Introduction to Partial Differential Equations

Lecture Notes

Fall 2018

Preliminary version – Do not distribute

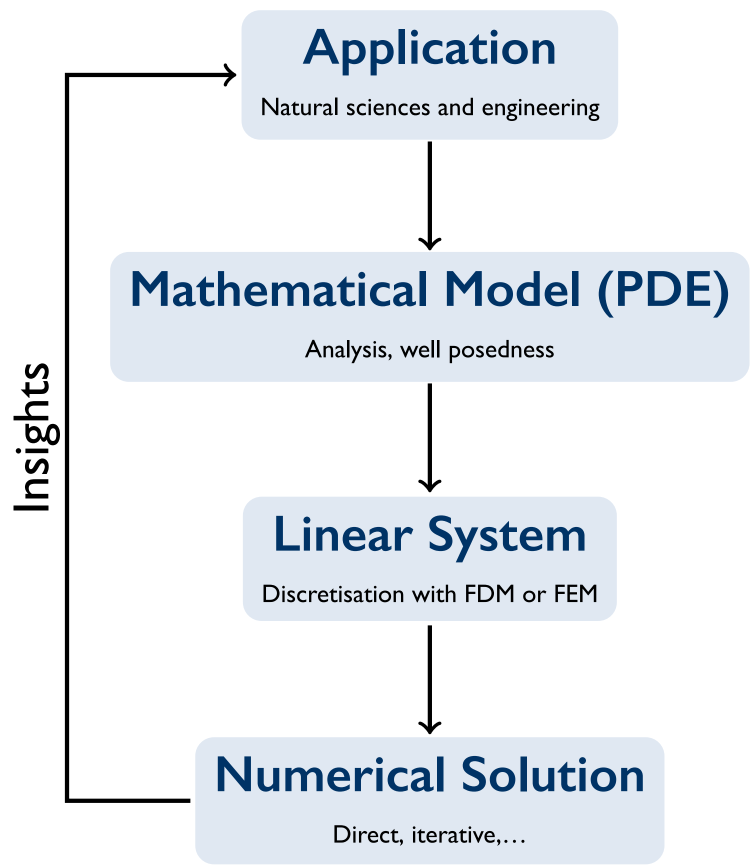
Michael Multerer

December 7, 2018

Introduction

Habt Euch vorher wohl präpariert,
Paragraphos wohl einstudiert,
Damit Ihr nachher besser seht,
Daß er nichts sagt, als was im Buche steht;
Doch Euch des Schreibens ja befleißt,
Als diktiert, Euch der Heilig Geist!

–J. W. v. Goethe, Faust I



Contents

1. Partial Differential Equations

In what follows, let $\Omega \subset \mathbb{R}^d$ for $d \in \{1, 2, 3, 4\}$ denote a connected, open set with Lipschitz continuous boundary $\Gamma := \partial D$. We call Ω a *domain*. In this course, we focus on linear second order partial differential equations of the form

$$(\mathcal{L}u)(\boldsymbol{x}) := -\sum_{i,j=1}^{d} a_{i,j}(\boldsymbol{x}) u_{x_i x_j}(\boldsymbol{x}) + \sum_{i=1}^{d} b_i(\boldsymbol{x}) u_{x_i}(\boldsymbol{x}) + c(\boldsymbol{x}) u(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \tag{1.1}$$

where we use the abbreviations

$$u_{x_i} := \frac{\partial u}{\partial x_i} \quad \text{and} \quad u_{x_i x_j} := \frac{\partial^2 u}{\partial x_i \partial x_j}.$$

Often, equation (1.1) is rewritten in vector notation according to

$$-\operatorname{div}\big(\boldsymbol{A}(\boldsymbol{x}) \nabla u(\boldsymbol{x})\big) + \boldsymbol{b}^{\intercal}(\boldsymbol{x}) \nabla u(\boldsymbol{x}) + c(\boldsymbol{x}) u(\boldsymbol{x}) = f(\boldsymbol{x})$$

with

$$\boldsymbol{A}(\boldsymbol{x}) := \begin{bmatrix} a_{1,1}(\boldsymbol{x}) & \cdots & a_{1,d}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ a_{d,1}(\boldsymbol{x}) & \cdots & a_{d,d}(\boldsymbol{x}) \end{bmatrix} \quad \text{and} \quad \boldsymbol{b}(\boldsymbol{x}) := \begin{bmatrix} \tilde{b}_1(\boldsymbol{x}) \\ \vdots \\ \tilde{b}_d(\boldsymbol{x}) \end{bmatrix}.$$

In this context, for a given given scalar field $v \colon \mathbb{R}^d \to \mathbb{R}$ and a vector field $\boldsymbol{f} \colon \mathbb{R}^d \to \mathbb{R}^d$, we set

$$\nabla v := \big[v_{x_1}, \ldots, v_{x_d}\big]^{\intercal} \quad \text{and} \quad \operatorname{div} \boldsymbol{f} := \sum_{i=1}^{d} \frac{\partial f_i}{\partial x_i}.$$

Definition 1.1

Multi indices We call $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_d]^{\intercal} \in \mathbb{N}$ a *multi index* and set

$$\partial^{\boldsymbol{\alpha}} := \frac{\partial^{|\boldsymbol{\alpha}|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}, \quad \text{where } |\boldsymbol{\alpha}| := \alpha_1 + \ldots + \alpha_d.$$

Moreover, we set

$$\boldsymbol{x}^{\boldsymbol{\alpha}} := x_1^{\alpha_1} \cdots x_d^{\alpha_d}$$

and

$$\boldsymbol{\alpha}! := \alpha_1! \cdots \alpha_d!, \quad \binom{\boldsymbol{\alpha}}{\boldsymbol{\beta}} := \binom{\alpha_1}{\beta_1} \cdots \binom{\alpha_d}{\beta_d}.$$

Classical Function Spaces Let $\Omega \subset \mathbb{R}^d$ denote an open set. For $s \in \mathbb{N}$, we define the spaces

$$C^s(\Omega) := \{v \colon \Omega \to \mathbb{R} : \partial^{\boldsymbol{\alpha}} v \text{ is continuous for all } |\boldsymbol{\alpha}| \leq s\}.$$

If Ω is bounded, we can further define the norms

$$\|v\|_{C^s(\overline{\Omega})} := \max_{|\boldsymbol{\alpha}| \leq s, \boldsymbol{x} \in \overline{\Omega}} |\partial^{\boldsymbol{\alpha}} v(\boldsymbol{x})|$$

with the corresponding spaces

$$C^s(\overline{\Omega}) := \{v \colon \overline{\Omega} \to \mathbb{R} : v \text{ is continuous for all } |\boldsymbol{\alpha}| \leq s \text{ and } \|v\|_{C^s(\overline{\Omega})} < \infty\}.$$

These spaces are complete with respect to the norm $\|\cdot\|_{C^s(\overline{\Omega})}$ and hence *Banach spaces*.

Space of Test Functions We further define

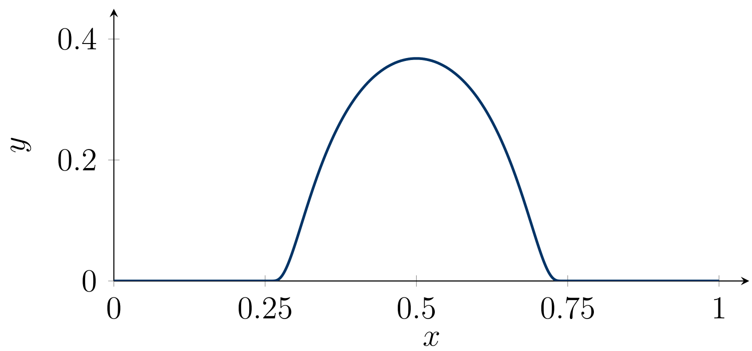
$$C^\infty(\Omega) := \bigcap_{s \in \mathbb{N}} C^s(\Omega).$$

Then, the *space of test functions* is defined as

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

where $\operatorname{supp}(v) := \overline{\{\boldsymbol{x} \in \Omega : v(\boldsymbol{x}) \neq 0\}}$ is the *support* of the function $v \colon \Omega \to \mathbb{R}$.

Example 1.2



The function

$$\phi(x) := \begin{cases} e^{\frac{-1}{1-(4x-2)^2}}, & 0.25 < x < 0.75, \\ 0, & \text{otherwise}, \end{cases}$$

is in $C_0^\infty(0,1)$, where $\operatorname{supp}(\phi) = [0.25, 0.75]$. △

In what follows, we make the assumption $u \in C^2(\Omega)$. Therefore, Schwartz's theorem yields $u_{x_i x_j} = u_{x_j x_i}$. Hence, without loss of generality the matrix $\boldsymbol{A}(\boldsymbol{x})$ is symmetric, i.e. $a_{i,j}(\boldsymbol{x}) = a_{j,i}(\boldsymbol{x})$, and has only real eigen values. Depending on the eigen values of $\boldsymbol{A}$, we distinguish three different types of partial differential equations that have to be supplemented by appropriate boundary- and/or initial- conditions in order to obtain a *well posed problem*, i.e. the problem exhibits a unique solution which depends continuously on the given data.

Definition 1.3 The differential operator $\mathcal{L}$ from (1.1) is called...

- ...*elliptic* at $\boldsymbol{x} \in \Omega$, iff all eigen values of $\boldsymbol{A}(\boldsymbol{x})$ are positive.
- ...*parabolic* at $\boldsymbol{x} \in \Omega$, iff $d-1$ eigen values of $\boldsymbol{A}(\boldsymbol{x})$ are positive, one eigen value is zero and $\operatorname{rank}[\boldsymbol{A}(\boldsymbol{x}), \boldsymbol{b}(\boldsymbol{x})] = d$.
- ...*hyperbolic* at $\boldsymbol{x} \in \Omega$, iff $d-1$ eigen values of $\boldsymbol{A}(\boldsymbol{x})$ are positive and one eigen value is negative.

The differential operator $\mathcal{L}$ is called *elliptic/parabolic/hyperbolic*, iff it is *elliptic/parabolic/hyperbolic* for every $\boldsymbol{x} \in \Omega$. Accordingly, equation (1.1) is called *elliptic/parabolic/hyperbolic*, iff the underlying differential operator exhibits this property.

Next, we shall consider three different examples, which are prototypes for the three different types of partial differential equations.

1.1 The Plateau Problem

We consider a soap film that is supported by a wireframe. This wireframe shall be represented by a smooth and closed curve in $\mathbb{R}^3$. We assume that its parallel projection onto the (x, y)-plane has no double points. Then, the shape of the soap film can be described by the graph of a function $u\colon \Omega \to \mathbb{R}$, while the wireframe is the graph of a function $g\colon \Gamma \to \mathbb{R}$. Due to surface tension, the soap film minimises its surface area

$$\int_\Omega \sqrt{1 + u_x^2 + u_y^2}\,\mathrm{d}x\,\mathrm{d}y \to \min.$$

To solve this nonlinear variational problem approximately, we employ the first order Taylor expansion $\sqrt{1+z} = 1 + \frac{z}{2} + \mathcal{O}(z^2)$. Hence, for small values of u_x and u_y, we may replace the integrand by a quadratic expression. We arrive at the minimisation problem

$$F(u) := \frac{1}{2}\int_\Omega u_x^2 + u_y^2\,\mathrm{d}x\,\mathrm{d}y \to \min. \tag{1.2}$$

The values of u at Γ are prescribed by the position of the wireframe, i.e. $u|_\Gamma = g$.
Let $u \in C^2(\Omega) \cap C(\overline{\Omega})$, $u|_\Gamma = g$ be a solution to (1.2). Then, for any $v \in C^1(\Omega) \cap C(\overline{\Omega})$ with $v|_\Gamma = 0$, the Gâteaux derivative satisfies

$$0 = \lim_{\varepsilon \to 0} \frac{F(u + \varepsilon v) - F(u)}{\varepsilon} = \int_\Omega u_x v_x + u_y v_y\,\mathrm{d}x\,\mathrm{d}y = \int_\Omega \langle \nabla u, \nabla v \rangle\,\mathrm{d}\boldsymbol{x}, \tag{1.3}$$

where $\langle \cdot, \cdot \rangle\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ denotes the dot product in $\mathbb{R}^d$.
On the other hand, we have for a continuously differentiable vector field $\boldsymbol{f}$ by the *divergence theorem*

$$\int_\Omega \operatorname{div} \boldsymbol{f}\,\mathrm{d}\boldsymbol{x} = \int_\Gamma \langle \boldsymbol{f}, \boldsymbol{n} \rangle\,\mathrm{d}\sigma,$$

where $\boldsymbol{n}$ denotes the outward pointing normal vector. Hence, for $\boldsymbol{f} := \nabla u v$, the multivariate product rule yileds

$$\int_\Omega \Delta u v\,\mathrm{d}\boldsymbol{x} + \int_\Omega \langle \nabla u, \nabla v \rangle\,\mathrm{d}\boldsymbol{x} = \int_\Omega \operatorname{div} \boldsymbol{f}\,\mathrm{d}\boldsymbol{x} = \int_\Gamma \langle \boldsymbol{f}, \boldsymbol{n} \rangle\,\mathrm{d}\sigma = \int_\Gamma v \langle \nabla u, \boldsymbol{n} \rangle\,\mathrm{d}\sigma = 0,$$

since $v|_\Gamma = 0$. This can be rewritten as

$$\int_\Omega -\Delta u v \, \mathrm{d}\boldsymbol{x} = \int_\Omega \langle \nabla u, \nabla v \rangle \, \mathrm{d}\boldsymbol{x}.$$

Inserting the latter into (1.3), leads to

$$\int_\Omega -\Delta u v \, \mathrm{d}\boldsymbol{x} = 0.$$

Since this equation holds for any *test function* $v \in C^1(\Omega) \cap C(\overline{\Omega})$ with $v|_\Gamma = 0$, the *fundamental lemma of calculus of variations* yields *Laplace's equation*

$$\Delta u(\boldsymbol{x}) = 0 \quad \text{for } \boldsymbol{x} \in \Omega.$$

Summary The first order approximation of the Plateau problem leads to Laplace's equation

$$\Delta u(\boldsymbol{x}) = 0 \quad \text{for } \boldsymbol{x} \in \Omega.$$

This equation is of the form (1.1) with

$$\boldsymbol{A}(\boldsymbol{x}) := \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \in \mathbb{R}^{d \times d}, \quad \boldsymbol{b}(\boldsymbol{x}) = \boldsymbol{0} \in \mathbb{R}^d, \quad c(\boldsymbol{x}) = 0$$

and thus elliptic.

In order to obtain a well posed problem, the equation has to be supplemented by boundary conditions. Let g, κ be continuous functions. Common boundary conditions are ...

- **...Dirichlet conditions**

$$u = g \quad \text{on } \Gamma.$$

- **...Neumann conditions**

$$\langle \boldsymbol{n}, \boldsymbol{A} \nabla u \rangle = g \quad \text{on } \Gamma.$$

- **...Robin conditions**

$$\langle \boldsymbol{n}, \boldsymbol{A} \nabla u \rangle + \kappa u = g \quad \text{on } \Gamma.$$

1.2 The Heat Equation

Let $u \colon [0, \infty) \times \Omega \to \mathbb{R}$ denote the distribution of temperature in an object. Moreover, we denote by $f \in C\big([0, \infty) \times \Omega\big)$ a heat source inside the object. We consider the balance of heat in a control volume $V \subset \Omega$. The energy principle states that the rate of change in the total energy in V is comprised of the inflow of heat via the surface ∂V and the heat injection f. Hence, there holds for the energy $E(t, \boldsymbol{x})$ that

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_V E \, \mathrm{d}\boldsymbol{x} = -\int_{\partial V} \langle \boldsymbol{j}, \boldsymbol{n} \rangle \, \mathrm{d}\sigma + \int_V f \, \mathrm{d}\boldsymbol{x},$$

where the vector field $\boldsymbol{j}(t, \boldsymbol{x})$ denotes the heat flux. Now, the application of the divergence theorem yields

$$\int_V \frac{\partial E}{\partial t} + \operatorname{div} \boldsymbol{j} - f \,\mathrm{d}\boldsymbol{x} = 0 \quad \text{for } t > 0$$

and hence, since $V \subset \Omega$ can be chosen arbitrarily

$$\frac{\partial E}{\partial t}(t, \boldsymbol{x}) + \operatorname{div} \boldsymbol{j}(t, \boldsymbol{x}) = f(t, \boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \Omega,\ t > 0.$$

In accordance with *Fourier's law*, we assume that

$$\boldsymbol{j} = -\kappa \nabla u,$$

where κ is a material dependent diffusion constant. Making further the assumption that the energy depends linearly on the temperature, i.e. $E = E_0 + \lambda u$ for some $\lambda \in \mathbb{R}$, we obtain the *heat equation*

$$\frac{\partial u}{\partial t} - \frac{\kappa}{\lambda} \Delta u = \frac{f}{\lambda} \quad \text{for } \boldsymbol{x} \in \Omega,\ t > 0.$$

Summary Based on the conservation of energy, we have derived the heat equation

$$\frac{\partial u}{\partial t} - \kappa \Delta u = f \quad \text{for } \boldsymbol{x} \in \Omega,\ t > 0.$$

This equation is of the form (1.1) with

$$\boldsymbol{A}(\boldsymbol{x}) := \kappa \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{bmatrix} \in \mathbb{R}^{\tilde{d} \times \tilde{d}}, \quad \boldsymbol{b}(\boldsymbol{x}) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^{\tilde{d}}, \quad c(\boldsymbol{x}) = 0, \quad \tilde{d} := d + 1$$

and thus parabolic. The additional dimension denotes the time.

In order to obtain a well posed problem, the equation has to be supplemented by boundary conditions and an initial condition at $t = 0$.

1.3 The Wave Equation

The motion of molecules in an ideal gas is described by three laws. In what follows, we denote the velocity by $\boldsymbol{v}$, the density by ρ, and the pressure by p.

Due to the conservation of mass, the change of mass, i.e. $\int_V \frac{\partial \rho}{\partial t} \,\mathrm{d}\boldsymbol{x}$ in a volume $V \subset \Omega$ equals the flow through its surface, i.e. $\int_{\partial V} \rho \langle \boldsymbol{v}, \boldsymbol{n} \rangle \,\mathrm{d}\sigma$. As in the derivation of the heat equation, we arrive at the *continuity equation*

$$\frac{\partial \rho}{\partial t} = -\rho_0 \operatorname{div} \boldsymbol{v},$$

where ρ is approximated by the fixed density ρ_0.

By *Newton's third law*, the pressure gradient induces an accelaration of the molecules, i.e.

$$\rho_0 \frac{\partial \boldsymbol{v}}{\partial t} = -\nabla p.$$

Finally, the *ideal gas law* states that the pressure is proportional to the density for constant temperature, i.e.

$$p = c^2 \rho, \quad c > 0.$$

Combining these three laws yields

$$\frac{\partial^2 p}{\partial t^2} = c^2 \frac{\partial^2 \rho}{\partial t^2} = c^2 \frac{\partial}{\partial t}\big(-\rho_0 \operatorname{div} \boldsymbol{v}\big) = -c^2 \operatorname{div}\left(\rho_0 \frac{\partial \boldsymbol{v}}{\partial t}\right) = c^2 \operatorname{div}(\nabla p) = c^2 \Delta p.$$

Summary Based on the conservation of mass, Newton's third law and the ideal gas law, we have derived the *wave equation*

$$\frac{\partial^2 u}{\partial t^2} - c^2 \Delta u = 0 \quad \text{for } \boldsymbol{x} \in \Omega,\ t > 0.$$

This equation is of the form (1.1) with

$$\boldsymbol{A}(\boldsymbol{x}) := \begin{bmatrix} c^2 & & & \\ & \ddots & & \\ & & c^2 & \\ & & & -1 \end{bmatrix} \in \mathbb{R}^{\tilde{d}\times\tilde{d}}, \quad \boldsymbol{b}(\boldsymbol{x}) = \boldsymbol{0} \in \mathbb{R}^{\tilde{d}}, \quad c(\boldsymbol{x}) = 0, \quad \tilde{d} := d+1$$

and thus hyperbolic. The additional dimension denotes the time.

In order to obtain a sensible problem, the equation has to be supplemented by boundary conditions and two initial conditions at $t = 0$, one for u and one for $\partial u/\partial t$.

1.4 Analytical Solutions in One Dimension

Poisson's Equation

We consider the one dimensional *Poisson's equation* with Dirichlet boundary conditions, i.e.

$$\begin{aligned} -u''(x) &= f(x) \quad \text{for } x \in \Omega := (0,1), \\ u(0) &= u_0,\ u(1) = u_1. \end{aligned}$$

The application of the *Fundamental theorem of calculus* yields

$$u'(x) = -\int_0^x f(s)\,\mathrm{d}s + \alpha$$

and hence

$$u(x) = \int_0^x u'(y)\,\mathrm{d}y + \beta = -\int_0^x \int_0^y f(s)\,\mathrm{d}s\,\mathrm{d}y + \alpha x + \beta$$

for some constants $\alpha, \beta \in \mathbb{R}$.

By considering $x = 0$ and $x = 1$, the constants can be determined according to

$$\alpha = u_1 - u_0 + \int_0^1 \int_0^y f(s)\,\mathrm{d}s\,\mathrm{d}y, \quad \beta = u_0.$$

By differentiation it can be seen that α and β and therefore $u(x)$ are uniquely determined. Moreover, the solution depends continuously on the data. Hence, the one dimensional Dirichlet problem is well posed.

The Heat Equation

We consider the one dimensional heat equation with an initial condition and Dirichlet boundary conditions, i.e.

$$\begin{aligned} u_t(t,x) &= u_{xx}(t,x) \quad \text{for } x \in \Omega := (0,1), \\ u(t,0) &= u(t,1) = 0 \\ u(0,x) &= f(x). \end{aligned}$$

Let the initial values be given by a Fourier series

$$f(x) = \sum_{k=1}^{\infty} a_k \sin(k\pi x) \quad \text{for } x \in \Omega$$

with coefficients $\{a_k\}_{k=1}^{\infty} \subset \mathbb{R}$. The functions

$$e^{-(k\pi)^2 t} \sin(k\pi x) \quad \text{for } k \in \mathbb{Z}$$

satisfy the homogenous heat equation $u_t = u_{xx}$ with homogenous Dirichlet boundary conditions. Therefore, the function

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

solves the initial boundary value problem at hand.

In the case $\Omega = \mathbb{R}$, the boundary condition drops out and we face a pure initial value problem. It is called the *Cauchy problem* for the one dimensional heat equation. Given a bounded and continuous initial condition f, we can represent the solution by means of Fourier integrals instead of Fourier series. It holds

$$u(t,x) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{\infty} e^{\frac{-\xi^2}{4t}} f(x-\xi)\,\mathrm{d}\xi.$$

We remark that the solution in (t,x) depends on the knowledge of f in the entire domain. This means that the propagation of data is performed with infinite speed. The solution is unique and depends continuously on the initial condition. Hence, the Cauchy problem for the one dimensional heat equation is well posed.

The Wave Equation

We consider the pure initial value problem for the the one dimensional wave equation, which reads

$$u_{tt}(t,x) = u_{xx}(t,x) \quad \text{for } \Omega := \mathbb{R},$$
$$u(0,x) = f(x), \; u_t(0,x) = g(x).$$

To solve this equation, we consider the change of variables

$$\xi = x + t, \quad \eta = x - t.$$

By the chain rule, we obtain

$$u_x = u_\xi \frac{\partial \xi}{\partial x} + u_\eta \frac{\partial \eta}{\partial x},$$

et cetera, we obtain

$$u_x = u_\xi + u_\eta, \quad u_{xx} = u_{\xi\xi} + 2u_{\xi\eta} + u_{\eta\eta},$$
$$u_t = u_\xi - u_\eta, \quad u_{tt} = u_{\xi\xi} - 2u_{xi\eta} + u_{\eta\eta}.$$

Therefore, the wave equation in the new coordinates reads

$$u_{\xi\eta} = 0.$$

Its general solution is

$$u(\xi,\eta) = \phi(\xi) + \psi(\eta) = \phi(x+t) + \psi(x-t),$$

where the functions ϕ and ψ have to be determined from the initial conditions. It holds

$$\phi(x) + \psi(x) = f(x), \quad \phi'(x) - \psi'(x) = g(x).$$

By differentiating the first equation, we obtain two equations for ϕ' and ψ', which read

$$\phi' = \frac{1}{2}(f' + g), \quad \phi(\xi) = \frac{1}{2} f(\xi) + \frac{1}{2}\int_{x_0}^{\xi} g(s)\,\mathrm{d}s$$
$$\psi' = \frac{1}{2}(f' - g), \quad \psi(\eta) = \frac{1}{2} f(\eta) - \frac{1}{2}\int_{x_0}^{\eta} g(s)\,\mathrm{d}s.$$

Hence, we obtain

$$u(t,x) = \phi(\xi) + \psi(\eta) = \frac{1}{2}\big(f(x+t) + f(x-t)\big) + \frac{1}{2}\int_{x-t}^{x+t} g(s)\,\mathrm{d}s.$$

Note that the solution $u(t,x)$ only depends on the initial values between $x-t$ and $x+t$. This corresponds to the fact that the underlying physical system propagates changes in the data only with finite velocity. Further, we remark that the solution is unique and depends continuously on the data. Therefore, the initial value problem for the one dimensional wave equation is well posed.

1.5 The Maximum Principle

An important tool into show the uniqueness and continuous dependency of the data of the solution to elliptic and parabolic partial differential equations is the maximum principle. Its discrete analogue also is also used in the analysis of finite difference methods. We consider here the version for the elliptic case.

In what follows, let

$$(\mathcal{L}u)(\boldsymbol{x}) := -\sum_{i,j=1}^{d} a_{i,j}(\boldsymbol{x}) u_{x_i x_j}(\boldsymbol{x}) \tag{1.4}$$

be an elliptic differential operator and let Ω be a bounded domain.

Theorem 1.4 (Maximum principle) For $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$, let

$$\mathcal{L}u = f \leq 0 \quad \text{in } \Omega.$$

Then u attains its maximum at the boundary Γ.

Proof. Without loss of generality, we assume that the matrix $\boldsymbol{A}(\boldsymbol{x})$ is diagonal, i.e. $a_{i,j}(\boldsymbol{x}) = 0$ if $i \neq j$, see e.g. [Braess].

First, we carry out the proof under the stronger assumption $f < 0$. Suppose that u attains its maximum at $\boldsymbol{x}_0 \in \Omega$, i.e.

$$u(\boldsymbol{x}_0) = \sup_{\boldsymbol{x} \in \Omega} u(\boldsymbol{x}) > \max_{\boldsymbol{x} \in \Gamma} u(\boldsymbol{x}).$$

Since $\boldsymbol{x}_0$ is a maximum, the gradient of u vanishes at $\boldsymbol{x}_0$ and the Hessian is negative definite, i.e.

$$\nabla u(\boldsymbol{x}_0) = \boldsymbol{0} \quad \text{and} \quad u_{x_i x_i} \leq 0.$$

Due to the ellipticity of $\mathcal{L}$, there further holds $a_{i,i}(\boldsymbol{x}) > 0$ and hence

$$(\mathcal{L}u)(\boldsymbol{x}_0) = -\sum_{i=1}^{d} a_{i,i}(\boldsymbol{x}_0) u_{x_i x_i}(\boldsymbol{x}_0) \geq 0.$$

This is a contradiction to the assumption $(\mathcal{L}u)(\boldsymbol{x}_0) = f(\boldsymbol{x}_0) < 0$.

Now, let $f \leq 0$ and again suppose there exists an $\boldsymbol{x}_0 \in \Omega$ such that $u(\boldsymbol{x}_0) > \max_{\boldsymbol{x} \in \Gamma} u(\boldsymbol{x})$. We introduce the auxiliary function $h(\boldsymbol{x}) := \|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2$, which is bounded on Γ. Hence, for some $\delta > 0$ sufficiently small, the function

$$w = u + \delta h$$

still attains its maximum at a point $\tilde{\boldsymbol{x}}_0 \in \Omega$. Since $h_{x_i x_j} = 2\delta_{i,j}$, we obtain

$$(\mathcal{L}w)(\boldsymbol{x}) = (\mathcal{L}u)(\boldsymbol{x}) + \delta(\mathcal{L}h)(\boldsymbol{x}) = f(\boldsymbol{x}) - 2\delta \sum_{i=1}^{d} a_{i,i}(\boldsymbol{x}) < 0$$

for all $x \in \Omega$. This yields a contradiction as in the first part of the proof. $\square$

Several simple but important consequences can be derived from the maximum principle.

Corollary 1.5 Let the conditions of Theorem 1.4 be satisfied.

1. **Minimum principle** If $\mathcal{L}u \geq 0$ in Ω, then u attains its minimum at the boundary Γ.
2. **Comparison principle** Assume $v \in C^2(\Omega) \cap C^0(\overline{\Omega})$ and
$$\begin{aligned} \mathcal{L}u \leq \mathcal{L}v \quad & \text{in } \Omega, \\ u \leq v \quad & \text{on } \Gamma. \end{aligned}$$
Then, there holds $u \leq v$ in Ω.
3. **Continuous dependency on the boundary data** The solution to the Dirichlet problem
$$\mathcal{L}u = f \text{ in } \Omega, \qquad u = g \text{ on } \Gamma$$
depends continuously on the boundary data: If $\mathcal{L}u_1 = \mathcal{L}u_2 = f$ and $u_1|_\Gamma = g_1, u_2|_\Gamma = g_2$, then
$$\sup_{\boldsymbol{x}\in\Omega} |u_1(\boldsymbol{x}) - u_2(\boldsymbol{x})| = \max_{\boldsymbol{z}\in\Gamma} |g_1(\boldsymbol{z}) - g_2(\boldsymbol{z})|.$$
4. **Uniqueness of the solution** The solution to the Dirichlet problem is unique.
5. **Helmholtz terms** If $(\mathcal{L} + c)u \leq 0$ for $c(\boldsymbol{x}) \geq 0$, then
$$\sup_{x\in\Omega} u(\boldsymbol{x}) \leq \max \left\{0, \max_{\boldsymbol{z}\in\Gamma} u(\boldsymbol{z})\right\}.$$

Proof. 1. Apply the maximum principle to $v := -u$.

2. There holds $\mathcal{L}w := \mathcal{L}v - \mathcal{L}u \geq 0$ in Ω and $w := v - u \geq 0$ on Γ.
   Hence, the minimum principle yields $w(\boldsymbol{x}) \geq 0$ for $\boldsymbol{x} \in \Omega$.
3. We have $\mathcal{L}w = 0$ for $w := u_1 - u_2$. Hence, the maximum principle yields
$$w(\boldsymbol{x}) \leq \max_{\boldsymbol{z}\in\Gamma} w(\boldsymbol{z}) \leq \max_{\boldsymbol{z}\in\Gamma} |w(\boldsymbol{z})| \quad \text{for } \boldsymbol{x} \in \Omega.$$
On the other hand, the minimum principle gives us
$$w(\boldsymbol{x}) \geq \min_{\boldsymbol{z}\in\Gamma} w(\boldsymbol{z}) \geq -\max_{\boldsymbol{z}\in\Gamma} |w(\boldsymbol{z})| \quad \text{for } \boldsymbol{x} \in \Omega.$$
Consequently, there holds
$$|w(\boldsymbol{x})| \leq \max_{\boldsymbol{z}\in\Gamma} |w(\boldsymbol{z})| \quad \text{for all } \boldsymbol{x} \in \Omega.$$
From this, the assertion is obtained by the continuity of w.
4. This follows directly from the continuous dependency on the boundary data.
5. If $u(\boldsymbol{x}) \leq 0$ for all $\boldsymbol{x} \in \Omega$, there is nothing to show. Thus, suppose there exists $\boldsymbol{x}_0 \in \Omega$ with $u(\boldsymbol{x}_0) > 0$. Then, it holds $(\mathcal{L}u)(\boldsymbol{x}_0) \leq (\mathcal{L}u)(\boldsymbol{x}_0) + c(\boldsymbol{x}_0)u(\boldsymbol{x}_0) \leq 0$ and the maximum principle yields the assertion.

□

Under stronger assumptions on the differential operator $\mathcal{L}$ we can also show the continuous dependence of the solution on the right hand side.

Definition 1.6 An elliptic operator of the form (1.4) is called *uniformly elliptic* if there exists a constant $\alpha > 0$ such that

$$\boldsymbol{\xi}^\mathsf{T} \boldsymbol{A}(\boldsymbol{x})\boldsymbol{\xi} \geq \alpha \|\boldsymbol{\xi}\|_2^2$$

for all $\boldsymbol{\xi} \in \mathbb{R}^d$, $\boldsymbol{x} \in \Omega$. The largest such α is called *constant of ellipticity*.

Corollary 1.7 (Continuous dependence on the right hand side) Let $\mathcal{L}$ be uniformly elliptic in Ω. Then there exists a constant $c > 0$ which only depends on Ω and α such that

$$|u(\boldsymbol{x})| \leq \max_{\boldsymbol{z}\in\Gamma} |u(\boldsymbol{z})| + c \sup_{\boldsymbol{z}\in\Omega} |(\mathcal{L}u)(\boldsymbol{z})|$$

for every $\boldsymbol{x} \in \Omega$ and $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$.

Proof. Since Ω is bounded, there exists $R > 0$ such that $\Omega \subset B_R(\mathbf{0}) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 < R\}$. Let

$$w(\boldsymbol{x}) := R^2 - \|\boldsymbol{x}\|_2^2.$$

Since $w_{x_i x_j} = -2\delta_{i,j}$, we have

$$\mathcal{L}w \geq 2\alpha \quad \text{and} \quad 0 \leq w(\boldsymbol{x}) \leq R^2 \text{ for all } \boldsymbol{x} \in \Omega.$$

Therefore, the function

$$v(\boldsymbol{x}) := \max_{\boldsymbol{z}\in\Gamma} |u(\boldsymbol{z})| + w(\boldsymbol{x}) \frac{1}{2\alpha} \sup_{\boldsymbol{z}\in\Omega} |(\mathcal{L}u)(\boldsymbol{z})|$$

satisfies $\mathcal{L}v \geq |\mathcal{L}u|$ in Ω and $v \geq |u|$ on Γ.

Now, the comparison principle yields $-v(\boldsymbol{x}) \leq u(\boldsymbol{x}) \leq v(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Omega$. Since $w \leq R^2$, we arrive at the assertion with $c := R^2/(2\alpha)$. $\square$

2. The Finite Difference Method

2.1 The Poisson Problem

In what follows, we restrict ourselves to Poisson's equation with Dirichlet boundary conditions. Let $f \in C^0(\Omega)$ and $g \in C^0(\Gamma)$. We want to compute $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ such that

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

Definition 2.1 A solution $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ to Poisson's equation is called *classical solution*. In the case $f \equiv 0$, the solution u is called *harmonic*.

In order to compute the classical solution numerically, we discretise the partial derivatives occurring in the Laplacian.

Definition 2.2 For $v \in C^0(\mathbb{R}^d)$, a direction $\boldsymbol{e}_i \in \mathbb{R}^d$, i.e. $e_{i,j} = \delta_{i,j}$ and $h > 0$, we define the *forward difference*

$$(\partial_i^+ v)(\boldsymbol{x}) := \frac{v(\boldsymbol{x} + h\boldsymbol{e}_i) - v(\boldsymbol{x})}{h},$$

the *backward difference*

$$(\partial_i^- v)(\boldsymbol{x}) := \frac{v(\boldsymbol{x}) - v(\boldsymbol{x} - h\boldsymbol{e}_i)}{h}$$

and the *central difference*

$$(\partial_i^\bullet v)(\boldsymbol{x}) := \frac{v(\boldsymbol{x} + h\boldsymbol{e}_i) - v(\boldsymbol{x} - h\boldsymbol{e}_i)}{2h}.$$

For $v \in C^1(\mathbb{R}^d)$, obviously the limit $h \to 0$ exists and there holds

$$\lim_{h\to 0}(\partial_i^+ v)(\boldsymbol{x}) = \lim_{h\to 0}(\partial_i^- v)(\boldsymbol{x}) = \lim_{h\to 0}(\partial_i^\bullet v)(\boldsymbol{x}) = v_{x_i}(\boldsymbol{x}).$$

The next lemma tells us how good this approximation to the actual derivative is. The corresponding approximation order is called *consistency* of the difference operator.

Lemma 2.3 For $\boldsymbol{x}_0 \in \Omega$ let $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq h\} =: \overline{B_h(\boldsymbol{x}_0)} \subset \overline{\Omega}$ and suppose $v \in C^4(\overline{\Omega})$. Then, it holds

$$v_{x_i}(\boldsymbol{x}_0) = \partial_i^{\pm}(\boldsymbol{x}_0) + \mathcal{O}(h),$$
$$v_{x_i}(\boldsymbol{x}_0) = \partial_i^{\bullet}(\boldsymbol{x}_0) + \mathcal{O}(h^2)$$

and

$$v_{x_i x_i}(\boldsymbol{x}_0) = (\partial_i^- \partial_i^+ v)(\boldsymbol{x}_0) + \mathcal{O}(h^2) = \frac{v(\boldsymbol{x}_0 + h\boldsymbol{e}_i) - 2v(\boldsymbol{x}_0) + v(\boldsymbol{x}_0 - h\boldsymbol{e}_i)}{h^2} + \mathcal{O}(h^2).$$

Proof. Since we only consider directional derivatives, it is sufficient prove the assertions for $d = 1$. A *Taylor expansion* of v yields

$$v(x_0 \pm h) = v(x_0) \pm hv'(x_0) + \frac{h^2}{2}v''(\xi), \quad \xi \in (x_0, x_0 \pm h).$$

This directly yields the claim for $\partial_i^{\pm}$. For the central difference, it holds

$$\begin{aligned} v(x_0 + h) &= v(x_0) + hv'(x_0) + \frac{h^2}{2}v''(x_0) + \frac{h^3}{6}v'''(\xi_1), \quad \xi_1 \in (x_0, x_0 + h), \\ v(x_0 - h) &= v(x_0) - hv'(x_0) + \frac{h^2}{2}v''(x_0) - \frac{h^3}{6}v'''(\xi_2), \quad \xi_2 \in (x_0 - h, x_0). \end{aligned}$$

Subtracting these two equations yields

$$v(x + h) - v(x - h) = 2hv'(x_0) + \frac{h^3}{6}\big(v'''(\xi_1) + v'''(\xi_2)\big).$$

The assertion is hence obtained by dividing by $2h$. Finally, the claim on the second order derivative is obtained by adding up the following three equations

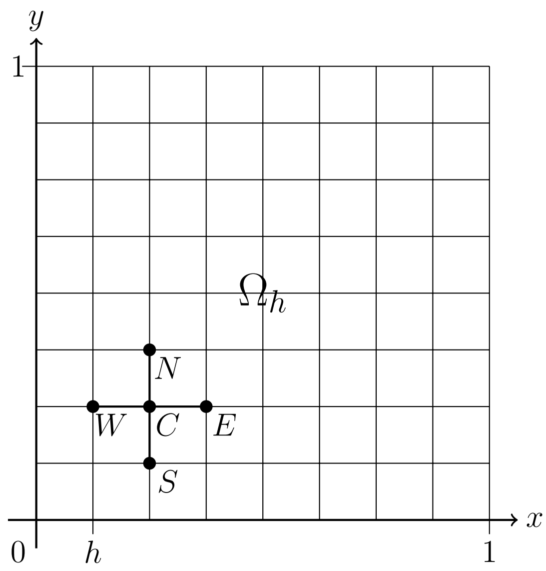
$$\begin{aligned} v(x_0 + h) &= v(x_0) + hv'(x_0) + \frac{h^2}{2}v''(x_0) + \frac{h^3}{6}v'''(x_0) + \frac{h^4}{24}v^{(4)}(\xi_1), \quad \xi_1 \in (x_0, x_0 + h), \\ -2v(x_0) &= -2v(x_0) \\ v(x_0 - h) &= v(x_0) - hv'(x_0) + \frac{h^2}{2}v''(x_0) - \frac{h^3}{6}v'''(x_0) + \frac{h^4}{24}v^{(4)}(\xi_2), \quad \xi_2 \in (x_0 - h, x_0), \end{aligned}$$

which yields

$$v(x_0 + h) - 2v(x_0) + v(x_0 - h) = h^2 v''(x_0) + \frac{h^4}{24}\big(v^{(4)}(\xi_1) + v^{(4)}(\xi_2)\big).$$

Division by h^2 yields the assertion. □

The discretisation by finite differences is based on a *mesh* with *meshwidth* h for Ω. For the sake of simplicity, we shall assume in what follows that the domain Ω is given by the hypercube, i.e. $\Omega := (0, 1)^d \subset \mathbb{R}^d$. In principle, it is also possible to consider the finite difference method on more general domains. Nevertheless, this leads to a cumbersome discussion of corner cases. We refer to [Braess,Hackbusch] for a more general discussion of this topic.



For $n \in \mathbb{N}^*$ let $h = 1/n$. A mesh with meshwidth h for Ω is defined by

$$\Omega_h := \{\boldsymbol{x} \in \Omega : \boldsymbol{x} = h\boldsymbol{k} \text{ for } \boldsymbol{k} \in \mathbb{N}^d\},$$
$$\Gamma_h := \{\boldsymbol{x} \in \Gamma : \boldsymbol{x} = h\boldsymbol{k} \text{ for } \boldsymbol{k} \in \mathbb{N}^d\}.$$

As in the continuous case, we set $\overline{\Omega}_h := \Omega_h \cup \Gamma_h$.
The previous definition of a mesh is also valid for the slightly more general case that Ω is comprised of cubes of edge length h.
For each mesh point $\boldsymbol{x} \in \Gamma_h$, we have $u(\boldsymbol{x}) = g(\boldsymbol{x})$. For each point $\boldsymbol{x} \in \Omega_h$, we obtain an equation for $u(\boldsymbol{x})$ by approximating Poisson's equation by means of finite differences. For each $\boldsymbol{x} \in \Omega_h$, we set

$$(\Delta_h u)(\boldsymbol{x}) := \sum_{i=1}^{d} (\partial_i^- \partial_i^+ u)(\boldsymbol{x}) = (\Delta u)(\boldsymbol{x}) + \mathcal{O}(h^2). \tag{2.1}$$

For $d = 2$, we can express the equation for $u(x, y)$ in terms of the *standard five-point stencil*

$$\begin{bmatrix} \alpha_{NW} & \alpha_N & \alpha_{NE} \\ \alpha_W & \alpha_C & \alpha_E \\ \alpha_{SW} & \alpha_S & \alpha_{SE} \end{bmatrix}_* = \frac{1}{h^2} \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}_* .$$

It holds

$$\alpha_C u_C + \alpha_E u_E + \alpha_S u_S + \alpha_W u + \alpha_N u_N = f(x, y) \quad \text{for } (x, y) \in \Omega_h$$

with $u_C := u(x, y), u_E = u(x+h, y), u_S = u(x, y-h), u_W = u(x-h, y), u_N = u(x, y+h)$.

Example 2.4

Poisson's Equation in One Dimension For $d = 1$, we have

$$-u'' = f \text{ in } (0, 1), \quad u(0) = \alpha, \; u(1) = \beta.$$

Setting $u_i := u(x_i)$ and $f_i := f(x_i)$ for $x_i = hi$, $i = 1, \ldots, n-1$ and $h = 1/n$, we end up with the linear system of equations

$$\underbrace{\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}}_{\boldsymbol{L}} \underbrace{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-2} \\ u_{n-1} \end{bmatrix}}_{\boldsymbol{u}} = \underbrace{\begin{bmatrix} f_1 + \alpha/h^2 \\ f_2 \\ \vdots \\ f_{n-2} \\ f_{n-1} + \beta/h^2 \end{bmatrix}}_{\boldsymbol{f}}.$$

Poisson's equation in d Dimensions With the aid of the aid of the matrix $\boldsymbol{L} \in \mathbb{R}^{(n-1)\times(n-1)}$ and the identity matrix $\boldsymbol{I} \in \mathbb{R}^{(n-1)\times(n-1)}$, the finite difference approximation of the d-dimensional Laplacian can be written as

$$\boldsymbol{L}^{(d)} := \sum_{i=1}^{d} \underbrace{\boldsymbol{I} \otimes \cdots \otimes \boldsymbol{I}}_{(i-1)\text{-times}} \otimes \boldsymbol{L} \otimes \underbrace{\boldsymbol{I} \otimes \cdots \otimes \boldsymbol{I}}_{(d-i)\text{-times}}. \tag{2.2}$$

Herein, for two matrices $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{m'\times n'}$, $(\boldsymbol{A} \otimes \boldsymbol{B}) \in \mathbb{R}^{(mm')\times(nn')}$ denotes their *Kronecker product*, which is defined according to

$$\boldsymbol{A} \otimes \boldsymbol{B} := \begin{bmatrix} a_{1,1}\boldsymbol{B} & \cdots & a_{1,n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m,1}\boldsymbol{B} & \cdots & a_{m,n}\boldsymbol{B} \end{bmatrix}.$$

The Kronecker product ist associative but not commutative.

We will also make use of the *column-wise vectorisation* of a matrix, which is defined as follows. Let $\boldsymbol{A} := [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] \in \mathbb{R}^{m\times n}$ with columns $\boldsymbol{a}_i \in \mathbb{R}^{m\times 1}$ for $i = 1, \ldots, n$. Then, we set

$$\operatorname{vec}(\boldsymbol{A}) := \begin{bmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix} \in \mathbb{R}^{nm}.$$

Now, for $d = 2$ we obtain for the Poisson problem $\Delta u = f$ in Ω, $u = 0$ on Γ the linear system of equations

$$(\boldsymbol{L} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{L}) \operatorname{vec}(\boldsymbol{U}) = \operatorname{vec}(\boldsymbol{F}),$$

where we set

$$\boldsymbol{U} := [u(jh, ih)]_{i,j=1}^{n-1} \quad \text{and} \quad \boldsymbol{F} := [f(jh, ih)]_{i,j=1}^{n-1}.$$

Neumann Boundary Conditions Often, instead of Dirichlet conditions, which fix the function values at Γ, information on the flux are available and shall be incorporated. This leads to Neumann boundary conditions

$$\frac{\partial u}{\partial \boldsymbol{n}}(\boldsymbol{x}) = g(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \Gamma.$$

To sustain the overall accuracy $\mathcal{O}(h^2)$ from the discretization of the Laplacian, we have to employ central differences for the discretisation of the gradient. We show the procedure here for $d = 1$.

At the point $x_0 = 1$ and $x_n = 1$, we obtain the equations

$$\frac{u_{n+1} - u_{n-1}}{2h} = g(x_n) \quad \text{and} \quad \frac{u_{-1} - u_1}{2h} = g(x_0),$$

respectively, where we set $x_{-1} := -h$ and $x_{n+1} = 1 + h$. Hence, there holds

$$u_{n+1} = u_{n-1} + 2hg(x_n) \quad \text{and} \quad u_{-1} = u_1 + 2hg(x_0).$$

Now we can insert these expressions into the discretisation of the Laplacian and end up with the equations

$$\Delta_h u(x_0) = \frac{2hg(x_0) - 2u_0 + 2u_1}{h^2} \quad \text{and} \quad \Delta_h u(x_n) = \frac{2hg(x_n) - 2u_n + 2u_{n-1}}{h^2}.$$

△

Next, we are concerned with the solvability and the rate of convergence of the finite difference method. To that end, we start with a discrete version of the maximum principle for the problem at hand. For a more general version, we refer to [Braess].

Theorem 2.5 (Discrete maximum principle) Let u_h be the solution to the discrete problem

$$-\Delta_h u_h = f \quad \text{in } \Omega_h \text{ with } f \le 0. \tag{2.3}$$

Then, there holds

$$\max_{\boldsymbol{x} \in \Omega_h} u_h(\boldsymbol{x}) \le \max_{\boldsymbol{z} \in \Gamma_h} u_h(\boldsymbol{z}).$$

Proof. For a proof of this theorem, see [Braess]. □

As a consequence, the properties derived in Corollary 1.4 directly transfer to u_h. This accounts particularly for the comparison principle and the continuous dependence on the data. Moreover, there holds the following

Corollary 2.6 The solution $\boldsymbol{u} \in \mathbb{R}^{(n-1)^d}$ to

$$\boldsymbol{L}^{(d)} \boldsymbol{u} = \boldsymbol{f},$$

cf. (2.2), is unique.

Proof. The solution of the homogenous system $\boldsymbol{L}^{(d)} \boldsymbol{u} = \boldsymbol{0}$ corresponds to the solution of (2.3) with $f \equiv 0$ and homogenous Dirichlet data. Hence, $\min_{\boldsymbol{x} \in \Omega_h} u_h(\boldsymbol{x}) = \max_{\boldsymbol{x} \in \Omega_h} u_h(\boldsymbol{x}) = 0$. Therefore, the homogenous system has only the trivial solution and the matrix $\boldsymbol{L}^{(d)}$ is nonsingular. □

2.2 Convergence of the Finite Difference Method

On Ω_h and $\overline{\Omega}_h$, we introduce the sup norm according to

$$\|v_h\|_{\Omega_h} := \max_{\boldsymbol{x}\in\Omega_h} |v_h(\boldsymbol{x})| \quad \text{and} \quad \|v_h\|_{\overline{\Omega}_h} := \max_{\boldsymbol{x}\in\overline{\Omega}_h} |v_h(\boldsymbol{x})|,$$

respectively.

Definition 2.7 Let $\mathcal{L}_h$ denote the finite difference approximation of the second order differential operator $\mathcal{L}$. The corresponding finite difference method is called ...

- ...*convergent* of order p, iff
$$\|u - u_h\|_{\overline{\Omega}_h} = \mathcal{O}(h^p).$$
- ...*consistent* of order p, iff
$$\|\mathcal{L}_h u - \mathcal{L}u\|_{\Omega_h} = \mathcal{O}(h^p).$$
- ...*stable*, iff there exists $C_s > 0$ such that for all $v_h\colon \Omega_h \to \mathbb{R}$ with $v_h|_{\Gamma_h} = 0$ holds
$$\|v_h\|_{\overline{\Omega}_h} \le C_s \|\mathcal{L}_h v_h\|_{\Omega_h}.$$

Note that there holds $\|\Delta_h u - \Delta u\|_{\Omega_h} = \mathcal{O}(h^2)$. Thus, the discretisation of the Laplacian is consistent of order 2.

The stability of the discretisation translates to $\|(\boldsymbol{L}^{(d)})^{-1}\|_\infty \le C_s$ independently of the mesh width $h > 0$: Let $\boldsymbol{v}, \boldsymbol{w}$ contain the values of a grid function $v_h\colon \Omega_h \to \mathbb{R}$ with $v_h|_{\Gamma_h} = 0$ and $\Delta_h v_h$, respectively. Hence, there holds $\boldsymbol{w} = \boldsymbol{L}^{(d)}\boldsymbol{v}$. Now, the stability condition translates to

$$\left\|\left(\boldsymbol{L}^{(d)}\right)^{-1}\boldsymbol{w}\right\|_\infty = \|\boldsymbol{v}\|_\infty = \|v_h\|_{\overline{\Omega}_h} \le C_s \|\Delta_h v_h\|_{\Omega_h} = C_s \|\boldsymbol{w}\|_\infty.$$

A similar consideration also applies for general finite difference approximations $\mathcal{L}_h$.

The next theorem tells us under which conditions a finite difference method is convergent.

Theorem 2.8 If a finite difference method is stable and consistent of order p, then it is also convergent of order p.

Proof. There holds

$$\|u - u_h\|_{\overline{\Omega}_h} \le C_s \|\mathcal{L}_h(u - u_h)\|_{\Omega_h} = C_s \|\mathcal{L}_h u - \mathcal{L}u\|_{\Omega_h} = \mathcal{O}(h^p),$$

since $(\mathcal{L}_h u_h)(\boldsymbol{x}) = f(\boldsymbol{x}) = (\mathcal{L}u)(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Omega_h$. □

The theorem tells us that, in order to show convergence of the finite element discretisation (2.1) of the Laplacian, it remains to show the stability, which also holds on more general domains.

Theorem 2.9 Assume $\Omega \subset B_R(\mathbf{0})$ and let $v_h \colon \Omega_h \to \mathbb{R}$, $v_h|_{\Gamma_h} = 0$. Then, it holds

$$\|v_h\|_{\overline{\Omega}_h} \leq \frac{R^2}{2d} \|\Delta_h v_h\|_{\Omega_h}.$$

Proof. Let $w(\boldsymbol{x}) := (R^2 - \|\boldsymbol{x}\|_2^2)/(2d)$. It holds

$$-\Delta_h w = -\Delta w = 1 \text{ in } \Omega_h, \quad w \geq 0 \text{ on } \Gamma_h.$$

Now, consider the solution u_h to $-\Delta_h u_h = 1$ in Ω_h, $u_h|_{\Gamma_h} = 0$. The discrete comparison principle gives us $u_h(\boldsymbol{x}) \leq w(\boldsymbol{x})$ for all $\boldsymbol{x} \in \overline{\Omega}_h$. Together with the minimum principle, we arrive at

$$0 \leq u_h(\boldsymbol{x}) \leq \frac{1}{2d}(R^2 - \|\boldsymbol{x}\|_2^2) \quad \text{for all } \boldsymbol{x} \in \overline{\Omega}_h. \tag{2.4}$$

From this, the stability is derived as follows: Let $v_h \colon \Omega_h \to \mathbb{R}$, $v_h|_{\Gamma_h} = 0$. It holds

$$-\big(\Delta_h(-u_h)\big)(\boldsymbol{x}) = -1 \leq -\frac{(\Delta_h v_h)(\boldsymbol{x})}{\|\Delta_h v_h\|_{\Omega_h}} \leq 1 = -(\Delta_h u_h)(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \Omega_h.$$

The discrete comparison principle yields

$$-u_h(\boldsymbol{x}) \leq \frac{v_h(\boldsymbol{x})}{\|\Delta_h v_h\|_{\Omega_h}} \leq u_h(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \overline{\Omega}_h$$

and hence due to (2.4) we arrive at

$$\frac{\|v_h\|_{\overline{\Omega}_h}}{\|\Delta_h v_h\|_{\Omega_h}} \leq \|u_h\|_{\overline{\Omega}_h} \leq \frac{R^2}{2d}.$$

□

3. Variational Formulation

3.1 Sobolev Spaces

In what follows, let $\Omega \subset \mathbb{R}^d$ denote a domain with piecewise smooth boundary. The function space $L^2(\Omega)$ consists of all equivalence classes of square integrable functions on Ω. Two functions $f, g\colon \Omega \to \mathbb{R}$ are identified, if $f(\boldsymbol{x}) = g(\boldsymbol{x})$ for almost every $\boldsymbol{x} \in \Omega$. Endowed with the inner product

$$(f, g)_{L^2(\Omega)} := \int_\Omega f(\boldsymbol{x}) g(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x},$$

$L^2(\Omega)$ becomes a *Hilbert space* with the induced norm

$$\|f\|_{L^2(\Omega)} := \sqrt{(f, f)_{L^2(\Omega)}}.$$

Note that two functions $f, g\colon \Omega \to \mathbb{R}$ are identified, iff $\|f - g\|_{L^2(\Omega)} = 0$.

Definition 3.1 The function $u \in L^2(\Omega)$ posseses the *weak derivative* $v = \partial^{\boldsymbol{\alpha}} u$ in $L^2(\Omega)$, iff $v \in L^2(\Omega)$ and

$$(v, \phi)_{L^2(\Omega)} = (-1)^{|\boldsymbol{\alpha}|}(u, \partial^{\boldsymbol{\alpha}}\phi) \quad \text{for all } \phi \in C_0^\infty(\Omega).$$

Example 3.2 Let $\Omega = (0, 1)$ and consider the function $u(x) = |x - 0.5|$. For $\phi \in C_0^\infty(0, 1)$, integration by parts yields

$$\begin{aligned}
-\int_0^1 u\phi'(x)\,\mathrm{d}x &= -\int_0^{0.5} (0.5 - x)\phi'(x)\,\mathrm{d}x - \int_{0.5}^1 (x - 0.5)\phi'(x)\,\mathrm{d}x \\
&= -[u(x)\phi(x)]_0^1 + \int_0^{0.5} -1 \cdot \phi(x)\,\mathrm{d}x + \int_{0.5}^1 1 \cdot \phi(x)\,\mathrm{d}x \\
&= \int v(x)\phi(x)\,\mathrm{d}x,
\end{aligned}$$

where v is the *heavyside function*, which is defined as

$$v(x) = \begin{cases} 1, & x \geq 0, \\ -1, & x < 0. \end{cases}$$

Hence, the weak derivative of u is given by $u' = v$. △

Remark Let $u \in C^1(\Omega)$ and $\phi \in C_0^\infty(\Omega)$, the divergence theorem yields

$$(u_{x_i}, \phi)_{L^2(\Omega)} + (u, \phi_{x_i})_{L^2(\Omega)} = \int_\Omega \frac{\partial}{\partial x_i}(u\phi)\,\mathrm{d}\boldsymbol{x} = \int_\Gamma u\phi n_i\,\mathrm{d}\sigma,$$

where $\boldsymbol{n} = [n_1, \ldots, n_d]^\intercal \in \mathbb{R}^d$ is the outward pointing normal vector. Consequently, it holds

$$(u_{x_i}, \phi)_{L^2(\Omega)} = -(u, \phi_{x_i})_{L^2(\Omega)} \quad \text{for all } \phi \in C_0^\infty(\Omega),$$

i.e. the weak derivative coincides with the classical derivative. △

Definition 3.3 Let $m \in \mathbb{N}$. The *Sobolev space* $H^m(\Omega)$ is defined according to

$$H^m(\Omega) := \{v \in L^2(\Omega) : \partial^{\boldsymbol{\alpha}} v \in L^2(\Omega) \text{ for all } |\boldsymbol{\alpha}| \leq m\}.$$

Theorem 3.4 The Sobolev space $H^m(\Omega)$ endowed with the inner product

$$(v, w)_{H^m(\Omega)} := \sum_{|\boldsymbol{\alpha}| \leq m} (\partial^{\boldsymbol{\alpha}} v, \partial^{\boldsymbol{\alpha}} w)_{L^2(\Omega)}$$

and the corresponding norm $\|v\|_{H^m(\Omega)} = \sqrt{(v,v)_{H^m(\Omega)}}$ is a Hilbert space.

Proof. Let $\{v_n\}_{n\in\mathbb{N}}$ denote a Cauchy sequence in $H^m(\Omega)$. Then, due to the definition of the $\|\cdot\|_{H^m(\Omega)}$-norm, the sequence $\{\partial^{\boldsymbol{\alpha}} v_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $L^2(\Omega)$ for every $|\boldsymbol{\alpha}| \leq m$. The completeness of $L^2(\Omega)$ implies that there exist the limits

$$\left\|\partial^{\boldsymbol{\alpha}} v_n - v^{(\boldsymbol{\alpha})}\right\|_{L^2(\Omega)} \to 0, \quad n \to \infty$$

for certain functions $v^{(\boldsymbol{\alpha})} \in L^2(\Omega)$. It remains to show $\partial^{\boldsymbol{\alpha}} v^{(\mathbf{0})} = v^{(\boldsymbol{\alpha})}$. To that end, let $\{w_n\}_{n\in\mathbb{N}}$ denote a Cauchy sequence in $L^2(\Omega)$ with limit $w \in L^2(\Omega)$. Due to the *Cauchy-Schwarz inequality*, it holds

$$\left|(w_n - w, \phi)_{L^2(\Omega)}\right| \leq \|w - w_n\|_{L^2(\Omega)} \|\phi\|_{L^2(\Omega)}$$

and hence $(w_n, \phi)_{L^2(\Omega)} \to (w, \phi)_{L^2(\Omega)}$ for every $\phi \in C_0^\infty(\Omega)$. Consequently, we have

$$(v^{(\boldsymbol{\alpha})}, \phi)_{L^2(\Omega)} = \lim_{n\to\infty} (\partial^{\boldsymbol{\alpha}} v_n, \phi)_{L^2(\Omega)} = \lim_{n\to\infty} (-1)^{|\boldsymbol{\alpha}|}(v_n, \partial^{\boldsymbol{\alpha}}\phi)_{L^2(\Omega)} = (-1)^{|\boldsymbol{\alpha}|}(v^{(\mathbf{0})}, \partial^{\boldsymbol{\alpha}}\phi)_{L^2(\Omega)}.$$

By the definition of the weak derivative, we arrive at $\partial^{\boldsymbol{\alpha}} v^{(\mathbf{0})} = v^{(\boldsymbol{\alpha})}$. □

Remark The smoothness of the functions of functions in $H^m(\Omega)$ in the classical sense is dependent on the spatial dimension: For $d = 1$, there holds $H^1(\Omega) \subset C^0(\Omega)$. For $d = 2$, functions in $H^1(\Omega)$ may exhibit point singularities. For example, it holds

$$v(r, \phi) = \log\left(\log\frac{2}{r}\right) \in H^1\big(B_1(\mathbf{0})\big).$$

More general, there holds

$$\|\boldsymbol{x}\|_2^{-\beta} \in H^1\big(B_1(\mathbf{0})\big) \quad \text{for } \beta < \frac{d-2}{2} \text{ and } d \geq 3.$$

$\triangle$

Note that it is also possible to introduce Sobolev spaces without the recourse to the concept of weak derivatives.

Theorem 3.5 Let $\Omega \subset \mathbb{R}^d$ denote a domain and let $m \geq 0$. Then $C^\infty(\Omega) \cap H^m(\Omega)$ is dense in $H^m(\Omega)$.

Proof. See for example [Alt,Wloka]. $\square$

The theorem tells us that $H^m(\Omega)$ is the completion of $C^\infty(\Omega) \cap H^m(\Omega)$ with respect to the $\|\cdot\|_{H^m(\Omega)}$-norm, i.e.

$$H^m(\Omega) = \overline{C^\infty(\Omega) \cap H^m(\Omega)}^{\|\cdot\|_{H^m(\Omega)}}.$$

Based on this fact, we introduce a corresponding generalisation for functions with zero boundary values.

Definition 3.6 We define the Sobolev spaces $H_0^m(\Omega)$ for $m \in \mathbb{N}$ as the completion of $C_0^\infty(\Omega)$ with respect to the $\|\cdot\|_{H^m(\Omega)}$-norm, i.e.

$$H_0^m(\Omega) := \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^m(\Omega)}}.$$

Note that $H_0^m(\Omega)$ is a closed subspace of $H^m(\Omega)$ and hence also a Hilbert space. Moreover, there holds $H_0^0(\Omega) = H^0(\Omega) = L^2(\Omega)$ such that we arrive at the following scheme

$$\begin{array}{ccccccccccc}
L^2(\Omega) & = & H^0(\Omega) & \supset & H^1(\Omega) & \supset & H^2(\Omega) & \supset & \dots & & \\
 & & \| & & \cup & & \cup & & & & \\
 & & H_0^0(\Omega) & \supset & H_0^1(\Omega) & \supset & H_0^2(\Omega) & \supset & \dots & \supset & C_0^\infty(\Omega).
\end{array}$$

In particular, $C_0^\infty(\Omega)$ is a dense subset of $L^2(\Omega)$.
The functional

$$|v|_{H^m(\Omega)} := \sqrt{\sum_{|\boldsymbol{\alpha}|=m} \|\partial^{\boldsymbol{\alpha}} v\|^2_{L^2(\Omega)}}$$

defines a *seminorm* on $H^m(\Omega)$, i.e. $|\cdot|_{H^m(\Omega)}$ satisfies all properties of a norm except for that it is not point-separating. Obviously, there holds $|v|_{H^m(\Omega)} = 0$, $m \geq 1$, for any constant function $v \in H^m(\Omega)$. However , the next theorem tells us that $|\cdot|_{H^m(\Omega)}$ is an equivalent norm on $H_0^m(\Omega)$.

Theorem 3.7 (Poincaré inequality) Let $\Omega \subset [0, s]^d$ for some $s > 0$. Then, it holds

$$\|v\|_{L^2(\Omega)} \leq s|v|_{H^1(\Omega)} \quad \text{for every } v \in H_0^1(\Omega).$$

Proof. Since $C_0^\infty(\Omega)$ is a dense subset of $H_0^1(\Omega)$, it is sufficient to show the result for $v \in C_0^\infty(\Omega)$. We set $v(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in [0, s]^d \setminus \Omega$. Then, the fundamental theorem of calculus yields

$$v(\boldsymbol{x}) = v(0, x_2, \dots, x_d) + \int_0^{x_1} v_{x_1}(y, x_2, \dots, x_d)\,\mathrm{d}y.$$

Since $v(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in [0, s]^d \setminus \Omega$, the first term vanishes. Now, the application of the Cauchy-Schwarz inequality yields

$$\begin{aligned}\big(v(\boldsymbol{x})\big)^2 &= \left(\int_0^{x_1} v_{x_1}(y, x_2, \dots, x_d)\,\mathrm{d}y\right)^2 \leq \int_0^{x_1} 1^2\,\mathrm{d}y \int_0^{x_1} \big(v_{x_1}(y, x_2, \dots, x_d)\big)^2\,\mathrm{d}y \\ &\leq s \int_0^s \big(v_{x_1}(y, x_2, \dots, x_d)\big)^2\,\mathrm{d}y.\end{aligned}$$

Since the right hand side is independent of x_1, integration with respect to x_1 yields

$$\int_0^s \big(v(\boldsymbol{x})\big)^2\,\mathrm{d}x_1 \leq s^2 \int_0^s \big(v_{x_1}(y, x_2, \dots, x_d)\big)^2\,\mathrm{d}y = s^2 \int_0^s \big(v_{x_1}(\boldsymbol{x})\big)^2\,\mathrm{d}x_1.$$

Finally, integrating with respect to the other coordinates yields

$$\begin{aligned}\|v\|_{L^2(\Omega)}^2 &= \int_\Omega \big(v(\boldsymbol{x})\big)^2\,\mathrm{d}\boldsymbol{x} = \int_{[0,s]^d} \big(v(\boldsymbol{x})\big)^2\,\mathrm{d}\boldsymbol{x} \\ &\leq s^2 \int_{[0,s]^d} \big(v_{x_1}(\boldsymbol{x})\big)^2\,\mathrm{d}\boldsymbol{x} = s^2 \int_\Omega \big(v_{x_1}(\boldsymbol{x})\big)^2\,\mathrm{d}\boldsymbol{x} \leq s^2|v|_{H^1(\Omega)}^2.\end{aligned}$$

□

Remark The Poincaré inequality also holds under the weaker assumption of homogenous boundary conditions on part $\Gamma_D \subset \Gamma$ of the boundary with $|\Gamma_D| > 0$. △

Corollary 3.8 Let $\Omega \subset [0, s]^d$ for some $s > 0$. Then, it holds

$$|v|_{H^m(\Omega)} \leq \|v\|_{H^m(\Omega)} \leq (1+s)^m |v|_{H^m(\Omega)} \quad \text{for every } v \in H_0^m(\Omega),\ m \geq 0.$$

Proof. The proof is by induction on m. See for example [Braess]. □

The boundary Γ of a domain $\Omega \subset \mathbb{R}^d$ is a null set with respect to the d-dimensional Lebesgue measure. Hence, functions $v \in L^2(\Omega)$ do not exhibit boundary values. The next theorem tells us that the situation is different for functions in $H^1(\Omega)$.

Theorem 3.9 (Trace theorem) Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with piecewise smooth boundary. In addition, let Ω satisfy the *cone condition*, i.e. there exist an angle $\alpha_0 > 0$ such that the interior angle at every corner of Ω is bigger than α_0. Then, there exists a continuous linear mapping

$$\gamma\colon H^1(\Omega) \to L^2(\Gamma), \quad \|\gamma(v)\|_{L^2(\Gamma)} \leq c\|v\|_{H^1(\Omega)}, \quad c > 0,$$

such that $\gamma(v) = v|_\Gamma$ for every $v \in C^1(\overline{\Omega})$.

Proof. See for example [Braess]. □

3.2 Variational Formulation of Dirichlet Problems

We consider the Dirichlet problem

$$\begin{aligned} -\operatorname{div}\big(\boldsymbol{A}(\boldsymbol{x})\nabla u(\boldsymbol{x})\big) + c(\boldsymbol{x})u(\boldsymbol{x}) &= f(\boldsymbol{x}), \quad && \boldsymbol{x} \in \Omega, \\ u(\boldsymbol{x}) &= 0, && \boldsymbol{x} \in \Gamma. \end{aligned} \tag{3.1}$$

In addition to the classical solution $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ we will derive in this section also *weak solutions*, which are obtained from the *variational formulation*.

Theorem 3.10 (Characterisation theorem) Let V denote a vector space and denote by

$$a\colon V \times V \to \mathbb{R}$$

a symmetric and positive bilinear form, i.e. $a(v, w) = a(w, v)$ and $a(v, v) > 0$ for all $v \in V \setminus \{0\}$. Moreover, let

$$\ell\colon V \to \mathbb{R}$$

be a linear functional. Then, the functional

$$J(v) := \frac{1}{2}a(v, v) - \ell(v)$$

attains its minimum at $u \in V$, iff

$$a(u, v) = \ell(v) \quad \text{for all } v \in V. \tag{3.2}$$

In addition, there exists at most one solution to (3.2).

Proof. Let $u, v \in V$ and $t \in \mathbb{R}$. It holds

$$\begin{aligned} J(u + tv) &= \frac{1}{2}a(u + tv, u + tv) - \ell(u + tv) \\ &= J(u) + t[a(u, v) - \ell(v)] + \frac{1}{2}t^2 a(v, v). \end{aligned} \tag{3.3}$$

Hence, if $u \in V$ satisfies (3.2), there holds

$$J(u+tv) = J(u) + \frac{1}{2}t^2 a(v,v) > J(u) \quad \text{for every } v \neq 0,\ t > 0.$$

Therefore, u is the unique minimum of J. Vice versa, if u is a minimum of J then the derivative of the function $t \mapsto J(u+tv)$ must vanish at $t = 0$. The derivative is according to (3.3) given by $a(u,v) - \ell(v)$, which directly yields (3.2). □

The next theorem tells us that every classical solution to (3.1) is also a minimum in the sense of the characterisation theorem.

Theorem 3.11 Every classical solution $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ to (3.1) is also a solution to the variational problem

$$J(v) := \int_\Omega \frac{1}{2}\big(\langle \boldsymbol{A}\nabla v, \nabla v\rangle + cv^2\big) - fv \,\mathrm{d}\boldsymbol{x} \to \min$$

among all functions $v \in C^2(\Omega) \cap C^0(\overline{\Omega})$ with $v|_\Gamma = 0$.

Proof. The divergence theorem yields

$$\int_\Omega \operatorname{div}(\boldsymbol{A}\nabla u)v + \langle \boldsymbol{A}\nabla u, \nabla v\rangle \,\mathrm{d}\boldsymbol{x} = \int_\Omega \operatorname{div}\big((\boldsymbol{A}\nabla u)v\big) \,\mathrm{d}\boldsymbol{x} = \int_\Gamma \langle \boldsymbol{A}\nabla u, \boldsymbol{n}\rangle v \,\mathrm{d}\sigma.$$

Hence, for functions v with $v|_\Gamma = 0$, we obtain

$$\int_\Omega \operatorname{div}(\boldsymbol{A}\nabla u)v \,\mathrm{d}\boldsymbol{x} = -\int_\Omega \langle \boldsymbol{A}\nabla u, \nabla v\rangle \,\mathrm{d}\boldsymbol{x}.$$

Next, we define

$$a(u,v) := \int_\Omega \langle \boldsymbol{A}\nabla u, \nabla v\rangle + cuv \,\mathrm{d}\boldsymbol{x}, \quad \ell(v) := (f,v)_{L^2(\Omega)}.$$

Then, it holds for every $v \in C^1(\Omega) \cap C^0(\overline{\Omega})$ with $v|_\Gamma = 0$ that

$$a(u,v) - \ell(v) = \int_\Omega \langle \boldsymbol{A}\nabla u, \nabla v\rangle + cuv - fv \,\mathrm{d}\boldsymbol{x} = \int_\Omega \big(-\operatorname{div}(\boldsymbol{A}\nabla u) + cu - f\big)v \,\mathrm{d}\boldsymbol{x} = 0,$$

if u is a classical solution to (3.1). The minimal property is now implied by the characterisation theorem. □

In a similar fashion, one shows that every solution $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ to the variational problem is also a classical solution to (3.1). However, a minimum of J does not necessarily exist in $C^2(\Omega) \cap C^0(\overline{\Omega})$.

Definition 3.12 Let H denote a Hilbert space with norm $\|\cdot\|_H$. A bilinear form $a\colon H \times H \to \mathbb{R}$ is *continuous*, iff there exists $c_S > 0$ such that

$$|a(v,w)| \leq c_S \|v\|_H \|w\|_H \quad \text{for all } v, w \in H.$$

The bilinear form is called *elliptic*, iff there exists $c_E > 0$ such that

$$a(v,v) \geq c_E \|v\|_H^2 \quad \text{for all } v \in H.$$

The *energy norm*

$$\|v\|_a := \sqrt{a(v,v)} \tag{3.4}$$

induced by a continuous and elliptic bilinear form a on a Hilbert space H is equivalent to the Hilbert space norm. Obviously, there holds

$$\sqrt{c_E}\|v\|_H \le \|v\|_a \le \sqrt{c_S}\|v\|_H \quad \text{for all } v \in H.$$

Theorem 3.13 (Lax-Milgram) Let $V \subset H$ be a closed subspace of the Hilbert space H with dual space V'. Moreover, let $a\colon H \times H \to \mathbb{R}$ be a continuous bilinear form that is elliptic on V. Then, for every $\ell \in V'$, the variational problem

$$J(v) := \frac{1}{2}a(v,v) - \ell(v) \to \min$$

exhibits a unique solution $u \in V$.

Proof. See for example [Braess]. □

Remark In the particular case $V = H$ and $a(v,w) = (v,w)_H$, the Lax-Milgram theorem yields the Riesz representation theorem: For every $\ell \in H'$ there exists a $u \in H$ such that

$$(u,v)_H = \ell(v) \quad \text{for all } v \in H.$$

△

In view of the Lax-Milgram theorem, we can now specify the notion of *weak solution.*

Definition 3.14 A function $u \in H_0^1(\Omega)$ is called *weak solution* to (3.1), iff there holds

$$a(u,v) = \ell(v) \quad \text{for all } v \in H_0^1(\Omega),$$

where a and ℓ are given as in the characterisation theorem.

The next theorem specifies under which conditions (3.1) exhibits a unique weak solution.

Theorem 3.15 Let $f \in L^2(\Omega)$ and

$$0 \le c(\boldsymbol{x}) \le \overline{c} < \infty, \quad 0 < \underline{\alpha}\|\boldsymbol{\xi}\|_2^2 \le \boldsymbol{\xi}^\intercal \boldsymbol{A}(\boldsymbol{x})\boldsymbol{\xi} \le \overline{\alpha}\|\boldsymbol{\xi}\|_2^2 < \infty$$

for all $\boldsymbol{x} \in \Omega$, $\boldsymbol{\xi} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$. Then, (3.1) exhibits a unique weak solution $u \in H_0^1(\Omega)$, which solves the minimisation problem

$$J(v) = \frac{1}{2}a(v,v) - \ell(v) \to \min.$$

Proof. Since

$$\begin{aligned} a(v,w) &= \int_\Omega \langle \boldsymbol{A}\nabla v, \nabla w\rangle + cvw \,\mathrm{d}\boldsymbol{x} \le \int_\Omega \overline{\alpha}\|\nabla v\|_2\|\nabla w\|_2 + \overline{c}|v||w| \,\mathrm{d}\boldsymbol{x} \\ &\le \overline{\alpha}\sqrt{\int_\Omega \|\nabla v\|_2^2 \,\mathrm{d}\boldsymbol{x}}\sqrt{\int_\Omega \|\nabla w\|_2^2 \,\mathrm{d}\boldsymbol{x}} + \overline{c}\sqrt{\int_\Omega v^2 \,\mathrm{d}\boldsymbol{x}}\sqrt{\int_\Omega v^2 \,\mathrm{d}\boldsymbol{x}} \\ &\le \max\{\overline{\alpha},\overline{c}\}\|v\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)}, \end{aligned}$$

the bilinear form a is continuous on $H^1(\Omega)$. The ellipticity of a follows from

$$a(v,v) = \int_\Omega \langle \boldsymbol{A}\nabla v, \nabla v\rangle + cv^2 \,\mathrm{d}\boldsymbol{x} \ge \underline{\alpha}\int_\Omega \|\nabla v\|_2^2 \,\mathrm{d}\boldsymbol{x} = \underline{\alpha}|v|_{H^1(\Omega)}^2,$$

since by the Poincaré inequality $|\cdot|_{H^1(\Omega)}$ is an equivalent norm on $H_0^1(\Omega)$. Finally, since

$$|\ell(v)| = \left|\int_\Omega fv \,\mathrm{d}\boldsymbol{x}\right| \le \|f\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)},$$

also the linear form is continuous. Therefore, the unique solvability follows from the Lax-Milgram theorem. □

Remark

1. The dual space of $H^{-1}(\Omega) := [H_0^1(\Omega)]'$ can be defined as completion of $L^2(\Omega)$ with respect to the *dual norm*

$$\|f\|_{H^{-1}(\Omega)} := \sup_{0\neq v\in H^1(\Omega)} \frac{|(f,v)_{L^2(\Omega)}|}{\|v\|_{H^1(\Omega)}}, \quad f \in L^2(\Omega).$$

   Hence, we can extend the linear form to $H^{-1}(\Omega)$ according to

$$|\ell(v)| = \|v\|_{H^1(\Omega)}\frac{|(f,v)_{L^2(\Omega)}|}{\|v\|_{H^1}(\Omega)} \le \|v\|_{H^1(\Omega)}\|f\|_{H^{-1}(\Omega)}.$$

   Since $H^{-1}(\Omega) \subset L^2(\Omega)$, this implies that the right hand side f in (3.1) does not even have to be square integrable.
2. The case of inhomogenous Dirichlet conditions $u = g \not\equiv 0$ at Γ is usually reduced to the homogenous case as follows: Determine $u_g \in H^1(\Omega)$ with $u_g = g$ on Γ in the sense of the trace theorem. Now, letting $u = u_0 + u_g$ leads to the variational problem:

$$\begin{aligned} &\text{find } u_0 \in H_0^1(\Omega) \text{ such that} \\ &\qquad a(u_0,v) = \ell(v) - a(u_g,v) \quad \text{for all } v \in H_0^1(\Omega). \end{aligned}$$

△

3.3 Variational Formulation of Neumann Problems

We consider the Neumann problem

$$\begin{aligned} -\operatorname{div}\big(\boldsymbol{A}(\boldsymbol{x})\nabla u(\boldsymbol{x})\big) + c(\boldsymbol{x})u(\boldsymbol{x}) &= f(\boldsymbol{x}), \quad \boldsymbol{x}\in\Omega, \\ \langle \boldsymbol{A}(\boldsymbol{x})\nabla u(\boldsymbol{x}), \boldsymbol{n}\rangle &= g(\boldsymbol{x}), \quad \boldsymbol{x}\in\Gamma. \end{aligned} \tag{3.5}$$

with a uniformly elliptic differential operator and a bounded, positive reaction term, i.e.

$$0 < \underline{\alpha}\|\boldsymbol{\xi}\|_2^2 \le \boldsymbol{\xi}^\intercal \boldsymbol{A}(\boldsymbol{x})\boldsymbol{\xi} \le \overline{\alpha}\|\boldsymbol{\xi}\|_2^2 < \infty, \quad 0 < \underline{c} \le c(\boldsymbol{x}) \le \overline{c} < \infty.$$

Multiplying (3.5) by $\phi \in C^\infty \cap H^1(\Omega)$ and integration yield

$$\int_\Omega \big(-\operatorname{div}(\boldsymbol{A}\nabla u) + cu\big)\phi \,\mathrm{d}\boldsymbol{x} = \int_\Omega \langle \boldsymbol{A}\nabla u, \nabla\phi\rangle + cu\phi \,\mathrm{d}\boldsymbol{x} - \int_\Gamma \langle \boldsymbol{A}\nabla u, \boldsymbol{n}\rangle\phi \,\mathrm{d}\sigma = \int_\Omega f\phi \,\mathrm{d}\boldsymbol{x}.$$

Hence, we obtain

$$a(v,w) = \int_\Omega \langle \boldsymbol{A}\nabla v, \nabla w\rangle + cvw \,\mathrm{d}\boldsymbol{x}, \quad \ell(v) = \int_\Gamma gv \,\mathrm{d}\sigma + \int_\Omega fv \,\mathrm{d}\boldsymbol{x} \quad \text{for } v,w \in H^1(\Omega).$$

The continuity of the bilinear form a follows as for the Dirichlet problem. In addition, there holds

$$a(v,v) \ge \underline{\alpha}|v|_{H^1(\Omega)}^2 + \underline{c}\|v\|_{L^2(\Omega)}^2 \ge \min\{\underline{\alpha}, \underline{c}\}\|v\|_{H^1(\Omega)}^2,$$

which shows the ellipticity of a on $H^1(\Omega)$. Moreover, for $f \in L^2(\Omega)$, $g \in L^2(\Gamma)$, we obtain

$$|\ell(v)| \le \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)}\|\gamma(v)\|_{L^2(\Gamma)} \le C\|v\|_{H^1(\Omega)}, \quad C > 0,$$

by the trace theorem. Hence, the linear form is continuous.

Theorem 3.16 Let $\Omega \subset \mathbb{R}^d$ satisfy the cone condition and let $f \in L^2(\Omega)$, $g \in L^2(\Gamma)$. Then, (3.5) exhibits a unique weak solution $u \in H^1(\Omega)$, which solves the minimisation problem

$$J(v) = \frac{1}{2}a(v,v) - \ell(v) \to \min.$$

In addition, it holds $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$, iff (3.5) has a classical solution.

Proof. The first part of the proof follows from the Lax-Milgram theorem. The second part follows by the fundamental lemma of calculus of variations. □

If $c \equiv 0$ in (3.5), then obviously $u + \eta$, $\eta \in \mathbb{R}$, is a solution to (3.5). In particular, the bilinear form a is not elliptic anymore. Let $V := \big\{v \in H^1(\Omega) : (v,1)_{L^2(\Omega)} = 0\big\} \subset H^1(\Omega)$ be the subspace of functions with vanishing *mean value*

$$\overline{v} := \frac{1}{|\Omega|}(v,1)_{L^2(\Omega)} = 0.$$

Since there holds a Poincaré inequality of the form

$$\|v\|_{L^2(\Omega)} \le C(|\overline{v}| + |v|_{H^1(\Omega)}), \quad C > 0,$$

in $H^1(\Omega)$, the bilinear form a is elliptic on V. Consequently, the Lax-Milgram theorem indicates the unique solvability of (3.5) in V in the case $c \equiv 0$.

Remark For $v \equiv 1$, the variational formulation yields the *compatibility condition*

$$\int_\Omega f \,\mathrm{d}\boldsymbol{x} + \int_\Gamma g \,\mathrm{d}\sigma = 0,$$

which is necessary to guarantee the existence of a solution. However, there holds

$$(f+\eta, v)_{L^2(\Omega)} = \int_\Omega f v \,\mathrm{d}\boldsymbol{x} + \eta \int_\Omega v \,\mathrm{d}\boldsymbol{x} = (f, v)_{L^2(\Omega)} \quad \text{for every } v \in V,\ \eta \in \mathbb{R}.$$

To enforce the compatibility condition, we set $\tilde{f} = f - \eta$ and make the ansatz

$$0 = \int_\Omega \tilde{f} \,\mathrm{d}\boldsymbol{x} + \int_\Gamma g \,\mathrm{d}\sigma = \int_\Omega f \,\mathrm{d}\boldsymbol{x} - \eta|\Omega| + \int_\Gamma g \,\mathrm{d}\sigma.$$

This yields

$$\eta := \frac{1}{|\Omega|}\Big(\int_\Omega f \,\mathrm{d}\boldsymbol{x} + \int_\Gamma g \,\mathrm{d}\sigma \Big).$$

△

3.4 Galerkin Methods

Let $a\colon V \times V \to \mathbb{R}$ denote a continuous, elliptic bilinear form on the Hilbert space V. In addition let $\ell \in V'$ denote a continuous linear form. Our goal is to compute the solution $u \in V$ to the variational problem

$$\begin{aligned} &\text{Find } u \in V \text{ such that} \\ &\quad a(u, v) = \ell(v) \quad \text{for all } v \in V. \end{aligned} \tag{3.6}$$

To that end, we restrict the variational problem to a finite dimensional subspace $V_h \subset V$. This yields the *Galerkin method*

$$\begin{aligned} &\text{Find } u_h \in V_h \text{ such that} \\ &\quad a(u_h, v_h) = \ell(v_h) \quad \text{for all } v_h \in V_h. \end{aligned} \tag{3.7}$$

Since V_h is a finite dimensional subspace of V, it is particularly closed, i.e. it is itself a Hilbert space. Hence, the existence and the uniqueness of the solution $u_h \in V_h$ are guaranteed by the Lax-Milgram theorem. In addition, there holds

$$c_E \|u_h\|_V^2 \le a(u_h, u_h) = \ell(u_h) \le \|\ell\|_{V'} \|u_h\|_V,$$

where $c_E > 0$ is the constant of ellipticity. From this, we directly infer the stability estimate

$$\|u_h\|_V \le \frac{1}{c_E} \|\ell\|_{V'}.$$

Remark If the bilinear form a from (3.6) is additionally symmetric, then, due to the Characterisation theorem 3.10, the variational problem (3.6) is equivalent to the minimisation problem

$$J(v) = \frac{1}{2} a(v, v) - \ell(v) \to \min, \quad v \in V.$$

Replacing V by V_h, we end up with the finite dimensional minimisation problem

$$J(v_h) = \frac{1}{2} a(v_h, v_h) - \ell(v_h) \to \min, \quad v_h \in V_h.$$

The latter procedure is called *Ritz-Galerkin method.* Herein, the solution $u_h \in V_h$ is also described by the Characterisation theorem 3.10. △

The Galerkin method leads to a linear system of equations in a straightforward manner: Let $\{\varphi_1, \dots, \varphi_n\}$ denote a basis of V_h. Then, due to linearity, (3.7) is equivalent to

$$\begin{aligned}&\text{Find } u_h \in V_h \text{ such that}\\ &\quad a(u_h, \varphi_i) = \ell(\varphi_i) \quad \text{for all } i = 1, \dots, n.\end{aligned}$$

Hence, making the ansatz

$$u_h = \sum_{i=1}^{n} u_i \varphi_i$$

results in the linear system

$$\sum_{j=1}^{n} a(\varphi_j, \varphi_i) u_j = \ell(\varphi_i) \quad \text{for all } i = 1, \dots, n.$$

In matrix notation, this system reads

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f},$$

where $\boldsymbol{A} := [a(\varphi_j, \varphi_i)]_{i,j=1}^{n}$, $\boldsymbol{u} := [u_i]_{i=1}^{n}$, $\boldsymbol{f} := [\ell(\varphi_i)]_{i=1}^{n}$.

The matrix $\boldsymbol{A}$ is referred to as *stiffness matrix*. In case of an elliptic, continuous and symmetric bilinear form a, the matrix $\boldsymbol{A}$ is symmetric and positive definite. There holds

$$\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{v} = \sum_{i,j=1}^{n} a(\varphi_j, \varphi_i) v_i v_j = a\bigg(\sum_{j=1}^{n} v_j \varphi_j, \sum_{i=1}^{n} v_i \varphi_i\bigg) \geq c_E \bigg\| \sum_{i=1}^{n} v_i \varphi_i \bigg\|_V^2 .$$

The latter norm only becomes zero, iff $\boldsymbol{v} = \boldsymbol{0}$.

The next lemma quantifies the quality of the approximation by the Galerkin method.

Theorem 3.17 (Céa's lemma) Let $a\colon V \times V \to \mathbb{R}$ denote an elliptic and continuous bilinear form. Moreover, let $u \in V$ and $u_h \in V_h \subset V$ denote the solutions to (3.6) and (3.7), respectively. Then, it holds

$$\|u - u_h\|_V \leq \frac{c_S}{c_E} \inf_{v_h \in V_h} \|u - v_h\|_V ,$$

where $c_E > 0$ is the constant of ellipticity and is $c_S > 0$ the constant of continuity associated to a.

Proof. By definition, there holds

$$a(u, v_h) = \ell(v_h) = a(u_h, v_h) \quad \text{for all } v_h \in V_h \subset V$$

and hence

$$a(u - u_h, v_h) = 0 \quad \text{for all } v_h \in V_h. \tag{3.8}$$

This property is called *Galerkin orthogonality* .

Now, suppose $v_h \in V_h$. Then (3.8) yields for $v_h - u_h \in V_h$ that

$$a(u - u_h, v_h - u_h) = 0$$

and consequently

$$\begin{aligned} c_E\|u-u_h\|_V^2 &\le a(u-u_h, u-u_h) \\ &= a(u-u_h, u-v_h) + a(u-u_h, v_h-u_h) \\ &\le c_S\|u-u_h\|_V\|u-v_h\|_V. \end{aligned}$$

Simplifying this inequality and taking into account that it holds for any $v_h \in V_h$ yields the assertion. $\square$

Corollary 3.18 Let $a\colon V\times V \to \mathbb{R}$ denote an elliptic, continuous and symmetric bilinear form. Moreover, let $u \in V$ and $u_h \in V_h \subset V$ denote the solutions to (3.6) and (3.7), respectively. Then, it holds

$$\|u-u_h\|_V \le \sqrt{\frac{c_S}{c_E}} \inf_{v_h\in V_h} \|u-v_h\|_V,$$

where $c_E > 0$ is the constant of ellipticity and is $c_S > 0$ the constant of continuity associated to a.

Proof. Due to the Galerkin orthogonality (3.8), the Cauchy-Schwarz inequality for the energy norm, cf. (3.4), reads

$$\|u-u_h\|_a^2 = a(u-u_h, u-u_h) = a(u-u_h, u-v_h) \le \|u-u_h\|_a\|u-v_h\|_a \quad \text{for all } v_h \in V_h.$$

This yields Céa's lemma in the energy norm

$$\|u-u_h\|_a \le \|u-v_h\|_a \quad \text{for all } v \in V_h.$$

The claim is now obtained via the inequality

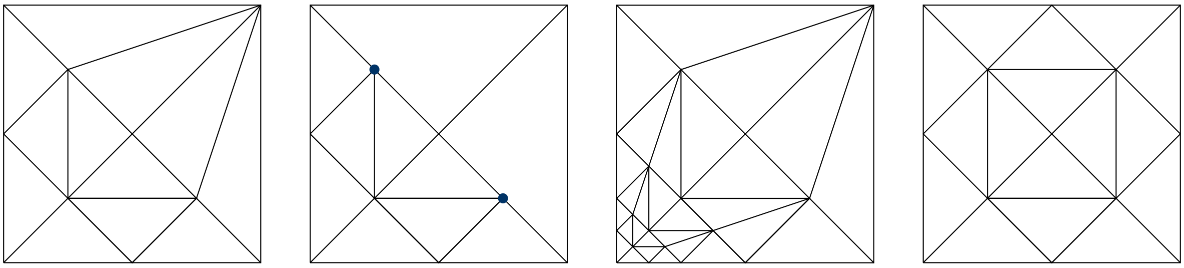
$$c_E\|u-u_h\|_V^2 \le \|u-u_h\|_a^2 \le \|u-v_h\|_a^2 \le c_S\|u-v_h\|_V^2.$$

$\square$

4. The Finite Element Method

Céa's lemma tells us that the quality of the approximation u_h to the solution u of (3.6) is governed by the approximation quality of the subspace V_h. Hence, the next step is to construct suitable approximation spaces V_h in a systematic fashion. For the sake of simplicity, we assume in what follows that $\Omega \subset \mathbb{R}^d$ is a polygonal domain.

4.1 Meshing



admissible non-admissible shape-regular uniform

Definition 4.1 A partition $\mathcal{T} = \{T_1, \dots, T_m\}$ of Ω into simplices is called *admissible*, iff the following two conditions are satisfied:

1. It holds $\overline{\Omega} = \cup_{i=1}^m T_i$.
2. If $T_i \cap T_j \neq \emptyset$ for $i \neq j$, then $T_i \cap T_j$ is either a point, an edge or a face of T_i as well as of T_j.

We write $\mathcal{T}_h$ instead of $\mathcal{T}$, iff $\operatorname{diam}(T_i) \leq 2h$ for $i = 1, \dots, m$.
A family $\{\mathcal{T}_h\}_{h>0}$ of triangulations is *shape-regular* , iff there exists $\kappa > 0$ such that

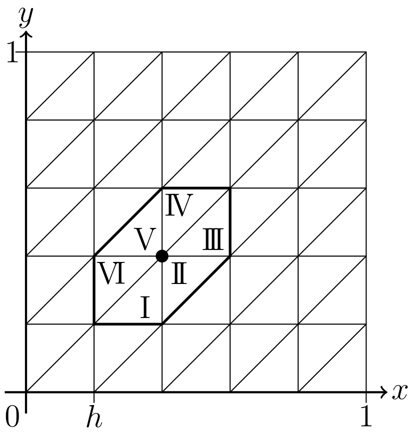
$$\rho_T \geq h_T/\kappa \tag{4.1}$$

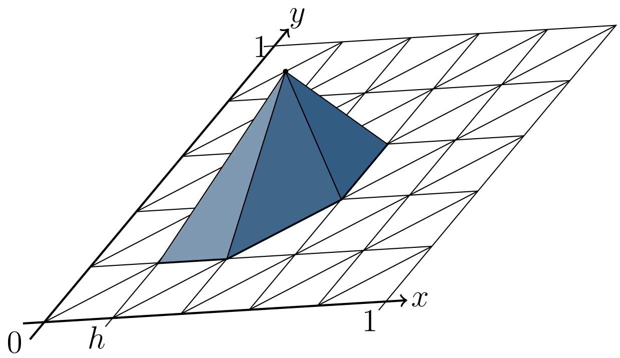
for all $T \in \mathcal{T}_h$, where ρ_T is the radius of the incircle of T and h_T is the diameter of T.
The family is called *uniform*, iff there exists $h > 0$ such that $h_T = h$ independently of $T \in \mathcal{T}_h$ in (4.1).

Example 4.2 (Courant 1943) We consider the Poisson's equation

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \Gamma,$$

where $\Omega := (0,1)^2$ is the unit square, which is triangulated with congruent triangles of mesh size $h > 0$.





The finite dimensional approximation space in the Galerkin method is given by

$$V_h := \{v \in C^0(\overline{\Omega}) : v|_T \text{ is linear for every } T \in \mathcal{T}_h \text{ and } v|_\Gamma = 0\} \subset H_0^1(\Omega).$$

For every $T \in \mathcal{T}_h$, a function $v \in V_h$ satisfies $v|_T(x, y) = a + bx + cy$ and is hence uniquely determined by its values at the vertices of T. Globally, $v \in V_h$ is determined by its values at the grid points $(x_i, y_i) \in \overline{\Omega}$. A suitable basis of V_h is given by the hat functions

$$\varphi_i(x_j, y_j) = \delta_{ij}.$$

Thus, if $h = 1/n$ for $n \in \mathbb{N}^*$ then $\dim V_h = (n-1)^2$. In the sequel, we compute the entries of the stiffness matrix $\boldsymbol{A}$. Let φ_Z denote the basis function. supported on the elements I–VI with maximum in (x_i, y_i). Analogously, we define$\varphi_N, \varphi_S, \varphi_W, \varphi_E$ via

$$\varphi_N(x_i, y_i + h) = 1,\ \varphi_S(x_i, y_i - h) = 1,\ \varphi_W(x_i - h, y_i) = 1,\ \varphi_E(x_i + h, y_i) = 1.$$

The derivatives of the basis function φ_Z are given by

	I	II	III	IV	V	VI	else
$\partial_x \varphi_Z$	0	$-1/h$	$-1/h$	0	$1/h$	$1/h$	0
$\partial_y \varphi_Z$	$1/h$	$1/h$	0	$-1/h$	$-1/h$	0	0

Exploiting symmetry, we obtain

$$\begin{aligned} a(\varphi_Z, \varphi_Z) &= \int_{\text{I–VI}} \langle \nabla\varphi_Z, \nabla\varphi_Z \rangle \,\mathrm{d}\boldsymbol{x} = 2\int_{\text{I–III}} (\partial_x\varphi_Z)^2 + (\partial_y\varphi_Z)^2 \,\mathrm{d}\boldsymbol{x} \\ &= 2\int_{\text{II–III}} (\partial_x\varphi_Z)^2 \,\mathrm{d}\boldsymbol{x} + 2\int_{\text{I–II}} (\partial_y\varphi_Z)^2 \,\mathrm{d}\boldsymbol{x} = \frac{2}{h^2}\int_{\text{II–III}} 1 \,\mathrm{d}\boldsymbol{x} + \frac{2}{h^2}\int_{\text{I–II}} 1 \,\mathrm{d}\boldsymbol{x} = 4. \end{aligned}$$

Moreover, we have

$$\begin{aligned} a(\varphi_Z, \varphi_S) = a(\varphi_S, \varphi_Z) &= \int_{\text{I–II}} \langle \nabla\varphi_Z, \nabla\varphi_S \rangle \,\mathrm{d}\boldsymbol{x} \\ &= \int_{\text{I–II}} \partial_y\varphi_Z \partial_y\varphi_S \,\mathrm{d}\boldsymbol{x} = \int_{\text{I–II}} \frac{1}{h}\Big(-\frac{1}{h}\Big) \,\mathrm{d}\boldsymbol{x} = -1. \end{aligned}$$

Accordingly, there holds due to symmetry

$$a(\varphi_Z, \varphi_S) = a(\varphi_Z, \varphi_N) = a(\varphi_Z, \varphi_W) = a(\varphi_Z, \varphi_E) = -1.$$

Finally, we infer

$$a(\varphi_Z, \varphi_{SW}) = \int_{\text{I+VI}} \langle \nabla\varphi_Z, \nabla\varphi_{SW} \rangle \,\mathrm{d}\boldsymbol{x} = 0 = a(\varphi_Z, \varphi_{NE}).$$

Consequently, as in the finite difference method, we end up with the standard five-point stencil

$$\begin{bmatrix} \alpha_{NW} & \alpha_N & \alpha_{NE} \\ \alpha_W & \alpha_C & \alpha_E \\ \alpha_{SW} & \alpha_S & \alpha_{SE} \end{bmatrix}_* = \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}_*$$

for the particular choice of the mesh and the basis functions. However, we remark that the scaling of the stencil is different here. △

As in the example, we typically choose *piecewise* polynomial basis functions. Conseqently, the functions $v \in V_h$ are piecewise polynomials, as well. Here and in what follows, we say that a function satisfies a given property piecewise, iff this property holds for all elements $T \in \mathcal{T}$. Hence, $v \in V_h$ satisfies $v|_T \in \mathcal{P}_m(T)$ for all $T \in \mathcal{T}_h$, where

$$\mathcal{P}_m(T) := \left\{ v\colon T \to \mathbb{R} : v(\boldsymbol{x}) = \sum_{0 \le |\boldsymbol{\alpha}| \le m} c_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}} \text{ with } c_{\boldsymbol{\alpha}} \in \mathbb{R} \text{ for } 0 \le |\boldsymbol{\alpha}| \le m \right\}.$$

The next theorem gives a sufficient criterion for $V_h \subset H^k(\Omega)$ for $k \ge 1$.

Theorem 4.3 Let $k \ge 1$ and assume $\Omega \subset \mathbb{R}^d$ is bounded. Suppose $v\colon \overline{\Omega} \to \mathbb{R}$ satisfies $V|_T \in C^\infty(T)$ for all $T \in \mathcal{T}$. Then, it holds $v \in H^k(\Omega)$, iff $v \in C^{k-1}(\overline{\Omega})$.

Proof. For a proof of this theorem, see [Braess]. □

4.2 The P1-Element

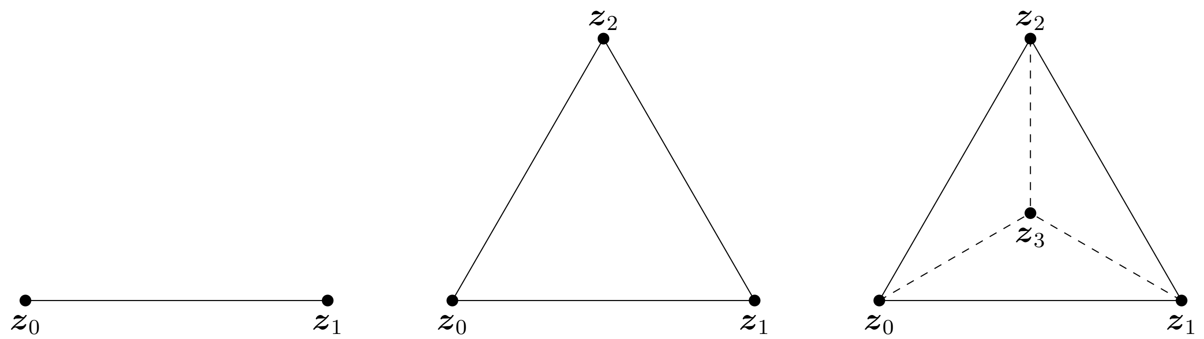
Although it is possible to construct finite elements also on quadrilaterals and hexahedrons and for higher order polynomials, we restrict ourselves in this lecture to the situation of finite elements of lowest order on triangles and tetrahedrons and refer to [Braess] for a more comprehensive discussion.

For historical reasons, we start by the definition of finite elements as introduced by Ciarlet.

Definition 4.4 A *finite element* is a triple (T, Π, Σ) with the following properties:

1. $T \subset \mathbb{R}^d$ is closed set called *element*.
2. $\Pi \subset C^0(T)$ is an s-dimensional subspace. The basis funcions $\{v_1, \dots v_s\} \subset \Pi$ are called *shape functions*.
3. Σ is a set of s bounded, linear functionals $\chi_1, \dots, \chi_s$ on Π and every $p \in \Pi$ is uniquely determined by the values $\chi_i(p)$, $i = 1, \dots, s$. (These functionals correspond to generalised interpolation conditions).

The dimension s corresponds the *number of local degrees of freedom*.



Definition 4.5 Let $\boldsymbol{z}_0, \ldots, \boldsymbol{z}_d \in \mathbb{R}^d$ be $(d+1)$ non-collinear vertices, i.e. the vectors $\boldsymbol{z}_1 - \boldsymbol{z}_0, \ldots, \boldsymbol{z}_d - \boldsymbol{z}_0$ are linearly independent. We set

$$T := \operatorname{conv}\{\boldsymbol{z}_0, \ldots, \boldsymbol{z}_d\} \subset \mathbb{R}^d,$$

i.e. T is the convex hull of the vertices. Moreover, we define $\Pi := \mathcal{P}_1(T)$ and

$$\Sigma := \{\chi_i \colon \Pi \to \mathbb{R} : \chi_i(v) := v(\boldsymbol{z}_i),\ i = 0, \ldots, d\}.$$

The triplet (T, Π, Σ) is called *P1-element*.

The P1-element satisfies the definition of a finite element.

Theorem 4.6 The P1-element is a finite element.

Proof. Obviously, T is a closed set. Moreover, it holds

$$\mathcal{P}_1(T) = \begin{cases} \operatorname{span}\{1, x_1\}, & d = 1, \\ \operatorname{span}\{1, x_1, x_2\}, & d = 2, \\ \operatorname{span}\{1, x_1, x_2, x_3\}, & d = 3. \end{cases}$$

Therefore, the dimension of $\mathcal{P}_1(T)$ is $s = d + 1$. In order to show the uniqueness of the representation, it suffices to show that from $\chi_i(q) = 0$ for $i = 0, \ldots, d$ already follows $q \equiv 0$. Let

$$q = c_0 + \sum_{i=1}^{d} c_i x_i \quad \text{for } c_0, \ldots, c_d \in \mathbb{R}.$$

Then, the coefficients are determined by the linear system

$$\begin{bmatrix} 1 & z_{0,1} & \cdots & z_{0,d} \\ 1 & z_{1,1} & \cdots & z_{1,d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{d,1} & \cdots & z_{d,d} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_d \end{bmatrix} = \mathbf{0}.$$

The non-collinearity of the points implies that the matrix is regular. Hence, $c_0 = c_1 = \ldots = c_d = 0$ is the only solution, which shows $q \equiv 0$. $\square$

Remark If $\Omega \subset \mathbb{R}^d$ is a bounded *Lipschitz domain*, i.e. Γ can locally be parametrised by a Lipschitz continuous function, then the Sobolev imbedding theorem guarantees $H^2(\Omega) \subset C^0(\overline{\Omega})$ for $d \leq 3$. Hence, the point evaluations in the definition of the P1-element are well defined on $H^2(\Omega)$. $\triangle$

Theorem 4.7 (Nodal Basis) Let (T, Π, Σ) be a finite element. There exists a basis $\{\varphi_1, \ldots, \varphi_s\} \subset \Pi$ such that

$$\chi_i(\varphi_j) = \delta_{ij} \quad \text{for } i, j = 1, \ldots, s.$$

Proof. Let $\{\psi_1, \ldots, \psi_s\}$ denote a basis of Π and define the matrix $\boldsymbol{A} := [\chi_i(\psi_j)]_{i,j=1}^s \in \mathbb{R}^{s \times s}$. This matrix is regular, since the ψ_j form a basis of Π and the functionals χ_i are linear. Now, let $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_s\} \subset \mathbb{R}^s$ denote the canonical basis. Then, the functions

$$\varphi_j := \sum_{i=1}^s c_{j,i}\psi_i, \quad \text{where } \boldsymbol{c}_j := \boldsymbol{A}^{-1}\boldsymbol{e}_j,$$

satisfy

$$\chi_\ell(\varphi_j) = \sum_{i=1}^s c_{j,i}\chi_\ell(\psi_i) = \sum_{i=1}^s c_{j,i}a_{\ell,i} = [a_{\ell,i}]_{\ell=1}^s \boldsymbol{A}^{-1}\boldsymbol{e}_j = \boldsymbol{e}_\ell^\intercal \boldsymbol{e}_j = \delta_{\ell j}$$

and form a basis of Π. The latter is verified by observing that $\psi_j = \sum_{i=1}^s \chi_i(\psi_j)\varphi_i$ for $i = 1, \ldots, s$. $\square$

Example 4.8 Let $\widehat{T} := \operatorname{conv}\{[0,0]^\intercal, [1,0]^\intercal, [0,1]^\intercal\} \subset \mathbb{R}^2$, then the nodal basis of the P1-element is given by

$$\varphi_1(x,y) = 1 - x - y, \ \varphi_2(x,y) = x, \ \varphi_3(x,y) = y.$$

Let $\widehat{T} := \operatorname{conv}\{[0,0,0]^\intercal, [1,0,0]^\intercal, [0,1,0]^\intercal, [0,0,1]^\intercal\} \subset \mathbb{R}^3$, then the nodal basis of the P1-element is given by

$$\varphi_1(x,y,z) = 1 - x - y - z, \ \varphi_2(x,y,z) = x, \ \varphi_3(x,y,z) = y, \ \varphi_4(x,y,z) = z.$$

$\triangle$

Definition 4.9 Given a finite element (T, Π, Σ) and a function $v \in H^2(T)$ the *nodal interpolant* $\mathcal{I}_T v \in \Pi$ is defined as

$$\mathcal{I}_T v := \sum_{i=1}^s \chi_i(v)\varphi_i,$$

where $\{\varphi_1, \ldots, \varphi_s\} \subset \Pi$ denotes the nodal basis. In particular, there holds

$$\chi_i(\mathcal{I}_T v) = \chi_i(v) \quad \text{for } i = 1, \ldots, s.$$

Remark The nodal interpolant is unique: Assume $\mathcal{I}_T v = \sum_{i=1}^{s} c_i \varphi_i$, then it holds

$$\chi_i(\mathcal{I}_T v) = \chi_i(v) = c_i$$

by the definition of the nodal interpolant. △

4.3 Affine Families

Although it is possible to compute the finite element stiffness matrix and the corresponding right hand side directly on the actual triangulation as in Example 4.2, it is much more convenient to perform the calculations on a reference configuration $(\widehat{T}, \widehat{\Pi}, \widehat{\Sigma})$. This accounts particularly, when numerical quadrature has to be employed, e.g. in the case of a non trivial diffusion coefficient $\boldsymbol{A}(\boldsymbol{x})$ in the variational formulation.

Definition 4.10 Let $\mathcal{T}_h$ denote a triangulation for $\Omega \subset \mathbb{R}^d$. An *affine family* is a familiy of finite elements

$$(T, \Pi_T, \Sigma_T)_{T \in \mathcal{T}_h}$$

such that each finite element (T, Π_T, Σ_T) is obtained by an affine transformation of a common *reference element* $(\widehat{T}, \widehat{\Pi}, \widehat{\Sigma})$, i.e. for each $T \in \mathcal{T}_h$ there exists an affine diffeomorphism Φ_T such that

$$T = \Phi_T(\widehat{T}),\ \Pi_T = \left\{\hat{p} \circ \Phi_T^{-1} : \hat{p} \in \widehat{\Pi}\right\},\ \Sigma_T = \left\{\hat{\chi} \circ \Phi_T^{-1} : \hat{\chi} \in \widehat{\Sigma}\right\}.$$

The next theorem tells us under which conditions there exists an affine diffeomorphism that can be used to construct an affine family and establishes norm "equivalences" for the Sobolev spaces defined on T and $\widehat{T}$, respectively.

Theorem 4.11 Let $\widehat{T} := \operatorname{conv}\{\boldsymbol{0}, \boldsymbol{e}_1, \dots \boldsymbol{e}_d\}$, with the canonical basis $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_d\} \subset \mathbb{R}^d$, denote the unit simplex. Assume $T = \operatorname{conv}\{\boldsymbol{z}_0, \dots, \boldsymbol{z}_d\} \subset \mathbb{R}^d$ is a non-degenerate simplex. Then, there exists a unique affine diffeomorphism $\Phi_T \colon \widehat{T} \to T$ with

$$\Phi_T(\boldsymbol{0}) = \boldsymbol{z}_0 \quad \text{and} \quad \Phi_T(\boldsymbol{e}_i) = \boldsymbol{z}_i \text{ for } i = 1, \dots, d.$$

In addition, there holds for every $v \in H^m(T)$ and $\hat{v} := v \circ \Phi_T \in H^m(\widehat{T})$ that

$$\begin{aligned} |v|_{H^k(T)} &\le c\rho_T^{-k} |\det \boldsymbol{B}|^{\frac{1}{2}} |\hat{v}|_{H^k(\widehat{T})}, \\ |\hat{v}|_{H^k(\widehat{T})} &\le c' h_T^k |\det \boldsymbol{B}|^{-\frac{1}{2}} |v|_{H^k(T)}, \end{aligned}$$

for $0 \le k \le m$ and constants $c, c' > 0$. Herein, the H^k-semi-norm is defined via

$$|v|_{H^k(\Omega)} := \left(\sum_{|\boldsymbol{\alpha}|=k} \|\partial^{\boldsymbol{\alpha}} v\|_{L^2(\Omega)}^2 \right)^{\frac{1}{2}}.$$

The constants are independent of ρ_T, h_T, cf. Definition 4.1, and of the Jacobian $\boldsymbol{B} \in \mathbb{R}^{d \times d}$ of Φ_T.

Proof. Let

$$\boldsymbol{B} := [\boldsymbol{z}_1 - \boldsymbol{z}_0, \dots, \boldsymbol{z}_d - \boldsymbol{z}_0] \in \mathbb{R}^{d\times d}.$$

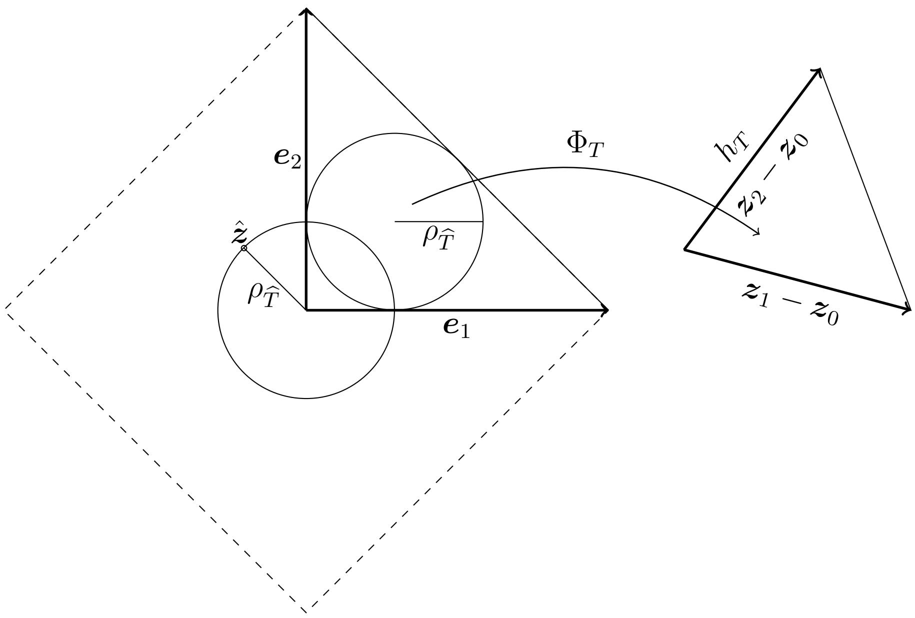
Due to the non-degeneracy of T, $\boldsymbol{B}$ is a regular matrix. Moreover, the transformation

$$\Phi_T \colon \widehat{T} \to \mathbb{R}^d,\ \hat{\boldsymbol{x}} \mapsto \boldsymbol{B}\hat{\boldsymbol{x}} + \boldsymbol{z}_0$$

satisfies $\Phi_T(\boldsymbol{0}) = \boldsymbol{z}_0$ and $\Phi_T(\boldsymbol{e}_i) = \boldsymbol{z}_i$ for $i = 1, \dots, d$. Since the image of a convex set under an affine transformation is convex, we have $\Phi_T(\widehat{T}) = T$.
The inverse of Φ_T is given by

$$\Phi_T^{-1} \colon T \to \widehat{T},\ \boldsymbol{x} \mapsto \boldsymbol{B}^{-1}(\boldsymbol{x} - \boldsymbol{z}_0).$$

Hence, Φ_T is an affine diffeomorphism with Jacobian $\boldsymbol{B}$.



Now, let $\hat{\boldsymbol{z}} \in \mathbb{R}^d$ with $\|\hat{\boldsymbol{z}}\|_2 = \rho_{\widehat{T}}$. Then, there exist $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}} \in \widehat{T}$ with $\hat{\boldsymbol{z}} = \hat{\boldsymbol{y}} - \hat{\boldsymbol{x}}$ and it holds $\boldsymbol{B}\hat{\boldsymbol{z}} = \boldsymbol{B}(\hat{\boldsymbol{y}} - \hat{\boldsymbol{x}})$ and $\|\boldsymbol{B}(\hat{\boldsymbol{y}} - \hat{\boldsymbol{x}})\|_2 \le h_T$. From this, we deduce

$$\|\boldsymbol{B}\|_2 = \frac{1}{\rho_{\widehat{T}}} \sup_{\|\hat{\boldsymbol{z}}\|_2 = \rho_{\widehat{T}}} \|\boldsymbol{B}\hat{\boldsymbol{z}}\|_2 = \frac{1}{\rho_{\widehat{T}}} \sup_{\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}} \in \widehat{T}} \|\boldsymbol{B}(\hat{\boldsymbol{y}} - \hat{\boldsymbol{x}})\|_2 = \frac{h_T}{\rho_{\widehat{T}}}.$$

By interchanging the roles of T and $\widehat{T}$, we find in analogously

$$\|\boldsymbol{B}^{-1}\|_2 \le \frac{h_{\widehat{T}}}{\rho_T}.$$

Next, let $v \in C^\infty(T) \cap H^k(T)$. By the chain rule, it holds

$$\nabla v = \nabla(\hat{v} \circ \Phi_T^{-1}) = \boldsymbol{B}^{-\intercal}\widehat{\nabla}\hat{v} \circ \Phi_T^{-1} \quad \text{or} \quad \widehat{\nabla v} = \boldsymbol{B}^{-\intercal}\widehat{\nabla}\hat{v}.$$

Multiplication with $\boldsymbol{B}^\intercal$ then gives

$$\widehat{\nabla}\hat{v} = \boldsymbol{B}^\intercal\widehat{\nabla v}.$$

With the previous estimate on the norm of $\boldsymbol{B}$, we obtain

$$\|\widehat{\nabla v}\|_2 \leq c\rho_T^{-1}\|\hat{\nabla}\hat{v}\|_2 \quad \text{and} \quad \|\hat{\nabla}\hat{v}\|_2 \leq c' h_T \|\widehat{\nabla v}\|_2$$

for some constants $c, c' > 0$. More generally, since the Jacobian of Φ_T is constant, we have for any multi index $\boldsymbol{\alpha} \in \mathbb{N}^d$ that

$$|\widehat{\partial^{\boldsymbol{\alpha}} v}| \leq c\rho_T^{-|\boldsymbol{\alpha}|} \max_{|\boldsymbol{\beta}|=|\boldsymbol{\alpha}|} |\hat{\partial}^{\boldsymbol{\beta}}\hat{v}| \quad \text{and} \quad |\hat{\partial}^{\boldsymbol{\alpha}}\hat{v}| \leq c' h_T^{|\boldsymbol{\alpha}|} \max_{|\boldsymbol{\beta}|=|\boldsymbol{\alpha}|} |\widehat{\partial^{\boldsymbol{\beta}} v}|$$

for some other constants $c, c' > 0$ which depend on k, d, where $\boldsymbol{\beta} \in \mathbb{N}^d$. From these expressions, it is easy to derive

$$\sum_{|\boldsymbol{\alpha}|=k} (\widehat{\partial^{\boldsymbol{\alpha}} v})^2 \leq c\rho_T^{-2k} \sum_{|\boldsymbol{\alpha}|=k} (\hat{\partial}^{\boldsymbol{\alpha}}\hat{v})^2 \quad \text{and} \quad \sum_{|\boldsymbol{\alpha}|=k} (\hat{\partial}^{\boldsymbol{\alpha}}\hat{v})^2 \leq c' h_T^{2k} \sum_{|\boldsymbol{\alpha}|=k} (\widehat{\partial^{\boldsymbol{\alpha}} v})^2$$

for some other constants $c, c' > 0$ which depend on k, d. For a given function $w \in L^1(T)$, the transformation formula yields with $\hat{w} := w \circ \Phi_T \in L^1(\widehat{T})$ that

$$\int_{\widehat{T}} \hat{w} \,\mathrm{d}\hat{\boldsymbol{x}} = \int_{\Phi_T(\widehat{T})} \hat{w} \circ \Phi_T^{-1} |\det \boldsymbol{B}^{-1}| \,\mathrm{d}\boldsymbol{x} = |\det \boldsymbol{B}^{-1}| \int_T w \,\mathrm{d}\boldsymbol{x}.$$

Hence, applying the transformation formula to $w = (\partial^{\boldsymbol{\alpha}} v)^2$ gives

$$\begin{aligned} |v|_{H^k(T)}^2 &= \sum_{|\boldsymbol{\alpha}|=k} \int_T (\partial^{\boldsymbol{\alpha}} v)^2 \,\mathrm{d}\boldsymbol{x} = |\det \boldsymbol{B}| \sum_{|\boldsymbol{\alpha}|=k} \int_{\widehat{T}} (\widehat{\partial^{\boldsymbol{\alpha}} v})^2 \,\mathrm{d}\hat{\boldsymbol{x}} \\ &\leq |\det \boldsymbol{B}| c\rho_T^{-2k} \sum_{|\boldsymbol{\alpha}|=m} \int_{\widehat{T}} (\hat{\partial}^{\boldsymbol{\alpha}}\hat{v})^2 \,\mathrm{d}\hat{\boldsymbol{x}} = c\rho_T^{-2k} |\det \boldsymbol{B}| |\hat{v}|_{H^k(\widehat{T})}. \end{aligned}$$

The second inequality is derived in a similar fashion. The assertion is then obtained by a density argument. □

4.4 Approximation Properties

In this section, we consider the global approximation error induced by the interpolation of a function $v \in H^2(\Omega)$ in the finite element space V_h.

Theorem 4.12 (Bramble-Hilbert Lemma) Let $F\colon H^m(T) \to \mathbb{R}$ denote a bounded and sublinear functional, i.e.

$$|F(v)| \leq c\|v\|_{H^m(\Omega)} \text{ for } c > 0 \quad \text{and} \quad |F(v+w)| \leq \big(|F(v)| + |F(w)|\big),$$

and assume $\mathcal{P}_{m-1}(T) \subset \ker(F)$. Then, there exists $C > 0$ such that

$$|F(v)| \leq C|v|_{H^m(T)} \quad \text{for all } v \in H^m(T).$$

Proof. For a proof of the Bramble-Hilbert lemma, see [Bartels]. □

Corollary 4.13 Let (T, Π, Σ) be a finite element with $\mathcal{P}_{m-1}(T) \subseteq \Pi$. Then, there exists $C > 0$ such that

$$\|v - \mathcal{I}_T v\|_{H^m(T)} \leq C|v|_{H^m(T)} \quad \text{for all } v \in H^m(T).$$

Proof. We set $F(v) = \|v - \mathcal{I}_T v\|_{H^m(T)}$ and note that F is sublinear. Moreover, with the definition of the interpolant $\mathcal{I}_T v = \sum_{i=1}^s \chi_i(v)\varphi_i$ and the continuity of the functionals χ_i, i.e. $|\chi_i(v)| \leq c\|v\|_{H^m(T)}$ for $i = 1, \dots, s$, we have

$$\begin{aligned}
|F(v)| &\leq \|v\|_{H^m(T)} + \|\mathcal{I}_T v\|_{H^m(T)} \leq \|v\|_{H^m(T)} + \sum_{i=1}^s |\chi_i(v)| \|\varphi_i\|_{H^m(T)} \\
&\leq \|v\|_{H^m(T)} + c\|v\|_{H^m} \sum_{i=1}^s \|\varphi_i\|_{H^m(T)} \leq C\|v\|_{H^m}
\end{aligned}$$

for some $C > 0$. In addition, there holds $F(v) = 0$ for all $v \in \Pi$. Hence, the conditions of the Bramble-Hilbert lemma are satisfied, which implies the assertion. □

Definition 4.14 Let $\mathcal{T}_h$ be a triangulation for $\Omega \subset \mathbb{R}^d$ and let $(T, \Pi_T, \Sigma_T)_{T \in \mathcal{T}_h}$ denote an affine family. The *global interpolant* $\mathcal{I}\colon H^m(\Omega) \to L^\infty(\Omega)$ is defined via

$$(\mathcal{I}v)|_T = \mathcal{I}_T(v|_T)$$

for all $T \in \mathcal{T}_h$. The affine family is called a C^k-*element*, iff $\mathcal{I}v \in C^k(\overline{\Omega})$ for all $v \in C^k(\overline{\Omega}) \cap H^m(\Omega)$.

In practice, for a given triangulation $\mathcal{T}_h$, one usually considers the *finite element spaces*

$$V_h = \{v \in C^0(\overline{\Omega}) : v|_T \in \mathcal{P}_m(T),\ T \in \mathcal{T}_h\} \subset H^1(\Omega), \tag{4.2}$$

which are based on a C^0-element. It is also possible to construct finite element spaces with higher global smoothness, e.g. C^1-elements for an H^2-conforming discretisation. However, they are much more difficult to construct.

In order to estimate the global approximation error of a function $v \in H^m$ by its interpolant $\mathcal{I}v \in V_h$, we introduce the mesh dependent norm

$$\|v\|_{m,h} := \left(\sum_{T \in \mathcal{T}_h} \|v\|^2_{H^m(T)} \right)^{\frac{1}{2}}.$$

This norm is a bit more general than the usual H^m-norm, since it does not require global smoothness of a function. However, it holds

$$\|v\|_{m,h} = \|v\|_{H^m(\Omega)} \quad \text{for all } v \in H^m.$$

We remark that, according to Theorem 4.3, this equality holds for the C^0-element for $m = 0, 1$.

Theorem 4.15 Let $(T, \Pi_T, \Sigma_T)_{T\in\mathcal{T}_h}$ denote an affine family with $\mathcal{P}_{m-1}(\widehat{T}) \subset \widehat{T}$, where $\mathcal{T}_h$ denotes a shape-regular triangulation for $\Omega \subset \mathbb{R}^d$. Moreover, let $m \geq 2$ and $0 \leq k \leq m$. Then, it holds

$$\|v - \mathcal{I}v\|_{k,h} \leq ch^{m-k}|v|_{H^m(\Omega)} \quad \text{for all } v \in H^m(\Omega)$$

with a constant $C > 0$, which only depends on Ω, κ and m.

Proof. From Theorem 4.11 we obtain for $T \in \mathcal{T}$ that

$$|v - \mathcal{I}_T v|_{H^\ell(T)} \leq c\left(\frac{h_T}{\kappa}\right)^{-\ell} |\det \boldsymbol{B}|^{\frac{1}{2}} |\hat{v} - \mathcal{I}_{\widehat{T}}\hat{v}|_{H^\ell(\widehat{T})} \leq Ch_T^{-\ell} |\det \boldsymbol{B}|^{\frac{1}{2}} |\hat{v} - \mathcal{I}_{\widehat{T}}\hat{v}|_{H^\ell(\widehat{T})}$$

for all $0 \leq \ell \leq k$ and some constants $c, C > 0$.
Moreover, due to Corollary 4.13, there holds

$$|\hat{v} - \mathcal{I}_{\widehat{T}}\hat{v}|_{H^\ell(\widehat{T})} \leq \|\hat{v} - \mathcal{I}_{\widehat{T}}\hat{v}\|_{H^m(\widehat{T})} \leq C|\hat{v}|_{H^m(\widehat{T})} \quad \text{for some } C > 0$$

and consequently

$$|v - \mathcal{I}_T v|_{H^\ell(T)} \leq ch_T^{-\ell} |\det \boldsymbol{B}|^{\frac{1}{2}} |\hat{v}|_{H^m(\widehat{T})}$$

for some other constant $c > 0$. Now, mapping the norm back to T results in

$$\begin{aligned} |v - \mathcal{I}_T v|_{H^\ell(T)} &\leq ch_T^{-\ell} |\det \boldsymbol{B}|^{\frac{1}{2}} |\hat{v}|_{H^m(\widehat{T})} \leq ch_T^{-\ell} |\det \boldsymbol{B}|^{\frac{1}{2}} c' h_T^m |\det \boldsymbol{B}|^{-\frac{1}{2}} |v|_{H^m(T)} \\ &\leq Ch_T^{m-\ell} |v|_{H^m(T)} \end{aligned}$$

for some constant $C > 0$. Thus, by summation we obtain

$$\|v - \mathcal{I}_T v\|^2_{H^k(T)} = \sum_{\ell=0}^{k} |v - \mathcal{I}_T v|^2_{H^\ell(T)} \leq C^2 |v|^2_{H^m(T)} \sum_{\ell=0}^{k} h_T^{2(m-\ell)} \leq ch_T^{2(m-k)} |v|_{H^m(T)}$$

for some constant $c > 0$. The assertion is obtained from this expression by taking the maximum over all diameters h_T and summation over all $T \in \mathcal{T}_h$. ☐

Consequently, for $v \in H^2(\Omega)$ and the P1-element, we obtain on a uniform triangulation the approximation estimates

$$\|v - \mathcal{I}v\|_{H^1(\Omega)} \leq ch|v|_{H^2(\Omega)} \quad \text{and} \quad \|v - \mathcal{I}v\|_{L^2(\Omega)} \leq ch^2|v|_{H^2(\Omega)}.$$

For functions $v \in V_h$, cf. (4.2), we can estimate a given Sobolev norm by a weaker one by introducing negative powers of $h > 0$.

Theorem 4.16 (Inverse Estimate) Let $\mathcal{T}_h$ denote a uniform triangulation of $\Omega \subset \mathbb{R}^d$. Moreover, let the finite element space V_h be given by (4.2), where the piecewise polynomial degree is $s \geq 0$. Then, there exists $c > 0$, which only depends on s, t, κ, such that

$$\|v\|_{m,h} \leq ch^{t-m}\|v\|_{t,h} \quad \text{for all } v \in V_h.$$

and $0 \leq t \leq m$

Proof. For a proof of this theorem, see [Braess]. □

From the theorem, we infer that, in the case of piecewise linear finite elements, there holds

$$\|v\|_{H^1(\Omega)} \leq ch^{-1}\|v\|_{L^2(\Omega)} \quad \text{for all } v \in V_h.$$

5. Error Estimates for Elliptic Problems

In the last paragraph, we have seen that we can derive approximation results based on interpolation, if the underlying function is in $H^2(\Omega)$, i.e.

$$\|v - \mathcal{I}v\|_{H^1(\Omega)} \leq ch|v|_{H^2(\Omega)} \quad \text{for all } v \in H^2(\Omega).$$

Hence, if the solution u to the elliptic variational problem

$$\begin{aligned} &\text{Find } u \in V \text{ such that} \\ &\qquad a(u,v) = \ell(v) \quad \text{for all } v \in V \end{aligned}$$

is contained in $H^2(\Omega)$, Céa's lemma yields

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{c_S}{c_E} ch|u|_{H^2(\Omega)}.$$

In this chapter, we shall investigate when this is the case for solutions to elliptic problems.

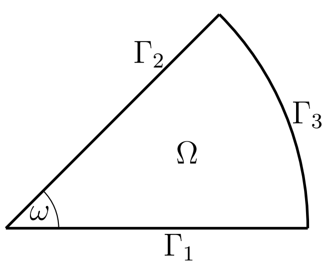
Definition 5.1 Let $H_0^1(\Omega) \subset V \subset H^1(\Omega)$ and $a\colon V \times V \to \mathbb{R}$ an elliptic bilinear form. The variational problem

$$\begin{aligned} &\text{Find } u \in V \text{ such that} \\ &\qquad a(u,v) = \ell(v) \quad \text{for all } v \in V \end{aligned}$$

is called $H^s(\Omega)$-*regular* for $s \geq 2$, iff there exists $c > 0$ such that

$$\|u\|_{H^s(\Omega)} \leq c\|f\|_{H^{s-2}(\Omega)} \quad \text{for all } f \in H^{s-2}(\Omega).$$

Example 5.2



We consider the domain

$$\Omega := \{(r\cos\alpha, r\sin\alpha) : 0 < r < 1,\ 0 < \alpha < \omega\}$$

with boundaries

$$\begin{aligned}\Gamma_1 &:= \{(r,0) : 0 \le r \le 1\},\\ \Gamma_2 &:= \{(r\cos\omega, r\sin\omega) : 0 \le r \le 1\},\\ \Gamma_3 &:= \{(\cos\alpha, \sin\alpha) : 0 < \alpha < \omega\}.\end{aligned}$$

The two dimensional Laplacian in polar coordinates is given by

$$\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} = \frac{1}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2}{\partial \alpha^2},$$

where $x(r,\alpha) = r\cos\alpha$ and $y(r,\alpha) = r\sin\alpha$.
The function

$$u(x,y) = \hat{u}(r,\alpha) = \left(r^2 - r^{\frac{\pi}{\omega}}\right)\sin\left(\frac{\pi}{\omega}\alpha\right)$$

satisfies

$$\Delta u = \left(4 - \frac{\pi^2}{\omega^2}\right)\sin\left(\frac{\pi}{\omega}\alpha\right).$$

Hence, u is the unique solution to the Dirichlet problem

$$-\Delta u = \left(\frac{\pi^2}{\omega^2} - 4\right)\sin\left(\frac{\pi}{\omega}\alpha\right) \text{ in } \Omega, \quad u = 0 \text{ on } \Gamma.$$

The function $r^{\pi/\omega}\sin(\pi\alpha/\omega)$ is in $H^2(\Omega)$, if $\pi/\omega \ge 1$, i.e. $\omega \le \pi$. Consequently, the same accounts for u. On the other hand, the right hand side is always in $L^2(\Omega)$. Hence, the problem is for $\omega > \pi$ not $H^2(\Omega)$-regular. △

Theorem 5.3 Let $\Omega \subset \mathbb{R}^d$ denote a convex and polygonal domain and let

$$a(v,w) := \int_\Omega \langle \boldsymbol{A}\nabla v, \nabla w\rangle \,\mathrm{d}\boldsymbol{x}, \quad v, w \in H_0^1(\Omega)$$

denote an elliptic bilinear form with Lipschitz continuous coefficients $a_{i,j}(\boldsymbol{x})$. Then, the variational problem

$$\begin{aligned}&\text{Find } u \in H_0^1(\Omega) \text{ such that}\\ &\qquad a(u,v) = \ell(v) \quad \text{for all } v \in H_0^1(\Omega)\end{aligned}$$

is $H^2(\Omega)$-regular.

Proof. For a proof of this theorem, see [Grisvard]. □

Based on the preceding regularity result, we can now state the convergence result for the finite element method in the elliptic case.

Theorem 5.4 Let $\Omega \subset \mathbb{R}^d$ denote a convex and polygonal domain and let $\mathcal{T}_h$ denote a uniform triangulation for Ω. Then, if $f \in L^2(\Omega)$, the piecewise linear Galerkin approximation $u_h \in V_h$ to the elliptic variational problem

$$\begin{aligned}&\text{Find } u \in H_0^1(\Omega) \text{ such that}\\ &\qquad a(u,v) = \ell(v) \quad \text{for all } v \in H_0^1(\Omega)\end{aligned}$$

satisfies the error estimate

$$\|u - u_h\|_{H^1(\Omega)} \leq ch\|f\|_{L^2(\Omega)} \quad \text{for some } c > 0.$$

Proof. According to Theorem 5.3, the underlying variational problem is $H^2(\Omega)$-regular. Hence, we have $\|u\|_{H^2(\Omega)} \leq c\|f\|_{L^2(\Omega)}$ for some constant $c > 0$. Consequently, we obtain by Theorem 4.15 that

$$\|u - \mathcal{I}u\|_{H^1(\Omega)} \leq ch|u|_{H^2(\Omega)} \leq ch\|u\|_{H^2(\Omega)} \leq Ch\|f\|_{L^2(\Omega)}$$

for some constants $c, C > 0$. Finally, observing $\mathcal{I}u \in V_h$, Céa's lemma yields

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{c_S}{c_E} \inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leq \frac{c_S}{c_E}\|u - \mathcal{I}u\|_{H^1(\Omega)} \leq \frac{c_S}{c_E} Ch\|f\|_{L^2(\Omega)}.$$

□

Remark In the case that *P2-elements*, i.e. piecewise quadratic finite elements are considered, Theorem 4.15 provides a higher approximation order. This results in an overall higher convergence rate given that the problem at hand is $H^3(\Omega)$-regular. In general, a $H^3(\Omega)$-regular solution is only obtained on a smooth, curved domain, which cannot be decomposed into simplices. △

The rate of convergence of the finite element is increased, if the error is measured with respect to $L^2(\Omega)$.

Theorem 5.5 (Aubin-Nitsche Lemma) Let $\Omega \subset \mathbb{R}^d$ denote a convex and polygonal domain and let $\mathcal{T}_h$ denote a uniform triangulation for Ω. Then, if $f \in L^2(\Omega)$, the piecewise linear Galerkin approximation $u_h \in V_h$ to the elliptic variational problem

$$\begin{aligned}&\text{Find } u \in H_0^1(\Omega) \text{ such that}\\ &\qquad a(u,v) = \ell(v) \quad \text{for all } v \in H_0^1(\Omega)\end{aligned}$$

satisfies the error estimate

$$\|u - u_h\|_{L^2(\Omega)} \leq ch^2\|f\|_{L^2(\Omega)} \quad \text{for some } c > 0.$$

Proof. We consider the *dual problem*

$$\begin{aligned}&\text{Find } \varphi_g \in H_0^1(\Omega) \text{ such that}\\ &\qquad a(w,\varphi_g) = (g,w)_{L^2(\Omega)} \quad \text{for all } w \in H_0^1(\Omega),\end{aligned}$$

which obviously exhibits a unique weak solution due to the Lax-Milgram lemma. It holds for $w := u - u_h$ that

$$(g, u - u_h)_{L^2(\Omega)} = a(u - u_h, \varphi_g) = a(u - u_h, \varphi_g - v_h) \leq c_S \|u - u_h\|_{H^1(\Omega)} \|\varphi_g - v_h\|_{H^1(\Omega)}$$

for all $v_h \in V_h$ due to the Galerkin orthogonality $a(u - u_h, v_h) = 0$.
By the Riesz representation theorem, there holds

$$\|w\|_{L^2(\Omega)} = \sup_{0 \neq g \in L^2(\Omega)} \frac{(g, w)_{L^2(\Omega)}}{\|g\|_{L^2(\Omega)}}.$$

Consequently, we infer

$$\|u - u_h\|_{L^2(\Omega)} = \sup_{0 \neq g \in L^2(\Omega)} \frac{(g, u - u_h)_{L^2(\Omega)}}{\|g\|_{L^2(\Omega)}} \leq c_S \|u - u_h\|_{H^1(\Omega)} \sup_{0 \neq g \in L^2(\Omega)} \frac{\|\varphi_g - v_h\|_{H^1(\Omega)}}{\|g\|_{L^2(\Omega)}}.$$

Since the choice of $v_h \in V_h$ in the previous estimate is arbitrary, employing Theorem 5.4 for $\varphi_g \in H_0^1(\Omega)$ results in

$$\|\varphi_g - \varphi_{g,h}\|_{H^1(\Omega)} \leq ch \|g\|_{L^2(\Omega)}$$

for some constant $c > 0$, where $\varphi_{g,h} \in V_h$ is the Galerkin approximation to φ_g. Inserting this estimate in the preceding one, results in

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch \|u - u_h\|_{H^1(\Omega)}$$

for some constant $C > 0$. The assertion is now obtained by another application of Theorem 5.4. □

For the sake of completeness, we also state an error bound on the finite element approximation with respect to the $L^\infty(\Omega)$.

Theorem 5.6 For the discretisation of an $H^2(\Omega)$-regular, elliptic variational problem by piecewise linear finite elements on a uniform triangulation $\mathcal{T}_h$, there holds the error estimate

$$\|u - u_h\|_{L^\infty(\Omega)} \leq ch^2 |\log h|^{\frac{3}{2}} \max_{|\boldsymbol{\alpha}| \leq 2} \|\partial^{\boldsymbol{\alpha}} u\|_{L^\infty(\Omega)}$$

for some constant $c > 0$.

Proof. For a proof of this theorem, see [Ciarlet]. □

6. Implementation

For the numerical realisation of the finite element method the following steps have to be implemented

1. Mesh generation
2. Assembly of stiffness matrix and right hand side
3. Solution of the linear system
4. A-posteriori error analysis (if the solution is insufficent, mark elements to be refined and go back to 1).
5. Visualisation of the solution and quantities of interest.

Mesh generation: For the generation of the initial triangulation, there exist several possibilities, for example they might manually be provided or by an automatic meshing tool. Based on this initial triangulation, one might either consider uniform or shape regular refinements. Next, we introduce an algorithm that is suitable for adaptive mesh refinement. It starts from all elements being colored "red".

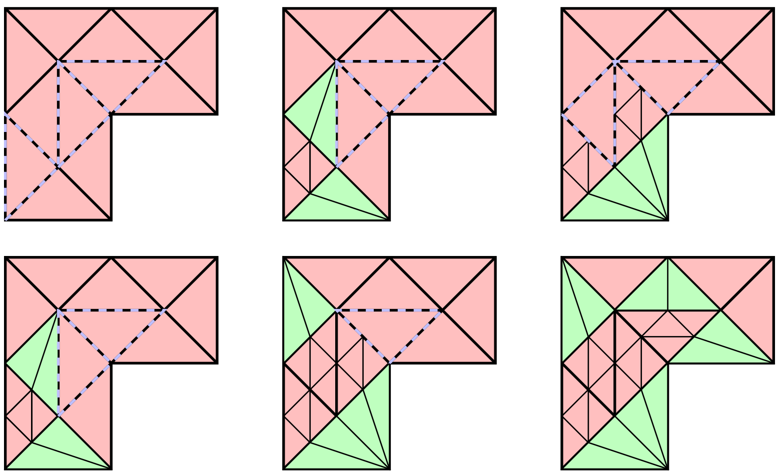
Algorithm 6.1 (Red-green refinement)

input: admissible partition and marked elements
output: admissible, refined partition
while there are marked elements **do**

① choose a marked element and refine it uniformly

② color all "red" elements with a hanging node "green" and bisect them

③ color all "green" elements with a hanging node "red", undo the bisection and mark them



Note that in this procedure, bisected "green" elements are still considered as a single element. The same procedure also works in three spatial dimensions but is more involved. Finally, we remark that a shape-regular partition stays shape-regular since all angles are at most bisected.

Assembly: Given an affine family $(T, \Pi_T, \Sigma_T)_{T\in\mathcal{T}_h}$ of P1- or P2-elements, it is convenient to assemble the stiffness matrix with respect to the nodal basis $\{\hat{\varphi}_i\}_{i=1}^s$ for $\Pi_{\widehat{T}}$. Then, obviously $\varphi_i := \hat{\varphi}_i \circ \Phi_T^{-1}$ for $i = 1, \ldots, s$ is the nodal basis for Π_T, where

$$\Phi_T \colon \widehat{T} \to T, \quad \hat{\boldsymbol{x}} \mapsto \boldsymbol{B}_T\hat{\boldsymbol{x}} + \boldsymbol{z}_{0,T}.$$

In particular, we can also introduce a nodal basis $\{\psi_i\}_{i=1}^N$ for V_h with respect to the nodes $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of the mesh such that $\psi_i(\boldsymbol{x}_j) = \delta_{ij}$. In particular, we have $\psi_i|_T = \varphi_j$ for some $1 \le j \le s$, Now, we may consider an element based assembly. Let

$$a(\psi_j, \psi_i) = \int_\Omega \langle \boldsymbol{A}\nabla\psi_j, \nabla\psi_i\rangle \,\mathrm{d}\boldsymbol{x}.$$

It holds

$$\begin{aligned} a(\psi_j, \psi_i) &= \sum_{T\in\mathcal{T}_h} \int_T \langle \boldsymbol{A}\nabla\psi_j, \nabla\psi_i\rangle \,\mathrm{d}\boldsymbol{x} \\ &= \sum_{T\in\mathcal{T}_h} \int_{\widehat{T}} \langle (\boldsymbol{A}\circ\Phi_T)\boldsymbol{B}_T^{-\intercal}\hat{\nabla}\hat{\psi}_j, \boldsymbol{B}_T^{-\intercal}\hat{\nabla}\hat{\psi}_i\rangle |\det \boldsymbol{B}_T| \,\mathrm{d}\hat{\boldsymbol{x}} =: \sum_{T\in\mathcal{T}_h} a_T(\hat{\psi}_j, \hat{\psi}_i). \end{aligned}$$

From this, we derive the element stiffness matrix for $T \in \mathcal{T}_h$, which is given by

$$\boldsymbol{A}_T := \begin{bmatrix} a_T(\hat{\varphi}_1, \hat{\varphi}_1) & \cdots & a_T(\hat{\varphi}_s, \hat{\varphi}_1) \\ \vdots & \ddots & \vdots \\ a_T(\hat{\varphi}_1, \hat{\varphi}_s) & \cdots & a_T(\hat{\varphi}_s, \hat{\varphi}_s) \end{bmatrix}. \tag{6.1}$$

In order to compute the element stiffness matrix (6.1) numerically, one usually employs numerical quadrature. Common quadrature rules for $d = 2$ are denoted in the subsequent table.

quadrature points	weights	exactness
$[1/3, 1/3]$	$1/2$	$\mathcal{P}_1$
$[1/2, 1/2], [1/2, 0], [0, 1/2]$	$1/6$	$\mathcal{P}_2$
$[1/3, 1/3]$	$-27/96$	
$[1/5, 1/5], [1/5, 3/5], [3/5, 1/5]$	$25/96$	$\mathcal{P}_3$
$[1/3, 1/3]$	$9/80$	
$\left[\frac{6+\sqrt{15}}{21}, \frac{6+\sqrt{15}}{21}\right], \left[\frac{9-2\sqrt{15}}{21}, \frac{6+\sqrt{15}}{21}\right], \left[\frac{6+\sqrt{15}}{21}, \frac{9-2\sqrt{15}}{21}\right]$	$\frac{155+\sqrt{15}}{2400}$	
$\left[\frac{6-\sqrt{15}}{21}, \frac{6-\sqrt{15}}{21}\right], \left[\frac{9+2\sqrt{15}}{21}, \frac{6-\sqrt{15}}{21}\right], \left[\frac{6-\sqrt{15}}{21}, \frac{9+2\sqrt{15}}{21}\right]$	$\frac{155-\sqrt{15}}{2400}$	$\mathcal{P}_5$

A very elegant way to assemble the right hand side is based on the finite element mass matrix

$$\boldsymbol{M} := [(\psi_i, \psi_j)_{L^2(\Omega)}]_{i,j=1}^N,$$

which can be assembled in a similar fashion as the stiffness matrix. Then, assuming

$$f(\boldsymbol{x}) \approx \sum_{i=1}^{N} f(\boldsymbol{x}_i)\psi_i(\boldsymbol{x})$$

yields

$$\ell(\psi_j) = \int_\Omega f\psi_i \,\mathrm{d}\boldsymbol{x} \approx \sum_{i=1}^{N} f(\boldsymbol{x}_i) \int_\Omega \psi_i\psi_j \,\mathrm{d}\boldsymbol{x}.$$

Consequently, we obtain

$$\begin{bmatrix} \ell(\psi_1) \\ \vdots \\ \ell(\psi_N) \end{bmatrix} \approx \boldsymbol{M} \begin{bmatrix} f(\boldsymbol{x}_1) \\ \vdots \\ f(\boldsymbol{x}_N) \end{bmatrix}.$$

On the other hand, if f is only in $L^2(\Omega)$ and, hence, does not allow for pointwise evaluations, the right hand side has to be assembled in an element based fashion, as before.

Dirichlet boundary conditions: The numerical treatment of homogenous Dirichlet problems is straightforward for $V_h \subset H_0^1(\Omega)$. For the case of Dirichlet boundary conditions $g \not\equiv 0$, we may also consider the interpolation of the Dirichlet data in the boundary nodes:

$$g_h(\boldsymbol{x}) := \sum_{i=1}^{N_D} g(\boldsymbol{x}_i^D)\varphi_i^D(\boldsymbol{x})\big|_\Gamma \approx g \quad \text{on } \Gamma.$$

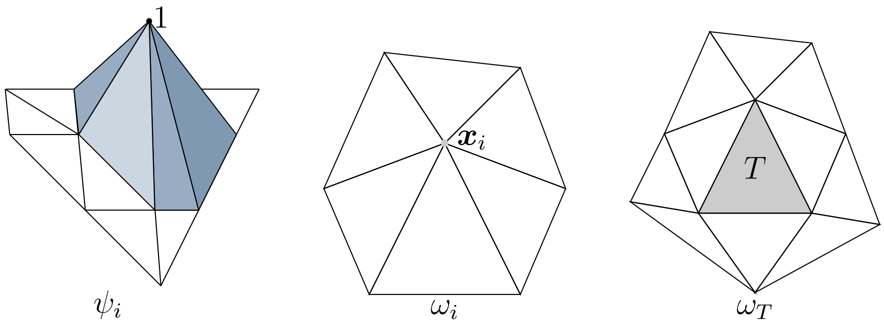
Obviously, this function satisfies $g_h \in H^1(\Omega)$. Hence, it suffices to find $u_h \in H_0^1(\Omega)$ such that

$$a(u_h, v_h) = \ell(v_h) - a(g_h, v_h) \quad \text{for all } v_h \in V_h.$$

However, we remark that the extension g_h of the boundary values is not uniformly stable in $H^1(\Omega)$ with respect to the mesh width $h > 0$. In the case of a non-linear partial differential equation, this effects the rate of convergence for numerical solution methods.

Remaining steps: The numerical solution of the emerging linear system can either be performed by a sparse direct solver or an iterative solver. In particular, the iterative multigrid method is of linear cost in terms of degrees of freedom, independently on the mesh size $h > 0$. In the sequel, we shall consider suitable adaptive error estimators.

7. Error Control and Adaptivity



The interpolation operator in Theorem 4.15 requires functions in $H^2(\Omega)$. In the sequel, we consider an approximation method that also works for $H^1(\Omega)$ functions. It is based on an idea von Philippe Clément. To that end, let $\mathcal{T}_h$ denote a shape-regular triangulation of $\Omega \subset \mathbb{R}^2$. For each node $\boldsymbol{x}_i$, we define the union of the adjacent elements

$$\omega_i := \operatorname{supp} \psi_i,$$

where $\{\psi_i\}_i$ are the nodal basis functions of finite element space based on the P1-element. Accordingly, we define the patch ω_T to be the union of all elements that share a lower dimensional face with T, i.e.

$$\omega_T := \bigcup_{\boldsymbol{x}_i \in T} \omega_i.$$

For shape-regular triangulations, there obviously holds

$$|\omega_T| \leq C|T| \leq C h_T^2,$$

where the constant $C > 0$ depends on κ.

Definition 7.1 The *Clément operator* is given according to

$$C_h \colon H^1(\Omega) \to V_h, (C_h v)(\boldsymbol{x}) := \sum_{i=1}^{N} (Q_i v) \psi_i(\boldsymbol{x}),$$

where $Q_i \colon L^2(\omega_i) \to \mathcal{P}_0(\omega_i)$ is the L^2-projection, i.e. $(v - Q_i v, 1)_{L^2(\omega_i)} = 0$.

In particular, there holds

$$\|v - Q_i v\|_{L^2(\omega_i)} \leq C \operatorname{diam}(\omega_i) \|v\|_{H^1(\omega_i)} \quad \text{for some } C > 0,$$

which can be easily derived as in the proof of the Poincaré inequality.

Theorem 7.2 (Clément) Let $\mathcal{T}_h$ denote a shape-regular triangulation of $\Omega \subset \mathbb{R}^2$. Then, the Clément operator C_h is well defined, linear and has the following properties: There exists $C > 0$ which depends on $\mathcal{T}_h$ such that for $k = 0, 1$ there holds

$$\begin{aligned} \|v - C_h v\|_{H^k(T)} &\leq C h_T^{1-k} \|v\|_{H^1(\omega_T)}, \\ \|v - C_h v\|_{L^2(e)} &\leq C h_T^{1/2} \|v\|_{H^1(\omega_T)} \end{aligned}$$

for all $T \in \mathcal{T}_h$, $v \in H^1(\Omega)$, where e denotes an arbitrary edge of T.

Proof. For a proof of this theorem, see [P. Clément. Approximation by finite element functions using local regularization]. □

In the sequel, we restrict ourselves to the homogenous Dirichlet problem

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \Gamma.$$

If we insert the Galerkin solution $u_h \in V_h$ into this equation, we obtain a residual. Moreover, the derivatives at the element boundaries of u_h exhibit jumps. Hence, we consider the area based residuals

$$R_T := R_T(u_h) := f + \Delta u_h \quad \text{for } T \in \mathcal{T}_h$$

and edge based jumps

$$R_e := R_e(u_h) := \left[\!\left[\frac{\partial u_h}{\partial \boldsymbol{n}} \right]\!\right] := \frac{\partial u_h}{\partial \boldsymbol{n}}\bigg|_{T_1} - \frac{\partial u_h}{\partial \boldsymbol{n}}\bigg|_{T_2} \quad \text{for } e = T_1 \cap T_2 \in \mathcal{E},$$

where $\mathcal{E}$ denotes the set of interior edges in $\mathcal{T}_h$. Moreover, we will also refer to the set of edges of T by $\mathcal{E}_T$. Based on these quantities, we define the local error estimators

$$\eta_T^2 := h_T^2 \|R_T\|_{L^2(T)}^2 + \frac{1}{2} \sum_{e \in \mathcal{E}_T} h_e \|R_e\|_{L^2(e)}^2.$$

From these, we obtain the global error estimator

$$\eta^2 := \sum_{T \in \mathcal{T}_h} \eta_T^2 = \sum_{T \in \mathcal{T}_h} h_T^2 \|R_T\|_{L^2(T)}^2 + \sum_{e \in \mathcal{E}} h_e \|R_e\|_{L^2(e)}^2.$$

Theorem 7.3 Let $\mathcal{T}_h$ denote a shape-regular triangulation. Then, there exists $c > 0$ depending on Ω and κ, such that

$$\|u - u_h\|_{H^1(\Omega)} \leq c\eta.$$

Proof. For $v \in H_0^1(\Omega)$, there holds by the divergence theorem

$$\begin{aligned}
\ell(v) &:= \big(\nabla(u-u_h), \nabla v\big)_{L^2(\Omega)} \\
&= (f,v)_{L^2(\Omega)} - \sum_{T\in\mathcal{T}_h} (\nabla u_h, \nabla v)_{L^2(\Omega)} \\
&= (f,v)_{L^2(\Omega)} - \sum_{T\in\mathcal{T}_h} \Big[-(\Delta u_h, v)_{L^2(\Omega)} + \sum_{e\in\mathcal{E}_T} \Big(\frac{\partial u_h}{\partial \boldsymbol{n}}, v\Big)_{L^2(e)} \Big] \\
&= \sum_{T\in\mathcal{T}_h} (\Delta u_h + f, v)_{L^2(T)} + \sum_{e\in\mathcal{E}} \Big(\Big[\!\Big[\frac{\partial u_h}{\partial \boldsymbol{n}} \Big]\!\Big], v \Big)_{L^2(e)} \\
&= \sum_{T\in\mathcal{T}_h} \Big[(R_T, v)_{L^2(T)} + \frac{1}{2} \sum_{e\in\mathcal{E}_T} (R_e, v)_{L^2(e)} \Big].
\end{aligned}$$

Now, let $v_h := C_h v$. Due to Galerkin orthogonality, it holds

$$\big(\nabla(u-u_h), \nabla v_h\big)_{L^2(\Omega)} = 0 \quad \text{for all } v_h \in V_h$$

and hence

$$\ell(v) = \ell(v - v_h) \le \sum_{T\in\mathcal{T}_h} \Big[\|R_T\|_{L^2(T)} \|v - v_h\|_{L^2(T)} \frac{1}{2} \sum_{e\in\mathcal{E}_T} \|R_e\|_{L^2(e)} \|v - v_h\|_{L^2(e)} \Big].$$

Since $\cup_{T\in\mathcal{T}_h} \omega_T$ is only a finite covering of Ω, Theorem 7.2 yields

$$\begin{aligned}
\ell(v) &\le c \sum_{T\in\mathcal{T}_h} \Big[h_T \|R_T\|_{L^2(T)} + \frac{1}{2} \sum_{e\in\mathcal{E}_T} h^{\frac{1}{2}} \|R_e\|_{L^2(e)} \Big] \|v\|_{H^1(\omega_T)} \\
&\le c \sum_{T\in\mathcal{T}_h} \eta_T \|v\|_{H^1(\omega_T)} \le c \Big(\sum_{T\in\mathcal{T}_h} \eta_T^2 \Big)^{\frac{1}{2}} \Big(\sum_{T\in\mathcal{T}_h} \|v\|^2_{H^1(\omega_T)} \Big)^{\frac{1}{2}} \le C\eta \|v\|_{H^1(\Omega)}
\end{aligned}$$

for some constants $c, C > 0$. The proof is completed the duality argument

$$|u - u_h|_{H^1(\Omega)} = \sup_{v\in H_0^1(\Omega), \|v\|_{H_0^1(\Omega)}=1} \big(\nabla(u-u_h), \nabla v\big)_{L^2(\Omega)} = \sup_{v\in H_0^1(\Omega), \|v\|_{H_0^1(\Omega)}=1} \ell(v).$$

□

Remark In practice, the area based residual is hard to compute if f cannot be integrated exactly. Hence, one applies the splitting $f = f_h + f - f_h$, such that $\Delta u_h + f_h$ can be calculated exactly. In this case, one ends up with the bound

$$\|u - u_h\|_{H^1(\Omega)} \le c \Big[\eta + \Big(\sum_{T\in\mathcal{T}_h} h_T^2 \|f - f_h\|^2_{L^2(T)} \Big)^{\frac{1}{2}} \Big].$$

In this estimate, the newly added term is called *data oscillation.*
In particular, one can show the lower bound

$$\eta^2 \le c' \Big[\|u - u_h\|^2_{H^1(\Omega)} + \sum_{T\in\mathcal{T}_h} h_T^2 \|f - f_h\|^2_{L^2(T)} \Big]$$

for some $c' > 0$. Therefore, the error estimator η is equivalent to the true error up to the data oscillation and the error is actually localised in the estimators η_T. In practice, one computes η_T for all elements and marks a certain fraction of the elements that provide the largest error contributions for refinement. △

8. The Heat Equation

In this section, we consider the heat equation

$$\frac{\partial u}{\partial t}(t,\boldsymbol{x}) - \Delta u(t,\boldsymbol{x}) = f(t,\boldsymbol{x}), \quad (t,\boldsymbol{x}) \in [0,T] \times \Omega \tag{8.1}$$

for a timepoint $T > 0$ and a domain $\Omega \subset \mathbb{R}^d$ as an example for parabolic equations. For the sake of simplicity, we restrict ourselves to homogenous Dirichlet conditions, i.e.

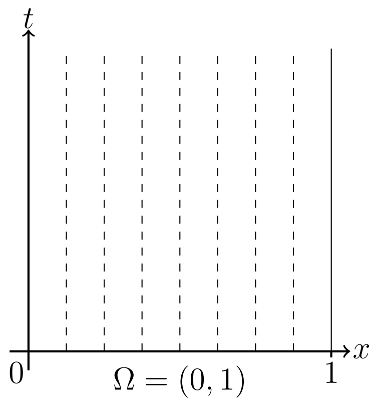
$$u(t,\boldsymbol{x}) = 0 \quad \text{for all } [0,T] \times \Gamma.$$

This corresponds to fixing the temperature for all times to zero degrees at the boundary. To end up with a well posed problem, we also prescribe the initial temperature distribution

$$u(0,\boldsymbol{x}) = g(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \Omega.$$

In the stationary case $\partial u/\partial t \equiv 0$, equation (8.1) reduces to the well known Poisson's equation.

8.1 The Method of Lines



In Order to compute a solution to the heat equation, we first apply a semi-discretisation with respect to the spatial variable. To that end, we consider the variational formulation

Find $u(t) \in H_0^1(\Omega)$ such that
$$\frac{\partial}{\partial t}\big(u(t),v\big)_{L^2(\Omega)} + \big(\nabla u(t),\nabla v\big)_{L^2(\Omega)} = \big(f(t),v\big)_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega).$$

The restriction of the variational formulation to the finite element space $V_h \subset H_0^1(\Omega)$ hence reads

Find $u_h(t) \in V_h$ such that

$$\frac{\partial}{\partial t}\big(u_h(t), v_h\big)_{L^2(\Omega)} + \big(\nabla u_h(t), \nabla v_h\big)_{L^2(\Omega)} = \big(f(t), v_h\big)_{L^2(\Omega)} \quad \text{for all } v_h \in V_h.$$

This translates to the linear system

$$\boldsymbol{M}\frac{\partial}{\partial t}\boldsymbol{u}(t) + \boldsymbol{A}\boldsymbol{u}(t) = \boldsymbol{f}(t), \quad \boldsymbol{u}(0) = \boldsymbol{g}. \tag{8.2}$$

Herein,

$$\boldsymbol{M} = [(\psi_i, \psi_j)_{L^2(\Omega)}]_{i,j}$$

denotes the finite element mass matrix, while

$$\boldsymbol{A} = [(\nabla\psi_j, \nabla\psi_i)_{L^2(\Omega)}]_{i,j}$$

is the finite element stiffness matrix. Note that these matrices are independent of the time variable t. This is in contrast to the coefficients of the discrete solution $\boldsymbol{u}(t) = [u_i(t)]_i$ and the coefficients of the right hand side $\boldsymbol{f}(t) = \big[\big(f(t), \varphi_i\big)_{L^2(\Omega)}\big]_i$, which both depend on t.
The finite element solution is of the form

$$u_h(t, \boldsymbol{x}) = \sum_i u_i(t)\psi_i(\boldsymbol{x}).$$

In view of the vector valued function $t \mapsto \boldsymbol{u}(t)$, the semi-discretisation (8.2) is referred to as *method of lines*. For each $t \geq 0$, we obtain a vector that represents the function $u_h(t, \boldsymbol{x})$ with respect to the triangulation $\mathcal{T}_h$.

8.2 The θ-Scheme

Next, we introduce an appropriate discretisation for the time. To that and, we consider a subdivision of $[0, T]$ into M sub-intervals $[t_i, t_{i+1}]$, $i = 0, \ldots, M-1$ and set $k_i := t_{i+1} - t_i$ to the length of each sub-interval.
Then, the *forward Euler method* yields

$$\boldsymbol{M}\frac{\boldsymbol{u}_{i+1} - \boldsymbol{u}_i}{k_i} + \boldsymbol{A}\boldsymbol{u}_i = \boldsymbol{f}(t_i) \quad \text{or} \quad \boldsymbol{M}\boldsymbol{u}_{i+1} = (\boldsymbol{M} - k_i\boldsymbol{A})\boldsymbol{u}_i + k_i\boldsymbol{f}(t_i), \tag{8.3}$$

while the *backward Euler method* results in

$$\boldsymbol{M}\frac{\boldsymbol{u}_{i+1} - \boldsymbol{u}_i}{k_i} + \boldsymbol{A}\boldsymbol{u}_{i+1} = \boldsymbol{f}(t_{i+1}) \quad \text{or} \quad (\boldsymbol{M} + k_i\boldsymbol{A})\boldsymbol{u}_{i+1} = \boldsymbol{M}\boldsymbol{u}_i + k_i\boldsymbol{f}(t_{i+1}). \tag{8.4}$$

In both cases, the initial value is given by $\boldsymbol{u}_0 = \boldsymbol{g}$.
The idea of the θ-*scheme* is now to combine the two equations (8.3) and (8.4) with respect to the parameter $\theta \in [0, 1]$. We obtain

$$\boldsymbol{M}\frac{\boldsymbol{u}_{i+1} - \boldsymbol{u}_i}{k_i} + (1-\theta)\boldsymbol{A}\boldsymbol{u}_i + \theta\boldsymbol{A}\boldsymbol{u}_{i+1} = (1-\theta)\boldsymbol{f}(t_i) + \theta\boldsymbol{f}(t_{i+1}) \tag{8.5}$$

or

$$(\boldsymbol{M} + k_i\theta \boldsymbol{A})\boldsymbol{u}_{i+1} = \big(\boldsymbol{M} - k_i(1-\theta)\boldsymbol{A}\big)\boldsymbol{u}_i + k_i(1-\theta)\boldsymbol{f}(t_i) + k_i\theta \boldsymbol{f}(t_{i+1}),$$

respectively. In particular, it holds

$$\theta = \begin{cases} 0, & \text{forward Euler method} \\ 1/2, & \text{trapezoidal rule} \\ 1, & \text{backward Euler method} \end{cases}$$

In the context of parabolic equations, the trapezoidal rule is also called *Crank-Nicolson method*. In view of the local truncation error, the θ-scheme is consistent of order one, for $\theta = 0.5$, it is even consistent of order two.

Theorem 8.1 The θ-scheme for the heat equation is stable for $1/2 \leq \theta \leq 1$, i.e.

$$\begin{aligned} \|u_{h,M}\|_{L^2(\Omega)}^2 + \sum_{i=0}^{M-1} \Big[k_i|u_{h,i+\theta}|_{H^1(\Omega)}^2 + (2\theta-1)\|u_{h,i+1} - u_{h,i}\|_{L^2(\Omega)}^2\Big] \\ \leq \|u_{h,0}\|_{L^2(\Omega)}^2 + c\sum_{i=0}^{M-1} k_i\|f_{i+\theta}\|_{L^2(\Omega)}^2, \end{aligned}$$

where we set

$$u_{h,i+\theta} := (1-\theta)u_{h,i} + \theta u_{h,i+1}, \qquad f_{i+\theta} := (1-\theta)f(t_i) + \theta f(t_{i+1}).$$

Proof. In view of (8.5), inserting $u_{h,i+\theta}$ as a test function into the corresponding variational formulation yields

$$(u_{h,i+1} - u_{h,i}, u_{h,i+\theta})_{L^2(\Omega)} + k_i(\nabla u_{h,i+\theta}, \nabla u_{h,i+\theta})_{L^2(\Omega)} = k_i(f_{i+\theta}, u_{h,i+\theta})_{L^2(\Omega)}. \tag{8.6}$$

On the other hand, there holds

$$\begin{aligned} &(u_{h,i+1} - u_{h,i}, u_{h,i+\theta})_{L^2(\Omega)} \\ &\quad = \big(u_{h,i+1} - u_{h,i}, (1-\theta)u_{h,i} + \theta u_{h,i+1}\big)_{L^2(\Omega)} \\ &\quad = \left(u_{h,i+1} - u_{h,i}, \frac{1}{2}u_{h,i+1} + \frac{1}{2}u_{h,i} + \left(\theta - \frac{1}{2}\right)(u_{h,i+1} - u_{h,i})\right)_{L^2(\Omega)} \\ &\quad = \frac{1}{2}\|u_{h,i+1}\|_{L^2(\Omega)}^2 - \frac{1}{2}\|u_{h,i}\|_{L^2(\Omega)}^2 + \left(\theta - \frac{1}{2}\right)\|u_{h,i+1} - u_{h,i}\|_{L^2(\Omega)}^2. \end{aligned}$$

The combination of (8.6), the Cauchy-Schwarz inequality and the Poincaré inequality yields

$$\begin{aligned} &\|u_{h,i+1}\|_{L^2(\Omega)}^2 - \|u_{h,i}\|_{L^2(\Omega)}^2 + (2\theta-1)\|u_{h,i+1} - u_{h,i}\|_{L^2(\Omega)}^2 + 2k_i|u_{h,i+\theta}|_{H^1(\Omega)}^2 \\ &\quad = 2k_i(f_{i+\theta}, u_{h,i+\theta})_{L^2(\Omega)} \\ &\quad \leq 2k_i c\|f_{i+\theta}\|_{L^2(\Omega)}|u_{h,i+\theta}|_{H^1(\Omega)} \\ &\quad \leq k_i c^2\|f_{i+\theta}\|_{L^2(\Omega)}^2 + k_i|u_{h,i+\theta}|_{H^1(\Omega)}^2 \end{aligned}$$

for some constant $c > 0$. From this, we infer

$$\|u_{h,i+1}\|^2_{L^2(\Omega)} - \|u_{h,i}\|^2_{L^2(\Omega)} + (2\theta - 1)\|u_{h,i+1} - u_{h,i}\|^2_{L^2(\Omega)} \\ + k_i|u_{h,i+\theta}|^2_{H^1(\Omega)} \leq k_i c^2 \|f_{i+\theta}\|^2_{L^2(\Omega)}.$$

Now, summation with respect to i yields the assertion. □

Consequently, the θ-scheme is stable for all $\theta \in [0.5, 1]$. This means that errors will not be amplified exponentially. For $\theta < 0.5$, the method is only stable if $k_i \sim h^2$. This is the so called *CFL-condition* for parabolic problems. Herein, CFL stands for Courant, Friedrichs and Lewy. Given that $\theta > 0.5$, high frequency contributions to the solution are exponentially dampened. Accordingly, local perturbations in the data $\boldsymbol{f}(t_i)$ and $\boldsymbol{g}$ are substantially reduced.

Remark The Crank-Nicolson method is the simplest second order method. Hence, it is rather popular. However, since errors do not get dampened exponentially, it might produce unphysical oscillations. Hence, it is preferable to choose $\theta = 1/2 + \varepsilon$ for a small quantity $\varepsilon > 0$. △

8.3 Error Analysis

In the sequel, we consider the error analysis for the backward Euler method, i.e. $\theta = 1$. For the sake of simplicity, we slightly modify the right hand side according to

$$\overline{f}(t_{i+1}) := \frac{1}{k_i}\int_{t_i}^{t_{i+1}} f(t)\,\mathrm{d}t = f(t_{i+1}) + \mathcal{O}(k_i)$$

In order to estimate the error, we introduce the discrete semi-norm

$$\|u\|_{h,\infty} := \max_{i=1}^{M} \|u(t_i)\|_{L^2(\Omega)}.$$

For functions $u_h \in V_h \times \{t_1, \ldots, t_M\}$, which are discrete with respect to time, this is even a norm.

Theorem 8.2 Let Ω denote a convex, polygonal domain and $\mathcal{T}_h$ a shape-regular mesh, which is fixed with respect to time.
Assume that the continuous solution to the heat equation satisfies $u \in H^1(0,T) \otimes H^2(\Omega)$. Then the backward Euler method together with a piecewise linear finite element discretisation satisfies the error estimate

$$\|u - u_h\|_{h,\infty} \leq c\left[\sqrt{T}h^2 k^{-1/2}\|\Delta u\|_{h,\infty} + \left(\sum_{i=0}^{M-1} k_i^2 \int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t\right)^{1/2}\right],$$

given that $\min_{i=1}^{M}\{k_i\} \geq k$.

Proof. Let $P_h u_i$ denote the Galerkin projection of $u_i := u(t_i)$ onto V_h. It is given via the Galerkin method accoring to

$$\begin{aligned}&\text{Find } P_h u_i \in V_h \text{ such that}\\ &\qquad (\nabla P_h u_i, \nabla v_h)_{L^2(\Omega)} = (\nabla u_i, \nabla v_h)_{L^2(\Omega)} \quad \text{for all } v_h \in V_h.\end{aligned}$$

We split the error $u_i - u_{h,i}$ into $\xi_i = (P_h - I)u_i$ and $\eta_i = u_{h,i} - P_h u_i \in V_h$. According to Theorem 5.5, see also Theorem 5.4, with respect to ξ_i there holds the error estimate

$$\|\xi_i\|_{L^2(\Omega)} = \|(P_h - I)u_i\|_{L^2(\Omega)} \leq ch^2 |u_i|_{H_2(\Omega)}. \tag{8.7}$$

From this, we derive

$$\|(P_h - I)u\|_{h,\infty} \leq ch^2 \|\Delta u\|_{h,\infty}.$$

For the error contribution η_{i+1}, we employ the identity

$$\|\eta_{i+1}\|^2_{L^2(\Omega)} - \|\eta_i\|^2_{L^2(\Omega)} = 2(\eta_{i+1} - \eta_i, \eta_{i+1})_{L^2(\Omega)} - \|\eta_{i+1} - \eta_i\|^2_{L^2(\Omega)}. \tag{8.8}$$

We compute a bound for the term $(\eta_{i+1} - \eta_i, \eta_{i+1})_{L^2(\Omega)}$. Remind that the Galerkin formulation for the backward Euler method reads at time t_{i+1}

$$\begin{aligned}&\text{Find } u_{h,i+1} \in V_h \text{ such that}\\ &\qquad (u_{h,i+1} - u_{h,i}, v_h)_{L^2(\Omega)} + k_i(\nabla u_{h,i+1}, \nabla v_h)_{L^2(\Omega)} = k_i(f_{i+1}, v_h)_{L^2(\Omega)} \quad \text{for all } v_h \in V_h.\end{aligned}$$

Hence, it holds due to $\eta_{i+1} \in V_h$ and the definition of the Galerkin projection that

$$\begin{aligned}
&(\eta_{i+1} - \eta_i, \eta_{i+1})_{L^2(\Omega)}\\
&\quad = k_i(f_{i+1}, \eta_{i+1})_{L^2(\Omega)} - k_i(\nabla u_{h,i+1}, \nabla \eta_{i+1})_{L^2(\Omega)} - (P_h u_{i+1} - P_h u_i, \eta_{i+1})_{L^2(\Omega)}\\
&\quad = k_i(f_{i+1}, \eta_{i+1})_{L^2(\Omega)} - k_i\big(\nabla(\eta_{i+1} + u_{i+1}), \nabla \eta_{i+1}\big)_{L^2(\Omega)} - (u_{i+1} - u_i, \eta_{i+1})_{L^2(\Omega)}\\
&\qquad - (P_h u_{i+1} - u_{i+1} - P_h u_i + u_i, \eta_{i+1})_{L^2(\Omega)}\\
&\quad = k_i(f_{i+1}, \eta_{i+1})_{L^2(\Omega)} - k_i\big(\nabla(\eta_{i+1} + u_{i+1}), \nabla \eta_{i+1}\big)_{L^2(\Omega)} - (u_{i+1} - u_i, \eta_{i+1})_{L^2(\Omega)}\\
&\qquad - (\xi_{i+1} - \xi_i, \eta_{i+1})_{L^2(\Omega)}.
\end{aligned}$$

The first three terms on the right hand side of the previous identity can be bounded by using the heat equation according to

$$\begin{aligned}
E_1 &:= k_i(f_{i+1}, \eta_{i+1})_{L^2(\Omega)} - k_i\big(\nabla(\eta_{i+1} + u_{i+1}), \nabla \eta_{i+1}\big)_{L^2(\Omega)} - (u_{i+1} - u_i, \eta_{i+1})_{L^2(\Omega)}\\
&= \int_{t_i}^{t_{i+1}} \left(f - \frac{\partial u}{\partial t}, \eta_{i+1}\right)_{L^2(\Omega)} \mathrm{d}t - k_i(\nabla u_{i+1}, \nabla \eta_{i+1})_{L^2(\Omega)} - k_i \|\nabla \eta_{i+1}\|^2_{L^2(\Omega)}\\
&= \int_{t_i}^{t_{i+1}} \big(\nabla(u - u_{i+1}), \nabla \eta_{i+1}\big)_{L^2(\Omega)} \mathrm{d}t - k_i \|\nabla \eta_{i+1}\|^2_{L^2(\Omega)}.
\end{aligned}$$

Now, the fundamental theorem of calculus yields for a differentiable function g that

$$\int_x^y \{g(z) - g(x)\}\,\mathrm{d}z = \int_x^y \int_x^z g'(t)\,\mathrm{d}t\,\mathrm{d}z.$$

Since

$$\{(t,z) : x \leq z \leq y,\ x \leq t \leq z\} = \{(t,z) : x \leq t \leq y,\ t \leq z \leq y\},$$

we end up with the identity

$$\int_x^y \{g(z)-g(x)\}\,\mathrm{d}z = \int_x^y \int_t^y g'(t)\,\mathrm{d}z\,\mathrm{d}t = \int_x^y g'(t)(y-t)\,\mathrm{d}t.$$

Inserting this into the representation of E_1 gives

$$\begin{aligned}
E_1 &= \int_{t_i}^{t_{i+1}} (t_i - t)\left(\nabla\frac{\partial u}{\partial t}, \nabla\eta_{i+1}\right)_{L^2(\Omega)} \mathrm{d}t - k_i\|\nabla\eta_{i+1}\|^2_{L^2(\Omega)} \\
&\le \int_{t_i}^{t_{i+1}} (t-t_i)\left|\frac{\partial u}{\partial t}\right|_{H^1(\Omega)} \|\nabla\eta_{i+1}\|_{L^2(\Omega)}\,\mathrm{d}t - k_i\|\nabla\eta_{i+1}\|^2_{L^2(\Omega)} \\
&\le \left(\frac{k_i^2}{2}\int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t\right)^{1/2} \left(2\int_{t_i}^{t_{i+1}} \|\nabla\eta_{i+1}\|^2_{L^2(\Omega)}\,\mathrm{d}t\right)^{1/2} - k_i\|\nabla\eta_{i+1}\|^2_{L^2(\Omega)}.
\end{aligned}$$

Now, estimating the geometric average by the arithmetic one, i.e. $\sqrt{ab} \le (a+b)/2$ for all $a, b > 0$, we arrive at

$$E_1 \le \frac{k_i^2}{4}\int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t.$$

Additionally, it holds

$$\begin{aligned}
E_2 &:= (\xi_{i+1} - \xi_i, \eta_{i+1})_{L^2(\Omega)} \\
&= (\xi_{i+1}, \eta_{i+1})_{L^2(\Omega)} - (\xi_i, \eta_i)_{L^2(\Omega)} + (\xi_i, \eta_i - \eta_{i+1})_{L^2(\Omega)} \\
&\le (\xi_{i+1}, \eta_{i+1})_{L^2(\Omega)} - (\xi_i, \eta_i)_{L^2(\Omega)} + \|\eta_i - \eta_{i+1}\|_{L^2(\Omega)}\|\xi_i\|_{L^2(\Omega)} \\
&\le (\xi_{i+1}, \eta_{i+1})_{L^2(\Omega)} - (\xi_i, \eta_i)_{L^2(\Omega)} + \frac{1}{2}\|\eta_i - \eta_{i+1}\|^2_{L^2(\Omega)} + ch^4|u_i|^2_{H^2(\Omega)},
\end{aligned}$$

where we employed (8.7) end the estimate on the geometric and arithmetic average in the last step.

Inserting the equality $(\eta_{i+1} - \eta_i, \eta_{i+1})_{L^2(\Omega)} = E_1 + E_2$ into (8.8) yields

$$\begin{aligned}
\|\eta_{i+1}\|^2_{L^2(\Omega)} - \|\eta_i\|^2_{L^2(\Omega)} &= 2E_1 + 2E_2 - \|\eta_{i+1} - \eta_i\|^2_{L^2(\Omega)} \\
&\le 2(\xi_{i+1}, \eta_{i+1})_{L^2(\Omega)} - 2(\xi_i, \eta_i)_{L^2(\Omega)} + \frac{k_i^2}{2}\int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t + ch^4|u_i|^2_{H^2(\Omega)}.
\end{aligned}$$

Now, summation with respect to i results in

$$\begin{aligned}
\|\eta_M\|^2_{L^2(\Omega)} \le \|\eta_0\|^2_{L^2(\Omega)} &+ 2\underbrace{(\xi_M, \eta_M)_{L^2(\Omega)}}_{\le(\sqrt{2}\|\xi_M\|_{L^2(\Omega)})(\|\eta_M\|_{L^2(\Omega)}/\sqrt{2})} - 2\underbrace{(\xi_0, \eta_0)_{L^2(\Omega)}}_{\le(\sqrt{2}\|\xi_0\|_{L^2(\Omega)})(\|\eta_0\|_{L^2(\Omega)}/\sqrt{2})} \\
&+ \sum_{i=0}^{M-1} \frac{k_i^2}{2}\int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t + ch^4\sum_{i=0}^{M-1} |u_i|^2_{H^2(\Omega)} \\
\le \|\eta_0\|^2_{L^2(\Omega)} &+ 2\|\xi_M\|^2_{L^2(\Omega)} + \frac{1}{2}\|\eta_M\|^2_{L^2(\Omega)} + 2\|\xi_0\|^2_{L^2(\Omega)} + \frac{1}{2}\|\eta_0\|^2_{L^2(\Omega)} \\
&+ \sum_{i=0}^{M-1} \frac{k_i^2}{2}\int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t + ch^4\sum_{i=0}^{M-1} k_i|k_i^{-1/2}u_i|^2_{H^2(\Omega)} \\
\le \frac{3}{2}\|\eta_0\|^2_{L^2(\Omega)} &+ 2\|\xi_M\|^2_{L^2(\Omega)} + \frac{1}{2}\|\eta_M\|^2_{L^2(\Omega)} + 2\|\xi_0\|^2_{L^2(\Omega)} \\
&+ \sum_{i=0}^{M-1} \frac{k_i^2}{2}\int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t + ch^4T\|k^{-1/2}\Delta u\|^2_{h,\infty}.
\end{aligned}$$

Projecting the initial value $u_0 = g$ via a suitable projection $Q_h\colon H_0^1(\Omega) \to V_h$ yields

$$\|\eta_0\|_{L^2(\Omega)} = \|Q_h u_0 - P_h u_0\|_{L^2(\Omega)} \le \|(I - Q_h)u_0\|_{L^2(\Omega)} + \|(I - P_h)u_0\|_{L^2(\Omega)} \le ch^2 |u_0|_{H^2(\Omega)}.$$

Thus, we arrive at

$$\begin{aligned} \frac{1}{2}\|\eta_M\|^2_{L^2(\Omega)} \le 2\|\xi_M\|^2_{L^2(\Omega)} + 2\|\xi_0\|^2_{L^2(\Omega)} + ch^4(T+1)\|k^{-1/2}\Delta u\|^2_{h,\infty} \\ + \sum_{i=0}^{M-1} \frac{k_i^2}{2} \int_{t_i}^{t_{i+1}} \left|\frac{\partial u}{\partial t}\right|^2_{H^1(\Omega)} \mathrm{d}t. \end{aligned}$$

Since $\|u - u_h\|_{h,\infty} \le \|\xi\|_{h,\infty} + \|\eta\|_{h,\infty}$, we finally obtain the assertion. □

Remark In the case of uniform time steps $k \le ck_i$, the error estimate from Theorem 8.2, indicates that

$$\|u - u_h\|_{h,\infty} = \mathcal{O}(h^2 k^{-1/2} + k).$$

In view of the factor $k^{-1/2}$ is not optimal. However, under the condition $h \le ck^{3/4}$ we obtain the optimal order of convergence $\mathcal{O}(k)$. Employing the Crank-Nicolson method, the discretisation error becomes $\mathcal{O}(k^2 + h^2)$. This implies quadratic convergence of order $\mathcal{O}(k^2)$, if $h \sim k$ is chosen. △

9. The Eikonal Equation

We start from the wave equation, cp. Paragraph 1.3,

$$\frac{1}{v^2(\boldsymbol{x})}\frac{\partial^2 u}{\partial t^2}(t,\boldsymbol{x}) = \Delta u(t,\boldsymbol{x}) \quad \text{for } \boldsymbol{x}\in\mathbb{R}^d,\ t>0,$$

where we assume that the propagation speed $v(\boldsymbol{x})$ is spatially dependent. Consider the *plane wave*, i.e. a wave with parallel wavefronts (surfaces of constant phase),

$$P(t,\boldsymbol{x}) = P_0 e^{-\mathrm{i}\omega[t-T(\boldsymbol{x})]}$$

with constant amplitude $P_0\in\mathbb{R}$ and a given angle $\omega\in[0,2\pi)$. The function $T(\boldsymbol{x})>0$ describes the travel time from $\mathbf{0}$ to $\boldsymbol{x}$. It holds

$$\begin{aligned}
P_{xx}(t,\boldsymbol{x}) &= -P_0\big[\omega^2 T_x^2(\boldsymbol{x}) - \mathrm{i}\omega T_{xx}(\boldsymbol{x})\big]e^{-\mathrm{i}\omega[t-T(\boldsymbol{x})]},\\
P_{yy}(t,\boldsymbol{x}) &= -P_0\big[\omega^2 T_y^2(\boldsymbol{x}) - \mathrm{i}\omega T_{yy}(\boldsymbol{x})\big]e^{-\mathrm{i}\omega[t-T(\boldsymbol{x})]},\\
P_{zz}(t,\boldsymbol{x}) &= -P_0\big[\omega^2 T_z^2(\boldsymbol{x}) - \mathrm{i}\omega T_{zz}(\boldsymbol{x})\big]e^{-\mathrm{i}\omega[t-T(\boldsymbol{x})]},
\end{aligned}$$

while

$$P_{tt} = -P_0\omega^2 e^{-\mathrm{i}\omega[t-T(\boldsymbol{x})]}.$$

Hence, by inserting P into the wave equation, we obtain in view of these expressions

$$\frac{\omega^2}{v^2(\boldsymbol{x})} = \omega^2\big[T_x^2(\boldsymbol{x}) + T_y^2(\boldsymbol{x}) + T_z^2(\boldsymbol{x})\big] - \mathrm{i}\omega\big[T_{xx}(\boldsymbol{x}) + T_{yy}(\boldsymbol{x}) + T_{zz}(\boldsymbol{x})\big].$$

Since the function on the left-hand side is real valued, the imaginary part on the right-hand side has to vanish, i.e. T has to be a harmonic function. Alternatively, let $\omega\gg 1$. In both cases, we obtain

$$\frac{1}{v^2(\boldsymbol{x})} = \|\nabla T(\boldsymbol{x})\|_2^2 \quad \text{or} \quad \frac{1}{v(\boldsymbol{x})} = \|\nabla T(\boldsymbol{x})\|_2,$$

which is called the *Eikonal equation*. It gives the travel time $T(\boldsymbol{x})$ for a ray starting at $\mathbf{0}$ and passing through $\boldsymbol{x}$, where the speed of the ray is given by v. The solution $T(\boldsymbol{x})$ is also an approximation for the solution to the wave equation in the high-frequency regime $\omega\to\infty$.

Let $\Omega\subset\mathbb{R}$ denote a polygonal domain. Then, the solution to the Dirichlet problem

$$\begin{aligned}
\|\nabla u(\boldsymbol{x})\|_2 &= \frac{1}{f(\boldsymbol{x})} \quad \text{for } \boldsymbol{x}\in\Omega,\\
u(\boldsymbol{x}) &= 0 \quad \text{on } \Gamma,
\end{aligned} \tag{9.1}$$

is the shortest travel time from the boundary Γ to the point $\boldsymbol{x} \in \Omega$, where $f(\boldsymbol{x})$ is the speed at $\boldsymbol{x}$.

The Eikonal equation may also be used to model evolving interfaces: Let the boundary Γ describe the position and the shape of a given interface at time $t = 0$. Then, for $t > 0$, the position of the interface is given by

$$\Gamma(t) := \{\boldsymbol{x} \in \mathbb{R}^d : u(\boldsymbol{x}) = t\},$$

where $u\colon \mathbb{R}^d \to \mathbb{R}$ is the solution to (9.1). In this context, the scalar field $f > 0$ denotes the speed with which the boundary Γ moves with respect to the direction of the exterior normal.

For the particular choice $f \equiv 1$, the solution u is the *signed distance function*

$$u(\boldsymbol{x}) := \begin{cases} -\inf_{\boldsymbol{y}\in\Gamma} \|\boldsymbol{x} - \boldsymbol{y}\|_2, & \text{if } \boldsymbol{x} \in \Omega, \\ \inf_{\boldsymbol{y}\in\Gamma} \|\boldsymbol{x} - \boldsymbol{y}\|_2, & \text{if } \boldsymbol{x} \in \mathbb{R}^d \setminus \Omega. \end{cases}$$

Signed distance functions are for example used in ray-marching algorithms in computer vision.

Example 9.1 Let $\Omega := B_1(\mathbf{0}) \subset \mathbb{R}^2$ denote the unit disc. It holds $x(r, \alpha) = r\cos\alpha$ and $y = r\sin\alpha$. Hence, given a function $f(r, \alpha) = f(x, y)$, it holds due to the chain rule

$$\begin{aligned} \partial_r f &= \partial_x f \partial_r x + \partial_y f \partial_r y = [\cos\alpha, \sin\alpha]\nabla_{\boldsymbol{x}} f, \\ \partial_\alpha f &= \partial_x f \partial_\alpha x + \partial_y f \partial_\alpha y = [-r\sin\alpha, r\cos\alpha]\nabla_{\boldsymbol{x}} f. \end{aligned}$$

Consequently, with

$$\boldsymbol{B} := \begin{bmatrix} \cos\alpha & -r\sin\alpha \\ \sin\alpha & r\cos\alpha \end{bmatrix} \quad \text{and} \quad \boldsymbol{B}^{-1} = \frac{1}{r}\begin{bmatrix} r\cos\alpha & r\sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix},$$

it holds

$$\nabla_{[r,\alpha]} f = \boldsymbol{B}^\intercal \nabla_{\boldsymbol{x}} f \quad \text{or} \quad \nabla_{\boldsymbol{x}} f = \boldsymbol{B}^{-\intercal}\nabla_{[r,\alpha]} f$$

Thus, for $u(r, \alpha) = r - 1 = \|\boldsymbol{x}\|_2 - 1$, it holds

$$\nabla_{\boldsymbol{x}} u = \boldsymbol{B}^{-\intercal}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix} \quad \text{and thus } \|\nabla_{\boldsymbol{x}} u\|_2 = 1.$$

Moreover, the function u vanishes on $\partial B_1(\mathbf{0})$. Hence, u solves the Eikonal equation (9.1) with $f \equiv 1$ and is hence the signed distance function for the unit disc. △

9.1 Viscosity Solutions

We remark that the solution to the Eikonal equation is not unique. If u is a solution, so is $-u$. Under certain assumptions there exist unique, so called *viscosity solutions* to (9.1).

Definition 9.2 A continuous function u is a *viscosity subsolution* of (9.1), iff $u \leq 0$ on Γ and if for any for any $\varphi \in C^2(\Omega)$ which satisfies $u \leq \varphi$ in Ω, $u(\boldsymbol{x}_0) = \varphi(\boldsymbol{x}_0)$ for some $\boldsymbol{x}_0 \in \Omega$, there holds

$$f(\boldsymbol{x}_0)\|\nabla\varphi(\boldsymbol{x}_0)\|_2 \leq 1.$$

We say that u is a *viscosity supersolution* of (9.1), iff $u \geq 0$ on Γ and if for any for any $\varphi \in C^2(\Omega)$ which satisfies $u \geq \varphi$ in Ω, $u(\boldsymbol{x}_0) = \varphi(\boldsymbol{x}_0)$ for some $\boldsymbol{x}_0 \in \Omega$, there holds

$$f(\boldsymbol{x}_0)\|\nabla\varphi(\boldsymbol{x}_0)\|_2 \geq 1.$$

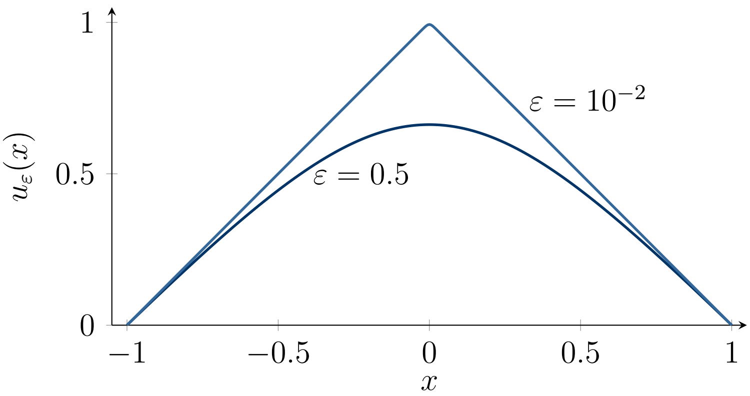
The function u is a *viscosity solution* of (9.1), iff it is a viscosity subsolution as well as a viscosity supersolution.

The existence viscosity solution has originally be shown by the *method of vanishing viscosity*: Equation (9.1) is modified by a small, additional viscosity term $-\varepsilon\Delta u$, which vanishes in the limit $\varepsilon \to 0$. We obtain

$$\begin{aligned} -\varepsilon\Delta u_\varepsilon + \|\nabla u_\varepsilon\|_2 &= \frac{1}{f}, \quad \text{in } \Omega \\ u_\varepsilon &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{9.2}$$

Moreover, it can be shown that the viscosity solution to (9.1), if $f(\boldsymbol{x}) > 0$ is a Lipschitz continuous function.

Example 9.3



Let $\Omega := (-1, 1)$. We consider the Dirichlet problem

$$-\varepsilon u_\varepsilon''(x) + \big(u_\varepsilon'(x)\big)^2 - 1 = 0 \quad x \in \Omega, \quad u_\varepsilon(x) = 0, \quad x \in \Gamma.$$

It can be shown that if this problem exhibits a classical solution, it is also unique. Presuming, that this solution is symmetric, we can also assume $u_\varepsilon'(0) = 0$. Now, setting $v_\varepsilon := u_\varepsilon'$ yields the ordinary differential equation

$$-\varepsilon v_\varepsilon'(x) + v_\varepsilon^2(x) = 1, \quad v_\varepsilon(0) = 0.$$

Since

$$\frac{\mathrm{d}}{\mathrm{d}x}\tanh(x) = 1 - \tanh^2(x),$$

the function $v_\varepsilon = -\tanh(x/\varepsilon)$ solves the ordinary differential equation above. Integration of v_ε yields the anti-derivative

$$u_\varepsilon(\varepsilon) = -\varepsilon \log\left(\frac{\cosh(x/\varepsilon)}{\cosh(1/\varepsilon)}\right) = -\varepsilon \log\cosh(x/\varepsilon) + \varepsilon \log\cosh(1/\varepsilon),$$

where the constant is fixed by the boundary values. Now, by L'Hospital's rule, we obtain

$$\lim_{\varepsilon\to 0} \varepsilon \log\cosh(x/\varepsilon) = x,$$

while $x < 0$ yields

$$\lim_{\varepsilon\to 0} \varepsilon \log\cosh(x/\varepsilon) = -x.$$

Hence, the function $u(x) = 1 - |x|$ is the viscosity solution to the Dirichlet problem under consideration. △

Example 9.4 Let $\Omega := (-1, 1)$. We consider the Dirichlet problem

$$|u'(x)| = 1 \quad x \in \Omega, \quad u(x) = 0, \quad x \in \Gamma.$$

We show that $u(x) = 1 - |x|$ satisfies Definition 9.2.

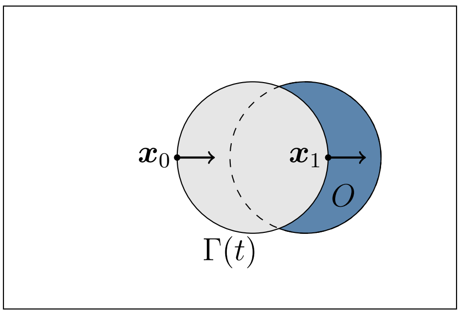
If a test function φ touches u at a point x_0, where u is differentiable, the derivatives of u and φ have to coincide, i.e. $|\varphi'(x_0)| = |u(x_0)| = 1$. Hence, we have only to check $x_0 = 0$. Suppose $\varphi \geq u$ in Ω and $\varphi(0) = u(0)$. Then the slope of φ is bounded by the slopes of u according to

$$\lim_{x\downarrow 0} u'(x) = -1 \leq \varphi'(0) \leq 1 = \lim_{x\uparrow 0} u'(x).$$

Consequently, $|\varphi'(0)| \leq 1$ and u is a viscosity subsolution.

Now, let $\varphi \leq u$ in Ω and $\varphi(0) = u(0)$. Then, due to the shape of u, it is not possible that φ is differentiable at $x_0 = 0$. Hence, it cannot be in $C^2(\Omega)$. Therefore, the function u is also a viscosity supersolution and hence the viscosity solution. △

9.2 Application to Image Segmentation



In order to apply the Eikonal equation to image segmentation, we return to the level set perspective. To that end, we consider the function $v(t, \boldsymbol{x}) := u(\boldsymbol{x}) - t$, where $u(\boldsymbol{x})$ solves (9.1). Then, obviously

$$\frac{\partial v}{\partial t} = f(\boldsymbol{x})\|\nabla v\|_2$$

and

$$\Gamma(t) = \{\boldsymbol{x} \in \Omega : u(\boldsymbol{x}) = t\} = \{\boldsymbol{x} \in \Omega : v(t, \boldsymbol{x}) = 0\}$$

is the *zero level set* of v, while f still describes the speed in normal direction.

Now, let I denote an image with white background. Inside this image is a blue object O. Let $\chi_O(\boldsymbol{x})$ denote the characteristic function of the object O, i.e.

$$\chi_O(\boldsymbol{x}) := \begin{cases} 1, & \text{if the pixel is blue,} \\ 0, & \text{if the pixel is white.} \end{cases}$$

We denote by $\Omega(t)$ the open set that is bounded by $\Gamma(t)$. Our goal is to move the curve $\Gamma(t)$ such that it coincides with the boundary of the object O. To that end, we define the mean intensity of the pixels in $\Omega(t)$ and in its complement $I \setminus \Omega(t)$ according to

$$m_1(t) := \frac{1}{|\Omega(t)|} \int_{\Omega(t)} \chi_O(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \quad \text{and} \quad m_2(t) := \frac{1}{|I| - |\Omega(t)|} \int_{I \setminus \Omega(t)} \chi_O(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}.$$

Based on these quantities, we define the, now time dependent, speed in normal direction as

$$f(t, \boldsymbol{x}) := \big(\chi_O(\boldsymbol{x}) - m_2(t)\big)^2 - \big(\chi_O(\boldsymbol{x}) - m_1(t)\big)^2.$$

To understand how this approach works, we focus on the figure above: Here, we have $m_1(t) \approx 1$, while $m_2(t) \approx 0$ since $\Omega(t)$ already covers a large portion of O. Hence, in view of the point $\boldsymbol{x}_0$, we obtain

$$f(t, \boldsymbol{x}_0) \approx (0 - 0)^2 - (0 - 1)^2 = -1 < 0.$$

Hence, the point $\boldsymbol{x}_0$ is moved to the right. On the other hand, we have for $\boldsymbol{x}_1$ that

$$f(t, \boldsymbol{x}_1) \approx (1 - 0)^2 - (1 - 1)^2 = 1 > 0.$$

Therefore, the point $\boldsymbol{x}_1$ is moved to the right as well.

Remark Numerically, the simulation of the moving boundary can efficiently be realised by means of the so called *fast marching method.* △

Recommended Literature

[Alt] H.W. Alt Linear Functional Analysis, Springer.

[Braess] D. Braess *Finite Elements*, 3rd edition, Cambridge University Press.

[Hackbusch] W. Hackbusch *Elliptic Partial Differential Equations*, 2nd edition, Springer.

[Larsson] S. Larsson and V. Thomée *Partial Differential Equations with Numerical Methods*, Springer.

[Quarteroni] A. Quarteroni *Numerical models for differential problems*, 4th edition, Springer.

[Salsa] S. Salsa *Partial Differential Equations in action, from Modelling to Theory*, Springer.

[Wloka] J. Wloka *Partielle Differentialgleichungen*, Teubner.

Index