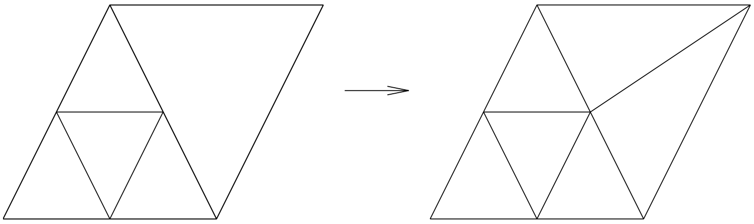


Figure 6.6: irregular closure

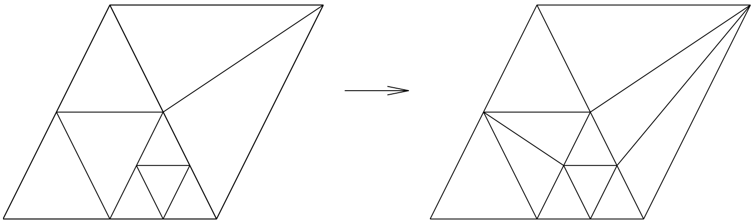


Figure 6.7: degenerate triangles by multiple irregular closures

Refinement algorithm (cf. fig. 6.8)
Given the triangulation $\mathcal{T}_k$, $k \geq 0$,

1) If $k > 0$, delete all irregular closures
2) Mark a subset of $\mathcal{T}_h$ for refinement
3) Refine all marked triangles regularly
4) Refine all triangles with more than one refined edge or with a twice refined edge regularly
5) Refine all triangles with refined neighbor irregular

This algorithms terminates (exercise). Of course, one wants to mark the triangles such that

the best possible approximation accuracy
with the lowest possible number of nodes

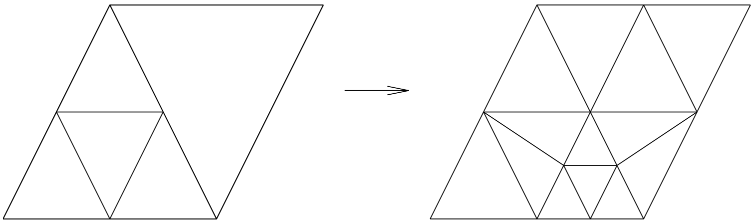


Figure 6.8: refinement without multiple irregular closures

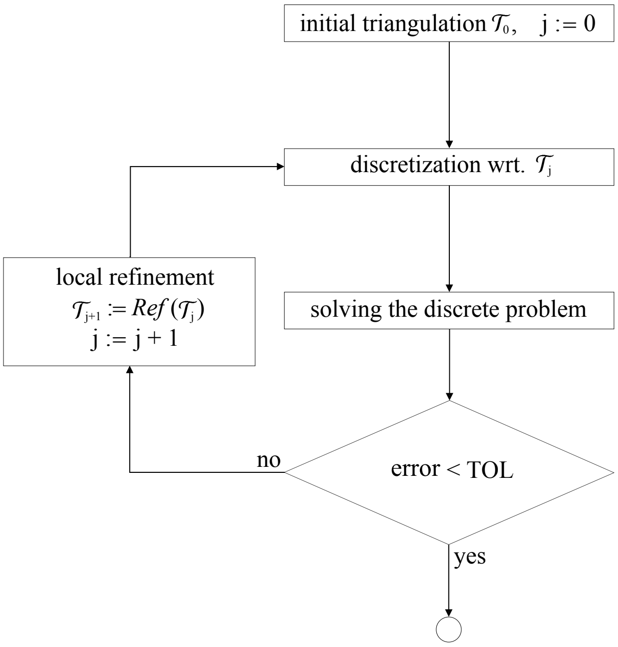

initial triangulation $\mathcal{T}_0$, j := 0
discretization wrt. $\mathcal{T}_j$
solving the discrete problem
local refinement
$\mathcal{T}_{j+1} := Ref(\mathcal{T}_j)$
j := j + 1
error < TOL
no
yes

Figure 6.9: adaptive finite element method

is reached. Thus, we want to refine there where it pays off.
To choose the corresponding triangles we need suitable *refinement indicators*. One possible indicator is the

$$\text{local error: } \|u - \tilde{u}\|_{1,t},$$

where $\tilde{u} \in S_k$ typically represents an approximation of $u_h \in S_k$, for example by approximation the linear system. Of course, the exact local error isn't accessible numerically, and has to be replaced by suitable a posteriori estimates, which can be calculated on the basis of $\tilde{u}$. We refer to chapter III, paragraph 7 in the book of Braess [6]. The whole adaptive solving process is illustrated in figure 6.9.

6.3 Condition Number of the Stiffness Matrix and Fourier Method

We again consider the model problem

$$u \in H_0^1(\Omega): \quad a(u,v) = l(v) \qquad \forall v \in H_0^1(\Omega) \quad , \tag{6.10}$$

and assume that $l \in (H_0^1(\Omega))'$ and $a(\cdot,\cdot)$ is elliptic and *symmetric*, i.e.

$$a(v,w) = \int_\Omega \alpha \nabla v \cdot \nabla w \; dx \quad .$$

Discretization with *linear* finite elements with respect to a sequence of *regular* triangulations $(\mathcal{T}_h)_{h\in\mathcal{H}}$ leads to the following discrete problem

$$u_h \in S_h : \quad a(u_h, v) = l(v) \qquad \forall v \in S_h \quad . \tag{6.11}$$

This can be written as a linear equation system

$$AU = b \tag{6.12}$$

for the unknown coefficient-vector

$$U = (u_p)_{p\in\mathcal{N}_h}$$

with

$$u_h = \sum_{p\in\mathcal{N}_h} u_p\lambda_p \quad .$$

Thereby $\Lambda_h = \{\lambda_p,\, p \in \mathcal{N}_h\}$ is the nodal basis. As we know stiffness matrix A and right hand side b are given by

$$A = (a(\lambda_p, \lambda_q))_{p,q\in\mathcal{N}_h} \ , \ b = (l(\lambda_p))_{p\in\mathcal{N}_h} \quad .$$

From here on we use the canonical isomorphism

$$v = \sum_{p\in\mathcal{N}_h} v_p\lambda_p \ \rightarrow \ \underline{v} = (v_p)_{p\in\mathcal{N}_h} \in \mathrm{I\!R}^{n_h}$$

and denote the Euclidean scalar product by

$$\langle \underline{v}, \underline{w} \rangle = \sum_{p\in\mathcal{N}_h} v_p w_p \quad .$$

Then the stiffness matrix A has the property

$$a(w, v) = \langle A\underline{w}, \underline{v} \rangle \qquad \forall v \in S_h$$

and $b \in \mathrm{I\!R}^{n_h}$ is the corresponding representation of l

$$l(v) = \langle b, \underline{v} \rangle \quad \forall v \in S_h \quad .$$

If there is no danger of confusion, we forego the underline under the vectors $\underline{v} \in \mathrm{I\!R}^{n_h}$.

Goal: Determine *with the least amount of computer operations* and *the least amount of memory* a $\tilde{u} \in S_h$ or, equivalent, a $\underline{\tilde{u}} \in \mathrm{I\!R}^{n_h}$, such that

$$\|\tilde{u} - u_h\|_1 \leq c\,h \quad . \tag{6.13}$$

Note:
An optimal procedure would reach (6.13) after $\mathcal{O}(n_h)$ operations using $\mathcal{O}(n_h)$ memory, as this would be the effort for a diagonal matrix. ◁

Note:
Naïve use of the Gaussian algorithm needs $\mathcal{O}\left(n_h^3\right)$ operations. ◁

Before we tackle the development of special adapted linear solvers, we want to study some basic properties of A. In advance we remind ourselves some basis linear algebra.

Lemma 6.13 *Let $A \in \mathrm{I\!R}^{n,n}$ be symmetric and positive definite (s.p.d.). Then there exists an orthogonal basis of $\mathrm{I\!R}^n$ consisting of eigenvectors e_i with positive eigenvalues μ_i, $i = 1, \ldots, n$. Furthermore there exists a s.p.d. matrix $A^{\frac{1}{2}} \in R^{n,n}$ with the property*

$$A^{\frac{1}{2}} A^{\frac{1}{2}} = A \quad .$$

Proof:
We only show the existence of $A^{\frac{1}{2}}$. Let $T = (e_1 \ldots e_n)$ (column-wise). Then follows

$$D = T^{-1}AT, \qquad D = \mathrm{diag}(\mu_1, \ldots, \mu_n), \qquad \mu_i > 0,\ i = 1, \ldots, n \quad .$$

We set $D^{\frac{1}{2}} := (\mu_1^{\frac{1}{2}}, \ldots, \mu_n^{\frac{1}{2}})$. Obviously $D^{\frac{1}{2}} D^{\frac{1}{2}} = D$ holds. Finally we set

$$A^{\frac{1}{2}} = TD^{\frac{1}{2}}T^{-1}$$

and the claim follows. □

Lemma 6.14 *Let $B \in \mathrm{I\!R}^{n,n}$ be s.p.d. Then*

$$\langle v, w \rangle_B = \langle Bv, w \rangle$$

defines a scalar product on $\mathrm{I\!R}^n$. If furthermore $A \in \mathrm{I\!R}^{n,n}$ is regular and symmetric with respect to $\langle \cdot, \cdot \rangle_B$, i.e.

$$\langle Av, w \rangle_B = \langle v, Aw \rangle_B \quad ,$$

we have

$$\mu_{\max}(A) = \max_{v \neq 0} \frac{\langle Av, v \rangle_B}{\langle v, v \rangle_B}$$

$$\mu_{\min}(A) = \min_{v \neq 0} \frac{\langle Av, v \rangle_B}{\langle v, v \rangle_B} \quad .$$

Proof:
The first statement is trivial.
We first consider the case $B = I$ (unity matrix). Using the orthonormal basis representation

$$v = \sum_{i=1}^{n} v_i e_i$$

from Lemma 6.13 we get

$$\frac{\langle Av, v \rangle}{\langle v, v \rangle} = \frac{\sum_{i=1}^{n} \mu_i v_i^2 \langle e_i, e_i \rangle}{\langle v, v \rangle} \leq \mu_{\max}(A) \quad .$$

The same follows for $v = e_{\max}$. The representation of $\mu_{\min}(A)$ follows from

$$\frac{\langle Av, v \rangle}{\langle v, v \rangle} = \frac{\langle w, w \rangle}{\langle A^{-1}w, w \rangle} \geq \frac{1}{\mu_{\max}(A^{-1})} = \mu_{\min}(A)$$

with $w = A^{\frac{1}{2}}v$.
Let now B be s.p.d. From the symmetry of a A w.r.t. $\langle \cdot, \cdot \rangle_B$ follows

$$\langle B^{\frac{1}{2}} A B^{-\frac{1}{2}} v, w \rangle = \langle A B^{-\frac{1}{2}} v, B^{-\frac{1}{2}} w \rangle_B = \langle B^{-\frac{1}{2}} v, A B^{-\frac{1}{2}} w \rangle_B = \langle v, B^{\frac{1}{2}} A B^{-\frac{1}{2}} w \rangle \quad ,$$

thus the symmetry of $B^{\frac{1}{2}} A B^{-\frac{1}{2}}$ w.r.t. the Euclidean scalar product. This matrix has the same Eigenvalues as A and

$$\frac{\langle Av, v \rangle_B}{\langle v, v \rangle_B} = \frac{\langle B^{\frac{1}{2}} A B^{-\frac{1}{2}} w, w \rangle}{\langle w, w \rangle}$$

with $w = B^{\frac{1}{2}}v$. Now follow the claims from the corresponding statements in the case $B = I$. □

Note:
The quotient

$$\frac{\langle Av, v\rangle_B}{\langle v, v\rangle_B}$$

is called *Rayleigh Quotient.* ◁

From now on A denotes the stiffness matrix again (w.r.t. the nodal basis).

Theorem 6.15 *The stiffness matrix A is* symmetric *and* positive definite. *The number of from non-zero elements in each row of A is uniformly bounded by h (A is* sparse*).*

Proof:
Exercise. □

Conclusion 6.16 *The evaluation of Av needs $\mathcal{O}(n_h)$ dot-operations.*

Theorem 6.17 *The condition number $\kappa(A)$ satisfies the estimates*

$$\boxed{\frac{1}{o(1)} \leq \kappa(A) = \frac{\mu_{\max}(A)}{\mu_{\min}(A)} \leq ch^{-2} \quad .} \tag{6.14}$$

Proof:

1) It holds

$$\mu_{\max}(A) = \max_{\underline{v}\in\mathbb{R}^{n_h}} \frac{\langle A\underline{v}, \underline{v}\rangle}{|\underline{v}|^2} = \max_{v\in S_h} \frac{a(v,v)}{|\underline{v}|^2} \leq \Gamma \frac{\|v\|_1^2}{|\underline{v}|^2} \leq c \quad ,$$

because of

$$\|v\|_{1,t}^2 \leq c\,\|\hat{v}\|_{1,T}^2 \leq C(v^2(p_1) + v^2(p_2) + v^2(p_3)) \quad ,$$

where $t = (p_1, p_2, p_3)$. Summing up gives (due to the regularity of $\mathcal{T}_h$!) the above estimate.

2) It holds

$$\mu_{\min}(A) = \min_{\underline{v}\in\mathbb{R}^{n_h}} \frac{\langle A\underline{v}, \underline{v}\rangle}{|\underline{v}|^2} = \min_{v\in S_h} \frac{a(v,v)}{|\underline{v}|^2} \geq \gamma \frac{\|v\|_1^2}{|\underline{v}|^2} \quad .$$

For each $t = (p_1, p_2, p_3) \in \mathcal{T}$ we have

$$v^2(p_1) + v^2(p_2) + v^2(p_3) \leq c\,\|\hat{v}\|_{L^2(T)}^2 \leq c\,h^{-2}\,\|v\|_{L^2(t)}^2 \leq C\,h^{-2}\,\|v\|_{1,t}^2 \quad ,$$

and summing up gives

$$|\underline{v}|^2 \leq c\,h^{-2}\,\|v\|_1^2 \quad ,$$

so

$$\mu_{\min}(A) \geq ch^2 \quad .$$

3) As contradiction to the proposition we assume there exists a $\varepsilon > 0$ such that

$$a(v,v) \geq \varepsilon\,|\underline{v}|^2 \qquad \forall v \in S_h\ ,\ \forall h \in \mathcal{H}$$

holds. Then we choose $\varphi \in C_0^\infty(\Omega) \cap H_0^1(\Omega)$ such that for a circle $K \subset \Omega$ with positive radius

$$\varphi(x) \equiv 1 \quad \forall x \in K$$

holds and set $v := I_h\varphi \in S_h$ (interpolation). Then

$$|\underline{v}|^2 \geq |\mathcal{N}_h \cap K| \to \infty \quad \text{für} \quad h \to \infty \quad ,$$

but

$$a(v,v) \leq \Gamma\,\|I_h\varphi\|_1^2 \leq 2\Gamma\,(\|I_h\varphi - \varphi\|_1^2 + \|\varphi\|_1^2) \leq \text{const.} \quad \forall h \in \mathcal{H} \quad .$$

which is a contradiction. □

Note:
The above estimate (6.14) is sharp. Below we will provide an example for which $\kappa(A) = \mathcal{O}(h^{-2})$ holds. ◁

Note:
Theorem 6.17 implies particularly that when solving our equation system with $h \to 0$ the rounding errors can become arbitrary big. This poses serious difficulties to directly solving (6.12). ◁

It is of vital importance that the condition number of the stiffness matrix depends on the choice of the basis of S_h.

Theorem 6.18 *Let $\hat{\Lambda}_h = \{\hat{\lambda}_p \,|\, p \in \mathcal{N}_h\}$ be a basis of S_h and*

$$v = \sum_{p\in\mathcal{N}_h} \hat{v}_p\hat{\lambda}_p \to \underline{\hat{v}} = (\hat{v}_p)_{p\in\mathcal{N}_h}$$

the corresponding isomorphism, $\hat{A}$ the stiffness matrix corresponding to $\hat{\Lambda}_h$, i.e.

$$a(v,w) = \langle \hat{A}\underline{\hat{v}}, \underline{\hat{w}}\rangle \qquad \forall v \in S_h \quad .$$

Let T^T be transformation matrix from Λ_h to $\hat{\Lambda}_h$, i.e.

$$\hat{\lambda}_p = \sum_{q\in\mathcal{N}_h} T_{qp}\lambda_q \quad .$$

Then we have

$$\hat{A} = T^T AT \quad .$$

Proof:
Inserting gives

$$\begin{aligned} v = \sum_{p\in\mathcal{N}_h} \hat{v}_p\hat{\lambda}_p &= \sum_{p\in\mathcal{N}_h} \hat{v}_p \sum_{q\in\mathcal{N}_h} T_{qp}\lambda_q \\ &= \sum_{q\in\mathcal{N}_h} \left(\sum_{p\in\mathcal{N}_h} T_{qp}\hat{v}_p\right)\lambda_q \quad , \end{aligned}$$

so

$$\underline{v} = T\underline{\hat{v}} \quad .$$

By the representation of $\hat{A}$ follows

$$\langle \hat{A}\underline{\hat{v}}, \underline{\hat{w}}\rangle = a(v,w) = \langle A\underline{v}, \underline{w}\rangle = \langle T^T AT\underline{\hat{v}}, \underline{\hat{w}}\rangle$$

and thereby the claim. □

As A is symmetric and positive definite, there exists an orthonormal basis of eigenvectors $\underline{e}_p$, to the eigenvalues $\mu_p, p \in \mathcal{N}_h$ of A. Due to

$$a(e_p, e_q) = \langle A\underline{e}_p, \underline{e}_q\rangle = \mu_p\langle \underline{e}_p, \underline{e}_q\rangle$$

the corresponding functions $e_p \in S_h$, $p \in \mathcal{N}_h$ are orthogonal respective to the energy scalar product $a(\cdot,\cdot)$, or short, a-orthogonal.

Corollary 6.19 *Choosing the a-orthonormal basis of S_h*

$$\hat{\lambda}_p = \frac{1}{\sqrt{\mu_p}} e_p$$

for the representation of $u_h \in S_h$, one gets

$$\boxed{\hat{A} = I \quad .}$$

Proof:
It holds

$$\underline{a}_{p,q} = a(\hat{\lambda}_p, \hat{\lambda}_q) = \langle \frac{1}{\sqrt{\mu_p}} A \underline{e}_p, \frac{1}{\sqrt{\mu_q}} \underline{e}_q \rangle = \delta_{p,q} \quad .$$

In special cases the eigenvalues μ_p and eigenvectors e_p can be calculated explicitly. Therefore we consider

$$a(v, w) = \int\limits_{\Omega} \nabla v \cdot \nabla w \; dx$$

with $\Omega = (0,1) \times (0,1)$ and $\mathcal{T}_h$ as in figure 6.10. Step-size $h = \frac{1}{m}$, number of unknowns $n_h = (m-1)^2$.

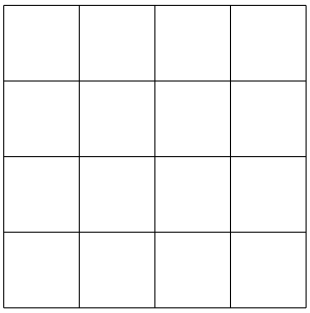


Figure 6.10: unit square Ω with tensor grid $\mathcal{T}_h$

We enumerate the nodal points

$$p_{ij} = (ih, jh)$$

row-wise. Therefore the unknown vector U becomes

$$U = (u_p)_{p \in \mathcal{N}_h} = (u_{p_{ij}})_{i,j=1 \ldots m-1} = \begin{pmatrix} U_1 \\ \vdots \\ U_{m-1} \end{pmatrix}$$

with vectors $U_j = (u_{p_{ij}})_{i=1,\ldots,m-1}$.
The stiffness matrix A has a block-tridiagonal shape

$$A = \begin{pmatrix} A_m & -I_m & & & \\ -I_m & A_m & -I_m & & \\ & & \ddots & & \\ & & -I_m & A_m & -I_m \\ & & & -I_m & A_m \end{pmatrix} \tag{6.15}$$

with the $(m-1)\times(m-1)$-matrices

$$A_m = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & & \ddots & & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix} \qquad I_m = \begin{pmatrix} 1 & 0 & & & \\ 0 & 1 & 0 & & \\ & & \ddots & & \\ & & 0 & 1 & 0 \\ & & & 0 & 1 \end{pmatrix}$$

Theorem 6.20 *The eigenvectors e_{ij} and the corresponding eigenvalues μ_{ij} of A are given by*

$$\begin{aligned}(e_{ij})_{lk} &= \sin(i\pi lh)\cdot\sin(j\pi kh)\\ \mu_{ij} &= 4\left(\sin^2(i\frac{\pi}{2}h)+\sin^2(j\frac{\pi}{2}h)\right) \quad .\end{aligned}$$

Proof:
Let $\omega^2=-1$. Using the de Moivre's formula we get

$$\begin{aligned}\sin\varphi &= \frac{1}{2\omega}\left(e^{\omega\varphi}-e^{-\omega\varphi}\right)\\ \cos\varphi &= \frac{1}{2}\left(e^{\omega\varphi}+e^{-\omega\varphi}\right)\end{aligned}$$

Furthermore

$$1-\cos\varphi = 2\sin^2\frac{\varphi}{2} \quad .$$

Because of

$$\begin{aligned}-e^{\omega i\pi(l-1)h}+2e^{\omega i\pi lh}-e^{\omega i\pi(l+1)h} &= e^{\omega i\pi lh}\left(-e^{-\omega i\pi h}+2-e^{\omega i\pi h}\right)\\ &= 2e^{\omega i\pi lh}\left(1-\cos(i\pi h)\right)\\ &= 4e^{\omega i\pi lh}\sin^2(i\frac{\pi}{2}h)\end{aligned}$$

it holds that

$$-\sin(i\pi(l-1)h)+2\sin(i\pi lh)-\sin(i\pi(l+1)h) = 4\sin^2(i\frac{\pi}{2}h)\sin(i\pi lh) \quad .$$

Particularly are

$$e_i = \left(\sin(i\pi lh)\right)_{l=1,\dots,m-1}\in\mathbb{R}^{m-1}, \qquad \mu_i = 4\sin^2(i\frac{\pi}{2}h)$$

eigenvectors and eigenvalues of the $(m-1)\times(m-1)$-matrix B

$$B = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \quad .$$

We write e_{ij} as a block vector

$$e_{ij} = \begin{pmatrix} E_1 \\ \vdots \\ E_{m-1}\end{pmatrix} \quad , \quad E_k = \sin(j\pi kh)\cdot e_i \quad .$$

Block-wise evaluation of Ae_{ij} then gives for the k-th block

$$\begin{aligned}(Ae_{ij})_k &= -E_{k-1}+A_mE_k-E_{k+1}\\ &= -E_{k-1}+2E_k+BE_k-E_{k+1}\\ &= \left(-\sin(j\pi(k-1)h)+2\sin(j\pi kh)-\sin(j\pi(k+1)h)\right)e_i+\sin(j\pi kh)Be_i\\ &= 4\sin^2(j\frac{\pi}{2}h)\sin(j\pi kh)e_i+4\sin^2(i\frac{\pi}{2}h)\sin(j\pi kh)e_i\\ &= 4\left(\sin^2(i\frac{\pi}{2}h)+\sin^2(j\frac{\pi}{2}h)\right)E_k \quad .\end{aligned}$$

What is exactly the claim. □

Note:

The condition number of the stiffness matrix A from (6.15) is

$$\boxed{\kappa(A) = \frac{\mu_{\max}}{\mu_{\min}} = \frac{\mu_{m-1,m-1}}{\mu_{1,1}} = \mathcal{O}(h^{-2}) \quad .}$$

Note:

Theorem 6.20 gets intuitively accessible by the fact that the eigenfunctions of the Laplace operator

$$-\Delta\varphi_{ij} = \left((i\pi)^2 + (j\pi)^2\right)\varphi_{ij}$$

on $\Omega = (0,1) \times (0,1)$ with zero-boundary conditions are given by

$$\varphi_{ij}(x,y) = \sin(i\pi x) \cdot \sin(j\pi y) \qquad i,j \in \mathbb{N}$$

These eigenvectors of A are in this case exactly the point-wise restrictions of the corresponding eigenfunctions.

Furthermore, it holds:

$$\frac{1}{h^2}\mu_{ij} \to (i\pi)^2 + (j\pi)^2 \quad \text{für} \quad h \to 0 \quad .$$

Proof as an exercise. ◁

Note:

Note that the eigenfunctions differ strongly in their *frequency*. There are *low-frequency* eigenvectors $(i,j \ll \frac{m}{2})$ and *high-frequency* eigenvectors $(i,j \gg \frac{m}{2})$ and everything in between. ◁

Using theorem 6.20 we can state the exact solution u_j directly.

Theorem 6.21 *Let $b = (b_{ij})_{i,j=1,\dots,m-1}$ be the right-hand side of our equation system. Then we have*

$$\boxed{u_{p_{ij}} = \sum_{l,k=1}^{m-1} \frac{(e_{lk})_{ij}}{|e_{lk}|}\mu_{lk}^{-1}\hat{b}_{lk} \quad , \qquad \hat{b}_{lk} := \sum_{r,s=1}^{m-1} \frac{(e_{lk})_{rs}}{|e_{lk}|} b_{rs} \quad .}$$

Proof:

By composing the transformation matrix T from the column-wise entries of the normed eigenvectors $\frac{(e_{lk})}{|e_{lk}|}$, it follows

$$\left(T^T A T\right) T^T U = T^T b \quad .$$

Therefore by $T^T A T = D := \operatorname{diag}(\mu_{ij})$

$$U = T D^{-1} T^T b \quad .$$

This is the matrix representation of the above equation. □

Note:
For the evaluation of the solution according to Theorem 6.21 using a naïve implementation of the summation ,one needs $\mathcal{O}\left(n_h^2\right)$ operations. That is because the matrix $TD^{-1}T^T$ is fully populated. But the sums from theorem 6.21 can be calculated with only $\mathcal{O}\left(\log n_h\right)$ operations using the the *fast Fourier transform* (FFT). Note for example the difference between $n_h = 10^6$ and $\log n_h = 6$. For the details we refer to [17]. Thereby we get a total effort of

$$\boxed{\text{Number of point operations} = \mathcal{O}\left(n \log n\right)}$$

to solve the equation system. The just described solution method is called *Fourier method.*◁

Note:
Under the assumption, that *the secondary diagonal blocks are exchangeable with the main diagonal blocks*, the Fourier method can be extended to other block-tri-diagonal equation systems. Unfortunately this isn't the case for non-constant coefficients $\alpha(x)$. ◁

So in general we don't know the eigenvectors $\underline{e}_p$ of A. Regarding our model problem we saw that the **a-orthogonal finite element functions $\mathbf{e_p} \in \mathbf{S_h}$ represent a scale of frequencies**. This is, as one can proof mathematically and understand physically, always the case. Thus, being interested in a-orthogonal or at least "almost" a-orthogonal functions, one should try those functions which cover a preferably big scale of frequencies.
This is the main idea of multi-grid methods.

6.4 Multi-Grid Methods

We still consider the model problem (6.10) with the *symmetric* bilinear form

$$a(v, w) = \int_\Omega \alpha \nabla v \cdot \nabla w \, dx \quad .$$

Grid hierarchy. We assume about the discretization

$$u_h \in S_h : \quad a(u_h, v) = l(v) \qquad \forall v \in S_h \tag{6.16}$$

that the triangulation $\mathcal{T}_h = \mathcal{T}_{h_j}$ was formed by j refining steps from a starting triangulation $\mathcal{T}_{h_0}$. For simplicity we assume a regular refinement. Therefore

$$h_j = \mathcal{O}(2^{-j}), \qquad h_{k-1} = 2h_k \quad k = 1, \ldots, j \quad . \tag{6.17}$$

We write shortly

$$\mathcal{T}_k = \mathcal{T}_{h_k}, \qquad k = 0, \ldots, j \quad .$$

Corresponding to the nested sequence of triangulations

$$\mathcal{T}_0 \subset \mathcal{T}_1 \subset \cdots \subset \mathcal{T}_j$$

with node set

$$\mathcal{N}_0 \subset \mathcal{N}_1 \subset \cdots \subset \mathcal{N}_j$$

is the nested sequence of finite element spaces

$$S_0 \subset S_1 \subset \cdots \subset S_j = S_{h_j}$$

with node bases

$$\Lambda_k = \{\lambda_p^{(k)} \mid p \in \mathcal{N}_k\}$$

on each refinement level $k = 0, \ldots, j$. Due to the uniform refinement the number of nodes $n_k = \#\mathcal{N}_k$ grows geometrically, i.e. there exists a $q > 1$ with

$$n_k \geq q n_{k-1} \qquad k = 1, \ldots, j. \tag{6.18}$$

As is well known, the discrete problem (6.16) is equivalent to the minimization problem

$$u_j \in S_j : \qquad J(u_j) \leq J(v) \qquad \forall v \in S_j \tag{6.19}$$

with

$$J(v) = \tfrac{1}{2} a(v,v) - l(v) \qquad v \in S_j \quad .$$

and $u_j = u_{h_j}$.

Subspace correction methods. We now want to construct iterative solution methods to (6.19). Therefore we replace the "large" minimization problem (6.19) by a sequence of "smaller" minimization problems. For this purpose we choose a partition

$$S_j = V_0 + V_1 + \cdots + V_m \tag{6.20}$$

of subspaces

$$V_k \subset S_j, \qquad k = 0, \ldots, m \quad .$$

Successive minimization of the energy J leads to the following iterative method to calculate the new iterate $u_j^{\nu+1} \in S_j$ from a given iterate $u_j^\nu \in S_j$.

Algorithm 6.1 (Successive Minimization)
given: $w_{-1} = u_j^\nu \in S_j$.

for $k = 0, \ldots, m$ *solve:*

$$v_k \in V_k : \quad J(w_{k-1} + v_k) \leq J(w_{k-1} + v) \;\; \forall v \in V_k, \qquad w_k = w_{k-1} + v_k, \tag{6.21}$$

new iterate: $u_j^{\nu+1} = w_m$.

Thus each partition (6.20) directly gives a corresponding iterative method. We will now try to choose the partition (6.20) such that the corresponding iterative method converges as quickly as possible and can be implemented with effort $\mathcal{O}(n_j)$.

Gauss–Seidel method. Obviously one can solve (6.21) without difficulties exactly, when choosing one dimensional subspaces

$$V_l = \text{spann}\{\lambda_l\}, \qquad l = 1, \dots, m,$$

A natural partition then is

$$S_j = \sum_{l=1}^{n_j} V_l, \qquad V_l = \text{spann}\{\lambda_{p_l}^{(j)}\}, \qquad l = 1, \dots, n_j. \tag{6.22}$$

The resulting iterative method is

$$\boxed{u_j^{\nu+1} = u_j^\nu + \sum_{l=1}^{n_j} v_l, \qquad v_l = \frac{l(\lambda_{p_l}^{(j)}) - a(w_{l-1}, \lambda_{p_l}^{(j)})}{a(\lambda_{p_l}^{(j)}, \lambda_{p_l}^{(j)})} \lambda_{p_l}^{(j)},} \tag{6.23}$$

where $w_0 = u_j^\nu$ and $w_l = w_{l-1} + v_l$, $l = 1, \dots, n_j - 1$.

Theorem 6.22 *The iterative method* (6.23) *is the Gauss–Seidel method. The correction*

$$v^{(j)} = v_1 + \dots + v_{n_j} = u_j^{\nu+1} - u_j^\nu$$

is the solution of the variational problem

$$v^{(j)} \in S_j : \qquad b_j(v^{(j)}, v) = l(v) - a(u_j^\nu, v) \qquad \forall v \in S_j \tag{6.24}$$

where

$$b_j(v, w) = \sum_{\substack{i,l=1 \\ i \le l}}^{n_j} v(p_i) a\big(\lambda_{p_i}^{(j)}, \lambda_{p_l}^{(j)}\big) w(p_l) \ , \qquad v, w \in S_j \ , \tag{6.25}$$

Proof:
Let $l = 1, \dots, n_j$ be chosen arbitrarily. From $w_{l-1} = u_j^\nu + \sum_{i=1}^{l-1} v_i$ in (6.23) follows directly

$$\begin{aligned} l(\lambda_{p_l}^{(j)}) - a\big(u_j^\nu, \lambda_{p_l}^{(j)}\big) &= v_l(p_l) a\big(\lambda_{p_l}^{(j)}, \lambda_{p_l}^{(j)}\big) + \sum_{i=1}^{l-1} a\big(v_i, \lambda_{p_l}^{(j)}\big) \\ &= \sum_{i=1}^{l} v^{(j)}(p_i) a\big(\lambda_{p_i}^{(j)}, \lambda_{p_l}^{(j)}\big) \\ &= b_j\big(v^{(j)}, \lambda_{p_l}^{(j)}\big) \quad . \end{aligned}$$

As l was arbitrary, the claim follows.
The matrix representation of (6.24) is

$$B\underline{v}^\nu = b - A\underline{u}^\nu \quad . \tag{6.26}$$

Thereby

$$A = (a_{p_i,p_l})_{i,l=1}^{n_j}, \ \ a_{p_i,p_l} = a\big(\lambda_{p_i}^{(j)}, \lambda_{p_l}^{(j)}\big), \qquad b = \big(l(\lambda_{p_i}^{(j)})\big)_{i=1}^{n_j}$$

and the matrix

$$B = L + D$$

results from the partition of $A = L + D + R$ in a sub-diagonal part L, a diagonal D and a super diagonal part R. Hence (6.26) is the known Gauss–Seidel method (cf i.e. Braess [6, chapter 4, paragraph 1]). □

The convergence rate of the Gauss–Seidel method degenerates exponentially with j:

Theorem 6.23 *It holds for all $u_j^0 \in S_j$*

$$\boxed{\|u_j - u_j^{\nu+1}\| \le (1 - ch^2)\|u_j - u_j^\nu\| \quad \forall \nu \ge 0 \quad .}$$

with the energy norm $\| \cdot \| = a(\cdot,\cdot)^{1/2}$ and a j-independent constant $c > 0$.

Proof:
Most textbooks usually only treat the application to the discrete model problem (6.15). A general proof can be found in Kornhuber [15][theorem 7.32]. □

Multilevel Gauss–Seidel. According to Theorem 6.23 the Gauss–Seidel method is globally convergent, but the rate of convergence decreases quickly with increasing amount of unknowns.
Why is that? The answer follows from the following remark.

Theorem 6.24 *Let* (6.20) *be an a-orthogonal partition, i.e.*

$$a(v, w) = 0, \qquad \forall v \in V_i, w \in V_k, \quad i \neq k \quad ,$$

then the corresponding algorithm 6.1 gives for each starting value u_j^0 the exact solution u_j after one step.

Proof:
Exercise. □

Knowing the eigenvectors $\underline{e}_p$ of the stiffness matrix A,

$$V_l = \text{spann}\{e_{p_l}\} \quad l = 1, \ldots, n$$

would be an optimal, because a-orthogonal choice of the subspaces. We remember: e_{p_l} covers a big scale of frequencies. Conversely the functions λ_l, which cover a big scale of frequencies, could lead to a fast convergent method. Anyhow, the nodal basis elements of S_j *don't* have this property: all $\lambda_p^{(j)}$ have a "high frequency" (Support with diameter $\mathcal{O}(h_j)$).
Contrary one can classify the nodal basis elements of coarser grids as having a "low frequency" (see figure 6.11).
We therefore define the so called *Multi-level nodal basis* Λ,

$$\Lambda = \bigcup_{k=0}^{j} \Lambda_k = \{\lambda_l \mid l = 1, \ldots, m_S\}, \qquad m_S = n_0 + \cdots + n_j \ ,$$

as a union of all nodal basis functions of all refinement levels. The enumeration $\lambda_l = \lambda_{p_l}^{(k_l)}$ goes from fine to coarse, i.e. from $l > l'$ follows $k_l \le k_{l'}$. Thus, given our *purely heuristic point of view*, the corresponding partition

$$S_j = \sum_{l=1}^{m_S} V_l, \qquad V_l = \text{spann}\{\lambda_l\}, \qquad l = 1, \ldots, m_S, \tag{6.27}$$

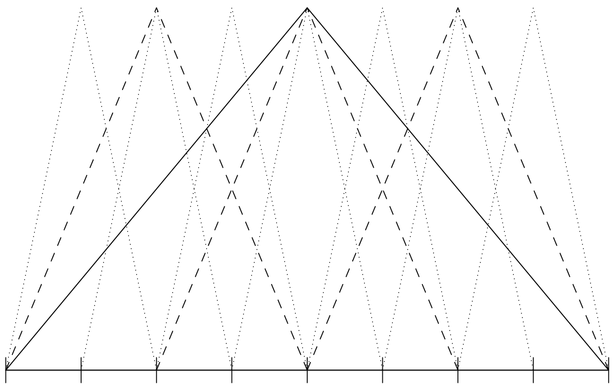


Figure 6.11: multi-level nodal basis as a scale of frequencies

covers a scale of frequencies, beginning with low frequency functions $\lambda_l \in \Lambda_0$ and reaching to high frequency functions $\lambda_l \in \Lambda_j$. The extended partition (6.27) therefore according to algorithm 6.1 gives the so called *multi-level Gauss–Seidel algorithm*

$$u_j^{\nu+1} = u_j^{\nu} + \sum_{l=1}^{m_S} v_l, \qquad v_l = \frac{l(\lambda_l) - a(w_{l-1}, \lambda_l)}{a(\lambda_l, \lambda_l)} \lambda_l \quad , \tag{6.28}$$

where $w_0 = u_j^{\nu}$ and $w_l = w_{l-1} + v_l$, $l = 1, \ldots, m_S - 1$.

Multi-grid V-cycle. The algorithm (6.28) can be rewritten equivalently, condensing all corrections on level k to one correction $v^{(k)}$. Thereby one gets the following *multi-grid method.*

Algorithm 6.2 (Multi-grid V-cycle)

```
given: $u_j^{\nu}$
initialize: $r_j = \ell - a(u_j^{\nu}, \cdot), \quad a_j(\cdot,\cdot) = a(\cdot,\cdot)$

for $k = j, \ldots, 1$ do:
        {
          solve:
          $v^{(k)} \in S_k : \quad b_k(v^{(k)}, v) = r_k(v) \quad \forall v \in S_k$        (pre-smoothing)

          $r_k = r_k - a_k(v^{(k)}, \cdot)$        (update the residual)

          $r_{k-1} = r_k|_{S_{k-1}}$
          $a_{k-1}(\cdot,\cdot) = a_k(\cdot,\cdot)|_{S_{k-1} \times S_{k-1}}$        (canonical restriction)
        }

solve:
$v^{(0)} \in S_0 : \quad b_0(v^{(0)}, v) = r_0(v) \quad \forall v \in S_0$        (approximate coarse-grid solution)

for $k = 1, \ldots, j$ do:
    {
        $v^{(k)} = v^{(k)} + v^{(k-1)}$                    (canonical interpolation)
    }
new iterate: $u_j^{\nu+1} = u_j^{\nu} + v^{(j)}$
```

The bilinear forms

$$b_k(v,w) = \sum_{\substack{i,l=1\\ i\le l}}^{n_k} v(p_i)a\big(\lambda_{p_i}^{(k)}, \lambda_{p_l}^{(k)}\big)w(p_l)\ , \qquad v,w \in S_k\ . \tag{6.29}$$

correspond to a Gauß–Seidel step on level k. The nodes $p_l \in \mathcal{N}_k$ need to be traversed in the same order as the subspaces $V_l = \text{spann}\{\lambda_{p_l}^{(k_l)}\}$ on level $k_l = k$ in the formulation (6.1). As these Gauss–Seidel steps reduce the high frequent parts on S_k quickly, one speaks of a *smoother.*

The *canonical restrictions* of the residual $r_k \in S_k'$ and the bilinear form $a_k(\cdot,\cdot)$ on S_j are defined by

$$r_{k-1}(v) = r_k(v), \quad a_{k-1}(v,w) = a_k(v,w), \qquad v,w \in S_{k-1} \subset S_k \quad .$$

The term *V-cycle* is motivated by the procedure of first calculating descending corrections $v^{(k)}$ from the fine to the coarse grid, and then collecting them ascending from the coarse to the fine grid. On this way, according to (6.18), each iterative step needs at most $\mathcal{O}(n_j)$ point operations.

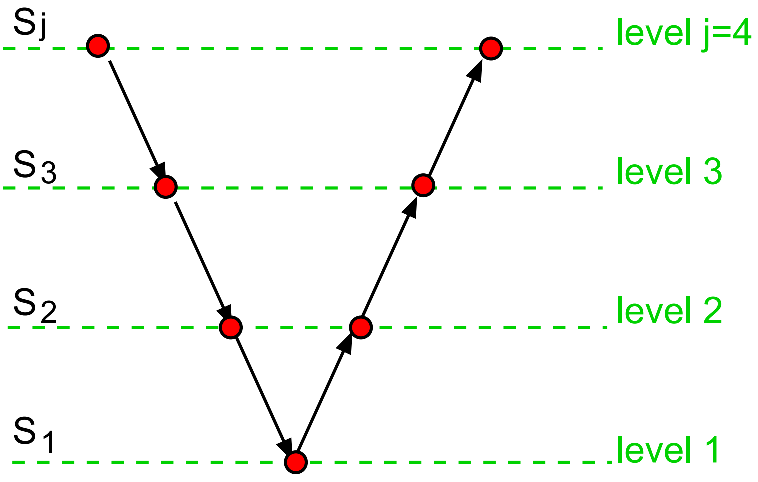


Figure 6.12: de- and ascend through the grids in the V-cycle for $j = 3$.

Implementation. Before implementing the method, one usually has to express the finite element functions $v \in S_k$ by vectors $\underline{v} \in \mathbb{R}^{n_k}$ and the bilinear forms $a_k(\cdot,\cdot)$, $b_k(\cdot,\cdot)$ by matrices $A_k, B_k \in \mathbb{R}^{n_k,n_k}$. Residues $r \in S_k'$ are represented by their values in the nodal basis, so

$$\underline{r} = \big(r(\lambda_{p_i}^{(k)})\big)_{i=1}^{n_k} \in \mathbb{R}^{n_k} \quad .$$

The canonical interpolation from S_{k-1} to S_k becomes a matrix $I_k \in \mathbb{R}^{n_k,n_{k-1}}$ and the restriction of the residuals from S_k to S_{k-1} can be represented by $R_k = I_k^T \in \mathbb{R}^{n_{k-1},n_k}$. However in the practical implementation one doesn't calculate these matrices explicitly. For the interpolation one uses the relation

$$\boxed{v^{(k-1)}(p) = \sum_{q \in \mathcal{N}_{k-1}} v^{(k-1)}(q)\lambda_q^{(k-1)}(p) \quad \forall p \in \mathcal{N}_k}$$

and for the restriction the relation

$$\boxed{r_k(\lambda_p^{(k-1)}) = \sum_{q \in \mathcal{N}_k} \lambda_p^{(k-1)}(q) r_k(\lambda_q^{(k)}) \quad \forall p \in \mathcal{N}_{k-1} \quad .}$$

directly. This way one gets, similar to the difference methods, local interpolation- and restriction-spaces (exercise).

Convergence. Our heuristic motivation is justified by the following convergence result.

Theorem 6.25 *There exists a $\varrho < 1$, depending only on the regularity of $\mathcal{T}_0$ and the ellipticity constants γ, Γ of $a(\cdot,\cdot)$, such that for all $u_j^0 \in S_j$ the error estimate*

$$\boxed{\|u_j - u_j^{\nu+1}\| \le \varrho \|u_j - u_j^\nu\| \qquad \forall \nu \ge 0} \tag{6.30}$$

holds.

The proof would exceed the scope of this lecture. We refer to the renowned outline articles of Xu [22] and Yserentant [23] or (as a reading aid) to the script of Kornhuber [15]. A very elegant presentation of the classical multi-grid convergence theory is given by Braess [6][chapter V]. The above result (asymmetric Gauß–Seidel methods as smoother) follows by results of Neuss [16].

Nested iterations. In the practical application of multi-grid methods we are interested in approximating the exact finite element approximation u_j in the energy norm up to a error of the order $\mathcal{O}(h_j)$. The therefore necessary amount of iteration steps of course depends greatly on the starting iterate. Thus it is standing to reason, to calculate a good starting iterate for the iteration on $\mathcal{T}_{k+1}$ by a specific amount ν^* of multi-grid steps on $\mathcal{T}_k$. This procedure is called *nested iteration* (see Hackbusch [12]).

Algorithm 6.3 (Multi-grid with nested iteration)

```
given: $\tilde{u}_0 \in S_0$

for $k = 1, \dots, j$ do:
    {
    starting iterate: $u_k^0 = \tilde{u}_{k-1}$
    $\nu^*$ multi-grid steps: $\tilde{u}_k = u_k^{\nu^*}$
    }

result: $\tilde{u}_j$
```

Now the question arises how to choose ν^*. The answer is given by the following theorem.

Theorem 6.26 *We assume that the finite element approximations u_k satisfy the estimates*

$$\|u - u_k\| \leq c_1 h_k\ , \qquad k = 0, 1, \dots\ , \tag{6.31}$$

with a k-independent constant c_1, furthermore

$$\tilde{u}_0 = u_0 \tag{6.32}$$

and finally ν^* is chosen such that the stopping criterion

$$\|u_k - \tilde{u}_k\| \leq \tfrac{\sigma}{2}\|u_k - u_k^0\|\ , \qquad k = 1, 2, \dots\ , \tag{6.33}$$

is satisfied with a k-independent $\sigma < 1$. Then there exists a j-independent C, such that

$$\boxed{\|u - \tilde{u}_j\| \leq C h_j\ .} \tag{6.34}$$

Proof:

Using the preconditions (6.33), (6.31), (6.32) and (6.17) one calculates

$$\begin{aligned}
\|u - \tilde{u}_j\| &\leq \|u - u_j\| + \|u_j - \tilde{u}_j\| \leq \|u - u_j\| + \tfrac{1}{2}\sigma\|u_j - \tilde{u}_{j-1}\| \\
&\leq (1 + \tfrac{1}{2}\sigma)\|u - u_j\| + \tfrac{1}{2}\sigma\|u - \tilde{u}_{j-1}\| \\
&\leq (\tfrac{\sigma}{2})^j\|u - u_0\| + (1 + \tfrac{1}{2}\sigma)\sum_{i=0}^{j-1}(\tfrac{\sigma}{2})^i\|u - u_{j-i}\| \\
&\leq (1 + \tfrac{1}{2}\sigma)c_1\sum_{i=0}^{j}\sigma^i h_j \\
&\leq \tfrac{3}{2}c_1(1 - \sigma)^{-1}h_j \quad .
\end{aligned}$$

(6.34) follows with $C = \frac{3}{2}c_1(1 - \sigma)^{-1}$. □

Of course in general c_1 and thereby also C depend on u. As is generally known the regularity assumption $u \in H \cap H^2(\Omega)$ is sufficient for the discretization accuracy (6.31). The exact solution u_0 on the (hopefully) coarse grid $\mathcal{T}_0$ is calculated with a directly. To check the stopping criterion (6.33), we need an *a posteriori estimate on the algebraic error* $\|u_k - u_k^\nu\|$. By the triangle inequality follows with ϱ from theorem 6.25 (Exercise)

$$(1+\varrho)^{-1}\|u_k^{\nu+1} - u_k^\nu\| \leq \|u_k - u_k^\nu\| \leq (1-\varrho)^{-1}\|u_k^{\nu+1} - u_k^\nu\| \; . \tag{6.35}$$

The multi-grid corrections therefore give lower and upper bounds for the algebraic error at the same time.

Corollary 6.27 *For the calculation of an approximation $\tilde{u}_j$ with accuracy*

$$\|u - \tilde{u}_j\| = \mathcal{O}(h_j)$$

the multi-grid with nested iteration method needs (see theorem 6.26)

$$\boxed{\mathcal{O}(n_j) \text{ point operations.}}$$

Proof:
The number of point operations on a fixed level k is bounded by cn_k, with c independent of k. As the convergence rate ϱ is independent of j, the number ν^* of multi-grid steps needed to satisfy the stopping criterion (6.33) is also independent of j. Particularly one can calculate for each given $\sigma < 1$ the necessary number of multi-grid steps ν^* by $\varrho^{\nu^*} < \frac{1}{2}\sigma$. In view of (6.18) the total effort of the multi-grid with nested iteration method is thus bounded by

$$c\nu^* \sum_{k=1}^{j} n_k \leq c\nu^*(1-q^{-1})^{-1} n_j = \mathcal{O}(n_j) \; .$$

Note:
Our main goal is the approximation of the unknown function u up to a given tolerance TOL. Therefore the amount j of refinement steps should not be fixed a priori, but be determined *a posteriori* by an adaptive multilevel method. Therefore one needs corresponding a posteriori estimates for the discretization error. ◁

Note:
In practice, multi-grid methods are used mostly as *preconditioner* for the *conjugate gradient method* (cg-method). This way one gets an additional increase in the convergence rate. For details on the cg-methods and preconditioning we refer to Deuflhard and Hohmann [8, chapter 8] or Braess [6, chapter IV]. ◁

7 Parabolic Differential Equations

7.1 Classical Solutions

Details to the following two sections can be found in F. John, Partial Differential Equations, Springer, chapter 7.

7.1.1 The Cauchy-Problem for the Heat Equation

We consider the heat equation

$$\begin{aligned} u_t &= \alpha u_{xx} \quad x \in \mathbb{R},\ t \in (0,1] \\ u(x,0) &= u_0(x) \quad x \in \mathbb{R} \quad , \end{aligned} \tag{7.1}$$

where we begin by setting $\alpha \equiv 1$. We look for a *classical* solution u, i.e.

$$u \in C(\overline{Q}) \quad , \quad u_{xx}, u_t \in C(Q) \quad \text{with} \quad Q = \mathbb{R} \times (0,T] \quad .$$

By "qualified guessing" we want to develop a closed solution for (7.1). Therefore we define for $v \in C_0^\infty(\mathbb{R})$ the *Fourier transform*

$$\boxed{\hat{v}(\xi) = (2\pi)^{-\frac{1}{2}} \int\limits_{\mathbb{R}} e^{-ix\xi}\, v(x)\, dx \quad \text{with} \quad i^2 = -1 \quad .} \tag{7.2}$$

In a slightly bigger space then $C_0^\infty(\mathbb{R})$ (the space of "rapidly decreasing" C^∞-functions) the Fourier transform $v \mapsto \hat{v}$ happens to be bijective, and the inverse relation holds

$$v(x) = (2\pi)^{-\frac{1}{2}} \int\limits_{\mathbb{R}} e^{ix\xi}\, \hat{v}(\xi)\, d\xi \quad . \tag{7.3}$$

A fundamental property of the Fourier transform is converting derivatives into multiplications. Using the product rule (with $|v(x)| \to 0$ for $|x| \to \infty$) we have

$$\begin{aligned} i\xi\hat{v}(\xi) &= -(2\pi)^{-\frac{1}{2}} \int\limits_{\mathbb{R}} \frac{\partial}{\partial x}(e^{-ix\xi})\, v(x)\, dx \\ &= (2\pi)^{-\frac{1}{2}} \int\limits_{\mathbb{R}} e^{-ix\xi}\, \frac{dv}{dx}(x)\, dx \quad , \end{aligned}$$

so

$$\boxed{i\xi\hat{v} = \widehat{\frac{dv}{dx}} \quad .}$$

Using this property, our problem (7.1) can be simplified and solved by further steps. Therefore we apply the Fourier transform to the space part of u in the heat equation and get for each $t > 0$

$$\hat{u}_t = \widehat{u_t} = \widehat{u_{xx}} = (i\xi)^2\hat{u} = -\xi^2\hat{u} \quad .$$

The initial condition becomes

$$\hat{u}(\xi, 0) = \widehat{u_0} \quad .$$

The solution of this ordinary differential equation in the Fourier space is given by

$$\hat{u}(\xi, t) = \widehat{u_0}(\xi)e^{-\xi^2 t} \quad . \tag{7.4}$$

Now we have to transform back. This could be done with the above mentioned inverse of the Fourier transform. But it is more effective using the following formula, which again applies for "rapidly decreasing" C^∞-functions v and w:

$$(2\pi)^{-\frac{1}{2}}\widehat{v * w} = \hat{v}\hat{w} \quad . \tag{7.5}$$

Here $v * w$ is the convolution of v and w, defined by

$$(v * w)(x) = \int\limits_{\mathbb{R}} v(y)w(x-y)\,dy \quad \forall x \in \mathbb{R} \quad .$$

The Fourier transform thus transforms convolutions to products of functions. To apply (7.5) to (7.4) we have to identify the function $w : \xi \mapsto e^{-\xi^2 t}$ as a Fourier transform. Indeed this function is, up to a variable transformation, its own Fourier transform. Namely $x \mapsto e^{-x^2/2}$ is an eigenfunction of the Fourier transform and thus w is the Fourier transform of

$$x \mapsto (2t)^{-\frac{1}{2}}e^{-x^2/4t}$$

(proof as exercise). Thus, together with (7.5), as solution to our Cauchy problem we get

$$u(x, t) = (4\pi t)^{-\frac{1}{2}} \int\limits_{\mathbb{R}} e^{-\frac{(x-\xi)^2}{4t}} u_0(\xi)\,d\xi \quad . \tag{7.6}$$

We now see that our transformations give us a closed solution u of the Cauchy problem (7.1) (at least for the here not further defined "rapidly decreasing" C^∞-initial data u_0). We will now show that (7.6) is a solution for every continuous and bounded u_0, which furthermore depends continuously on u_0. Therefore we define, analogously to the elliptic case, the *Greens function G* by

$$\boxed{G(x, \xi, t) = (4\pi t)^{-\frac{1}{2}}\, e^{-\frac{(x-\xi)^2}{4t}} \quad , \quad t > 0\ ,\ x, \xi \in \mathbb{R} \quad .}$$

Theorem 7.1 *For each initial condition $u_0 \in C(\mathbb{R})$ with $\|u_0\|_\infty = \sup_{x\in\mathbb{R}} |u_0(x)| < \infty$,*

$$\boxed{u(x, t) = \int\limits_{\mathbb{R}} G(x, \xi, t)u_0(\xi)\,d\xi \quad , \ x \in \mathbb{R}\ ,\ t > 0}$$

defines a classical solution for the Cauchy problem (7.1)*. Furthermore holds for every* $T > 0$

$$\max_{t\in[0,T]} \|u(\cdot,t)\|_\infty \le \|u_0\|_\infty \quad . \tag{7.7}$$

Proof:
It holds that

$$\begin{aligned}\frac{\partial}{\partial t}G(x,\xi,t) &= \frac{1}{t\sqrt{4\pi t}}\,e^{-\frac{(x-\xi)^2}{4t}}\left[-\frac{1}{2}+\frac{(x-\xi)^2}{4t}\right] \quad ,\\ \frac{\partial}{\partial x}G(x,\xi,t) &= -\frac{(x-\xi)}{2t\sqrt{4\pi t}}\,e^{-\frac{(x-\xi)^2}{4t}} \quad ,\\ \frac{\partial^2}{\partial x^2}G(x,\xi,t) &= \frac{1}{t\sqrt{4\pi t}}\,e^{-\frac{(x-\xi)^2}{4t}}\left[\frac{(x-\xi)^2}{4t}-\frac{1}{2}\right] \quad ,\end{aligned}$$

so that

$$u_t(x,t) = \int\limits_{\mathbb{R}} \frac{\partial}{\partial t}G(x,\xi,t)u_0(\xi)\,d\xi = \int\limits_{\mathbb{R}} \frac{\partial^2}{\partial x^2}G(x,\xi,t)u_0(\xi)\,d\xi = u_{xx}(x,t) \quad .$$

Thereby one can exchange integration and differentiation, as for $t > 0$ the partial derivatives of G in direction x or t are equicontinuous with respect to ξ and converge uniformly and strong enough in a neighbourhood of x/t for $\xi \to \infty$ to 0. Concrete: Let $\varepsilon > 0$. We consider $\frac{\partial}{\partial t}f(t,\xi)$ for $f(t,\xi) = G(x,\xi,t)u_0(\xi)$, $x \in \mathbb{R}$, and a $t > 0$. Applying the mean value theorem we get

$$\left|\frac{f(t+h,\xi)-f(t,\xi)}{h} - \frac{\partial}{\partial t}f(t,\xi)\right| \le \varepsilon$$

if $h \le \delta$ for a $\delta > 0$, chosen independently of ξ. Hereby follows

$$\frac{\partial}{\partial t}\int\limits_{I_n} f(t,\xi)d\xi = \int\limits_{I_n} \frac{\partial}{\partial t}f(t,\xi)d\xi$$

for each interval $I_n = [-n,n]$, $n \in \mathbb{N}$ (Why?). Furthermore for a specific neighbourhood U_t of t and a $N \in \mathbb{N}$

$$\left|\int\limits_{\mathbb{R}\setminus I_n} \frac{\partial}{\partial t'}f(t',\xi)d\xi\right| \le \varepsilon$$

holds for $n \ge N$ uniformly in $t' \in U_t$. As f is also improperly integrable, one can deduce the exchangeability of the derivative and integral over the whole unbounded set of integration $\mathbb{R}$. (How?)
Now we examine the limit of $u(\cdot,t)$ for $t \to 0$. Evidently we have

$$x \ne \xi \quad\Longrightarrow\quad \lim_{t\to 0} G(x,\xi,t) = 0$$

uniformly for $|x-\xi| \ge \delta > 0$ and

$$G(x,\xi,t) > 0 \quad \forall x,\xi \in \mathbb{R}\,,\ t > 0 \quad .$$

Together this gives for every $\delta > 0$

$$\left|\int\limits_{|x-\xi|\ge\delta} G(x,\xi,t)u_0(\xi)\,d\xi\right| \le \|u_0\|_\infty \int\limits_{|x-\xi|\ge\delta} G(x,\xi,t)\,d\xi \to 0 \quad , \quad t \to 0 \quad .$$

Furthermore because of

$$\int\limits_{\mathbb{R}} G(x,\xi,t)\,d\xi = 1 \qquad x \in \mathbb{R}\,,\ t > 0 \tag{7.8}$$

(Recalculate!) the following estimate is correct:

$$\begin{aligned}\left| u_0(x) - \int\limits_{|x-\xi|\le\delta} G(x,\xi,t)u_0(\xi)\,d\xi \right| &\le |u_0(x)| \int\limits_{|x-\xi|\ge\delta} G(x,\xi,t)\,d\xi + \int\limits_{|x-\xi|\le\delta} G(x,\xi,t)\,|u_0(x)-u_0(\xi)|\,d\xi \\ &\le \|u_0\|_\infty \int\limits_{|x-\xi|\ge\delta} G(x,\xi,t)\,d\xi + \max_{|x-\xi|\le\delta} |u_0(x)-u_0(\xi)| \to 0 \\ &\qquad\qquad \text{for } \ t,\delta\to 0 \quad .\end{aligned}$$

As u_0 is uniformly continuous on compact subsets of IR, the above estimates imply the uniform convergence of $u(\cdot,t) \to u_0$ for $t\to 0$ on compact subsets of IR and thereby $u \in C(\overline{Q})$. (Why doesn't the point-wise convergence suffice?)
The a-priori-estimate (7.7) follows directly from the mean value property (7.8). □

Note:

1) The property

$$G(x,\xi,t) > 0 \quad \forall x,\xi \in \mathrm{IR}\,,\ t>0$$

means that a perturbation of u_0 in an arbitrary point x_0 leads to perturbations in the solution $u(x,t)$ in *every* $x\in$ IR for *every* time $t>0$. Thereby the *domain of dependence* of each $(x,t)\in Q$ is all IR.

In other words:

Perturbations in the initial conditions spread with *infinite speed.*

2) Independently of the choice of initial conditions $u_0 \in C(\mathrm{IR})$, we have $u(\cdot,t)\in C^\infty(\mathrm{IR})$ for each $t>0$. (Why?) This smoothing property is typical for parabolic equations.

3) u even suffices the *maximum principle*

$$\inf_{y\in\mathrm{IR}} u_0(y) \le u(x,t) \le \sup_{y\in\mathrm{IR}} u_0(y) \quad \forall x\in\mathrm{IR}\ ,\ t>0 \quad .$$

4) The solution of the Cauchy problem *is not unique.* In particular there exist "unphysical" solutions with $|u(x,t)|\to\infty$ for $t\to 0$. A construction of such solutions can be found in [13, S. 211].

Sufficient conditions for uniqueness are for example

$$\begin{aligned} |u(x,t)| \le M\,e^{\alpha x}\ , &\quad x\in\mathrm{IR},\ 0<t<T \quad \text{(John [13, S. 217])}\ , \\ u(x,t)\ge 0 \quad\ , &\quad x\in\mathrm{IR},\ 0<t<T \quad \text{(John [13, S. 222])}\ . \end{aligned}$$

7.1.2 Initial–Boundary–Value Problems

We consider on a bounded domain $\Omega\subset\mathrm{IR}^n$ the initial–boundary–value problem for the heat equation

$$\begin{aligned} u_t - \Delta u &= f && \text{in } Q = \Omega\times(0,T] \\ u(x,0) &= u_0(x) && x\in\overline{\Omega} \\ u(x,t) &= g(x,t) && (x,t)\in\partial\Omega\times(0,T] \quad . \end{aligned} \tag{7.9}$$

To ensure $u \in C(\overline{\Omega})$, the data f, u_0 and g must be required as continuous and to satisfy the consistency condition

$$u_0(x) = g(x,0) \quad \forall x \in \partial\Omega \quad .$$

An important tool to the proof of uniqueness and a priori estimates is the following maximum principle.

Theorem 7.2 *Let $u \in C(\overline{Q})$, $u_t, \frac{\partial^2 u}{\partial x_i^2} \in C(Q)$, $i = 1, \ldots, n$, and*

$$u_t - \Delta u \ \leq \ 0 \qquad \text{in } \ Q \quad .$$

Then follows

$$\max_{(x,t)\in\overline{Q}} u(x,t) \leq \max_{(x,t)\in\partial Q\backslash\Gamma_T} u(x,t) \quad ,$$

where $\Gamma_T := \{(x,t) \,|\, x \in \Omega \ \text{ and } \ t = T\}$.

Proof:
Choose $\varepsilon > 0$ and consider the function

$$v(x,t) = u(x,t) - \varepsilon\, t \quad .$$

Assume that v achieves its maximum in (x_0,t_0). We will show $(x_0,t_0) \notin Q \cup \Gamma_T$.
If this was the case, by the necessary conditions for a maximum we would have

$$v_t(x_0,t_0) \geq 0 \quad , \quad \Delta v(x_0,t_0) \leq 0 \quad .$$

This is a contradiction to

$$v_t(x_0,t_0) = u_t(x_0,t_0) - \varepsilon \leq \Delta u(x_0,t_0) - \varepsilon = \Delta v(x_0,t_0) - \varepsilon \leq -\varepsilon < 0 \quad .$$

If now u wouldn't achieve its maximum in $\Gamma_D := \partial Q\backslash\Gamma_T$, there would be a $(x_0,t_0) \in Q \cup \Gamma_T$ with

$$u(x_0,t_0) > u(x,t) \qquad \forall (x,t) \in \Gamma_D \quad ,$$

and thereby an $\varepsilon > 0$ with

$$u(x_0,t_0) - \varepsilon\, t_0 > u(x,t) - \varepsilon\, t \qquad \forall\, (x,t) \in \Gamma_D$$

contradictory to the above shown. □

Theorem 7.3 *The initial value problem (7.9) has at most one classical solution $u \in C(\overline{Q}) \cap C^2(Q)$.*

Proof:
Assume u_1, u_2 were two different classical solutions of (7.9), then $w = u_1 - u_2$ would be the solution of the homogeneous problem (with $u_0 = 0$ and $g = 0$) and

$$w_t - \Delta w \leq 0 \quad .$$

Thus we have

$$\max_{(x,t)\in\overline{Q}} w(x,t) \leq \max_{(x,t)\in\Gamma_D} w(x,t) = 0 \quad .$$

Analogously follows from $(-w)_t - (-w)_{xx} \leq 0$ directly $-w \leq 0$. This shows $u_1 = u_2$. □

Theorem 7.4 *A classical solution of (7.9), if existent, depends continuously on the data u_0 and g with respect to the $\|\cdot\|_\infty$-norm. It even holds that*

$$\sup_{0<t\leq T} \|u(\cdot,t)\|_\infty = \max\{\|u_0\|_\infty \,,\, \sup_{0<t\leq T} \|g(\cdot,t)\|_\infty\} \quad .$$

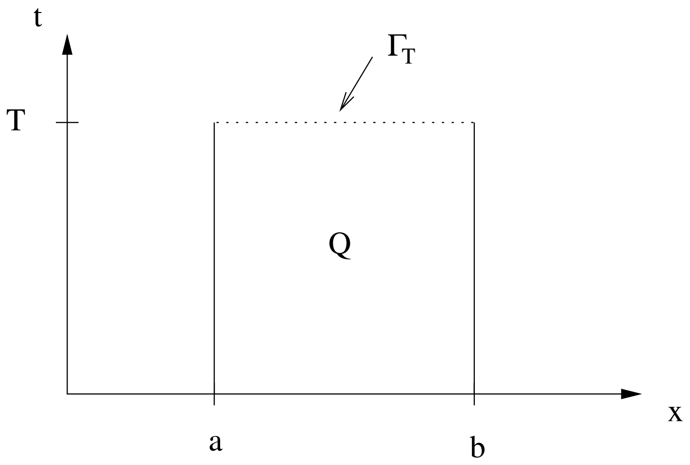


Figure 7.1: domain of calculation Q

Proof:
Exercise. □

We want to proof the existence of a solution in the homogeneous case $f = 0$ on $\Omega = (0, 1)$. For this, we make use of the *Fourier method.* The basic idea is separation of variables

$$\boxed{u(x,t) = v(x)w(t) \quad .}$$

Inserting into the differential equation gives

$$\frac{w'}{w} = \frac{v''}{v} = \text{const.} =: \lambda$$

and thereby the *eigenvalue problem*

$$\begin{aligned} v'' &= -\lambda v \\ w' &= -\lambda w \quad . \end{aligned} \tag{7.10}$$

Solution of $v'' = -\lambda v$: Ansatz: $v(x) = \sin(\lambda^{\frac{1}{2}} x)$
The ansatz suffices the differential equation, and by $v(0) = v(1) = 0$ follows

$$\lambda_k = (k\pi)^2 \quad , \quad k = 0, 1, 2, \dots \quad .$$

Solution of $w' = -\lambda_k w$:

$$w(t) = a_k e^{-\lambda_k t} \quad .$$

Solution of the heat equation with zero-boundary data:

$$u_k(x,t) = a_k e^{-(k\pi)^2 t} \sin(k\pi x) \qquad k = 0, 1, 2, \dots \quad .$$

Due to the linearity of the differential equation,

$$u_N(x,t) = \sum_{k=0}^{N} u_k(x,t) = \sum_{k=0}^{N} a_k e^{-(k\pi)^2 t} \sin(k\pi x)$$

is a solution for each N, which satisfies the boundary condition.
Continuity up the the initial condition:
Determine the coefficients a_k such that

$$u_0(x) = \sum_{k=0}^{\infty} a_k \sin(k\pi x) \quad .$$

Extending u_0 to an uneven function on $[-1,1]$ and developing its Fourier series, one gets the *Fourier coefficients*

$$a_k = 2 \int_0^1 u_0(\xi) \sin(k\pi\xi)\, d\xi \quad .$$

If u_0 is continuous and piecewise continuously differentiable, the Fourier series converges point-wise to u_0 (see Endl/Luh: Analysis II, theorem 4.5.2).
As solution to the differential equation one now suspects

$$u(x,t) = \sum_{k=0}^{\infty} a_k e^{-(k\pi)^2 t} \sin(k\pi x) \quad . \tag{7.11}$$

To justify this approach we have to answer the following questions:

(i) Does the series (7.11) converge for each $(x,t) \in \Omega$?

(ii) Can we differentiate term-wise? (Only then u suffices the differential equation.)

(iii) Is the solution continuous up to the initial condition?

Under specific assumptions all the questions can be answered with “yes”:

(i) Let $\sup_{x\in(0,1)} |u_0(x)| \leq M$. Then follows

$$|a_k| \leq 2M \quad ,$$

and thus for each $t \geq \delta > 0$

$$\left| a_k e^{-(k\pi)^2 t} \sin(k\pi x) \right| \leq 2M\ e^{-(k\pi)^2 \delta} \quad , \quad k = 0,1,2,\ldots \quad .$$

The comparison test even gives uniform convergence for each fixed $t > 0$.

(ii) Alike one sees that

$$2M(k\pi)^2\, e^{-\delta\pi^2}\, e^{-k^2} \quad , \quad k = 0,1,2,\ldots$$

dominates the term-wise, in t-direction differentiated series. Thus the term-wise differentiated series converges uniformly and coincides with the corresponding derivative of the series (see Endl/Luh, Analysis II, thm. 3.7.2). The analogue holds for all other partial derivatives.

(iii) Let

$$\sum_{k=1}^{\infty} |a_k| < \infty \quad . \tag{7.12}$$

Then by the comparison test the series

$$\sum_{k=1}^{\infty} a_k e^{-(k\pi)^2 t} \sin(k\pi x)$$

converges uniformly in $x \in [0,1]$ and $t \geq 0$. Thus the asymptote is continuous in $x \in [0,1]$, $t \geq 0$ (see Endl/Luh Analysis II, thm. 3.7.1). Furthermore we can exchange limit and summation and get

$$\lim_{t \to 0} u(x,t) = \sum_{k=0}^{\infty} \lim_{t \to 0} a_k e^{-(k\pi)^2 t} \sin(k\pi x) = u_0(x) \qquad \forall x \in [0,1] \quad .$$

Sufficient for (7.12) is u_0 being continuous, piecewise differentiable and $u(0,t) = u(1,t) = 0$ (cf. Heuser, Lehrbuch zur Analysis, part 2, 1981, theorem 136.5).

We now have shown (together with theorem 7.3) the following existence theorem

Theorem 7.5 *Let u_0 be continuous, piecewise differentiable and $u(0,t) = u(1,t) = 0$. Then the initial–boundary–value problem* (7.9) *has a unique solution u given by*

$$\boxed{u(x,t) = \sum_{k=1}^{\infty} a_k e^{-(k\pi)^2 t} \sin(k\pi x)}$$

with the Fourier coefficients

$$\boxed{a_k = 2 \int_0^1 u_0(\xi) \sin(k\pi\xi) \, d\xi \quad .}$$

Note:
Inserting the formula for the Fourier series in the series (7.11), due to the uniform convergence of the resulting series for $t > 0$, one can exchange integration and summation and thus gets the representation

$$u(x,t) = \int_0^1 G(x,\xi,t) u_0(\xi) \, d\xi$$

with the Greens function

$$G(x,\xi,t) = 2 \sum_{k=1}^{\infty} \sin(k\pi x) \sin(k\pi\xi) e^{-(k\pi)^2 t} \quad .$$

Note for $x \in (0,1)$

$$G(x,x,t) = 2\sum_{k=1}^{\infty}(\sin(k\pi x))^2 e^{-(k\pi)^2 t} \to \infty, \quad t \to 0$$

and thereby the solution of the Cauchy problem in Theorem 7.1. ◁

7.2 Weak Solutions

On the bounded domain $\Omega \subset \mathbb{R}^n$ we consider the initial–value–boundary problem

$$\begin{array}{rcll} u_t &=& \operatorname{div}(\alpha(x)\nabla u) + f(x,t) \ , & x \in \Omega,\ t > 0 \\ u(x,0) &=& u_0(x) \ , & x \in \Omega \\ u(x,t) &=& 0 \ , & (x,t) \in \partial\Omega \times (0,T] \quad . \end{array} \tag{7.13}$$

where $0 < \alpha_0 \le \alpha(x) < \alpha_1$ almost everywhere in Ω. As with elliptic problems, we cannot expect classical solutions, if $\alpha \notin C^1(\Omega)$. We therefore want to expand the notation of weak solutions from elliptic (= stationary parabolic) to parabolic problems.

As in the case of elliptic partial differential equations, we can find a weak formulation by multiplying the differential equation (7.13) for fixed $t > 0$ with test functions $v \in H_0^1(\Omega)$, then integrating, using the greens formula and exchanging the derivative w.r.t. t with the integration.

Thus we get for every $t \in (0,T)$ the variational problem:

Find $u: \ (0,T) \to H_0^1(\Omega)$ such that

$$\frac{d}{dt}(u(t),v) + a(u(t),v) = (f(t),v) \qquad \forall v \in H_0^1(\Omega) \tag{7.14}$$

if $t \in (0,T)$ and $u(0) = u_0$.

We used the abbreviations

$$(u(t),v) = \int\limits_{\Omega} u(x,t)v(x)\,dx$$

$$a(v,w) = \int\limits_{\Omega} \alpha v_x w_x\,dx$$

$$(f(t),v) = \int\limits_{\Omega} f(x,t)v(x)\,dx$$

To give sense to this kind of formulation we have to discuss the following questions:

1) Which maps $u: \ (0,T) \to H_0^1(\Omega)$ are possible?

2) How should the initial condition be attained?

3) How should the derivative $\frac{d}{dt}$ be interpreted?

We first tackle the first question and therefore consider the space $C([0,T],W)$ of all continuous functions v

$$[0,T] \ni t \mapsto v(t) \in W \quad .$$

Here let W be a Hilbert space with scalar product $(\cdot,\cdot)_W$ and corresponding norm $\|\cdot\|_W$. Equipped with the norm

$$\|v\|_{C([0,T],W)} := \max_{t\in[0,T]} \|v(t)\|_W \quad ,$$

$C([0,T],W)$ becomes a Banach space.

Example:
$W = \mathrm{I\!R} \quad \Rightarrow \quad C([0,T],\mathrm{I\!R}) = C([0,T])$ ◁

Example:
$W = L^2(\Omega)$: Note that the resulting space $C([0,T],L^2(\Omega))$ on the one hand is a space of continuous maps, but on the other hand consists of functions $v(x,t)$, which don't need to be continuous in x-direction (equivalence classes). For example is

$$v(x,t) = v_1(x)v_2(t) \in C([0,T],L^2(\Omega))$$

if $v_1(x) \in L^2(\Omega)$ and $v_2(t) \in C([0,T])$. ◁

With the scalar product

$$(v,w)_{L^2((0,T),W)} := \int_0^T (v(t),w(t))_W \, dt \tag{7.15}$$

the linear space

$$L_C((0,T),W) := \{v \in C([0,T],W) \mid \int_0^T \|v(t)\|_W^2 \, dt < \infty\}$$

becomes pre-Hilbert. We denote the completion with respect to the by (7.15) induced norm with

$$L^2((0,T),W) \quad .$$

Hereby, as with the special case

$$L^2((0,T),\mathrm{I\!R}) = L^2(0,T)$$

we are dealing with equivalence classes $[v]$ of functions $v:\,(0,T) \to W$, which may differ on subsets of $(0,T)$ with Lebesgue measure zero.
As solution space we now take

$$C([0,T],L^2(\Omega)) \,\cap\, L^2((0,T),H_0^1(\Omega)) \quad .$$

The condition $u \in C([0,T],L^2(\Omega))$ thereby ensures that the initial value $u_0 \in L^2(\Omega)$ can be attained continuously.

$$u(x,0) = u_0 \qquad x \in \Omega$$

means in particular that

$$\lim_{t\to 0} \|u(\cdot,t) - u_0\|_{L^2(\Omega)} = 0 \quad . \tag{7.16}$$

This solves question 2).
We now come to question 3). The property $u \in L^2((0,T), H_0^1(\Omega))$ gives

$$(u(\cdot), v)\ ,\ a(u(\cdot), v) \in L^2(0,T) \qquad \forall v \in H_0^1(\Omega) \quad , \tag{7.17}$$

as

$$\int_0^T a(u(t),v)^2\, dt \le \alpha_1^2 \int_0^T \|u(t)\|^2_{H^1(\Omega)}\, \|v\|^2_{H^1(\Omega)}\, dt < \infty \quad .$$

The same accounts for the case $f \in L^2((0,T), L^2(\Omega))$

$$(f(t), v)\ \in\ L^2(0,T)$$

for all $v \in L^2(\Omega)$, because

$$\int_O^T (f(t),v)^2\, dt\ \le\ \int_0^T \|f(t)\|^2_{L^2(\Omega)}\, \|v\|^2_{L^2(\Omega)}\, dt < \infty \quad .$$

The time derivative $\frac{d}{dt}$ can thus be interpreted in the weak sense we know, but from now on on $L^2(0,T)$.

Reminder: $\frac{dw}{dt} \in L^2(0,T)$ is called weak derivative of $w \in L^2(0,T)$ if

$$\int_0^T w(t)\varphi'(t)\, dt = -\int_0^T \frac{dw}{dt}\, \varphi(t)\, dt$$

holds for all $\varphi \in C_0^\infty(0,T)$.
The weak formulation (7.14) of (7.13) thus is:

Find $u \in L^2((0,T), H_0^1(\Omega))$, such that

$$-\int_0^T (u(t),v)\varphi'(t)\, dt = \int_0^T \left(-a(u(t),v) + (f(t),v)\right)\varphi(t)\, dt \tag{7.18}$$

holds for all $v \in H_0^1(\Omega)$ and all $\varphi \in C_0^\infty(0,T)$.

Let from now on $u_0 \in L^2(\Omega)$ and $f \in L^2((0,T), L^2(\Omega))$.

Definition 7.6 $u \in C([0,T], L^2(\Omega)) \cap L^2((0,T), H_0^1(\Omega))$ *is called weak solution of the initial–boundary–value problem* (7.13) *if* u *suffices the variational equality*

$$\frac{d}{dt}(u(t),v) + a(u(t),v) = (f(t),v) \qquad \forall v \in H_0^1(\Omega) \tag{7.19}$$

for almost all $t \in (0,T)$ *in the sense of* (7.18) *and it achieves the initial condition* $u(0) = u_0$ *in the sense of* (7.16).

As preparation for the proof of existence and uniqueness we need

Theorem 7.7 *There exist functions $\varphi_k \in H_0^1(\Omega)$, $k = 1, 2, \ldots$, and corresponding $\mu_k \in \mathbb{R}$ with*

$$0 < \mu_1 \le \mu_2 \le \ldots \quad , \quad \lim_{k\to\infty} \mu_k = \infty \quad ,$$

such that

$$a(\varphi_k, v) = \mu_k(\varphi_k, v) \qquad \forall v \in H_0^1(\Omega) \tag{7.20}$$

holds and

$$\{\varphi_k\}_{k\in\mathbb{N}} \quad \text{is an orthonormal basis of } L^2(\Omega) \quad . \tag{7.21}$$

Proof:
As $a(\cdot,\cdot)$ is continuous and $H_0^1(\Omega)$-elliptic, the variational problem

$$u \in H_0^1(\Omega) : \quad a(u, v) = (f, v) \quad \forall v \in H_0^1(\Omega)$$

has, according to theorem 5.31, a unique solution $Lu \in H_0^1(\Omega)$ for every $f \in L^2(\Omega)$. The stability estimates give the continuity of the solution operator $L : L^2(\Omega) \to H_0^1(\Omega)$. We now equip $H_0^1(\Omega)$ with the new scalar product $a(\cdot,\cdot)$ and consider the operator $T : H_0^1(\Omega) \to H_0^1(\Omega)$, defined by $Tv = Lv \; \forall v \in H_0^1(\Omega)$. By the embedding theorem of Rellich (see theorem 5.28) follows with the continuity of L the compactness of T. Due to the symmetry respectively the $H_0^1(\Omega)$-ellipticity of $a(\cdot,\cdot)$, T is self-adjoint and positive on $(H_0^1(\Omega), a(\cdot,\cdot))$ (the latter says $a(Tv, v) > 0 \; \forall v \in H_0^1(\Omega)$, $v \ne 0$) The spectral theorem for compact operators in Hilbert spaces (see Werner, Funktionalanalysis, 2. edition, theorem VI.3.2) now gives an $a(\cdot,\cdot)$-orthonormal basis of eigenvectors $\{\psi_k\}_{k\in\mathbb{N}}$ of T in $(H_0^1(\Omega), a(\cdot,\cdot))$ to a zero sequence of positive eigenvalues λ_k. By $\mu_k := \lambda_k^{-1}$ and $\varphi_k := \lambda_k^{1/2}\psi_k$ follows the claim. □

Note:

1) The statement of theorem 7.7 remains true if we replace $H_0^1(\Omega)$ by another, compactly in $L^2(\Omega)$ embedded Hilbert space H with an H-elliptic bilinear form $a(\cdot,\cdot)$.

2) The line (7.21) particularly implies

$$v = \sum_{k=1}^{\infty} (v, \varphi_k)\varphi_k \qquad \text{(Fourier expansion)}$$

$$\|v\|^2_{L^2(\Omega)} = \sum_{k=1}^{\infty} (v, \varphi_k)^2 \qquad \text{(Parseval's equality)}$$

for all $v \in L^2(\Omega)$.

3) In the case $\alpha \equiv 1$ and $\Omega = (0, 1)$, we have

$$\varphi_k(x) = \sqrt{2}\sin(k\pi x) \quad , \quad x \in \Omega = (0,1)$$
$$\mu_k = (k\pi)^2$$

for $k = 1, 2, \ldots$ (see [2, p. 228]).

Note the eigenfunctions and eigenvalues of the Laplace operator in our model problem on $\Omega = (0, 1) \times (0, 1)$ (see theorem 6.20). One can associate the eigenvalues μ_k with a scale of frequencies of the eigenfunctions also in more general cases. ◁

Now we can prove the following representation theorem.

Theorem 7.8 *Let* $u \in C([0,T], L^2(\Omega)) \cap L^2((0,T), H_0^1(\Omega))$ *be a solution of the variational problem* (7.19)*. Then* u *has the form*

$$u(t) = \sum_{k=1}^{\infty} \left((u_0, \varphi_k) e^{-\mu_k t} + \int_0^t (f(s), \varphi_k) e^{-\mu_k (t-s)} \, ds \right) \varphi_k \quad .$$

Proof:
So let u be a solution. Then, in particular

$$u(t) \in L^2(\Omega) \qquad \forall t > 0$$

and thus

$$u(t) = \sum_{k=1}^{\infty} (u(t), \varphi_k) \, \varphi_k \qquad \forall t > 0 \quad . \tag{7.22}$$

Inserting into the variational equality gives, with the use of (7.19) and (7.20) together with continuity of $(\cdot, \cdot)$ and $a(\cdot, \cdot)$ in $L^2(\Omega)$,

$$\frac{d}{dt} \sum_{k=1}^{\infty} (u(t), \varphi_k)(\varphi_k, v) + \sum_{k=1}^{\infty} (u(t), \varphi_k) \mu_k (\varphi_k, v) = (f(t), v) \qquad \forall v \in H_0^1(\Omega) \quad .$$

In the next step we set $v = \varphi_k$, $k = 1, 2 \ldots$, and achieve the ordinary differential equations

$$u_k'(t) + \mu_k u_k(t) = (f(t), \varphi_k) \tag{7.23}$$

for the (yet) unknown Fourier coefficients

$$u_k(t) = (u(t), \varphi_k) \quad , \quad k = 1, 2, \ldots \quad .$$

The initial conditions

$$u_k(0) = (u_0, \varphi_k) \tag{7.24}$$

follow by Fourier expansion of u_0 and equating coefficients.
The solution of the initial value problem (7.23), (7.24) is

$$u_k(t) = (u_0, \varphi_k) e^{-\mu_k t} + \int_0^t (f(s), \varphi_k) e^{-\mu_k (t-s)} \, ds \quad .$$

This formula, which can be easily verified for C^1-solutions u_k, can be derived for continuous right-hand side in (7.23) by *variation of constants.* Please note that this initial value problem can be solved for arbitrary right-hand side in $L^2(0,T)$ in the weak sense.
The proof for a unique solution of (7.23), (7.24) in this sense, given by the above formula, is left as an exercise for the reader.
Finally, inserting the formula for the u_k in the Fourier expansion (7.22) proves the claim. □

Conclusion 7.9 *The solution of the variational problem is determined uniquely as we know (in the case of existence) the Fourier coefficients.*

The goal of our investigations is the following concluding existence theorem.

Theorem 7.10 *The initial–boundary–value problem* (7.13) *has the unique solution* $u \in C([0,T], L^2(\Omega)) \cap L^2((0,T), H_0^1(\Omega))$ *given by*

$$u(t) = \sum_{k=1}^{\infty} \left((u_0, \varphi_k) e^{-\mu_k t} + \int_0^t (f(s), \varphi_k) e^{-\mu_k (t-s)} \, ds \right) \varphi_k \quad . \tag{7.25}$$

Proof:

1. Step 1

   We solve the problem for the approximation

$$u_{0,m} = \sum_{k=1}^{m} (u_0, \varphi_k)\varphi_k$$

   of u_0. The solution u_m is then given by

$$u_m(t) = \sum_{k=1}^{m} \left((u_0, \varphi_k) e^{-\mu_k t} + \int\limits_0^t (f(s), \varphi_k) e^{-\mu_k (t-s)} \, ds \right) \varphi_k \quad . \tag{7.26}$$

   This is shown by simply inserting and the use of the linearity of the weak derivative. Again we can regard u_m as a weak solution of the initial value problem.

2. Step 2

   We let $m \to \infty$ in: a) $C([0,T], L^2(\Omega))$; b) $L^2((0,T), H_0^1(\Omega))$.

   a) We show that u_m is a Cauchy sequence in $C([0,T], L^2(\Omega))$.

   Therefore, let $m < n$ arbitrarily but fixed. Then by the Parseval's equality and the Cauchy–Schwarz inequality, we get for each $t \in [0,T]$

$$\begin{aligned} \|u_n(t) - u_m(t)\|_{L^2(\Omega)} &\le \left(\sum_{k=m+1}^{n} (u_0, \varphi_k)^2 \right)^{\frac{1}{2}} + \left(\sum_{k=m+1}^{n} \left(\int\limits_0^t (f(s), \varphi_k) e^{-\mu_k (t-s)} \, ds \right)^2 \right)^{\frac{1}{2}} \\ &\le \left(\sum_{k=m+1}^{n} (u_0, \varphi_k)^2 \right)^{\frac{1}{2}} + \left(\sum_{k=m+1}^{n} \frac{1}{2\mu_k} \int\limits_0^T (f(s), \varphi_k)^2 \, ds \right)^{\frac{1}{2}} \quad . \end{aligned}$$

   For $n, m \to \infty$ the right hand side goes to 0 (and this uniformly in $t \in [0,T]$), as we have

$$\|u_0\|_{L^2(\Omega)} = \left(\sum_{k=1}^{\infty} (u_0, \varphi_k)^2 \right)^{\frac{1}{2}}$$

   and

$$\begin{aligned} \|f\|_{L^2((0,T), L^2(\Omega))} &= \left(\int\limits_0^T \|f(t)\|_{L^2(\Omega)}^2 \, dt \right)^{\frac{1}{2}} \\ &= \left(\int\limits_0^T \sum_{k=1}^{\infty} (f(t), \varphi_k)^2 \, dt \right)^{\frac{1}{2}} \quad . \end{aligned}$$

   Hereby, we use the Lebesque theorem of dominated convergence. Altogether we see that u_m is a Cauchy sequence in $C([0,T], L^2(\Omega))$, thus

$$u_m \to u_1^* \quad \text{in} \quad C([0,T], L^2(\Omega)) \quad . \tag{7.27}$$

   b) We show that u_m is a Cauchy sequence in $L^2((0,T), H_0^1(\Omega))$.

   Inserting of the representation (7.26) gives

$$a(u_n(t) - u_m(t), u_n(t) - u_m(t)) = \sum_{k=m+1}^{n} \mu_k \left((u_0, \varphi_k) e^{-\mu_k t} + \int\limits_0^t (f(s), \varphi_k) e^{-\mu_k (t-s)} \, ds \right)^2 \quad .$$

   Using the $H_0^1(\Omega)$-ellipticity of $a(\cdot,\cdot)$ and $(a+b)^2 \le 2(a^2+b^2)$, we get for a $c > 0$ and all $t \in [0,T]$

$$\begin{aligned} \|u_n(t) - u_m(t)\|_{H^1(\Omega)}^2 &\le \frac{1}{c} a(u_n(t) - u_m(t), u_n(t) - u_m(t)) \\ &\le \frac{2}{c} \sum_{k=m+1}^{n} \mu_k \left((u_0, \varphi_k)^2 e^{-2\mu_k t} + \left(\int\limits_0^t (f(s), \varphi_k) e^{-\mu_k (t-s)} \, ds \right)^2 \right) \quad . \end{aligned}$$

We have to integrate this estimate in t over $[0,T]$. For a further estimate on the upper bound we calculate

$$\mu_k \int_0^T e^{-2\mu_k t}\,dt = \frac{1}{2}(1-e^{-2\mu_k T}) < \frac{1}{2}$$

and

$$\begin{aligned}\mu_k \int_0^T \left(\int_0^t (f(s),\varphi_k) e^{-\mu_k(t-s)}\,ds \right)^2 dt &\le \mu_k \int_0^T \int_0^t e^{-2\mu_k(t-s)}\,ds \int_0^t (f(s),\varphi_k)^2\,ds\,dt \\ &\le \frac{T}{2} \int_0^T (f(t),\varphi_k)^2\,dt \quad .\end{aligned}$$

Thereby follows

$$\int_0^T \|u_n(t)-u_m(t)\|^2_{H^1(\Omega)}\,dt \le \frac{1}{c} \sum_{k=m+1}^{n} \left((u_0,\varphi_k)^2 + T\int_0^T (f(t),\varphi_k)^2\,dt \right) \quad .$$

As above the right-hand side of this estimate vanishes for $n,m\to\infty$. Thus u_m is a Cauchy sequence in $L^2((0,T),H_0^1(\Omega))$, and by the completeness of the space follows the convergence

$$u_m \to u_2^* \quad \text{in} \quad L^2((0,T),H_0^1(\Omega)) \quad . \tag{7.28}$$

3. Step 3 $(u_1^* = u_2^*)$
   By (7.27) follows
   $$u_m \to u_1^* \quad \text{in} \quad L^2((0,T),L^2(\Omega)) \quad ,$$
   and (7.28) gives
   $$u_m \to u_2^* \quad \text{in} \quad L^2((0,T),L^2(\Omega)) \quad .$$
   So $u_1^* = u_2^* = u^* \;\in\; L^2((0,T),H_0^1(\Omega)) \;\cap\; C([0,T],L^2(\Omega))$ must hold.
4. Step 4 (u^* is solution)
   The derivation of (7.28) gives for all $t\in[0,T]$ the convergence
   $$u_m(t)\to u^*(t) \quad \text{in} \quad H_0^1(\Omega) \quad .$$
   Thereby we deduce, as above, the uniform convergence of
   $$a(u_m(t),v)\to a(u^*(t),v)$$
   and
   $$(u_m(t),v)\to(u^*(t),v)$$
   in $t\in[0,T]$ for all $v\in H_0^1(\Omega)$. But as
   $$\frac{d}{dt}(u_m(t),v)+a(u_m(t),v)=(f(t),v)$$
   holds for all $m\in\mathbb{N}$ and all $v\in H_0^1(\Omega)$, by known convergence results follows that also u^* satisfies the variational equality in the sense of (7.18).
   To the initial condition: From (7.27) one gets
   $$u_m(0)\to u^*(0) \quad \text{in} \quad L^2(\Omega) \quad .$$
   But on the other hand
   $$u_m(0)=\sum_{k=1}^{m}(u_0,\varphi_k)\varphi_k \to u_0 \quad \text{in} \quad L^2(\Omega) \quad .$$
   Thus $u^*(0)=u_0$ must hold. Thereby $u=u^*$ is the solution of our problem. □

From the representation (7.25) of the solution follows directly the continuous dependence on the input data u_0 and f. This results from the following a priori estimate.

Theorem 7.11 *It holds for all* $t \in [0,T]$

$$\|u(t)\|_{L^2(\Omega)} \leq \|u_0\|_{L^2(\Omega)} \, e^{-\mu_1 t} + \int\limits_0^t \|f(s)\|_{L^2(\Omega)} \, e^{-\mu_1 (t-s)} \, ds \quad .$$

Proof:
For exercise. □

8 Numerical Methods for Parabolic Problems

8.1 Time Integration

We consider the initial value problem

$$u \in C^1([0,T],\mathbb{R}) \;:\quad u' = \varphi(t,u) \quad \text{für} \quad t \in (0,T] \quad , \quad u(0) = u_0 \quad .$$

For the simplicity of the discretization we assume an equidistant mesh

$$t_i \;=\; i\Delta t \qquad i = 0,\dots,N$$

We now consider the one-step method

$$\frac{1}{\Delta t}(u_{i+1} - u_i) \;=\; \Theta\,\varphi(t_{i+1},u_{i+1}) \;+\; (1-\Theta)\,\varphi(t_i,u_i)$$

for $\Theta \in [0,1]$.
Intuitively we replace the exact solution between u_i and u_{i+1} by a straight line with the mean gradient $\Theta u'(t_{i+1}) + (1-\Theta)u'(t_i)$.

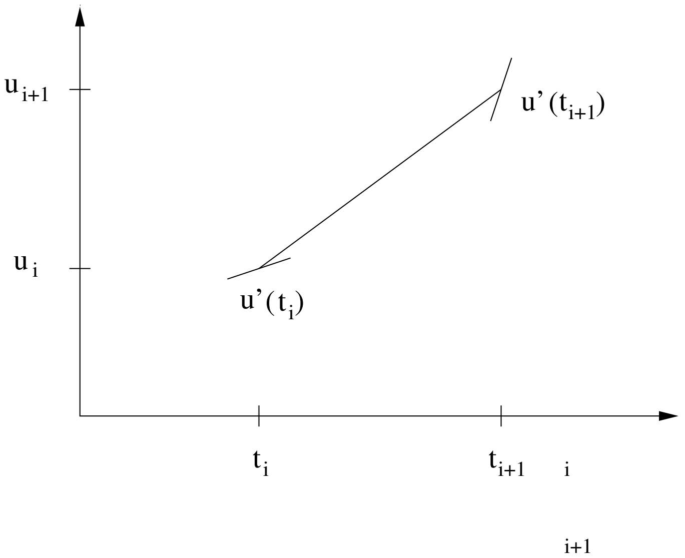


Figure 8.1: integration step

The values $\Theta = 0, \frac{1}{2}, 1$ are of specific interest. One gets:

$\Theta = 0$	explicit Euler method
$\Theta = \frac{1}{2}$	midpoint rule
$\Theta = 1$	implicit Euler method

Note that for $\Theta = \frac{1}{2}, 1$ the calculation of u_{i+1} might imply solving an in some cases nonlinear equation. Such methods are called *implicit*. For $\Theta = 0$, u_{i+1} is given *explicitly*.
We examine the *consistency* of these methods (as we did for the difference methods in the elliptic case) by looking at the truncation error. The truncation error is given by inserting the exact solution into the difference equation.
So we get by Taylor expansion

$$\begin{aligned}
\tau(t_i) &= \frac{1}{\Delta t}(u(t_{i+1}) - u(t_i)) - \Theta\,\varphi(t_{i+1}, u_{i+1}) \;-\; (1-\Theta)\,\varphi(t_i, u_i) \\
&= u'(t_i) + \frac{\Delta t}{2} u''(t_i) + \mathcal{O}\left((\Delta t)^2\right) - \big(\Theta(u'(t_{i+1}) - u'(t_i)) + u'(t_i)\big) \\
&= (\frac{1}{2} - \Theta)\,\Delta t\, u''(t_i) + \mathcal{O}\left((\Delta t)^2\right) \\
&= \begin{cases} \mathcal{O}\left(\Delta t\right) & \Theta \neq \frac{1}{2} \\ \mathcal{O}\left((\Delta t)^2\right) & \Theta = \frac{1}{2} \end{cases} \quad .
\end{aligned}$$

Thus the method is of first order for $\Theta \neq \frac{1}{2}$ and of second order for $\Theta = \frac{1}{2}$.
We consider the *stability* of the methods. Corresponding to the stability concept of Dahlquist (see [7, Kap. 6, 7]), we investigate the initial value problem

$$u' = -\lambda u \quad t > 0\ ,\ \lambda > 0 \quad \text{with} \quad u(0) = u_0 \tag{8.1}$$

with the solution $u(t) = u_0\, e^{-\lambda t}$ and demand that the corresponding approximations decay like u or at least do not grow. The use of our discretization leads to

$$\frac{1}{\Delta t}(u_{i+1} - u_i) = \Theta(-\lambda u_{i+1}) + (1-\Theta)(-\lambda u_i) \quad ,$$

so

$$\begin{aligned}
u_{i+1} &= \frac{1 - (1-\Theta)\,\Delta t\,\lambda}{1 + \Theta\,\Delta t\,\lambda}\, u_i \\
u_{i+1} &= \left(\frac{1 - (1-\Theta)\,\Delta t\,\lambda}{1 + \Theta\,\Delta t\,\lambda}\right)^i u_0 \quad .
\end{aligned}$$

Boundedness of $|u_i|$ is thus equivalent to

$$|R(\lambda\,\Delta t)| := \left|\frac{1 - (1-\Theta)\,\Delta t\,\lambda}{1 + \Theta\,\Delta t\,\lambda}\right| \leq 1 \quad , \tag{8.2}$$

and the "<"-sign takes care of the decay.
In the case $0 \leq \Theta < \frac{1}{2}$, (8.2) is equivalent to the step-size restriction

$$\lambda\,\Delta t \leq \frac{2}{1 - 2\Theta} \quad , \quad 0 \leq \Theta < \frac{1}{2} \quad ,$$

whereas (8.2) for $\frac{1}{2} \leq \Theta \leq 1$ always holds. With that we get

$\Theta = 0$	explicit Euler, stable if $\Delta t \leq \frac{2}{\lambda}$
$\Theta = \frac{1}{2}, 1$	midpoint rule, implicit Euler, always stable

There still is a slight difference between $\Theta = \frac{1}{2}$ and $\Theta = 1$. Obviously we have

$$\lim_{\Delta t \to 0} R(\Delta t\, \lambda) = |1 - \Theta^{-1}| \begin{cases} < 1 & \text{if } \Theta > \frac{1}{2} \\ \geq 1 & \text{if } \Theta \leq \frac{1}{2} \end{cases} \quad .$$

Hence the implicit Euler acts damping even for arbitrarily big time-steps. Such methods are called *strongly stable.* The midpoint rule doesn't have this property and thus, as we will see, is not so useful for long term calculations where many errors can accumulate.

8.2 Semidiscrete Methods

8.2.1 Method of Lines

Analogously to the Galerkin method for elliptic problems we now choose a sequence of finite dimensional subspaces

$$S_h \subset H_0^1(\Omega) \quad .$$

The discretization parameter $h = h_j$, $j = 0, 1, \ldots$ stands for example for the maximal diameter of all triangles of the triangulation.
We now want to approximate the continuous solution $u(t)$ by a $u_h(t) \in S_h$ for all $t \in [0, T]$. This corresponds to a discretization in space. In analogy to the continuous problem we now formulate the corresponding semidiscrete problem:
Find $u_h \in C([0, T], S_h)$, such that

$$\begin{aligned} \frac{d}{dt}(u_h(t), v) + a(u_h(t), v) = (f(t), v) \qquad \forall v \in S_h \\ u_h(0) = u_{0,h} \in S_h \quad . \end{aligned} \tag{8.3}$$

Note that the norms $\|\cdot\|_{L^2(\Omega)}$ and $\|\cdot\|_{H^1(\Omega)}$ are equivalent on the finite dimensional space S_h, i.e. there exists a $c > 0$ such that

$$c\,\|v\|_{H^1(\Omega)} \leq \|v\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)} \qquad \forall v \in S_h \quad .$$

Thus

$$C([0, T], S_{h,L^2(\Omega)}) = C([0, T], S_{h,H^1(\Omega)}) \subset L^2((0, T), S_{h,H^1(\Omega)}) \quad ,$$

where the additional index in S_h denotes the norm. Opposite to the continuous case (see theorem 7.10) the space $C([0, T], S_h)$ is already the solution space for the problem (8.3).

Lemma 8.1 *There exist functions $\varphi_{kh} \in S_h$, $k = 1, \ldots, N$, and corresponding $\mu_{k,h} \in \mathrm{I\!R}$ with*

$$0 < \mu_1 \leq \mu_{1,h} \leq \mu_{2,h} \leq \ldots \leq \mu_{N,h}$$

(where μ_1 comes from theorem 7.7), such that the following holds

$$\begin{aligned} a(\varphi_{k,h}, v) = \mu_{k,h}(\varphi_{k,h}, v) \quad \forall v \in S_h \quad , \\ \varphi_{k,h} \text{ is orthonormal basis of } S_{h,L^2(\Omega)} \quad . \end{aligned}$$

Proof:
See theorem 7.7 and the following remark 1). For the claim $\mu_1 \leq \mu_{1,h}$ see [18, p. 146]. □

Theorem 8.2 *The semidiscrete problem* (8.3) *has a uniquely determined solution* u_h, *and*

$$u_h(t) = \sum_{k=1}^{N} \left((u_{0,h}, \varphi_{k,h}) e^{-\mu_{k,h} t} + \int\limits_0^t (f(s), \varphi_{k,h}) e^{-\mu_{k,h}(t-s)}\, ds \right) \varphi_{k,h} \quad .$$

Proof:
Transcribe the proof to theorem 7.10. □

We want to show the convergence of u_h to u. An important step in that direction is the following a priori error estimate. To this end we define $C^1([0,T], H_0^1(\Omega))$ as the space of all $u \in C([0,T], H_0^1(\Omega))$, whose weak derivative $\frac{d}{dt}u(t) = u'(t)$ for each $t \in [0,T]$ appear as limit

$$\lim_{\Delta t \to 0} \frac{u(t + \Delta t) - u(t)}{\Delta t}$$

in the $H^1(\Omega)$-norm and which is in $C([0,T], H_0^1(\Omega))$. Therewith the in (8.3) appearing derivative is even classical for each fixed v (see (8.4) in the following proof).

Theorem 8.3 *Let* $u \in C^1([0,T], H_0^1(\Omega))$. *Then*

$$\begin{aligned} \|u_h(t) - u(t)\|_{L^2(\Omega)} \leq{}& \|u_{0,h} - P_h u_0\|_{L^2(\Omega)}\, e^{-\mu_1 t} + \|(I - P_h)u(t)\|_{L^2(\Omega)} \\ &+ \int\limits_0^t \left\|(I - P_h)\frac{du}{dt}(s)\right\|_{L^2(\Omega)} e^{-\mu_1(t-s)}\, ds \quad . \end{aligned}$$

holds for all $t \in [0,T]$. *Here* I *is the identity on* $H_0^1(\Omega)$ *and* $P_h : H_0^1(\Omega) \to S_h$ *the Ritz projection, hence*

$$P_h w \in S_h : \quad a(P_h w, v) = a(w, v) \qquad \forall v \in S_h$$

for all $w \in H_0^1(\Omega)$.

Proof:
Step 1: (defect problem)
As u and u_h fulfill the variational equality in (8.3) for all $S_h \subset H_0^1(\Omega)$, with the definition of P_h follows:

$$\frac{d}{dt}(u_h(t) - P_h\, u(t), v) + a(u_h(t) - P_h\, u(t), v) = \frac{d}{dt}(u(t) - P_h\, u(t), v) = \quad \forall v \in S_h \quad .$$

By the $H_0^1(\Omega)$-ellipticity of $a(\cdot,\cdot)$ follows the continuity of P_h, as P_h is a $a(\cdot,\cdot)$-orthogonal projection. Furthermore, due to $u \in C^1([0,T], H_0^1(\Omega))$, it holds

$$\begin{aligned} \left\|P_h\, u(t + \Delta t) - P_h\, u(t) - \Delta t\, P_h \frac{du}{dt}\right\|_{H^1(\Omega)} &\leq \text{const.} \left\|u(t + \Delta t) - u(t) - \Delta t \frac{du}{dt}\right\|_{H^1(\Omega)} \\ &= o(|\Delta t|) \end{aligned}$$

Thereby we have proven $P_h u \in C^1([0,T], H_0^1(\Omega))$ with

$$\left(\frac{d}{dt}P_h\right) u = P_h \frac{du}{dt} \quad .$$

Inserting gives

$$\frac{d}{dt}(u_h(t) - P_h\, u(t), v) + a(u_h(t) - P_h\, u(t), v) \;=\; (\frac{du}{dt} - P_h \frac{du}{dt}, v) \quad \forall v \in S_h \quad . \tag{8.4}$$

(How to justify $\frac{d}{dt}(u(t), v) = (\frac{du}{dt}, v)$?)
Moreover we have

$$u_h(0) - P_h u(0) = u_{0,h} - P_h u_0 \quad . \tag{8.5}$$

Step 2: (a priori estimate)
$u_h(t) - P_h u(t)$ is a solution of the defect problem (8.4), (8.5) and thus satisfies the a priori estimate from theorem 7.11. Thereby one first has to insert the discrete eigenvalue $\mu_{1,h}$ instead of μ_1, and by $\mu_1 \leq \mu_{1,h}$ follows the claim. □

Now we can proof the intended convergence theorem.

Theorem 8.4 *Let $u \in C^1([0,T], H_0^1(\Omega))$,*

$$\inf_{v_h \in S_h} \|v - v_h\|_{H^1(\Omega)} \to 0$$

for all $v \in H_0^1(\Omega)$ and also

$$\|u_{0,h} - u_0\|_{L^2(\Omega)} \to 0 \quad .$$

Then follows the convergence

$$u_h \to u \qquad \text{in } \; C([0,T], L^2(\Omega)) \quad .$$

Proof:
Step 1:
We consider the series $h_i \to 0$. For a fixed $v \in C([0,T], H_0^1(\Omega))$ we set

$$g_i := \|(I - P_{h_i})\, v(\cdot)\|_{H^1(\Omega)} \;\in\; C([0,T]) \quad .$$

We now want to show

$$\max_{t\in[0,T]} g_i(t) \to 0 \quad \text{für} \quad i \to \infty \quad . \tag{8.6}$$

The Céa-Lemma 5.16 gives for each fixed $t \in [0,T]$

$$\|(I - P_h)\, v(t)\|_{H^1(\Omega)} \leq \frac{\Gamma}{\gamma} \inf_{v_h \in S_h} \|v(t) - v_h\|_{H^1(\Omega)} \quad ,$$

thus by assumption

$$g_i(t) \to 0 \qquad \forall t \in [0,T] \quad .$$

As all P_h are $a(\cdot,\cdot)$-orthogonal projections, the estimate

$$\|P_h v(t)\|_{H^1(\Omega)} \;\leq\; \frac{\Gamma}{\gamma}\, \|v(t)\|_{H^1(\Omega)}$$

holds uniformly in h. This, together with the continuity of $v : [0,T] \to H_0^1(\Omega)$, gives the equicontinuity of $\{g_i \,|\, i \in \mathbb{N}\}$, i.e. for each $t^* \in [0,T]$ and each $\varepsilon > 0$ there exists a $\delta > 0$, such that

$$|t - t^*| < \delta \Longrightarrow |g_i(t) - g_i(t^*)| < \varepsilon \qquad \forall i \in \mathbb{N} \quad .$$

Now (8.6) follows by a simple contradiction or directly from the Arzelà-Ascoli theorem.

Step 2:
We use the result (8.6) in the case $v = u$ and $v = \frac{du}{dt}$ and get

$$\|u_{0,h} - P_h u_0\|_{L^2(\Omega)} \;\;\leq\;\; \|u_{0,h} - u_0\|_{L^2(\Omega)} + \|(I - P_h)u(0)\|_{L^2(\Omega)} \to 0 \quad ,$$

$$\max_{t\in[0,T]} \|(I - P_h)u(t)\|_{L^2(\Omega)} \;\;\leq\;\; \max_{t\in[0,T]} \|(I - P_h)u(t)\|_{H^1(\Omega)} \to 0 \quad ,$$

as well as

$$\max_{t\in[0,T]} \int_0^t \left\|(I-P_h)\frac{du}{dt}(s)\right\|_{L^2(\Omega)} e^{-\mu_1(t-s)}\,ds \;\le\; \frac{1}{\mu_1}\max_{s\in[0,T]}\left\|(I-P_h)\frac{du}{dt}(s)\right\|_{L^2(\Omega)} \to 0 \quad .$$

Now follows the claim from the a priori error estimate of theorem 8.3. □

The semidiscrete problem (8.3) is an initial value problem for a system of ordinary differential equations. Namely choosing a basis $\{\lambda_i\ ,\ i=1,\ldots,N\}$ of S_h, the representation

$$u_h(t) = \sum_{i=1}^N u_i(t)\lambda_i$$

leads to

$$\begin{array}{rcll} M\dot{U} + AU & = & F & t\in(0,T) \\ U(0) & = & U_0 & . \end{array} \tag{8.7}$$

Here we have

$$\begin{array}{rl} M = ((\lambda_i,\lambda_j))_{i,j=1}^N & \text{mass matrix} \\ A = (a(\lambda_i,\lambda_j))_{i,j=1}^N & \text{stiffness matrix} \\ F = ((f(t),\lambda_i))_{i=1}^N & \text{right-hand side} \\ U = (u_i(t))_{i=1}^N & \text{unknown-vector} \\ U_0 = (u_i(0))_{i=1}^N & u_{0,h} = \displaystyle\sum_{i=1}^N u_i(0)\lambda_i \quad . \end{array}$$

For solving problems of the form (8.7) there are many methods (and codes) available. However, due to the huge differences in the the decay of the components of the the solution (note the representation in theorem 8.2), this is a stiff system with certain stability problems.

Example:
Consider the problem (8.3) with the space S_h of the linear finite elements on $\Omega=(0,1)$ and $h=(N+1)^{-1}$ for some $N\in\mathbb{N}$. Constructing the system in (8.7) and joining the h-dependencies in the stiffness matrix by dividing through h, one obtains the $N\times N$-tridiagonal matrices

$$\frac{1}{h}M = \frac{1}{6}\begin{pmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & & \ddots & & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{pmatrix} \quad \text{und} \quad \frac{1}{h}A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \quad .$$

In practice the matrix M/h becomes the identity matrix by the so called “lumping”: Basically one simply sums up the row values and writes the sum to the diagonal. This is alike using the trapezoid rule as quadrature rule for the calculation of the values of M. Hereby one saves the construction of M^{-1}, and astonishingly, this simplification leads to the same approximation quality (see [21, ch. 15])! This is reasonable as the mass matrix M leads to a correlation of the time derivatives in different points—a phenomenon which occurs due to the discretization

and does not even exist in the continuous problem. In this sense, lumping can be regarded as a rectification of the discrepancy between the discrete and continuous problem.
The matrix A/h^2 has the eigenvalues $\lambda_i = 4h^{-2}\sin^2(i\frac{\pi}{2}h)$, $i = 1, \dots, N$ (see the proof of theorem 6.20), thus $\lambda_i \approx (i\frac{\pi}{2})^2$ for small h, i.e. a finer space discretization leads to a bigger range of eigenvalues. A numeric treatment of the arising initial value problem with explicit methods would now, like in section 8.1, require a step-size restriction in the time. Concretely, as λ_N is of the order of h^{-2}, one would have to attend the constraint

$$\Delta t \leq \frac{1}{2(1-\Theta)} h^2 \quad .$$

Thus one has to use implicit methods for such problems, where these stability assumptions do not have to be satisfied (which here is the case for the fully implicit method with $\Theta = 1$).
The need of a time-step restriction for $\Theta = 0$ (explicit) follows from a investigation of the *numerical domain of dependence.*

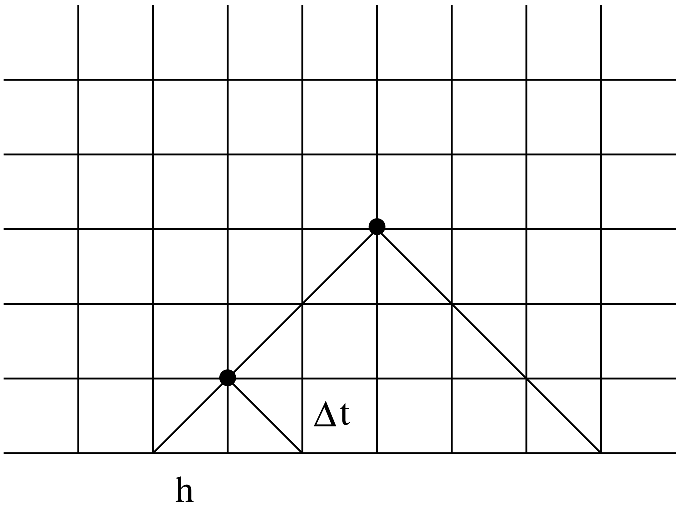


Figure 8.2: numerical domain of dependence

After the first time-step one would have to calculate the vector $U_1 = (U_{i,1})_{i=1}^N$ with

$$U_1 = U_0 - (\Delta t/h^2)AU_0 + \Delta t\, F(0)$$

as approximation to $(u_i(\Delta t))_{i=1}^N$.
But, as seen in section 7.1.2 (see the last remark: Greens function!), the domain of dependence of $u(x_i, \Delta t)$ consists of the whole interval $(0, 1)$, the numerical domain of dependence of $U_{i,1}$ is only $u_{i-1}(0)$, $u_i(0)$ and $u_{i+1}(0)$. Now we need $\frac{\Delta t}{h} \to 0$ to widen this cone asymptotically, which means Δt has to vanish faster then h. Otherwise one could change U_0 beyond $x_i \pm h$ $(x_i = i/N)$, without influencing the approximation $U_{i,1}$ of $u(x_i, \Delta t)$ Stability rule

(Courant, Friedrichs, Levi):
The numerical domain of dependence has to approximate the continuous one.

This CFL-condition is an important rule of thumb for the assessment of the stability properties of a numerical method.

The completely implicit method needs no time-step restriction. In this case the numerical domain of dependence consists of the whole previous space grid. In particular one has to solve the equation system

$$(I + (\Delta t/h^2)A)U_1 = U_0 + \Delta t\, F(0)$$

and one can show that the matrix $(I + (\Delta t/h^2)A)^{-1}$, which has to be applied to the previous time layer, is fully populated with positive coefficients. ◁

For further considerations regarding the stability of problems of the kind (8.7) we refer to the book of P. Deuflhard and F. Bornemann [7].

Note:
A change of the dimension of S_h, and thus of the accuracy of the approximation, changes the number of unknowns in (8.7) and requires a restart of the time integration. This imposes constraints on the *adaptive choice* of the space-discretization, which become painful in 2 and more space dimensions. ◁

8.2.2 Rothe's Method

In the proof of existence in section 7.1.2 as well as in the method of lines we replaced the initial–boundary–value problem for the heat equation with a initial value problem for a big system of ordinary differential equations. We now want to formulate the weak formulation of 7.13 (see (7.18)) directly as an initial value problem in an *infinite dimensional* function space. For this purpose we need some preparation.
Each $g \in L^2(\Omega)$ defines by $v \mapsto (g, v)$, $v \in H_0^1(\Omega)$ a bounded linear functional on $H_0^1(\Omega)$, thus in this sense

$$H_0^1(\Omega) \subset L^2(\Omega) \subset H_0^1(\Omega)' \ !$$

For each $w \in H_0^1(\Omega)$ is $a(w, \cdot) \in H_0^1(\Omega)'$! Thereby

$$Aw \in L^2(\Omega) : \quad (Aw, v) = a(w, v) \qquad \forall v \in H_0^1(\Omega) \tag{8.8}$$

defines a linear (unbounded!) mapping

$$A : \quad D(A) \subset H_0^1(\Omega) \to L^2(\Omega)$$

where

$$D(A) := \{w \in H_0^1(\Omega) \,|\, \text{ (8.8) has a solution } \} \quad .$$

A is called L^2*-representation* of the bilinear form $a(\cdot, \cdot)$.

Example:
In the case $a(v, w) = \int_\Omega \nabla v \, \nabla w \, dx$ it holds

$$D(A) = H^2(\Omega) \cap H_0^1(\Omega)$$

and

$$Aw = -\Delta w \quad , \quad w \in D(A) \quad .$$

By completion of $C^1([0,T],L^2(\Omega))$ with respect to the scalar product

$$(v,w)_{H^1((0,T),L^2(\Omega))} := \int_0^T (v(t),w(t))\,dt + \int_0^T (v'(t),w'(t))\,dt \quad ,$$

we obtain the space $H^1((0,T),L^2(\Omega))$. This space consists, as in the real valued case, of all (equivalence classes of) functions $v \in L^2((0,T),L^2(\Omega))$ whose weak derivative v' again is in $L^2((0,T),L^2(\Omega))$. Thereby v' is called weak derivative if

$$\int_0^T (v(t),\varphi'(t))\,dt = -\int_0^T (v'(t),\varphi(t))\,dt$$

holds for all $\varphi \in C_0^\infty((0,T),L^2(\Omega))$. (One defines the space $C_0^\infty((0,T),L^2(\Omega))$ analogous to the case of real valued functions)
We now consider the following initial value problem in $L^2(\Omega)$:
Find $u \in L^2((0,T),H_0^1(\Omega)) \cap H^1((0,T),L^2(\Omega))$ such that

$$\begin{aligned} u' - \Delta u &= f \\ u(0) &= u_0 \end{aligned} \tag{8.9}$$

holds in $L^2(\Omega)$. The initial value is attained continuously in $L^2(\Omega)$, since we have as with real valued functions

Lemma 8.5

$$H^1((0,T),L^2(\Omega)) \subset C([0,T],L^2(\Omega)) \quad .$$

The relationship to weak solution is explained by the following theorem.

Theorem 8.6 *Each solution of* (8.9) *is a weak solution of* (7.13)*. Conversely, each solution* $u \in H^1((0,T),L^2(\Omega))$ *of* (7.13) *a solution of* (8.9).

Proof:
Let u be a solution of (8.9). Multiplication with an arbitrary $v \in H_0^1(\Omega)$ and integration over Ω gives

$$(u'(t),v) + (Au(t),v) = (f(t),v) \qquad \forall t \in (0,T) \quad . \tag{8.10}$$

Let $\varphi \in C_0^\infty([0,T])$. Then $v\varphi \in C_0^\infty((0,T),L^2(\Omega))$ with $(v\varphi)' = v\varphi'$ holds (why?), and due to $u \in H^1((0,T),L^2(\Omega))$

$$\begin{aligned} \int_0^T (u'(t),v)\varphi(t)\,dt = \int_0^T (u'(t),v\varphi(t))\,dt &= -\int_0^T (u(t),v\varphi'(t))\,dt \\ &= -\int_0^T (u(t),v)\varphi'(t)\,dt \quad . \end{aligned}$$

Thus $(u(\cdot),v)$ is weakly differentiable with

$$(u'(t),v) = \frac{d}{dt}(u(t),v) \qquad \forall v \in H_0^1(\Omega) \quad . \tag{8.11}$$

By definition also

$$(Au, v) = a(u, v) \qquad \forall v \in H_0^1(\Omega) \quad .$$

Inserting into (8.10) leads to

$$\frac{d}{dt}(u(t), v) + (Au(t), v) = (f(t), v) \qquad \forall v \in (0, T) \quad .$$

Considering lemma 8.5, u is a weak solution.
Let conversely $u \in H^1((0,T), L^2(\Omega))$ be a weak solution, we again have (8.11) and thus

$$a(u(t), v) = (f(t) - u'(t), v) \qquad \forall v \in H_0^1(\Omega) \quad .$$

As $f(t) - u'(t) \in L^2(\Omega)$, $u(t) \in D(A)$ holds for all $t \in (0, T)$, and we get

$$(u'(t) + Au(t) - f(t), v) = 0 \qquad \forall v \in H_0^1(\Omega) \quad .$$

Finally, $H_0^1(\Omega) \subset L^2(\Omega)$ is dense and thus

$$u'(t) + Au(t) = f(t) \quad \text{f.ü. in } (0, T) \quad .$$

The formulation (8.9) is the origin of semigroup methods for evolutionary problems. We refer to [19, chapter 111] and the there cited literature.

Main idea of the Rothe method: (Original article [20])

a) Semi-discretization in t: Transfer known techniques for solving ordinary initial value problem in $\mathbb{R}^m$ to the initial value problem (8.9) (order and step-size control)

b) Discretization in space: The resulting space problems have to be solved so precisely that the properties of the time discretization survive.

Example: (Implicit Euler method for step-size Δt)
In each time-step one has to solve the space problem

$$U(t_i) - U(t_{i-1}) + \Delta t\, AU(t_i) = \Delta t\, f(t_i)$$

or equivalently the variational problem

$$(U(t_i), v) + \Delta t\, a(U(t_i), v) = (U(t_{i-1}) + \Delta t\, f(t_i), v) \qquad \forall v \in H_0^1(\Omega) \quad . \tag{8.12}$$

For solving (8.12) one now can use known methods (and codes) for elliptic problems. In particular one can have a different space grid for each time-step. To keep the accuracy requirements for the space discretization as low as possible, one should use preferably robust time integrators. For example, extrapolation methods are ruled out, as the underlying problem is badly conditioned for growing orders. Details can be found in the newest works of F. Bornemann [3, 4, 5]. ◁

Bibliography

[1] R.A. Adams. *Sobolev Spaces.* Academic Press, 1975.

[2] H. Alt. *Lineare Funktionalanalysis.* Springer, 2. edition, 1992.

[3] F. Bornemann. An adaptive multilevel approach to parabolic equations i: general theory and 1d implementation. *IMPACT Comput. Sci. Engrg.*, 2:279–317, 1990.

[4] F. Bornemann. An adaptive multilevel approach to parabolic equations ii: variable order time discretization based on a multiplicative error correction. *IMPACT Comput. Sci. Engrg.*, 3:93–122, 1991.

[5] F. Bornemann. An adaptive multilevel approach to parabolic equations iii: 2d error estimation and multilevel. *IMPACT Comput. Sci. Engrg.*, 4:1–45, 1992.

[6] D. Braess. *Finite Elemente.* Springer, 2. edition, 1997.

[7] P. Deuflhard and F. Bornemann. *Numerische Mathematik II.* de Gruyter, Berlin, 1994.

[8] P. Deuflhard and A. Hohmann. *Numerische Mathematik I.* de Gruyter, Berlin, 2. edition, 1993.

[9] P. Deuflhard and M. Weiser. *Numerische Mathematik 3: Adaptive Lösung partieller Differentialgleichungen.* de Gruyter, Berlin, 2011.

[10] P.L. George. *Automatic Mesh Generation.* Wiley, 1991.

[11] P.L. George. *Delaunay triangulation and meshing.* Hermes, Paris, 1998.

[12] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen.* Teubner, 1986.

[13] F. John. *Partial Differential Equations.* Springer, 4. edition, 1982.

[14] J. Jost. *Partielle Differentialgleichungen.* Springer, 1998.

[15] R. Kornhuber. Theorie und numerik partieller differentialgleichungen, skript ws 98/99. FU Berlin.

[16] N. Neuss. V-cycle convergence with unsymmetric smoothers and application to an anisotropic model problem. *SIAM J. Num. Anal.*, 35:1201–1212, 1998.

[17] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C.* Cambridge University Press, 1992.

[18] P.A. Raviart and J.M. Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles.* Masson, 1992.

[19] M. Renardy and R.C. Rogers. *An Introduction to Partial Differential Equations.* Springer, 1993.

[20] E. Rothe. Zweidimensionale randwertaufgaben als grenzfall eindimensionaler randwertaufgaben. *Math. Ann.*, 102:650–670, 1930.

[21] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems.* Springer, 1997.

[22] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 34:581–613, 1992.

[23] H. Yserentant. Old and new convergence proofs for multigrid methods. *Acta Numerica*, pages 285–326, 1993.