1 Classification of PDEs

Partial differential equations (PDEs) model a vast range of natural phenomena and industrial problems. Even though analytical solutions are not normally available, the equation itself may already reveal crucial information on some characteristic features of the (unknown) analytical solution. Our main objective is to design numerical approximations that reflect these properties accurately.

The theoretical part of this course is devoted to classifying PDEs based on such characteristic features.

1.1 Basic Properties

Often, we will focus on the most important special case of a (quasi-)linear, scalar PDE of second order in two variables:

$$a_{11}\frac{\partial^2 u}{\partial x_1^2} + 2a_{12}\frac{\partial^2 u}{\partial x_1 \partial x_2} + a_{22}\frac{\partial^2 u}{\partial x_2^2} + a_1\frac{\partial u}{\partial x_1} + a_2\frac{\partial u}{\partial x_2} + au = f \qquad \text{in } \Omega \subseteq \mathbb{R}^2, \tag{$\star$}$$

where not all of a_{11}, a_{12}, a_{22} are simultaneously equal to zero.

We also use the short operator notation for PDEs

$$Lu = f \qquad \text{in } \Omega.$$

1.1.1 Definition (Linear and Nonlinear PDEs) The PDE ($\star$) is said to be

Otherwise, if a second-order PDE

$$F(x, u, \nabla u, \nabla^2 u) = 0$$

cannot be written in the form ($\star$), then it is said to be *fully nonlinear*.

In practice, the domain Ω is usually bounded, or unbounded in one direction only (namely, the 'time-direction'). While the PDE describes the behaviour of a solution u in the interior of Ω, additional conditions prescribe certain values on the boundary $\partial\Omega$. The question what kind of boundary conditions are 'admissible' for a given PDE is not straightforward to answer. The answer will depend on the type and the exact form of the PDE. First, we clarify what 'admissible' means in this context:

1.1.2 Definition (Well-Posed Problem) A PDE equipped with boundary conditions is said to be *well-posed in the sense of* HADAMARD, if

In addition to these three properties, we are also interested in the regularity or smoothness of solutions, i.e. how many derivatives they possess.

1.2 Second-Order PDEs

There are three types of PDEs: *elliptic*, *parabolic* and *hyperbolic* equations. Strictly speaking, this classification only works for certain classes of PDEs, and therefore we continue to focus on the class of (quasi-)linear, scalar PDEs of second order in two variables ($\star$).

The classification of these PDEs into elliptic, parabolic and hyperbolic equations is based on a so-called CAUCHY *initial value problem*

$$a_{11}\frac{\partial^2 u}{\partial x_1^2} + 2a_{12}\frac{\partial^2 u}{\partial x_1 \partial x_2} + a_{22}\frac{\partial^2 u}{\partial x_2^2} + a_1\frac{\partial u}{\partial x_1} + a_2\frac{\partial u}{\partial x_2} + au = f \qquad \text{in } \Omega \tag{$\star$}$$

$$u = g \qquad \text{on } \Gamma \tag{C1}$$

$$\frac{\partial u}{\partial n} = h \qquad \text{on } \Gamma, \tag{C2}$$

where we assume that the 'initial curve' $\Gamma \subset \Omega$ is of the class C^∞, i.e. Γ has a parameterisation

$$x(\tau) = (x_1(\tau), x_2(\tau))$$

that is infinitely often differentiable.

By differentiating once, the CAUCHY condition (C1) gives the derivative $\frac{\partial u}{\partial t}$ in tangential direction along Γ. Therefore, (C1) and (C2) combined already prescribe both the function values u and the complete gradient $\nabla u = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}\right)^\top$ on the curve Γ.

Our objective is to find all second, third and higher partial derivatives of u on Γ, so that the TAYLOR series

$$u(x) = \sum_{k,l=0}^{\infty} \frac{(x_1 - x_1^*)^k (x_2 - x_2^*)^l}{(k+l)!} \frac{\partial^k}{\partial x_1^k} \frac{\partial^l}{\partial x_2^l} u(x^*) \tag{1.1}$$

around a point $x^* \in \Gamma$ would then give a solution to the CAUCHY problem for the PDE ($\star$) in a neighbourhood of Γ.

Since both first derivatives $\frac{\partial u}{\partial x_1}$ and $\frac{\partial u}{\partial x_2}$ are given on Γ, we differentiate in tangential direction:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau} \frac{\partial u}{\partial x_1} &= \frac{\partial^2 u}{\partial x_1^2} \frac{\mathrm{d}x_1}{\mathrm{d}\tau} + \frac{\partial^2 u}{\partial x_1 \partial x_2} \frac{\mathrm{d}x_2}{\mathrm{d}\tau} \\ \frac{\mathrm{d}}{\mathrm{d}\tau} \frac{\partial u}{\partial x_2} &= \frac{\partial^2 u}{\partial x_1 \partial x_2} \frac{\mathrm{d}x_1}{\mathrm{d}\tau} + \frac{\partial^2 u}{\partial x_2^2} \frac{\mathrm{d}x_2}{\mathrm{d}\tau}. \end{aligned}$$

Let us introduce some short-hand notations for the second partial derivatives of the solution u:

$$p := \frac{\partial^2 u}{\partial x_1^2} \qquad q := \frac{\partial^2 u}{\partial x_1 \partial x_2} \qquad r := \frac{\partial^2 u}{\partial x_2^2}.$$

Two equations for these three unknowns p, q, r are given by the differentiated CAUCHY conditions, and the PDE ($\star$) adds a third equation. This yields the following linear system:

$$\begin{pmatrix} \frac{\mathrm{d}x_1}{\mathrm{d}\tau} & \frac{\mathrm{d}x_2}{\mathrm{d}\tau} & 0 \\ 0 & \frac{\mathrm{d}x_1}{\mathrm{d}\tau} & \frac{\mathrm{d}x_2}{\mathrm{d}\tau} \\ a_{11} & 2a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} p \\ q \\ r \end{pmatrix} = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}\tau} \frac{\partial u}{\partial x_1} \\ \frac{\mathrm{d}}{\mathrm{d}\tau} \frac{\partial u}{\partial x_2} \\ f - a_1 \frac{\partial u}{\partial x_1} - a_2 \frac{\partial u}{\partial x_2} - au \end{pmatrix},$$

where the determinant of the coefficient matrix—let's call it B—is found to be

$$\det B = a_{11} \left(\frac{\mathrm{d}x_2}{\mathrm{d}\tau}\right)^2 - 2a_{12} \left(\frac{\mathrm{d}x_2}{\mathrm{d}\tau}\right) \left(\frac{\mathrm{d}x_1}{\mathrm{d}\tau}\right) + a_{22} \left(\frac{\mathrm{d}x_1}{\mathrm{d}\tau}\right)^2. \tag{1.2}$$

1st Case: $\det B \neq 0$ on all of Γ The linear system admits a unique solution. Hence, it determines the second derivatives p, q, r on Γ.

The third partial derivatives $\frac{\partial p}{\partial x_1}, \frac{\partial q}{\partial x_1}, \frac{\partial r}{\partial x_1}$ on Γ are found by differentiating the linear system with respect to x_1, which yields another linear system with exactly the same coefficient matrix B. The same holds true for the remaining third partial derivatives $\frac{\partial p}{\partial x_2}, \frac{\partial q}{\partial x_2}, \frac{\partial r}{\partial x_2}$ and then all higher derivatives. We have reached our objective to describe u in a neighbourhood of Γ through the TAYLOR series (1.1).

2nd Case: $\det B = 0$ in a point $x^* \in \Gamma$ The linear system does not admit a unique solution. Hence, the second partial derivatives of u in x^* cannot be determined from the CAUCHY conditions.

More specifically, the equation

$$a_{11}\left(\frac{\mathrm{d}x_2}{\mathrm{d}\tau}\right)^2 - 2a_{12}\left(\frac{\mathrm{d}x_2}{\mathrm{d}\tau}\right)\left(\frac{\mathrm{d}x_1}{\mathrm{d}\tau}\right) + a_{22}\left(\frac{\mathrm{d}x_1}{\mathrm{d}\tau}\right)^2 = 0 \tag{1.3}$$

determines the slope of curves $(x_1(\tau), x_2(\tau))$ through the point $x^* \in \Gamma$, along which the TAYLOR series approach breaks down. In other words, on these critical curves a.k.a. *characteristics* of the operator L (or, actually, its principal part L_0) the solution u cannot be derived from the given data. Instead, u or its derivatives may possess discontinuities along the characteristics.

This singular case could be ruled out by choosing an 'initial curve' Γ that is nowhere tangential to a characteristic, to avoid the situation where the tangential vectors $\frac{\mathrm{d}x}{\mathrm{d}\tau}$ of Γ and $\frac{\mathrm{d}x}{\mathrm{d}\tau}$ of a characteristic coincide.

To calculate the characteristics explicitly, we can re-arrange (1.3) (provided that $\frac{\mathrm{d}x_1}{\mathrm{d}\tau} \neq 0$) to obtain the quadratic equation

$$a_{11}\left(\frac{\mathrm{d}x_2}{\mathrm{d}x_1}\right)^2 - 2a_{12}\left(\frac{\mathrm{d}x_2}{\mathrm{d}x_1}\right) + a_{22} = 0$$

with solutions

$$\frac{\mathrm{d}x_2}{\mathrm{d}x_1} = \frac{a_{12}}{a_{11}} \pm \frac{\sqrt{a_{12}^2 - a_{11}a_{22}}}{a_{11}}.$$

Recall that (1.3) describes conic sections (ellipse, parabola and hyperbola), here in the variables $\frac{\mathrm{d}x_2}{\mathrm{d}\tau}$ and $\frac{\mathrm{d}x_1}{\mathrm{d}\tau}$. This motivates the following nomenclature:

1.2.1 Definition (Elliptic, Parabolic and Hyperbolic Equations) In a point $x^* \in \Omega$ the PDE $(\star)$ is said to be

elliptic if the discriminant $a_{12}^2 - a_{11}a_{22} < 0$, i.e. there are no characteristics through x^*,

parabolic if the discriminant $a_{12}^2 - a_{11}a_{22} = 0$, i.e. there is one characteristic through x^*,

hyperbolic if the discriminant $a_{12}^2 - a_{11}a_{22} > 0$, i.e. there are two characteristics through x^*.

It is important to note that this criterion gives a pointwise classification of a PDE, since the coefficients a_{11}, a_{12}, a_{22} generally depend on x and maybe even on the solution u itself. Hence, a single equation may be elliptic in some parts of the domain Ω, parabolic in other regions and hyperbolic elsewhere. In this course, we normally study equations that are uniformly (i.e. everywhere) elliptic, parabolic or hyperbolic, such as PDEs with constant coefficients.

1.2.2 Example (Prototypes of Elliptic, Parabolic and Hyperbolic Equations) The most fundamental representatives of linear, second-order PDEs are

- POISSON's equation $-\Delta u = f$, which is elliptic,
- the heat equation $\partial_t u - \partial_x^2 u = f$, which is parabolic,
- the wave equation $\partial_t^2 u - \partial_x^2 u = f$, which is hyperbolic.

1.3 Conservation Equations

Very often, PDE problems model physical, chemical or biological systems which are governed by certain *conservation laws*, such as

- conservation of mass
- conservation of energy
- conservation of momentum
- conservation of charge
- conservation of the number of individuals in a population

In mathematical terms, conservation of a (mass, energy, momentum ...) density u implies that if this quantity is transported through the domain with a flux F, the density increases inside any arbitrary control volume V if there is a net influx across the boundary of V and it decreases otherwise:

$$\int\limits_V \frac{\partial u}{\partial t}\,\mathrm{d}x = -\int\limits_{\partial V} F\cdot n\,\mathrm{d}s.$$

Provided that the flux F and the control volume V satisfy the assumptions of the divergence theorem, we obtain

$$\int\limits_V \frac{\partial u}{\partial t}\,\mathrm{d}x + \int\limits_V \operatorname{div} F\,\mathrm{d}x = 0.$$

Since this balance has to hold for arbitrary such volumes, we have derived the continuity equation

$$\frac{\partial u}{\partial t} + \operatorname{div} F = 0. \tag{1.4}$$

There are two particularly important fluxes:

- advective flux $F = ua$, with an advection (velocity) field a
- diffusive flux $F = -D\nabla u$, with a positive diffusivity D. More generally, D could also be a positive definite matrix, a so-called diffusion tensor.

Sometimes, the quantity u is not actually conserved, but sources or sinks such as chemical reactions or external forces give rise to an inhomogeneity r on the right-hand side of the equation.

Below, we list some prototypical conservation equations:

Unsteady advection-diffusion-reaction equation (parabolic)

$$\frac{\partial u}{\partial t} + \operatorname{div}(ua) - \operatorname{div}(D\nabla u) = r$$

Steady advection-diffusion-reaction equation (elliptic)

$$\operatorname{div}(ua) - \operatorname{div}(D\nabla u) = r$$

Unsteady advection equation (hyperbolic)

$$\frac{\partial u}{\partial t} + \operatorname{div}(ua) = 0$$

Steady advection equation (hyperbolic)

$$\operatorname{div}(ua) = 0$$

Note that all first-order PDEs are generally defined to be hyperbolic.

Many advection fields in applications are incompressible, which means that $\operatorname{div} a = 0$. Then the above advective terms can also be written as

$$\operatorname{div}(ua) = u \operatorname{div} a + a \cdot \nabla u = a \cdot \nabla u.$$

2 Second-Order Elliptic Equations

2.1 Characteristic Features

The prototypical representative of elliptic PDEs is the Poisson equation

$$-\mathit{\Delta} u = f, \tag{2.1}$$

which will serve as our elliptic model problem. This section is devoted to the most important properties of this problem, which shall be preserved under any ‘good’ numerical discretisation scheme.

Boundary Conditions

2.1.1 Definition (Boundary Conditions) Let $g, a : \partial\mathit{\Omega} \to \mathbb{R}$. A boundary condition of the form

(a)

is called Dirichlet *boundary condition* or *boundary condition of the first kind*,

(b)

is called Neumann *boundary condition* or *boundary condition of the second kind*,

(c)

is called Robin *boundary condition* or *boundary condition of the third kind.*

These boundary conditions are said to be *homogeneous* if $g \equiv 0$, otherwise *inhomogeneous*.

Existence of Strong Solutions

The *classical* or *strong formulation* of Poisson’s equation with Dirichlet boundary conditions reads: find $u \in C^2(\mathit{\Omega}) \cap C(\bar{\mathit{\Omega}})$ such that

$$-\mathit{\Delta} u = f \qquad \text{in } \mathit{\Omega} \tag{2.2a}$$

$$u = g \qquad \text{on } \partial\mathit{\Omega}. \tag{2.2b}$$

Proving existence of strong solutions of (2.2) on general domains is difficult and not exactly elegant. It also requires (often too) strong assumptions on the regularity of the domain $\mathit{\Omega}$ and the data f. We’ll skip the discussion and refer to the literature on PDEs—the keyword is Green*’s functions*.

Uniqueness of Strong Solutions

The classical uniqueness proof for strong solutions of POISSON's equation equipped with DIRICHLET boundary conditions relies on a *maximum principle* for elliptic equations:

2.1.2 Theorem (Elliptic Maximum Principle) *Let*

$$L = a_{11}\frac{\partial^2}{\partial x_1^2} + 2a_{12}\frac{\partial^2}{\partial x_1 \partial x_2} + a_{22}\frac{\partial^2}{\partial x_2^2} + a_1\frac{\partial}{\partial x_1} + a_2\frac{\partial}{\partial x_2}$$

be elliptic (note that $a \equiv 0$) and $u \in C^2(\Omega) \cap C(\bar{\Omega})$. Then

$$Lu \leq 0 \qquad \text{in } \Omega \qquad\qquad \Rightarrow \qquad\qquad \max_{x\in\bar{\Omega}} u(x) \leq \max_{x\in\partial\Omega} u(x),$$

i.e. the solution u assumes its maximum on the boundary.

Now the uniqueness of strong solutions to (2.2) is obtained as follows:

Strong vs Weak Solutions

A solution of the strong formulation of POISSON's equation with homogeneous DIRICHLET boundary conditions

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{P}$$

has to be twice continuously differentiable in the interior of the domain (so that the Laplacian can be applied to it) and it must be continuous up to the boundary (so that it actually approaches the boundary values from the interior).

However, asking for two continuous derivatives of the solution u is often too strong a condition. For example, if material parameters suddenly change across an interface that cuts through the domain Ω, then this will normally affect the smoothness of the solution u as well, for instance in the form of a 'kink' (a discontinuity in a first derivative).

It would therefore be desirable to have a relaxed formulation, which generalises the notion of solutions to the POISSON-DIRICHLET problem (P). The central idea behind *variational* or *weak formulations*:

The divergence theorem implies the identity

$$\int_{\Omega} (\operatorname{div} F) v \,\mathrm{d}x = \int_{\partial\Omega} (F \cdot n) v \,\mathrm{d}s - \int_{\Omega} F \cdot \nabla v \,\mathrm{d}x$$

which generalises integration by parts to higher dimensions, where $F : \Omega \to \mathbb{R}^d$ is a sufficiently regular vector field and $v : \Omega \to \mathbb{R}$ a scalar function on a domain $\Omega \subset \mathbb{R}^d$.

Applied to (P), we obtain

and if v vanishes on the boundary $\partial\Omega$ as well, then

Note that all the second derivatives have now disappeared. It is not even necessary to ask for *continuous* first derivatives of u and v. Instead, we require the following assumptions to ensure that the weak-formulation is well-defined:

Consequently, the spaces C^k of continuous(ly differentiable) functions are not well suited for weak formulations. Instead, we will use the so-called Lebesgue *spaces* L^p and Sobolev *spaces* H^k and $W^{k,p}$, which contain functions that meet certain integrability conditions.

2.1.3 Definition (Lebesgue Space of Square-Integrable Functions) For a domain $\Omega \subset \mathbb{R}^d$ we define

$$\|u\|_{L^2(\Omega)} = \left(\int_\Omega |u(x)|^2 \,\mathrm{d}x \right)^{1/2} .$$

The set

$$L^2(\Omega) = \left\{ u : \Omega \to \mathbb{R} \;\middle|\; \|u\|_{L^2(\Omega)} < \infty \right\}$$

is called the Lebesgue *space of order 2*.

2.1.4 Theorem (L^2 is a Hilbert Space) *The* Lebesgue *space L^2 of square-integrable functions is a* Hilbert *space with the scalar product*

$$(u, v)_{L^2} =$$

2.1.5 Remark The space $L^2(\Omega)$ is actually quite different from spaces of continuous functions, such as $C(\Omega)$:

-
- In Lebesgue spaces, we usually cannot assign point values to functions. The expression $u(x)$ is not meaningful. Consider for instance the two L^2-functions on $]-1,1[$

$$u_1(x) \equiv 1 \qquad \text{and} \qquad u_2(x) = \begin{cases} 1 & \text{for } x \neq 0 \\ -3 & \text{for } x = 0 \end{cases} .$$

  These two functions are not distinguishable in $L^2(]-1,1[)$ since

  even though $u_1(0) \neq u_2(0)$.

2.1.6 Definition (L^2-based Sobolev Spaces) For a domain $\Omega \subset \mathbb{R}^d$ and $k \in \mathbb{N}_0$ we define the Sobolev *space* $H^k(\Omega)$ as the set of all functions $u \in L^2(\Omega)$, of which all (weak) partial derivatives up to and including order k are in $L^2(\Omega)$ as well.

2.1.7 Theorem (H^k is a Hilbert Space) *The L^2-based* Sobolev *spaces H^k of square-integrable functions with square-integrable derivatives up to order k are* Hilbert *spaces with the scalar product*

$$(u, v)_{H^k} =$$

There are a few more mathematical technicalities involved when it comes to the definition of the homogeneous Dirichlet boundary conditions of our example problem ((P)). We will not go into further detail here, but refer to the *trace theorem for* Sobolev *spaces*, which can be found in the literature on functional analysis.

2.1.8 Definition (H_0^1 and H^{-1}) The space of all functions $u \in H^1(\Omega)$ with $u|_{\partial\Omega} = 0$ is denoted by $H_0^1(\Omega)$.

The dual space of $H_0^1(\Omega)$ is denoted by $H^{-1}(\Omega)$. That is,

$$H^{-1}(\Omega) = \big(H_0^1(\Omega)^*\big) = \big\{\, f : H_0^1(\Omega) \to \mathbb{R} \;\big|\; f \text{ is linear and continuous} \,\big\}\,.$$

For the *duality pairing* of these two spaces, i.e. for $f \in H^{-1}(\Omega)$ and a function $v \in H_0^1(\Omega)$ we normally use the symmetric notation $\langle f, v\rangle_{H^{-1},H_0^1}$ instead of $f(v)$.

Overall, we have $H_0^1(\Omega) \subset L^2(\Omega) \subset H^{-1}$. The space H^{-1} is actually bigger than L^2 and also contains abstract elements that are no functions on Ω. Depending on the dimension of Ω, this could for instance be objects like delta-‘functions’:

2.1.9 Example (Source Terms in H^{-1}) L^2-**functions** If the source term f of the PDE (P) is a given L^2-function, then

delta-‘function’ If $f = \delta_0$, i.e. the delta-‘function’ centred at the point $x = 0$

Note that this delta-‘function’ requires us to evaluate the test functions v at a point. Whenever this is not possible for all test functions in $H_0^1(\Omega)$—and this will depend on the dimension of Ω—the delta-‘function’ does not belong to the space $H^{-1}(\Omega)$.

We can finally write down a weak formulation of (P): given $f \in H^{-1}(\Omega)$, find a function $u \in H_0^1(\Omega)$ such that for all test functions $v \in H_0^1(\Omega)$

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \langle f, v\rangle_{H^{-1},H_0^1}. \tag{P'}$$

Due to the very different form of the problems (P) and (P'), the classical analytical toolkit required to derive these properties for strong solutions is very different from the functional analytical techniques for weak solutions.

Existence and Uniqueness of Weak Solutions

Two famous results from functional analysis immediately give the existence and uniqueness of weak solutions of (P'): the LAX-MILGRAM lemma together with the POINCARÉ inequality. In our numerical analysis, we will make use of these two results over and over again.

2.1.10 Lemma (POINCARÉ inequality) *There exists a constant $C > 0$ such that for all $u \in H_0^1(\Omega)$*

$$\|u\|_{L^2} \leq C\|\nabla u\|_{L^2}. \tag{2.3}$$

It is important to note that this inequality only holds for functions in H_0^1, not for all functions in H^1! A simple counterexample is given by

2.1.11 Theorem (LAX-MILGRAM Lemma) *Let V be a* HILBERT *space with scalar product $(\cdot,\cdot)_V$ and norm $\|\cdot\|_V = \sqrt{(\cdot,\cdot)_V}$. If $B : V \times V \to \mathbb{R}$ is a*

- continuous

- *and* coercive

bilinear form and $f \in V^*$, then the problem

$$B(u,v) = \langle f, v\rangle_{V^*,V}, \qquad \forall v \in V$$

admits a unique solution $u \in V$. This solution satisfies the a priori estimate

$$\|u\|_V \leq \frac{1}{c}\|f\|_{V^*}. \tag{2.4}$$

2.1.12 Example Consider the particular special case where the space V is the n-dimensional Euclidean space $\mathbb{R}^n$,

$$B(u,v) = v^\top A \cdot u \qquad f = b^\top$$

where A is a symmetric positive definite $n \times n$ matrix and $b \in \mathbb{R}^n$ a given (column) vector.

2.1.13 Corollary (Existence and Uniqueness of Solutions to (P')) *Given $f \in H^{-1}(\Omega)$, there is a unique solution $u \in H_0^1(\Omega)$ such that for all test functions $v \in H_0^1(\Omega)$*

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = \langle f, v \rangle_{H^{-1}, H_0^1}. \tag{P'}$$

Proof.

□

Note that this proof of existence and uniqueness is equally valid for strong solutions of (P), since

Continuous Dependence on Data

One more result follows immediately from the Lax-Milgram lemma: the continuous dependence of the unique solution u on the data f. If u is a strong or weak solution of the model problem

$$-\Delta u = f \qquad \text{in } \Omega \qquad\qquad u = 0 \qquad \text{on } \partial\Omega$$

and $\tilde{u}$ a strong or weak solution of the perturbed problem

$$-\Delta \tilde{u} = \tilde{f} \qquad \text{in } \Omega \qquad\qquad \tilde{u} = 0 \qquad \text{on } \partial\Omega,$$

then $u - \tilde{u}$ solves the problem

$$-\Delta(u - \tilde{u}) = f - \tilde{f} \qquad \text{in } \Omega \qquad\qquad u - \tilde{u} = 0 \qquad \text{on } \partial\Omega.$$

The estimate (2.4) from the Lax-Milgram lemma now gives

$$\|u - \tilde{u}\|_{H^1_0} \leq$$

Clearly, the perturbed solution $\tilde{u}$ must approach the unperturbed solution u in the H^1_0-norm as $\tilde{f} \to f$ in $H^{-1}(\Omega)$ (or in $L^2(\Omega)$, if both f's happen to be L^2-functions).

What about perturbations in the boundary data? So far, we have only considered homogeneous boundary conditions, in fact, without loss of generality. If $g \in H^1(\Omega)$, then its restriction to the boundary $\partial\Omega$ is a function that is less regular by 'half a derivative': $g|_{\partial\Omega} \in H^{1/2}(\partial\Omega)$. Every H^1-function in Ω has boundary values (aka 'trace') in $H^{1/2}(\partial\Omega)$ and every $H^{1/2}$-function on $\partial\Omega$ is the trace of a H^1-function in Ω*. Consequently, if boundary data $g \in H^{1/2}(\Omega)$ are prescribed,

$$-\,\Delta u = f \qquad \text{in } \Omega \qquad\qquad u = g \qquad \text{on } \partial\Omega \tag{2.5}$$

then this function g can be extended to a H^1-function on the entire domain. The weak formulation of (2.5) reads: find a solution u from the affine subspace $g + H^1_0(\Omega)$ of $H^1(\Omega)$ such that for all test functions $v \in H^1_0(\Omega)$

$$\int\limits_{\Omega} \nabla u \cdot \nabla v \,\mathrm{d}x = \langle f, v \rangle_{H^{-1},H^1_0}. \tag{2.6}$$

*Traces of H^2-functions are in $H^{3/2}(\partial\Omega)$, traces of H^3-functions in $H^{5/2}(\partial\Omega)$ etc. If $u \in H^1(\Omega)$, then its gradient is in $L^2(\Omega)$ and traces of the gradient, such as the normal derivative $\partial_n u$ on the boundary, are in $H^{-1/2}(\partial\Omega)$.

This problem is equivalent to the problem of finding a solution $w = u - g \in H_0^1(\Omega)$ such that for all test functions $v \in H_0^1(\Omega)$

$$\int\limits_\Omega \nabla w \cdot \nabla v \,\mathrm{d}x = \langle f, v\rangle_{H^{-1},H_0^1} - \int\limits_\Omega \nabla g \cdot \nabla v \,\mathrm{d}x, \tag{2.7}$$

another problem with homogeneous boundary conditions, but a modified right hand side instead. The inequality (2.4) from the LAX-MILGRAM lemma implies that w and hence also u depends continuously on the boundary data g as well.

Regularity

In the most general setting, weak solutions to an elliptic, linear, second-order PDE lie in the space $H^1(\Omega)$. In general, such solutions need not be continuous and higher derivatives may not exist. They may not even be bounded. Certain geometries and boundary conditions are particularly notorious for generating unbounded peaks in the solution.

We first introduce LEBESGUE and SOBOLEV spaces with exponents not necessarily equal to two.

2.1.14 Definition (General LEBESGUE Spaces) For a domain $\Omega \subset \mathbb{R}^d$ we define

$$\|u\|_{L^p(\Omega)} = \left(\int\limits_\Omega |u(x)|^p \,\mathrm{d}x\right)^{1/p} \qquad \text{for } 1 \le p < \infty$$

$$\|u\|_{L^\infty(\Omega)} = \operatorname*{ess\,sup}_{x\in\Omega} |u(x)| = \inf\left\{\, c \ge 0 \mid |u(x)| \le c \text{ almost everywhere in } \Omega \,\right\}.$$

For $1 \le p \le \infty$ the set

$$L^p(\Omega) = \left\{\, u : \Omega \to \mathbb{R} \;\middle|\; \|u\|_{L^p(\Omega)} < \infty \,\right\}$$

is called a LEBESGUE *space of order p*.

2.1.15 Definition (General SOBOLEV Spaces) For a domain $\Omega \subset \mathbb{R}^d$ and $k \in \mathbb{N}_0$ we define the SOBOLEV *space* $W^{k,p}(\Omega)$ as the set of all functions $u \in L^p(\Omega)$, of which all (weak) partial derivatives up to and including order k are in $L^p(\Omega)$ as well.

In many cases, the data of a PDE and its domain are smoother than in the most general setting. For example, we often have a proper function $f \in L^2(\Omega)$ instead of just $f \in H^{-1}(\Omega)$ on the right hand side of the PDE. Our hope is that then this higher regularity of f will propagate through to the solution u so that u will have one extra derivative as well. Provided that the domain is not too irregular, this is indeed the case:

2.1.16 Theorem (H^2-Regularity) *Let Ω be either a domain with C^2-boundary or a convex polygon. Let $u \in H_0^1(\Omega)$ be the unique weak solution to the elliptic problem*

$$Lu = f \quad \text{in } \Omega \qquad\qquad u = 0 \quad \text{on } \partial\Omega$$

where the linear operator L has coefficients

$$a_{ij} \in C^1(\bar{\Omega}), \qquad a_i, a \in L^\infty(\Omega) \qquad\qquad (i, j \in \{\, 1, 2 \,\})$$

and $f \in L^2(\Omega)$.

Then $u \in H^2(\Omega)$ and there exists a constant $C > 0$ such that

$$\|u\|_{H^2} \leq C\|f\|_{L^2}. \tag{2.8}$$

If the boundary, the coefficients of the differential operator and the data are even smoother, then even higher regularity may be deduced for u. SOBOLEV embeddings can be applied to investigate what other spaces such as L^p or C^k the solution belongs to.

In practical problems, however, domains often possess corners or edges. Then a prediction of H^2-regularity may already be the optimal smoothness result that the theory allows for. Domains that have corners and that are also non-convex are a classical source of trouble. Around *re-entrant corners* of the domain, i.e. corners where the interior angle is larger than π, solutions often tend to develop singularities that prevent them from being H^2 even if the data are arbitrarily smooth, as the following example shall demonstrate.

2.1.17 Example (Corner Singularities) On the circular sector, described in polar coordinates by $\Omega = \left\{ (r,\vartheta) \in \mathbb{R}^2 \,\middle|\, 0 < r < 1 \text{ and } 0 < \vartheta < \frac{3\pi}{2} \right\}$, we consider the problem

$$-\Delta u = 0 \qquad \text{in } \Omega$$

with boundary conditions

$$u(r,\vartheta) = 0 \qquad \text{for } \vartheta \in \left\{ 0, \frac{3\pi}{2} \right\} \qquad\qquad u(r,\vartheta) = \sin\frac{2\vartheta}{3} \qquad \text{for } r = 1.$$

The unique strong and weak solution is given by

$$u(r,\vartheta) = r^{2/3} \sin\frac{2\vartheta}{3}$$

and possesses a singularity at the origin. In this example $u \in C^2(\Omega) \cap C(\bar{\Omega})$, but $u \notin C^1(\bar{\Omega})$ and $u \notin H^2(\Omega)$.

This result can be generalised to circular sectors with angles $0 < \vartheta < \omega$, where $\omega \in]0, 2\pi]$. Then

$$u(r,\vartheta) = r^{\pi/\omega} \sin\frac{\vartheta\pi}{\omega}$$

solves the homogeneous POISSON equation with appropriate boundary conditions. As soon as $\omega > \pi$, a corner singularity arises in the solution. The extreme case $\omega = 2\pi$ describes a circular disk with a crack. Such problems are of great importance in mechanical and civil engineering.

2.2 Finite Differences for POISSON's Equation

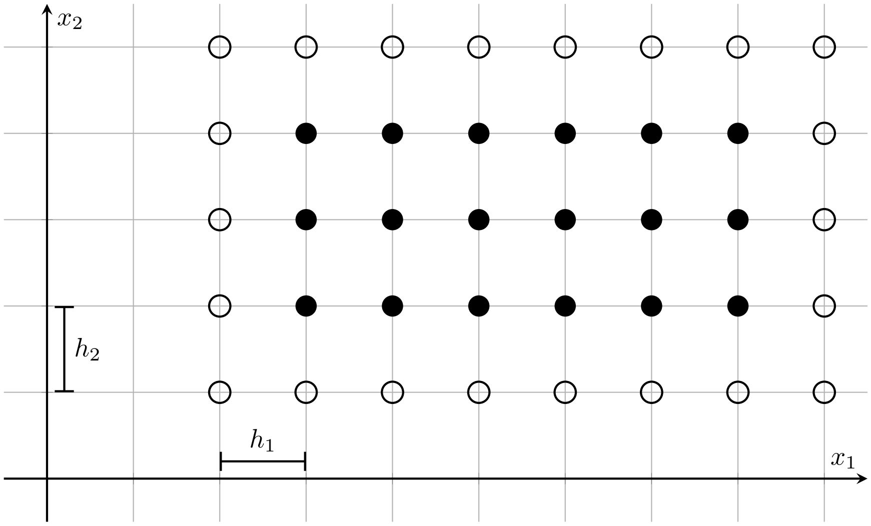
The finite difference method yields a discrete approximation of the *strong formulation* of a PDE. A problem is discretised in three steps:

1. mesh Ω with a (normally Cartesian) grid Ω^h
2. approximate all derivatives with difference quotients
3. set up a system of equations $L^h u^h = f^h$ for the unknown function values u^h on Ω^h

We consider the prototypical elliptic boundary value problem

$$-\Delta u = f \qquad \text{in } \Omega \qquad\qquad u = g \qquad \text{on } \partial\Omega \tag{2.9}$$

on a rectangular domain $\Omega \subset \mathbb{R}^2$ that is aligned with the coordinate axes and discretised by a Cartesian grid Ω^h. We assume that Ω^h is equidistant in each direction, with a constant grid spacing of $h_1 > 0$ in x_1-direction and a constant grid spacing of $h_2 > 0$ in x_2-direction. For the discrete domain with the boundary points included, we use the notation $\bar{\Omega}^h$.



Approximating Derivatives with Difference Quotients

Our objective is to find an approximation to $-\Delta u(x_1, x_2)$, where $(x_1, x_2) \in \Omega^h$. To this end, we approximate the first partial derivatives at two intermediate points by the central difference quotients

$$\frac{\partial u}{\partial x_1}\left(x_1 - \frac{h_1}{2}, x_2\right) \approx \frac{u(x_1, x_2) - u(x_1 - h_1, x_2)}{h_1}$$
$$\frac{\partial u}{\partial x_1}\left(x_1 + \frac{h_1}{2}, x_2\right) \approx \frac{u(x_1 + h_1, x_2) - u(x_1, x_2)}{h_1}$$

and another divided difference using these two quotients yields

$$\frac{\partial^2 u}{\partial x_1^2}(x_1, x_2) \approx \frac{u(x_1 - h_1, x_2) - 2u(x_1, x_2) + u(x_1 + h_1, x_2)}{h_1^2}.$$

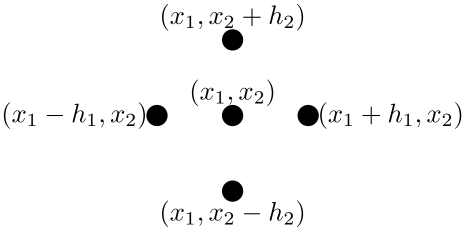
Hence,

$$-\frac{\partial^2 u}{\partial x_1^2}(x_1, x_2) \approx \frac{-u(x_1 - h_1, x_2) + 2u(x_1, x_2) - u(x_1 + h_1, x_2)}{h_1^2} \tag{2.10}$$

and analogously we obtain

$$-\frac{\partial^2 u}{\partial x_2^2}(x_1, x_2) \approx \frac{-u(x_1, x_2 - h_2) + 2u(x_1, x_2) - u(x_1, x_2 + h_2)}{h_2^2} \tag{2.11}$$

The last two equations show that the discrete Laplacian evaluated at the grid point (x_1, x_2) depends on the function values at the point (x_1, x_2) itself plus the four neighbouring grid points $(x_1 \pm h_1, x_2 \pm h_2)$. This dependency pattern is referred to as a *5-point stencil.*



For a computational implementation of the finite difference method, it is desirable to write the difference equations for the unknown function values of u^h on the grid points of Ω^h in matrix form.

If the points in Ω^h are ordered row-wise from bottom left to top right, then (2.10) leads to

$$\frac{1}{h_1^2}\begin{pmatrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{pmatrix}\begin{pmatrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{pmatrix} + \frac{1}{h_1^2}\begin{pmatrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{pmatrix}.$$

Similarly, (2.11) applied to all interior grid points results in

$$\frac{1}{h_2^2}\left(\qquad\qquad\qquad\qquad\right)\left(\qquad\right)+\frac{1}{h_2^2}\left(\qquad\right).$$

The complete linear system that represents the discrete counterpart of (2.9) is now given by

$$(L_1^h + L_2^h)u^h = f^h - (l_1^h + l_2^h) \tag{2.12}$$

where the entries of the vector f^h are the function values of the right hand side f evaluated on each grid point.

It is important to note that the discrete negative Laplacian $L^h = L_1^h + L_2^h$ is represented by a *sparse matrix*. If L^h is an $N \times N$ matrix, it contains a total number of N^2 entries. The 5-point stencil of the central differencing scheme reveals that out of the N entries in each row, at most 5 can be non-zero. For points near the boundary, the stencil includes known boundary values. Therefore, the corresponding rows of L^h have 3 or 4 non-zero entries only. Rows corresponding to grid points further away from the boundary contain 5 non-zero entries.

Consequently, only $O(N)$ out of the N^2 entries of L^h are non-zero. For this purpose, software packages for scientific computing usually offer a special data type for sparse matrices where only the non-zero entries and their indices are stored, but not all the $O(N^2)$ zero entries. Since PDE problems lead to systems with very large N ($N \sim 10^7 - 10^9$ for most industrial problems), the memory requirements for the full matrix would be excessive. Furthermore, the multiplication of a sparse matrix with a vector involves only $O(N)$ multiplications of non-zero numbers. If the sparse matrix had been stored fully, a computer would still carry out all N^2 multiplications of matrix entries with vector entries, despite the fact that almost all of these multiplications give zero. Sparse data formats hence allow for an economical use of both memory and CPU resources.

Fundamental Notions of the Numerical Analysis of PDEs

Three closely related notions describe a numerical scheme $T^h(u^h) = 0$ that discretises a problem $T(u) = 0$ (e.g. $T(u) = Lu - f$):

(1) *Consistency:* Is the discretisation scheme T^h a good approximation of the exact problem T?

(2) *Stability:* Is the discrete problem $T^h(u^h) = 0$ well-posed and does a small residual $T^h(v^h)$ imply a small error $v^h - u$?

(3) *Convergence:* Is the discrete solution u^h a good approximation of the exact solution u?

2.2.1 Definition (Consistency) The numerical scheme T^h is said to be

1. *consistent*, if for every solution of $T(u) = 0$

$$\|T^h(u) - T(u)\| \to 0 \qquad \text{as } h \to 0.$$

2. *consistent of order* $O(h^p)$, if additionally

$$\|T^h(u) - T(u)\| = O(h^p) \qquad \text{as } h \to 0.$$

Note that since $T(u) = 0$, we could have equally written $\|T^h(u) - T(u)\| = \|T^h(u)\|$, but the above notation is probably more illustrative.

2.2.2 Definition (Stability) The numerical scheme T^h is said to be *stable* (with respect to h) if there are constants $h_0 > 0$ and $C > 0$ such that $T^h(u^h) = 0$ has a unique solution for all $h \in \,]0, h_0]$ and if additionally the stability inequality

$$\|u^h - v^h\| \leq C\|T^h(u^h) - T^h(v^h)\|$$

holds for all discrete functions v^h and all $h \in \,]0, h_0]$.

Note that since $T^h(u^h) = 0$, we could have equally written $\|T^h(u^h) - T^h(v^h)\| = \|T^h(v^h)\|$, but the above notation is probably more illustrative.

2.2.3 Definition (Convergence) The numerical scheme T^h is said to be

1. *convergent*, if there exists a constant $h_0 > 0$ such that $T^h(u^h) = 0$ has a unique solution for all $h \in \,]0, h_0]$ and if with the solution of $T(u) = 0$

$$\|u^h - u\| \to 0 \qquad \text{as } h \to 0.$$

2. *convergent of order* $O(h^p)$, if additionally

$$\|u^h - u\| = O(h^p) \qquad \text{as } h \to 0.$$

2.2.4 Theorem (Consistency $\wedge$ Stability $\Rightarrow$ Convergence) *If the scheme T^h is*

(a) consistent and stable, then T^h is also convergent;

(b) consistent of order p and stable, then T^h is also convergent of order p.

Proof. For a scheme that is stable, we have existence and uniqueness of a discrete solution u^h on sufficiently fine grids along with the estimate

$$\|u^h - u\| \leq C\|T^h(u^h) - T^h(u)\| = \|T^h(u)\|$$

where the right hand side $\to 0$ (or $= O(h^p)$) if the scheme is consistent (of order $O(h^p)$). □

Consistency of Finite Difference Methods

So far, we have simply written down an ad hoc approximation of (2.9). We shall now (i) confirm that the approximations made do in fact lead to a consistent approximation and (ii) also find a more general, systematic approach for deriving a finite difference scheme. This approach relies on TAYLOR expansions.

For POISSON's equation over a one-dimensional domain Ω, we use the finite difference approximation

$$-u''(x) \approx \frac{-u(x-h) + 2u(x) - u(x+h)}{h^2}.$$

For a solution $u \in C^4(\bar{\Omega})$, TAYLOR expansion yields

$$u(x \pm h) = u(x) \pm hu'(x) + \frac{h^2}{2}u''(x) \pm \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\xi_\pm)$$

where the fourth derivative is evaluated at some point $\xi_+ \in]x, x+h[$ or $\xi_- \in]x-h, x[$, respectively. For the difference between the exact second derivative and its finite difference approximation, the so-called *truncation error*, we obtain

$$\begin{aligned}\left|-u''(x) - \frac{-u(x-h) + 2u(x) - u(x+h)}{h^2}\right| &= \left|-u''(x) - \frac{-\frac{h^2}{2}u''(x) - \frac{h^4}{24}u^{(4)}(\xi_-) - \frac{h^2}{2}u''(x) - \frac{h^4}{24}u^{(4)}(\xi_+)}{h^2}\right| \\ &= \left|\frac{h^2}{24}u^{(4)}(\xi_-) + \frac{h^2}{24}u^{(4)}(\xi_+)\right| \\ &\le \frac{h^2}{12}\max_{[x-h,x+h]}|u^{(4)}|.\end{aligned}$$

Note that $-u''(x) = f(x)$ and hence we have derived the following result:

$$\|T^h(u)\|_{C(\bar{\Omega})} = |L^h u|_{\Omega^h} - f^h|_\infty \le \frac{h^2}{12}\max_{\bar{\Omega}}|u^{(4)}|$$

i.e. second-order consistency, provided that the exact solution u of the problem is in $C^4(\bar{\Omega})$.

For a 2D domain, we follow the exact same steps, once for the partial derivatives in x_1-direction and once for the partial derivatives in x_2-direction. Then we add up the two truncation errors and obtain

2.2.5 Lemma (Consistency on Equidistant Grids) *Let $\Omega^h \subset \mathbb{R}^2$ be a rectangular grid with constant grid spacing h_1 and h_2 in x_1- and x_2-direction, respectively. Then the finite difference discretisation (2.12) for the* POISSON-DIRICHLET *problem is 2^{nd} order consistent, provided that the solution u of the continuous problem is in $C^4(\bar{\Omega})$:*

$$\left|L^h u|_{\Omega^h} - f^h\right|_\infty \le \frac{h^2}{6}\max_{\bar{\Omega}}\{\,|\partial_{x_1}^4 u|, |\partial_{x_2}^4 u|\,\} \tag{2.13}$$

with $h = \max\{\,h_1, h_2\,\}$.

Stability of Finite Difference Methods

For the finite difference method on equidistant grids, the discrete matrix possesses a lot of structure with repeating patterns. For the negative Laplacian on $\Omega =]0,1[$ with DIRICHLET boundary conditions, we have

$$L^h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & & \ddots & \\ & & -1 & 2 \end{pmatrix}$$

and is eigenvalues can be computed analytically:

$$\lambda_i = \frac{4}{h^2}\sin^2\left(\frac{N-i}{N}\frac{\pi}{2}\right) \qquad i = 1,\dots,N-1.$$

Here $N = 1/h$ is the number of subintervals.

First of all, we observe that all eigenvalues are positive. Consequently, the linear system of the discrete problem has a unique solution.

Furthermore, with any discrete function v^h we derive

$$\begin{aligned} & L^h(u^h - v^h) = f^h - L^h v^h \\ \Rightarrow \quad & u^h - v^h = (L^h)^{-1}(f^h - L^h v^h) \\ \Rightarrow \quad & |u^h - v^h| \le \|(L^h)^{-1}\||(f^h - L^h v^h)| \le \frac{1}{\lambda_{\min}}|(f^h - L^h v^h)| \end{aligned}$$

where

$$\lambda_{\min} = \lambda_{N-1} = \frac{4}{h^2}\sin^2\left(\frac{1}{N}\frac{\pi}{2}\right). \tag{2.14}$$

On fine grids with small h and large N,

$$\lambda_{\min} \approx \frac{4}{h^2}\left(\frac{1}{N}\frac{\pi}{2}\right)^2 = \pi^2, \tag{2.15}$$

so the eigenvalues remain bounded away from zero and we have derived the stability inequality

$$|u^h - v^h| \le C|(f^h - L^h v^h)| = C\|T^h(v^h)\|.$$

If we use the maximum norm $|\cdot|_\infty$ instead of the Euclidean norm $|\cdot|$, the constant C in this inequality may change.

On a rectangular domain in 2D, one may use separation of variables and then apply the same arguments.

2.2.6 Theorem (Stability Inequality) *For the discrete* POISSON-DIRICHLET *problem (2.12) we have the stability inequality*

$$|u^h - v^h|_\infty \le C|L^h v^h - f^h|_\infty \tag{2.16}$$

for all grid functions v^h and with a constant C that is independent of h.

Convergence of Finite Difference Methods

2.2.7 Theorem (Convergence on Equidistant Grids) *Let $\Omega^h \subset \mathbb{R}^2$ be a rectangular grid with constant grid spacing h_1 and h_2 in x_1- and x_2-direction, respectively. Then the finite difference discretisation (2.12) for the* POISSON-DIRICHLET *problem is 2^{nd} order convergent, provided that the solution u to the continuous problem is in $C^4(\bar{\Omega})$.*

Proof. Second-order consistency & stability ⇒ second-order convergence. □

2.2.8 Remark (C^4-Regularity up to the Boundary) The assumption that the analytical solution belongs to the space $C^4(\bar{\Omega})$ is not normally satisfied, as such a high regularity of the solution usually requires a sufficiently regular domain, with no corners. Even for the problem

$$-\Delta u = 1 \qquad \text{in } \Omega \qquad\qquad u = 0 \qquad \text{on } \partial\Omega$$

with C^∞-data, $u \notin C^4(\bar{\Omega})$ if Ω is, for example, the square $]0,1[^2$: assuming that u is four times continuously differentiable up to the boundary, then the PDE prescribes

$$-\Delta u|_{x=0} =$$

but the boundary condition implies

$$-\Delta u|_{x=0} =$$

Consequently, even with the perfectly smooth data in this example, the corners in the domain do not even allow for second derivatives that are continuous up to the boundary.

Discrete Maximum Principle

We will now have a closer look at the structure of the discretised POISSON-DIRICHLET (and other elliptic) problems. The better we understand the properties of the 'big linear system' $L^h u^h = f^h$, the better we will understand what characteristic features of the continuous problem $Lu = f$ are preserved under a 'suitable' discretisation scheme. Furthermore, we will later use all the information that we can possibly gather on the matrix L^h to select a numerical method for solving the discrete problem that is guaranteed to converge, and that additionally exploits all the structure of L^h to compute u^h as efficiently as somehow possible.

2.2.9 Definition (Diagonal Dominance) If in the i^{th} row of a matrix $A \in \mathbb{R}^{n \times n}$ the absolute value of the diagonal entry is

- greater than or equal to the sum of the absolute values of the off-diagonal terms

  then we say that A is *weakly diagonally dominant* in this row;
- greater than the sum of the absolute values of the off-diagonal terms

then we say that A is *strictly diagonally dominant* in this row.

A matrix A with the properties that

(i) A is weakly diagonally dominant in all rows

(ii) A is strictly diagonally dominant in at least one row

(iii) for all rows i_0 there exists a chain of indices $i_0 \to i_1 \to \cdots \to i_s$ to a strictly diagonally dominant row i_s such that all $a_{i_{l-1}i_l} \neq 0$ $(l = 1, \dots, s)$

is called *weakly chained diagonally dominant*. If, instead of (iii), there even holds that

(iv) for any two rows i_0, i_s there exists a chain of indices $i_0 \to i_1 \to \cdots \to i_s$ such that all $a_{i_{l-1}i_l} \neq 0$ and all $a_{i_l i_{l-1}} \neq 0$ $(l = 1, \dots, s)$

then A is called *irreducibly diagonally dominant*.

The two chain properties describe how data from the right hand side and information from each component of the solution propagate through the linear system. As can be seen from the conditions (iii) and (iv), they describe the structure of the sparsity pattern of A.

With the weaker chain property (iii), information from row i_s is referred to in row i_{s-1}. Then the equation in row i_{s-2} refers to the component i_{s-1} of the solution, and hence indirectly also to i_s. Finally, row i_0 directly or indirectly depends on the components $i_1, i_2, \dots, i_s$ of the solution and the right hand side. It is not required for (iii) that conversely row i_s also depends on row i_0.

This is the difference to the stronger property (iv). Here, information is shared globally and all rows directly or indirectly refer to themselves and all other rows.

2.2.10 Definition (Monotone Matrix) A matrix $A \in \mathbb{R}^{n \times n}$ is said to be *(inverse-)monotone* if for all vectors $u \in \mathbb{R}^n$

Recall that a matrix A is monotone if and only if A^{-1} exists and all its entries are positive or zero: $(A^{-1})_{ij} \geq 0, \forall i, j = 1, \dots, n$.

2.2.11 Definition (Z-, L- and M-Matrices) A matrix $A \in \mathbb{R}^{n \times n}$ is called

- *Z-matrix* or *L_0-matrix* if all off-diagonal entries are negative or zero:

- *L-matrix* if all off-diagonal entries are negative or zero and all diagonal entries are positive:

- *M-matrix*, if it is a monotone Z-matrix.

There are many characterisations of M-matrices. The following sufficient condition is the most illustrative one in the context of discretised elliptic PDEs:

2.2.12 Lemma (M-Criterion) *If $A \in \mathbb{R}^{n\times n}$ is a weakly chained diagonally dominant L-matrix, then A is also monotone and hence an M-matrix.*

2.2.13 Lemma (Discrete Laplacian is an M-Matrix) *The discretised elliptic operator L^h in the* Poisson-Dirichlet *problem (2.12) is an M-matrix, independent of h.*

Proof. The matrix L^h is

- strongly diagonally dominant in all rows corresponding to grid points adjacent to the boundary
- weakly, but not strongly diagonally dominant in all rows corresponding to the other grid points 'further away' from the boundary

The matrix L^h is irreducibly diagonally dominant[†], and in particular it is weakly chained diagonally dominant[‡]. Furthermore, all off-diagonal entries are non-negative while all diagonal entries are positive and hence L^h is an L-matrix. The M-criterion now implies that L^h is an M-matrix, as asserted. □

2.2.14 Theorem (Discrete Elliptic Maximum Principle) *Let the discretised elliptic operator $L^h \in \mathbb{R}^{n\times n}$ be a weakly chained diagonally dominant L-matrix. Then*

$$L^h u^h \le 0 \qquad \text{in } \Omega^h \qquad\qquad \Rightarrow \qquad\qquad \max_{x\in\bar{\Omega}^h} u^h(x) \le \max_{x\in\partial\Omega^h} u^h(x),$$

i.e. the discrete solution u^h assumes its maximum on the boundary.

Furthermore, the strong connectivity (irreducibility) of the system matrix L^h reveals that even a *local* change in only one component of the right hand side f^h immediately affects the entire solution u^h *globally*. This phenomenon of immediate global propagation of information through the entire domain is characteristic for elliptic equations.

2.2.15 Remark (Advantages and Disadvantages of Finite Difference Methods) $\oplus$ Finite difference methods are easy to set up and implement on equidistant, rectangular grids.

$\oplus$ In these cases, the discrete matrices possess a lot of structure, which can be exploited for efficient solution algorithms of the resulting linear systems.

$\oplus$ Some finite difference methods accurately reflect characteristic features of an elliptic PDE, e.g. well-posedness, maximum principle, symmetry, positive definiteness, global propagation of information.

$\ominus$ The equations become very complicated on non-equidistant grids or more complex domains. Then much of the structure of the discrete matrices is lost as well.

$\ominus$ Finite difference methods only approximate some point values of the solution, they do not return a function that is defined over the entire domain.

$\ominus$ C^4 regularity up to the boundary is required to guarantee 2nd order convergence for the Poisson equation. This is extremely unrealistic in practice.

[†] "One can walk from any interior grid point to any other interior grid point, taking only steps that are covered by the 5-point stencil."

[‡] "One can walk from any interior grid point to a point adjacent to the boundary where the matrix is strictly diagonally dominant."

2.3 Finite Elements for Poisson's Equation

The finite element method follows an approach which is completely different from the finite difference method. Instead of the strong formulation of a PDE, finite element discretisations start from the *weak formulation* of a PDE. For a linear elliptic equation in weak form, a solution $u \in V$ satisfies an expression of the form

$$B(u, v) = \langle f, v \rangle, \qquad \forall v \in V \tag{2.17}$$

with a bilinear form $B : V \times V \to \mathbb{R}$.

There are two common ways to arrive at a weak form of a given problem, one of which we already know:

We formulate a number of objectives that a 'good' finite element method should attain:

The central idea of finite element methods is:

To obtain a discrete problem

$$L^h u^h = f^h$$

from (2.17), we proceed in three steps:

1st Step Choose an N-dimensional subspace V^h and a basis $\left(\phi_i^h\right)_{i=1}^N$

2nd Step Write

$$u^h =$$

3d Step Substitute u^h in (2.17) to obtain the GALERKIN equations

Linear Finite Elements

To give an overview of a discretisation with finite elements for a simple example, we re-visit the elliptic model problem

$$-\varDelta u = f \qquad \text{in } \varOmega \qquad\qquad u = 0 \qquad \text{on } \partial\varOmega \tag{2.18}$$

in its weak form with a given right hand side $f \in L^2(\varOmega)$

$$\int\limits_{\varOmega} \nabla u \cdot \nabla v \,\mathrm{d}x = \int\limits_{\varOmega} f v \,\mathrm{d}x, \qquad \forall v \in H_0^1(\varOmega). \tag{2.19}$$

First of all, we have to fix a suitable subspace $V^h \subset H_0^1(\varOmega)$. To that end, we approximate the domain $\varOmega$ with a union of closed triangles T_i from a *regular triangulation* $\mathcal{T}^h = \{\, T_i \mid i = 1, \dots, n_{\mathrm{T}} \,\}$. 'Regular' means in this context: for any two triangles from $T, T' \in \mathcal{T}^h$ the intersection $T \cap T'$ is either empty, one common corner point of T and T' or one common edge of T and T'.

For the mesh size parameter we define $h = \max_{T \in \mathcal{T}^h} \operatorname{diam} T$.

On the discrete domain $\bar{\Omega}^h = \bigcup_{T \in \mathcal{T}^h} T$ we now introduce the space of linear finite elements:

A basis of this space is given by the functions $\phi_i^h \in V^h$ defined by

Moving on to the second step, we may now decompose any function $v^h \in V^h$ (including the discrete solution u^h) in this basis as

$$v^h = \sum_{j=1}^{N} v^h(p_j)\phi_j^h.$$

Obviously, v^h is fully determined by its function values $v^h(p_j)$ on the interior grid points p_j. Therefore, we also collect these nodal values in a column vector

$$\vec{v}^h = \begin{pmatrix} v^h(p_1) \\ \vdots \\ v^h(p_N) \end{pmatrix}$$

for an alternative representation of v^h.

Thirdly and lastly, the Galerkin equations for the model problem discretised with linear finite elements read

(2.20)

Note that due to the linearity of (2.19) in v, it is sufficient to only use the N basis functions ϕ_i as test functions instead of all infinitely many functions $v^h \in V^h$.

The Galerkin equations (2.20) can be written in matrix form with

2.3.1 Example (Linear Finite Elements in 1D) In one dimension, the model problem reads: find $u \in H_0^1(]0,1[)$ such that for all $v \in H_0^1(]0,1[)$:

$$\int\limits_0^1 u'v' \,\mathrm{d}x = \int\limits_0^1 fv \,\mathrm{d}x.$$

We discretise this problem on the equidistant grid

$$0, h, 2h, 3h, \ldots, (N-1)h, 1$$

with N subintervals and grid spacing $h = 1/N$.

Practical Implementation

To evaluate or approximate some integrals that arise in a finite-element discretisation, we often need quadrature formulae, i.e. numerical approximations of integrals. Quadrature formulae are of the form

$$\int_{\Omega} g(x)\,\mathrm{d}x \approx |\Omega| \sum_{i=1}^{n} w_i g(x^i)$$

where w_i are the weights and x^i the nodes of the quadrature formula.

2.3.2 Theorem (Quadrature Formulae on Intervals)

	Sketch	*Nodes*	*Weights*	*Error*
Midpoint Rule				
Trapezoidal Rule				
SIMPSON*'s Rule*				

At this stage, it is not quite clear yet which quadrature formula should be chosen for what problem. In our convergence analysis of finite element methods, we will derive a rule that will tell us how accurately we have to solve the integrals in the stiffness matrix and on the right hand side. Clearly, we would not want to worsen the convergence rate of the finite element method through too large integration errors. On another hand, we would not want to integrate ‘too accurately’ if there is a quadrature rule of lower order that already achieves the same convergence rate with less computational effort.

The midpoint rule and the trapezoidal rule extend naturally to higher spatial dimensions:

2.3.3 Theorem (Quadrature Formulae on Triangles)

	Sketch	*Nodes*	*Weights*	*Error*
Midpoint Rule				
Trapezoidal Rule				
Cubic Rule				

In practice, it is often easiest to assemble the stiffness matrix, the mass matrix, the right hand side and any other required discrete operators element-by-element. Note that we can split the integration over all of Ω^h into integrals over one triangle only:

We collect the (very few) nonzero contributions from a single element in an *element stiffness matrix* or *element mass matrix*.

2.3.4 Example (Element Mass Matrix) Let T be a triangle and $\phi_1^h, \phi_2^h, \phi_3^h$ the three hat functions that are nonzero on T.

2.3.5 Example (Element Stiffness Matrix) If T has the corner points p_1, p_2, p_3 with corresponding hat functions $\phi_1^h, \phi_2^h, \phi_3^h$ such that $\phi_i^h(p_j) = \delta_{ij}$, then

$$K_T^h = \frac{1}{4|T|} \begin{pmatrix} {d_{32}^h}^\top \\ {d_{13}^h}^\top \\ {d_{21}^h}^\top \end{pmatrix} \begin{pmatrix} d_{32}^h & d_{13}^h & d_{21}^h \end{pmatrix} = \frac{1}{4|T|} \begin{pmatrix} d_{32}^h \cdot d_{32}^h & d_{32}^h \cdot d_{13}^h & d_{32}^h \cdot d_{21}^h \\ d_{13}^h \cdot d_{32}^h & d_{13}^h \cdot d_{13}^h & d_{13}^h \cdot d_{21}^h \\ d_{21}^h \cdot d_{32}^h & d_{21}^h \cdot d_{13}^h & d_{21}^h \cdot d_{21}^h \end{pmatrix}$$

with the triangle edge vectors

$$d_{32}^h = p_3 - p_2 \qquad d_{13}^h = p_1 - p_3 \qquad d_{21}^h = p_2 - p_1$$

and the triangle area

Proof. An affine function u, i.e. a function of the form

$$u(x) = \alpha x_1 + \beta x_2 + \gamma$$

that is defined by its values on the three vertices $a = p_1, b = p_2, c = p_3 \in \mathbb{R}^2$ of a non-degenerate triangle T (they are not all on one line)

$$u(a) = u_a \qquad u(b) = u_b \qquad u(c) = u_c$$

satisfies the equation

$$\begin{aligned} u(x) = {} & \frac{b_1 c_2 - c_1 b_2 + (b_2 - c_2)x_1 + (c_1 - b_1)x_2}{2|T|} u_a \\ & + \frac{c_1 a_2 - a_1 c_2 + (c_2 - a_2)x_1 + (a_1 - c_1)x_2}{2|T|} u_b \\ & + \frac{a_1 b_2 - b_1 a_2 + (a_2 - b_2)x_1 + (b_1 - a_1)x_2}{2|T|} u_c . \end{aligned} \tag{2.21}$$

The gradient of u is constant:

$$\nabla u(x) = \frac{1}{4|T|} \begin{pmatrix} (b_2 - c_2)u_a + (c_2 - a_2)u_b + (a_2 - b_2)u_c \\ (c_1 - b_1)u_a + (a_1 - c_1)u_b + (b_1 - a_1)u_c \end{pmatrix} . \tag{2.22}$$

If u is a hat function (restricted to the triangle T), then one of u_a, u_b, u_c is one while the other two are zero. We may now evaluate the integrals

$$\int\limits_T \nabla \phi_i^h \cdot \nabla \phi_2^h \, \mathrm{d}x$$

e.g. with the midpoint rule (which is exact for constant functions) for all nine possible combinations of $i, j \in \{1, 2, 3\}$ and this yields the above element stiffness matrix. □

Software packages for finite element computations typically store the mesh data in three arrays:

- an array $P \in \mathbb{R}^{2 \times n_P}$ with the coordinates of all grid *points*, such that the k-th column of P defines coordinates p_k of the k-th point ($k = 1, \dots, n_P$)

- an array $E \in \{\,1, \dots, n_P\,\}^{2\times n_E}$ with the indices of points on the boundary *edges*, such that the k-th edge segment $e_{1k} \to e_{2k}$ coincides with the edge of a triangle ($k = 1, \dots, n_E$)
- an array $T \in \{\,1, \dots, n_P\,\}^{3\times n_T}$ defining the triangulation, such that t_{1k}, t_{2k}, t_{3k} are the indices of the corner points of the k-th *triangle* in anticlockwise order ($k = 1, \dots, n_T$).

Whenever DIRICHLET conditions are imposed on (parts of) the boundary, we want to eliminate these from the system of equations. This is easily achieved by means of projection matrices $P_f \in \{\,0, 1\,\}^{N\times n_P}$ and $P_{\mathrm{D}} \in \{\,0, 1\,\}^{(n_P-N)\times n_P}$. P_f projects a vector $\bar{\vec{u}}^h \in \mathbb{R}^{n_P}$ onto its N components $\vec{u}^h \in \mathbb{R}^N$ that are actual degrees of freedom, i.e. points in the interior and boundary points on which no DIRICHLET conditions are imposed. P_{D} projects a vector $\bar{\vec{u}}^h \in \mathbb{R}^{n_P}$ onto its components, for which boundary values are already prescribed.

If $\bar{K}^h, \bar{M}^h \in \mathbb{R}^{n_P\times n_P}$ are the stiffness and mass matrices and $\bar{f}^h$ the load vector on all of $\bar{\Omega}^h$, and $g^h \in \mathbb{R}^{n_P-N}$ are the DIRICHLET boundary values, then

$$\begin{aligned} \bar{\vec{u}}^h &= \\ \bar{K}^h \bar{\vec{u}}^h &= \\ \bar{M}^h \bar{\vec{u}}^h &= \end{aligned}$$

Recall that the test functions vanish on that part of the boundary, where DIRICHLET conditions apply. Therefore, we also have to delete the $n_P - N$ equations for these DIRICHLET points. For example, the finite element discretisation of the problem

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

reads

$$K^h \vec{u}^h = f^h - k_{\mathrm{D}}^h \tag{2.23}$$

with the (reduced) stiffness matrix and load vector

$$\begin{aligned} K^h &= \\ f^h &= \end{aligned}$$

and stiffness terms from any inhomogeneous DIRICHLET boundary values

$$k_{\mathrm{D}}^h =$$

As above, the boundary values from g can be added to the solution $\vec{u}^h$ by setting

$$\bar{\vec{u}}^h =$$

General Galerkin Methods

Besides subspaces of piecewise linear functions, there are many more types of finite-dimensional subspaces V^h of a function space V that one could use for a Galerkin approximation:

Spaces of Piecewise Polynomial Functions Such spaces V^h are defined based on a mesh of simple polytopes (called *cells*), e.g. intervals in 1D, triangles or quadrilaterals in 2D, tetrahedra or hexahedra in 3D. V^h is the set of all functions that belong to a certain class of polynomials when restricted to one cell, and which often have to meet additional conditions on continuity across cell interfaces. One may choose basis functions for V^h that only have local support.

Galerkin approximations with these discrete spaces of locally polynomial functions are called *finite element methods*.

Spaces of Polynomials An alternative approach would be to consider spaces of (globally) polynomial functions over a (hyper-)rectangular domain. On the unit square $\Omega =]0,1[^2$ these spaces are of the form

$$V^h = \left\{ v^h : \Omega \to \mathbb{R} \;\middle|\; v^h(x) = \sum_{i,j=0}^{m} a_{ij} x_1^i x_2^j \right\}.$$

Taking tensor products of the monomials, i.e. functions like $x_1^i x_2^j$ as basis vectors is not practicable, since no sparsity can be expected for the mass or stiffness matrix and—even worse—since the stiffness matrix is a Hilbert-type matrix. The condition of Hilbert matrices grows exponentially with their dimension N. Even for small values such as $N = 10$, the 10×10 Hilbert matrix already possesses a condition number of $> 10^{13}$. Numerical solutions to that poorly conditioned linear systems are worthless.

Tensor products of polynomials that are orthogonal in $L^2(]0,1[)$, such as the Chebychev or Legendre polynomials lead to well-conditioned mass and stiffness matrices with condition numbers 1 and $O(N)$.

Galerkin approximations with spaces of Chebychev or Legendre polynomials are sometimes classified as *spectral methods*.

Spaces of Trigonometric Functions Again on the unit square $\Omega =]0,1[^2$, these spaces contain truncated Fourier series. To approximate a problem with homogeneous Dirichlet boundary conditions, the space

$$V^h = \left\{ v^h : \Omega \to \mathbb{R} \;\middle|\; v^h(x) = \sum_{i,j=0}^{m} a_{ij} \sin(i\pi x_1)\sin(j\pi x_2) \right\}$$

is an admissible choice with basis vectors $\sin(i\pi x_1)\sin(j\pi x_2)$.

Galerkin approximations with trigonometric functions are called *spectral methods*.

2.3.6 Remark (Petrov-Galerkin Methods) The space used for approximating the solution u^h and the space of test functions do not necessarily have to be the same V^h. The generalisation of Galerkin methods with a space U^h of shape functions and another space V^h of test functions is known as a Petrov-Galerkin method.

2.3.7 Definition (Finite Element) A *finite element* (in the sense of Ciarlet) is defined as a triple (T, P, L) consisting of

- a bounded closed subset $T \subset \mathbb{R}^d$ with nonempty interior and piecewise smooth boundary
- a finite-dimensional space $P = P(T)$ of functions (normally polynomials) over T
- a set of degrees of freedom (or node functionals) $L = L(T)$ that forms a basis for the dual space P^*.

Recall that the dual space is the set of all linear and bounded functionals. Most commonly, these degrees of freedom are set by means of pointwise evaluations of function values or derivatives (and mappings $p \mapsto p(x_0)$, or $p \mapsto \nabla p(x_0)$ are indeed bounded linear functionals on P).

The property that the degrees of freedom must form a basis of P^*, i.e. that they uniquely determine every function in P, is also referred to as *unisolvence*.

2.3.8 Definition (Lagrange and Hermite Elements) If the degrees of freedom of a finite element (T, P, L) only evaluate function values of polynomials in P, then this element is called a Lagrange element.

If the degrees of freedom also include evaluations of derivatives of polynomials in P, then this element is called an Hermite element.

2.3.9 Definition (Interpolation with Finite Elements) Let (T, P, L) be a finite element. We introduce an interpolation operator

$$I^h : H^m(T) \to P(T), v \mapsto I^h v$$

where the interpolant $I^h v$ is defined by

$$\ell(v) = \ell(I^h v), \qquad \forall (\text{finitely many}) \; \ell \in L.$$

By combining the patches from (T, P, L) to a larger domain and composite functions, we obtain a finite-element space V^h. This space may possibly impose further restrictions on continuity or boundary conditions. This could be implemented by equating the degrees of freedom on neighbouring elements or prescribing their values on the boundary, respectively.

2.3.10 Definition (Conforming and Nonconforming Elements) A finite-element space V^h is said to be a *conforming* approximation of a function space V if $V^h \subset V$, otherwise the approximation is called *nonconforming*.

2.3.11 Example (Higher-Order Elements in 1D)

2.3.12 Example (Common Finite Elements in 2D) **Piecewise Constant**

T	P	L
triangle	$P_0(T)$, $\dim P_0(T) = 1$	
quadrilateral	$Q_0(T)$, $\dim Q_0(T) = 1$	

Piecewise Linear

T	P	L
triangle	$P_1(T)$, $\dim P_1(T) = 3$	
triangle	$P_1(T)$, $\dim P_1(T) = 3$	

Piecewise Quadratic

T	P	L
triangle	$P_2(T)$, $\dim P_2(T) = 6$	
triangle	$P_2(T)$, $\dim P_2(T) = 6$	
quadrilateral	$Q_1(T)$, $\dim Q_1(T) = 4$	

All of these elements possess natural generalisations on tetrahedrons or hexahedrons for three-dimensional problems.

Consistency

A discretisation with finite elements generally introduces errors from three distinct sources:

2.3.13 Theorem (Bramble-Hilbert Lemma) *Let $T \subset \mathbb{R}^d$ be a domain with* Lipschitz *boundary and let $F : H^{k+1}(T) \to \mathbb{R}$ be a bounded and sublinear functional that vanishes for all polynomials of degree $\leq k$:*

- $|F(v)| \leq c_1 \|v\|_{H^{k+1}(T)}$ *for all $v \in H^{k+1}(T)$*
- $|F(u) + F(v)| \leq |F(u)| + |F(v)|$ *for all $u, v \in H^{k+1}(T)$*
- $F(p) = 0$ *for all $p \in P_k(T)$.*

Then there exists a constant $c > 0$ such that

$$F(v) \leq c \|\nabla^{k+1} v\|_{L^2(T)}.$$

Proof. This result can be derived with a few technicalities on polynomial projections and a generalised version of Poincaré's inequality. The full proof is given on pp 224-225 in the book C Grossmann, HG Roos and M Stynes: *Numerical Treatment of Partial Differential Equations.* Springer, 2007. □

2.3.14 Theorem (Interpolation Error on Simplices) *Let (T, P, L) be a* Lagrange *or* Hermite *element, where $T \subset \mathbb{R}^d$ is a d-simplex (interval, triangle, tetrahedron, ...) with inradius r_T and diameter h_T and where $P = P_k(T)$. We denote the corresponding interpolation operator by $I^h : H^{k+1}(T) \to P_k(T)$.*

Then there exists an interpolation constant $c > 0$ depending only on the dimension d and the polynomial degree k such that for all functions $v \in H^{k+1}(T)$ and all $i \in \{ 0, \dots, k+1 \}$

$$\left\|\nabla^i \left(v - I^h v\right)\right\|_{L^2(T)} \leq c \frac{h_T^{k+1}}{r_T^i} \left\|\nabla^{k+1} v\right\|_{L^2(T)}.$$

Proof. We map a simplex T in the domain to a reference simplex $\hat{T}$, apply the Bramble-Hilbert lemma and map back to T. The mapping between $\hat{T}$ and T is affine:

$$F_T(\hat{x}) = A_T \hat{x} + b_T.$$

The change of variables leads to factors of $\det A_T$, $\|A_T\|$ and $\|A_T^{-1}\|$ in the above norms, which can be estimated in terms of h_T or r_T, respectively. The full details can be found pp 225-226 in the book of Grossmann, Roos, Stynes. □

2.3.15 Corollary (Interpolation Error with Finite Elements in 2D) *Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain and $\mathcal{T}^h$ a triangulation on Ω which satisfies the uniform (in h) shape regularity condition*

$$\max_{T \in \mathcal{T}^h} \frac{h_T}{r_T} \leq c. \tag{2.24}$$

Then there exists an interpolation constant $c > 0$ depending only on the regularity constant from (2.24), the dimension d and the polynomial degree k such that

$$\left\|\nabla^i \left(v - I^h v\right)\right\|_{L^2(\Omega)} \leq c h^{k+1-i} \left\|\nabla^{k+1} v\right\|_{L^2(\Omega)}.$$

Proof.

$$
\begin{aligned}
\left\|\nabla^i\left(v - I^h v\right)\right\|^2_{L^2(\Omega)} &= \int\limits_{\Omega} \left|\nabla^i(v - I^h v)\right|^2 \,\mathrm{d}x \\
&= \sum_{T\in\mathcal{T}^h} \int\limits_{T} \left|\nabla^i(v - I^h v)\right|^2 \,\mathrm{d}x \\
&\le \sum_{T\in\mathcal{T}^h} \left(c\frac{h_T^{k+1}}{r_T^i} \left\|\nabla^{k+1} v\right\|_{L^2(T)} \right)^2
\end{aligned}
$$

□

Note the assumption of a polygonal domain in Corollary 2.3.15. This ensures that this domain can be decomposed into triangles and there is no geometric approximation error: $\Omega^h = \Omega$. For curved boundaries, e.g. defined by cubic splines a.k.a. Bézier curves as they are typically used in computer aided design (CAD), an additional error arises from the approximation of Ω by Ω^h. This approximation has to be carried out carefully to not deteriorate the overall order of consistency of the discretisation.

In case of $V = H_0^1(\Omega)$ over a convex domain Ω that is approximated by a polygonal domain Ω^h such that all vertices of $\partial\Omega^h$ lie exactly on $\partial\Omega$, the resulting discrete space V^h is a conforming discretisation of V. If Ω is a non-convex domain with curved boundaries, then V^h would usually be non-conforming. This makes the analysis of the non-convex case significantly more difficult (and more suitable to an advanced finite-element course).

2.3.16 Lemma (Estimate on $\Omega \setminus \Omega^h$) *For a convex domain $\Omega \subset \mathbb{R}^2$ with C^2 boundary that is approximated with a polygonal domain Ω^h, we have*

$$\|w\|_{L^2(\Omega\setminus\Omega^h)} \leq Ch\|w\|_{H^1(\Omega)}$$

for all functions $w \in H^1(\Omega)$.

Proof. □

With this lemma, we conclude

$$\|I^h v - v\|_{L^2(\Omega\setminus\Omega^h)} =$$

$$\|\nabla(I^h v - v)\|_{L^2(\Omega\setminus\Omega^h)} =$$

Comparing this result with the interpolation error, we see that the polygonal approximation of Ω is 'good enough' for linear finite elements in the sense that the interpolation error doesn't decay at a higher rate than the geometric error: the polygonal boundary approximation does not slow down the rate of convergence for linear finite elements.

For quadratic and higher order elements, though, a boundary approximation that is only piecewise linear will reduce the order of consistency of the approximation. The most common technique to avoid this undesirable effect are isoparametric elements:

2.3.17 Definition (Parametric and Isoparametric Elements) A finite element (T, P, L) is said to be *parametric* if there exists a reference element $(\hat{T}, \hat{P}, \hat{L})$ and an invertible transformation $F_T : \hat{T} \to T$ such that

- $T = F_T(\hat{T})$
- for any function $p \in P$ there is a corresponding $\hat{p} \in \hat{P}$ such that $p = \hat{p} \circ F_T^{-1}$
- for any degree of freedom $\ell \in L$ there is a corresponding $\hat{\ell} \in \hat{L}$ such that $\ell(p) = \hat{\ell}(p \circ F_T)$ for all $p \in P$.

The element is said to be *isoparametric* if the d components of the transformation F_T are in $\hat{P}$, i.e. if F_T is a function of the same kind as the shape functions in the reference element.

Stability

We will derive three different stability estimates

We recall the abstract weak formulation of a linear PDE

its Galerkin approximation

and the property of Galerkin orthogonality

In the sequel we assume the bilinear form $B : V \times V \to \mathbb{R}$ to be continuous

and coercive

with respect to some norm $\|\cdot\|$ norm on the space V.

2.3.18 Lemma (Céa) *The error $e^h = u - u^h$ of the (conforming)* Galerkin *approximation satisfies the quasi-best approximation property*

$$\|e^h\| \leq \frac{C}{c} \inf_{v^h \in V^h} \|u - v^h\|.$$

Proof. Let $v^h \in V^h$ be arbitrary.

□

Note that if we choose the energy norm $\|u\| = \|u\|_B = \sqrt{B(u,u)}$, then $C = c = 1$ and CÉA's lemma gives the best approximation property.

Let us now move on to errors due to inexact numerical integration. Whenever we apply a quadrature formula to assemble e.g. the stiffness matrix, mass matrix or load vector, we obtain a perturbed bilinear form $\tilde{B}$ instead of B and a perturbed right hand side $\tilde{f}$ instead of f and we are solving the perturbed GALERKIN equations

$$\tilde{B}(u^h, v^h) = \langle \tilde{f}, v^h \rangle_{V^*,V}, \qquad \forall v^h \in V^h$$

instead. To make sure that this problem still possesses a unique solution, we need the additional assumptions that $\tilde{B}$ is continuous and coercive as well

$$\begin{aligned} |\tilde{B}(u^h, v^h)| &\le \tilde{C}\|u^h\|\|v^h\|, && \forall u^h, v^h \in V^h \\ \tilde{B}(u^h, u^h) &\ge \tilde{c}\|u^h\|^2, && \forall u^h \in V^h \end{aligned}$$

and that $\tilde{f}$ is continuous, just like their unperturbed counterparts B and f.

2.3.19 Lemma (STRANG's First Lemma) *The error $e^h = u^h - \bar{u}$ of the perturbed (but otherwise conforming) GALERKIN approximation satisfies the estimate*

$$\|e^h\| \le c \left(\inf_{v^h \in V^h} \left(\|v^h - \bar{u}\| + \|B(v^h, \cdot) - \tilde{B}(v^h, \cdot)\|_* \right) + \|f - \tilde{f}\|_* \right)$$

with a constant $c > 0$ that is independent of $\bar{u}, u^h$ and h.

Proof. Let $v^h \in V^h$ be arbitrary.

□

Finally, let us add a few remarks regarding nonconforming approximations. In this case we are solving the discrete problem

$$B^h(u^h, v^h) = \langle f^h, v^h\rangle_{V^{h*},V^h}, \qquad \forall v^h \in V^h$$

where $V^h \not\subset V$. We need the extra assumptions that B^h can be defined for arguments from V and conversely that B can be defined for arguments from V^h. Additionally we impose continuity and coercivity of B^h

$$\begin{aligned} |B^h(u, v)| &\leq C^h \|u\|_h \|v\|_h, && \forall u, v \in V + V^h \\ B^h(u^h, u^h) &\geq c^h \|u^h\|_h^2, && \forall u^h \in V^h \end{aligned}$$

where $\|\cdot\|_h$ is some norm on $V + V^h$, the space of all linear combinations $\lambda v + \mu v^h$ with $\lambda, \mu \in \mathbb{R}$, $v \in V$, $v^h \in V^h$. The inhomogeneity f^h is assumed to be continuous in this norm on the space V^h.

We will also need the corresponding operator norm, defined by

$$\|f^h\|_{h*} = \sup_{v^h \in V^h} \frac{|\langle f^h, v^h\rangle_{V^{h*},V^h}|}{\|v^h\|_h}.$$

2.3.20 Lemma (Strang's Second Lemma) *The error $e^h = u^h - \bar{u}$ of the possibly non-conforming finite element approximation satisfies the estimate*

$$\|e^h\|_h \leq c\left(\inf_{v^h \in V^h} \|v^h - \bar{u}\|_h + \|B^h(\bar{u}, \cdot) - f^h\|_{h*}\right)$$

with a constant $c > 0$ that is independent of $\bar{u}, u^h$ and h.

Proof. Let $v^h \in V^h$ be arbitrary.

□

Convergence

We will now move on to the error analysis of finite element discretisations. To obtain an error estimate in the energy norm $\|e^h\|_B$, we can simply apply the default strategy

$$\text{consistency of order } m \wedge \text{stability} \Rightarrow \text{convergence of order } m.$$

The error estimate in the L^2-norm $\|e^h\|_{L^2}$ can be derived from the energy estimate by applying the so-called AUBIN-NITSCHE trick. Estimates in the L^∞-norm require very different techniques and therefore we will present these without proof.

We consider a $H_0^1(\Omega)$-conforming discretisation of POISSON's equation with homogeneous DIRICHLET boundary conditions on a convex, polygonal domain $\Omega = \Omega^h \subset \mathbb{R}^2$.

2.3.21 Theorem (Convergence in the Energy Norm) *Let $V^h \subset V = H_0^1(\Omega)$ be a conforming space of linear finite elements space of piecewise polynomial functions of degree k. Then the error $e^h = u^h - \bar{u}$ for the approximation of the POISSON-DIRICHLET problem satisfies the a priori estimate*

$$\|e^h\|_B \leq ch\|\nabla^2\bar{u}\|_{L^2(\Omega)}$$

and here we even have

$$\|e^h\|_B \leq ch\|f\|_{L^2(\Omega)}$$

with the energy norm $\|e^h\|_B = \|\nabla e^h\|_{L^2}$.

Proof.

□

2.3.22 Theorem (Convergence in the L^2-Norm) *Let $V^h \subset V = H^1_0(\Omega)$ be a conforming finite element space of piecewise polynomial functions of degree k. Then the error $e^h = u^h - \bar{u}$ for the approximation of the* Poisson-Dirichlet *problem satisfies the a priori estimate*

$$\|e^h\|_{L^2} \leq ch^2 \|\nabla^2 \bar{u}\|_{L^2(\Omega)}$$

and we even have

$$\|e^h\|_{L^2} \leq ch^2 \|f\|_{L^2(\Omega)}.$$

Proof. This proof relies on the Aubin-Nitsche trick: consider the so-called dual problem

$$B(v,z) = \left\langle \frac{e^h}{\|e^h\|_{L^2}}, v \right\rangle_{L^2}, \qquad \forall v \in H^1_0(\Omega).$$

In our setting, the strong formulation of this problem reads

$$\begin{aligned} -\Delta z &= \frac{e^h}{\|e^h\|_{L^2}} \qquad \text{in } \Omega \\ z &= 0 \qquad \text{on } \partial\Omega \end{aligned}$$

and from (2.8) we conclude that the solution z belongs to $H^1_0(\Omega) \cap H^2(\Omega)$ with

$$\|z\|_{H^2} \leq c \left\| \frac{e^h}{\|e^h\|_{L^2}} \right\|_{L^2} = c.$$

Now, using the test function $v = e^h$ in the dual problem, we obtain

□

In many practical applications, estimates in the L^2-norm or in the energy norm would not be desirable, as they inherently include some averaging of the error over the entire domain. Even local singularities where the solution blows up may still give a finite error.

In structural engineering, for instance, a local spike in the stress acting on a building or a bridge may result in the failure of the structure. In such settings, pointwise estimates, just like for the finite difference method, are more desirable: