Numerical methods for partial differential equations

Lecture notes

written by Csaba Gáspár, 2020

Contents

1 Introduction

In the vast majority of scientific/engineering problems, the ordinary and/or partial differential equations play a special role. The rigorous description of these differential equations is very difficult in general. Moreover, for practical purposes, the computational solution of these differential equations is much more important than the theoretical investigation. However, the computational solutions are still difficult and require a lot of mathematical tools including special sections from analysis and computational methods i.e. solution of large linear systems of equations.

In this lecture notes, we try to give a compromise between the theoretical mathematics concerning differential equations and computational tools. We restrict ourselves to *partial* differential equations which describe steady-state phenomena, more procisely, to *elliptic* problems.

First, we overview the most important vector calculus which is essential in describing partial differential equations. After that, several concrete physical problems are shown that demonstrate how it is possible to derive partial differential equations based on physical laws. It is also pointed out how the boundary conditions come from the physics in a very natural way. Then the most popular method – the Finite Element Method – is outlined from theoretical and also from computational points of view. Finally, some other traditional and modern computational methods are outlined (Fourier's method, finite difference method, meshless methods, method of fundamental solutions).

The mathematical tools are rather difficult. We tried to show the mathematical background as detailed as possible, but we had to skip a lot of proofs, and we tried to concentrate on the computational aspects of the presented methods.

Should you observe some misprints, typos, or any other problems, please do not hesitate to contact the author:

Dr. Csaba Gáspár, Professor

Department of Mathematics and Computational Sciences
Széchenyi István University
Egyetem tér 1, H-9026 Győr, Hungary

gasparcs@math.sze.hu

2 Some vector calculus

Here we briefly summarize the main notations and concepts which are used in the rest of the lecture notes.

- $\mathbf{N}$: the set of natural numbers
- $\mathbf{Z}$: the set of integer numbers
- $\mathbf{R}$: the set of real numbers
- $\mathbf{R}^N$: the set of ordered real N-tuples (usually meant as column vectors)
- $\mathbf{C}$: the set of complex numbers

2.1 Differentiation in vector fields

Let $\Omega \subset \mathbf{R}^N$ be a bounded domain in the N-dimensional space (usually $N = 1$, $N = 2$ or $N = 3$) and let $u : \Omega \to \mathbf{R}$ be a multivariate scalar-valued function. Denote by $E : \Omega \to \mathbf{R}^N$ a multivariate vector-valued function (*vector field*),

$$E := (E_1, E_2, ..., E_N)$$

where E_j are the *components* of E.

Introduce the vector field

$$\operatorname{grad} u := \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, ..., \frac{\partial u}{\partial x_N} \right),$$

and also the scalar function

$$\operatorname{div} E := \frac{\partial E_1}{\partial x_1} + \frac{\partial E_2}{\partial x_2} + ... + \frac{\partial E_N}{\partial x_N}$$

$\operatorname{grad} u$ is called the *gradient* of u. $\operatorname{div} E$ is the *divergence* of E.

The following statement can be verified by direct calculations:

Proposition: Let u, v be sufficiently smooth scalar functions, and let E be a sufficiently smooth vector field. Then:

$$\operatorname{grad}(u \cdot v) = (\operatorname{grad} u) \cdot v + u \cdot (\operatorname{grad} v)$$

$$\operatorname{div}(u \cdot E) = \langle (\operatorname{grad} u), E \rangle + u \cdot (\operatorname{div} E)$$

where the symbol $\langle .,. \rangle$ means the scalar product (inner product) in $\mathbf{R}^N$: if $a = (a_1, a_2, ..., a_N)$, $b = (b_1, b_2, ..., b_N)$ are N-dimensional vectors, then

$$\langle a, b \rangle := \sum_{j=1}^{N} a_j \cdot b_j$$

Let u be a differentiable scalar function and denote by n an arbitrary unit vector. The expression

$$\frac{\partial u}{\partial n}(x) := \langle \operatorname{grad} u(x), n \rangle$$

is called the *derivative of u taken in the direction of n*. It is the limit of the quotient

$$\frac{f(x + \varepsilon \cdot n) - f(x)}{\varepsilon}$$

when $\varepsilon \to 0$.

If u is a (twice differentiable) scalar function, then define the *Laplacian of u* as follows:

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

The differential operator Δ is called *Laplace operator*.

Clearly:

$$\Delta u = \operatorname{div} \operatorname{grad} u$$

(check it!)

Remark: It is often convenient to introduce the symbolic vector ∇, the components of which are the partial differential operators with respect to the spatial variables (*nabla operator*):

$$\nabla := \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, ..., \frac{\partial}{\partial x_N} \right)$$

The gradient and the divergence can then be expressed with the nabla operator in the following form:

$$\operatorname{grad} u = \nabla u$$

$$\operatorname{div} E = \nabla \cdot E = \langle \nabla, E \rangle,$$

i. e. the divergence is the 'scalar product' of ∇ and the corresponding vector field.

2.2 Integrals

Denote by Γ the *boundary* of the domain Ω. This is a closed curve if $\Omega \subset \mathbf{R}^2$, and it is a closed surface if $\Omega \subset \mathbf{R}^3$. Recall that if u is a scalar-valued function defined on Ω, then the volume integral

$$\int_\Omega u \, d\Omega$$

is a real number. Its illustrative meaning is as follows. If Ω is subdivided into disjoint subdomains Ω_1, Ω_2, ..., Ω_n, then

$$\int_\Omega u \, d\Omega \approx \sum_{j=1}^{n} u_j \cdot |\Omega_j|$$

where $|\Omega_j|$ denotes the area (resp. the volume) of the subdomain Ω_j, while u_j means a value of u taken at an arbitrary point of Ω_j. Similarly the *line integral* (resp. *surface integral*)

$$\int_\Gamma u \, d\Gamma$$

is another real number; if Γ is subdivided into disjoint subsets Γ_1, Γ_2, ..., Γ_n, then

$$\int_\Gamma u \, d\Gamma \approx \sum_{j=1}^{n} u_j \cdot |\Gamma_j|$$

where $|\Gamma_j|$ denotes the length (resp. the area) of the subset Γ_j, and u_j is again a value of u taken at an arbitrary point of Γ_j. The approximate equalities become exact when the maximal volume/length of the subdivision tends to zero.

From physical point of view, the units of measurement of the above integrals are:

$$(unit\ of\ measurement\ of\ \int_\Omega u \, d\Omega) = (unit\ of\ measurement\ of\ u) \cdot meter^N$$

and

$$(unit\ of\ measurement\ of\ \int_\Gamma u \, d\Gamma) = (unit\ of\ measurement\ of\ u) \cdot meter^{N-1}$$

One of the most profound theorems in vector calculus is the *divergence theorem*:

Divergence theorem (Gauss): If $E : \Omega \to \mathbf{R}^N$ is a sufficiently smooth vector field and Ω is a sufficiently smooth domain, then:

$$\boxed{\int_\Omega \operatorname{div} E \, d\Omega = \int_\Gamma \langle E, n \rangle \, d\Gamma}$$

where n denotes the *outward normal unit vector* along the boundary Γ (i.e. n is orthogonal to Γ and its length equals to 1). Therefore $\langle E, n \rangle$ is the length of the component of E taken in the outward normal direction.

The mathematical meaning of the divergence theorem is that the integral of a derivative of a function (i.e. the divergence) equals to another integral taken on the boundary (which has one less dimension than the original domain). In this sense, the divergence theorem is a strong generalization of the well-known Newton-Leibniz theorem.

The integral $\int_\Gamma \langle E, n \rangle \, d\Gamma$ is often called the *flux of* E through the surface Γ and has some direct physical meaning. For example, if E is a stationary velocity field of a fluid e.g. water, then the physical dimension of $\int_\Gamma \langle E, n \rangle \, d\Gamma$ is $\frac{m}{sec} \cdot m^2 = \frac{m^3}{sec}$ i.e. *water discharge*. Since the water is incompressible, the mass conservation law says that the total flux through an arbitrary closed surface is always zero (provided that the domain contains no sources or sinks). Therefore, by virtue of the divergence theorem, E is necessarily divergence-free, i.e. $\operatorname{div} E \equiv 0$. This is a physical interpretation of the divergence theorem.

Now we give an illustration for the divergence theorem (which is, however, far from being a rigorous proof). In 2D, consider a vector field $E = (F, G)$, the values of which are known only in the centers of an orthogonal, uniform cell system (see Figure 1). Assume that the cell size h is small. In an elementary cell (denoted by C), let us approximate the left-hand side of the divergence theorem. Denote by the neighbouring cells of C by N, W, S and E. Since

$$(\operatorname{div} E)_C = \left(\frac{\partial F}{\partial x}\right)_C + \left(\frac{\partial G}{\partial y}\right)_C \approx \frac{F_E - F_W}{2h} + \frac{G_N - G_S}{2h},$$

the integral of the divergence of E on the cell C is approximately equal to:

$$\int_\Omega \operatorname{div} E \, d\Omega \approx \left(\frac{F_E - F_W}{2h} + \frac{G_N - G_S}{2h}\right) \cdot h^2 = \frac{h}{2} \cdot (F_E - F_W + G_N - G_S)$$

To compute the contour integral in the right-hand side of the divergence theorem, first divide the boundary of the cell C into four parts denoted by

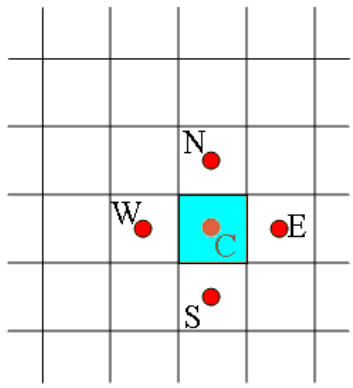

Figure 1: Cell system. A central cell and its neighbours

Γ_N, Γ_W, Γ_S and Γ_E (northern, western, southern and eastern parts). Along Γ_E, he outward normal unit vector is $n = (1, 0)$. Approximating the normal component of (F, G) (i.e. the function F) by the mean value $F|_{\Gamma_E} \approx \frac{F_E + F_C}{2}$, we have:

$$\int_{\Gamma_E} \langle E, n \rangle \, d\Gamma \approx \left(\frac{1}{2} F_E + \frac{1}{2} F_C \right) \cdot h$$

Similarly:

$$\int_{\Gamma_N} \langle E, n \rangle \, d\Gamma \approx \left(\frac{1}{2} G_N + \frac{1}{2} G_C \right) \cdot h \qquad \text{(here } n = (0, 1)\text{)}$$

$$\int_{\Gamma_W} \langle E, n \rangle \, d\Gamma \approx \left(-\frac{1}{2} F_W - \frac{1}{2} F_C \right) \cdot h \qquad \text{(here } n = (-1, 0)\text{)}$$

$$\int_{\Gamma_S} \langle E, n \rangle \, d\Gamma \approx \left(-\frac{1}{2} G_S - \frac{1}{2} G_C \right) \cdot h \qquad \text{(here } n = (0, -1)\text{)}$$

Summing up the above four (approximate) equalities:

$$\int_{\partial C} \langle E, n \rangle \, d\Gamma \approx$$

$$\approx \left(\frac{1}{2} F_E + \frac{1}{2} F_C + \frac{1}{2} G_N + \frac{1}{2} G_C - \frac{1}{2} F_W - \frac{1}{2} F_C - \frac{1}{2} G_S - \frac{1}{2} G_C \right) \cdot h =$$

$$= (F_E + G_N - F_W - G_S) \cdot \frac{h}{2} \approx \int_C \operatorname{div} E \, d\Omega.$$

Thus, the theorem is illustrated (not proved!) for an infinitesimal cell C.

If Ω is the union of finite number of infinitesimal cells C_1, C_2, ..., C_m, then:

$$\int_C \operatorname{div} E \, d\Omega = \sum_{k=1}^{m} \int_{C_k} \operatorname{div} E \, d\Omega \approx$$

$$\approx \sum_{k=1}^{m} \int_{\partial C_k} \langle E, n \rangle \, d\Gamma \approx$$

Let Γ_0 be a side of a cell lying in the interior of Ω. Then the integral $\int_{\Gamma_0} \langle E, n \rangle \, d\Gamma$ occurs in the above sum exactly twice with opposite normal vectors (see Figure 2). They cancel out, therefore the divergence theorem remains valid also in this case.

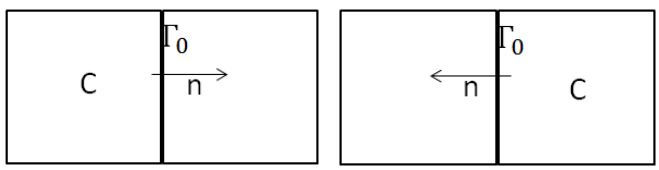

Figure 2: Two contour integrals appearing in neighboring interior cells

The divergence theorem implies several direct corollaries. The most important ones are collected in the following propositions. Let u and v be sufficiently smooth scalar functions (more precisely, they are supposed to be twice continuously differentiable in the closure of Ω). Then:

Proposition:

$$\int_\Omega \Delta u \, d\Omega = \int_\Gamma \frac{\partial u}{\partial n} \, d\Gamma$$

Proof: Obviously

$$\int_\Omega \Delta u \, d\Omega = \int_\Omega \operatorname{div} \operatorname{grad} u \, d\Omega$$

Applying the divergence theorem on the right-hand side:

$$\int_\Omega \Delta u \, d\Omega = \int_\Gamma \langle \operatorname{grad} u, n \rangle \, d\Gamma = \int_\Gamma \frac{\partial u}{\partial n} \, d\Gamma.$$

Green's first theorem:

$$\int_\Omega (\Delta u) \cdot v \, d\Omega = -\int_\Omega \langle \operatorname{grad} u, \operatorname{grad} v \rangle \, d\Omega + \int_\Gamma \frac{\partial u}{\partial n} \cdot v \, d\Gamma$$

Proof: First, since $\operatorname{div}\operatorname{grad} u = \Delta u$, we have:

$$\operatorname{div}\left((\operatorname{grad} u)\cdot v\right) = (\Delta u)\cdot v + \langle \operatorname{grad} u, \operatorname{grad} v\rangle$$

Integrating the left-hand side over Ω and applying the divergence theorem:

$$\int_\Omega \operatorname{div}\left((\operatorname{grad} u)\cdot v\right) d\Omega = \int_\Gamma \langle (\operatorname{grad} u)\cdot v, n\rangle \, d\Gamma = \int_\Gamma \frac{\partial u}{\partial n}\cdot v \, d\Gamma$$

Integrating the right-hand side over Ω:

$$\int_\Omega (\Delta u)\cdot v \, d\Omega + \int_\Omega \langle \operatorname{grad} u, \operatorname{grad} v\rangle \, d\Omega$$

from which the theorem follows.

Green's second theorem:

$$\int_\Omega (\Delta u)\cdot v \, d\Omega - \int_\Omega u\cdot \Delta v \, d\Omega = \int_\Gamma \frac{\partial u}{\partial n}\cdot v \, d\Gamma - \int_\Gamma u\cdot \frac{\partial v}{\partial n} \, d\Gamma -$$

Proof: Green's first theorem states that:

$$\int_\Omega (\Delta u)\cdot v \, d\Omega = -\int_\Omega \langle \operatorname{grad} u, \operatorname{grad} v\rangle \, d\Omega + \int_\Gamma \frac{\partial u}{\partial n}\cdot v \, d\Gamma$$

Swapping the roles of u and v:

$$\int_\Omega (\Delta v)\cdot u \, d\Omega = -\int_\Omega \langle \operatorname{grad} v, \operatorname{grad} u\rangle \, d\Omega + \int_\Gamma \frac{\partial v}{\partial n}\cdot u \, d\Gamma$$

Subtracting the last two equalities, we have the theorem.

3 From physical laws to partial differential equations

Now we will see through some examples, how a physical law generated a partial differential equation which describes the corresponding physical process. The key issue is the divergence theorem.

3.1 Heat conduction

Consider a 3D material in which there are heat sources and result in a nonuniform heat distribution. Suppose that the steady state has reached i.e. the temperature does not vary in time. Denote by $u(x)$ the temperature of the material at the spatial point x (measure unit: K). Let σ be the thermal conductivity of the material (physical dimension: $\frac{power}{length \cdot temperature}$, measure unit: $\frac{W}{m \cdot K}$), which characterizes the 'efficiency' of the heat transfer.

Let n be an arbitrary unit vector which defines a specified direction. The physical dimension of $\sigma \cdot \frac{\partial u}{\partial n}$ is $\frac{power}{length \cdot temperature} \cdot \frac{temperature}{length} = \frac{power}{length^2}$.

According to *Fourier's law*, the thermal flux is proportional to the derivative of the temperature taken in the direction n, and the coefficient of proportionality is (the negative) thermal conductivity.

Thus, if Γ_0 is a (closed or not closed) surface in the space, the integral

$$\int_{\Gamma_0} \sigma \frac{\partial u}{\partial n} \, d\Gamma$$

is the thermal energy across the surface Γ_0 per second, where n denotes the normal unit vector along Γ_0.

From the energy conservation law (the first law of thermodynamics), it follows that

$$\int_{\Gamma_0} \sigma \frac{\partial u}{\partial n} \, d\Gamma = 0$$

holds for every *closed* surface which surrounds a volume Ω_0. The divergence theorem implies that

$$\int_{\Omega_0} \operatorname{div}(\sigma \cdot \operatorname{grad} u) \, d\Omega = 0$$

for arbitrary subdomain Ω_0 contained in the domain Ω, the domain of the diffusion process. Consequently:

$$\operatorname{div}(\sigma \cdot \operatorname{grad} u) = 0$$

holds in Ω.

If the thermal conductivity σ is constant in the whole domain Ω, then the above partial differential equation has a simpler form:

$$\sigma \cdot \Delta u = 0$$

or simply

$$\Delta u = 0$$

in Ω.

3.2 Diffusion

Consider a fluid substance in which a molecular diffusion of some pollutant takes place. Suppose that the steady state has reached i.e. the physical characteristics of the process do not vary in time, Denote by $u(x)$ the concentration of the pollutant at the spatial point x (the physical dimension of u is $\frac{mass}{length^3}$). Let σ be the diffusion coefficient (diffusivity, physical dimension: $\frac{length^2}{time}$), which characterizes the 'speed' of diffusion (i.e. the mobility of the particles of the pollutant).

Let n be an arbitrary unit vector which defines a specified direction. The physical dimension of $\sigma \cdot \frac{\partial u}{\partial n}$ is $\frac{length^2}{time} \cdot \frac{\frac{mass}{length^3}}{length} = \frac{\frac{mass}{time}}{length^2}$.

According to *Fick's law*, the diffusion flux is proportional to the derivative of the concentration taken in the direction n, and the coefficient of proportionality is (the negative) diffusion coefficient.

Thus, if Γ_0 is a (closed or not closed) surface in the space, the integral

$$\int_{\Gamma_0} \sigma \frac{\partial u}{\partial n} \, d\Gamma$$

is the total mass of pollutant across the surface Γ_0 per second, where n denotes the normal unit vector along Γ_0.

From the mass conservation law, it follows that

$$\int_{\Gamma_0} \sigma \frac{\partial u}{\partial n} \, d\Gamma = 0$$

holds for every *closed* surface which surrounds a volume Ω_0. The divergence theorem implies that

$$\int_{\Omega_0} \operatorname{div}(\sigma \cdot \operatorname{grad} u) \, d\Omega = 0$$

for arbitrary subdomain Ω_0 contained in the domain Ω, the domain of the diffusion process. Consequently:

$$\operatorname{div}(\sigma \cdot \operatorname{grad} u) = 0$$

holds in Ω.

If the diffusivity σ is constant in the whole domain Ω, then the above partial differential equation has a simpler form:

$$\sigma \cdot \Delta u = 0$$

or simply

$$\Delta u = 0$$

in Ω.

3.3 Electric current in 3D materials

Consider an electrically conducting 3D material (electrolyte, biological tissues, soil etc.) in which a steady-state electric current takes place. Denote by $u(x)$ the electric potential at the point x (measure unit: V), and let σ be the specific conductance (conductivity, measure unit: $\frac{1}{ohm \cdot m}$). This is the reciprocal value of the specific resistance of the material.

The measure unit of $\sigma \frac{\partial u}{\partial n}$ is $\frac{1}{ohm \cdot m} \cdot \frac{V}{m} = \frac{A}{m^2}$ in every direction specified by the unit vector n. That is, $\sigma \frac{\partial u}{\partial n}$ means *current density* in the direction n.

According to *Ohm's law*, the current density is proportional to the derivative of the electric potential (voltage) taken in the direction n, and the coefficient of proportionality is the conductivity.

Thus, if Γ_0 is a (closed or not closed) surface in the space, the integral

$$\int_{\Gamma_0} \sigma \frac{\partial u}{\partial n} \, d\Gamma$$

is the total current through the surface Γ_0, where n denotes the normal unit vector along Γ_0. From the charge conservation law, it follows that

$$\int_{\Gamma_0} \sigma \frac{\partial u}{\partial n} \, d\Gamma = 0$$

holds for every *closed* surface which surrounds a volume Ω_0. The divergence theorem implies that

$$\int_{\Omega_0} \operatorname{div}(\sigma \cdot \operatorname{grad} u) \, d\Omega = 0$$

for arbitrary subdomain Ω_0 contained in the domain Ω, the domain of the electrical current. Consequently:

$$\operatorname{div}(\sigma \cdot \operatorname{grad} u) = 0$$

holds in Ω.

If the conductivity σ is constant in the whole domain Ω, then the above partial differential equation has a simpler form:

$$\sigma \cdot \Delta u = 0$$

or simply

$$\Delta u = 0$$

in Ω.

3.4 Seepage through porous medium

Consider a 3D porous material (soil, sand etc.) and the groundwater flow through the medium. Assume that the flow is steady-state. Denote by $u(x)$ the *hydraulic head* in the point x. Then

$$u = \frac{p}{\rho g} + z,$$

where p is the pressure, ρ is the density of water, g is the acceleration due to gravity, and z denotes the height above a reference level (i.e. the vertical component of the point x). Its physical dimension is length.

According to *Darcy's law*, the seepage velocity (apart from a dimensionless constant) is proportional to the derivative of the hydraulic head taken in the direction n, the direction of the velocity, and the (negative) coefficient of proportionality is the so-called *hydraulic conductivity* (or permeability) of the medium. This is denoted by K (physical dimension: $\frac{length}{time}$).

Thus, if Γ_0 is a (closed or not closed) surface in the space, the integral

$$\int_{\Gamma_0} K \cdot \frac{\partial u}{\partial n} \, d\Gamma$$

is the total discharge through the surface Γ_0, where n denotes the normal unit vector along Γ_0. From the mass conservation law, it follows that

$$\int_{\Gamma_0} K \cdot \frac{\partial u}{\partial n} \, d\Gamma = 0$$

holds for every *closed* surface which surrounds a volume Ω_0. The divergence theorem implies that

$$\int_{\Omega_0} \operatorname{div}(K \cdot \operatorname{grad} u)\, d\Omega = 0$$

for arbitrary subdomain Ω_0 contained in the seepage domain Ω. Consequently:

$$\operatorname{div}(K \cdot \operatorname{grad} u) = 0$$

holds in Ω.

If the hydraulic conductivity K is constant in the whole domain Ω, then the above partial differential equation has a simpler form:

$$K \cdot \Delta u = 0$$

or simply

$$\Delta u = 0$$

in Ω.

In all of the above four examples, the physical process (which is assumed to be steady-state) is described by a partial differential equation, which has the form:

$$\operatorname{div}(\sigma \cdot \operatorname{grad} u) = 0$$

where the a priori unknown function u is the physical quantity which characterizes the process, σ is a known multivariate, positive function, which characterizes some material property. In the presence of sources and sinks, the above partial differential equation has the form:

$$\operatorname{div}(\sigma \cdot \operatorname{grad} u) = f$$

where the known function f describes the density distribution of the sources.

The above equations are the simplest examples for second-order *elliptic partial differential equations*. They describe steady-state processes, when the physical quantities do not depend on time.

There are more general elliptic partial differential equations e.g. the convection-diffusion equation. The much more complicated equations which describe the fluid or gas motion are nonlinear. In this lecture notes, however, we restrict ourselves to the above simpler elliptic equations. Note that the corresponding time-dependent processes are described by similar, but more complicated partial differential equations (parabolic and hyperbolic equations) which contain the derivatives with respect to time as well.

An important special case when the material-characteristic function σ is constant in the whole domain Ω. In this case, the equation is simplified to the *Poisson equation*:

$$\Delta u = f$$

If the function f in the right-hand side is identically zero, we obtain the *Laplace equation*:

$$\Delta u = 0$$

The solutions of the Laplace equations are called *harmonic functions*.

A frequently appearing situation is that σ is a *piecewise constant function* (or, it is approximated by such a function). In this case, the differential equation is not more complicated than the Laplace equation, but an extra phenomenon appears at the inner surface that separates the subdomains where σ takes different values. As a model problem, consider the domain Ω, which consists of two subdomains, Ω_1 and Ω_2. Suppose that $\sigma \equiv \sigma_1 = const.$ in Ω_1 and $\sigma \equiv \sigma_2 = const.$ in Ω_2. Denote by Γ_0 the inner surface $\Gamma_0 := \partial\Omega_1 \cap \partial\Omega_2$, then Γ_0 separates Ω_1 and Ω_2. Consider a thin subdomain Ω_0 at the interface Γ_0 (see Figure 3. Let n be a normal unit vector pointing from Ω_1 into Ω_2. Let u be a solution of the partial differential equation

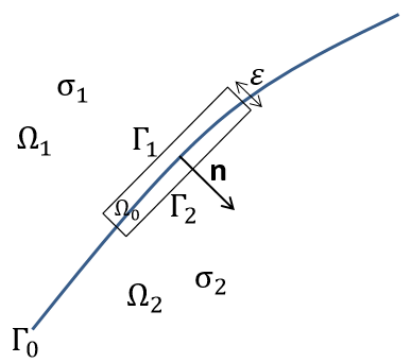

Figure 3: For the interface conditions.

$$\operatorname{div}(\sigma \cdot \operatorname{grad} u) = 0$$

and denote by $u_1 := u|_{\Omega_1}$ and $u_2 := u|_{\Omega_2}$, the restrictions of u to the subdomains Ω_1 and Ω_2. Then, *in the interior* of Ω_1 and Ω_2, respectively, the equations

$$\Delta u_1 = 0, \qquad \Delta u_2 = 0$$

are valid, but this is not the case along the interface Γ_0 (here u is *not* twice continuously differentiable). However, u is continuous (this is obvious from physical point of view), i.e.

$$u_1 = u_2 \quad \text{along } \Gamma_0$$

Integrating the function $\operatorname{div}(\sigma \cdot \operatorname{grad} u)$ over Ω_0, the integral is zero; on the other hand, from the divergence thorem, it follows that:

$$0 = \int_{\Omega_0} \operatorname{div}(\sigma \cdot \operatorname{grad} u)\, d\Omega = \int_{\partial\Omega_0} \sigma \cdot \frac{\partial u}{\partial n}\, d\Gamma \approx$$

$$\approx \int_{\Gamma_2} \sigma_2 \cdot \frac{\partial u_2}{\partial n}\, d\Gamma - \int_{\Gamma_1} \sigma_1 \cdot \frac{\partial u_1}{\partial n}\, d\Gamma,$$

since n points from Ω_1 into Ω_2, thus, n is the outward (resp. inward) normal unit vector along Γ_2 (resp. Γ_1). When $\varepsilon \to 0$, the approximate equality becomes an exact equality, and since Ω_0 was arbitrary, we obtain:

$$\sigma_1 \cdot \frac{\partial u_1}{\partial n} = \sigma_2 \cdot \frac{\partial u_2}{\partial n} \quad \text{along } \Gamma_0$$

along Γ_0. Note that this is also obvious from physical considerations: for instance, if u is electric potential, the equality $\sigma_1 \cdot \frac{\partial u_1}{\partial n}|_{\Gamma_0} = \sigma_2 \cdot \frac{\partial u_2}{\partial n}|_{\Gamma_0}$ means the conservation of charge.

In short, the value of the solution remain continuous along the interface Γ_0, but the normal derivative has a jump here.

So far, some partial differential equations have been introduced. It should be pointed out, that a physical process is *not* uniquely determined by the governing partial differential equation. From physical point of view, this is obvious: for instance, when an electric current is investigated which flows through a 3D material, the potential distribution highly depends on the incoming and outgoing current distributions. To properly describe the physical process, some extra information (*auxiliary condition*) is still needed.

4 Boundary conditions

As mentioned before, a differential equation does not describe the corresponding physical process uniquely. From pure mathematical point of view it means that a partial differential equation has infinitely many solutions. Indeed, considering the simplest 2D Laplace equation

$$\Delta u = 0,$$

the functions $u := 1$, $u := x$, $u := y$, $u := x^2 - y^2$, $u = x \cdot y$,... and all linear combinations of them satisfy the differential equation.

If one wants to find a specific solution (which describes the corresponding physical process), some additional information is needed. In the case of elliptic equations, this extra information is usually connected with the boundary of the domain. These *boundary conditions* have physical meaning.

Now we outline the usual boundary conditions. Without going into details we note that the above introduced elliptic differential equations supplied with one of the following boundary conditions have generally a unique solution in a well-defined function space.

Consider the elliptic partial differential equation:

$$\operatorname{div}(\sigma \cdot \operatorname{grad} u) = f \qquad \text{in } \Omega$$

where Ω is a bounded domain. Denote by Γ the boundary of the domain Ω.

First (or Dirichlet) boundary condition:

$$\boxed{u \text{ is prescribed along } \Gamma}$$

As an example, consider the heat conduction equation. Assume that a 3D object contains heat sources. If the object is immersed in melting ice, then a Dirichlet condition is enforced: the temperature on the boundary is 0 degrees Celsius (provided that the steady state has reached). In practice, this is rarely the case. A more probable situation is that measurements have been done along the boundary, so that we know the temperature distribution on the boundary by measurements.

Second (or Neumann) boundary condition:

$$\boxed{\sigma \cdot \frac{\partial u}{\partial n} \text{ is prescribed along } \Gamma}$$

where n denotes the outward normal unit vector along Γ. In practice, this is equivalent to the prescription of the normal derivative $\frac{\partial u}{\partial n}$ along the boundary Γ.

An important special case is when the normal derivative is identically zero along the boundary. This is the case in heat conduction problems, if the object is wrapped in a thermal insulator layer. In electrical current problems, this means that the material is electrically insulated from the outer materials.

Third (or Robin) boundary condition:

$$\boxed{\text{A linear combination of } u \text{ and } \frac{\partial u}{\partial n} \text{ is prescribed along } \Gamma}$$

At a first glance, this seems to be a pure mathematical generalization of the Dirichlet and Neumann boundary conditions. In fact, such a boundary condition appears in a natural way when solving the equation

$$\operatorname{div}(\sigma \cdot \operatorname{grad} u) = 0,$$

and the domain consists of two parts: an inner domain Ω and a thin layer Ω_0 which surrounds Ω.

Suppose that $\sigma \equiv const.$ in Ω, and $\sigma \equiv \sigma_0$ in Ω_0, where $0 < \sigma_0 \ll \sigma$, and the thickness of Ω_0 is much less that the characteristic size of Ω (see Figure 4). This is the case, when u means an electric potential and the layer Ω_0 is filled by a 'quasi-insulator' material i.e. its conductivity is much less that in the interior of Ω. (One can think also of a not perfect thermal insulator layer in heat conduction problems etc.)

Suppose that along Γ_0, the external boundary of Ω_0, the potential u_{ext} is given as a Dirichlet boundary condition. Then, along the interface Γ, we have (in accordance of the previous section):

$$\sigma \cdot \frac{\partial u}{\partial n} = \sigma_0 \cdot \frac{\partial u_0}{\partial n}$$

Since Ω_0 is thin, the derivative $\frac{\partial u_0}{\partial n}$ can be approximated by the difference quotient $\dfrac{u_0|_{\Gamma_0} - u_0|_{\Gamma}}{d} = \dfrac{u_{ext} - u}{d}$. This implies that (approximately):

$$\sigma \cdot \frac{\partial u}{\partial n} = \sigma_0 \cdot \frac{u_{ext} - u}{d}$$

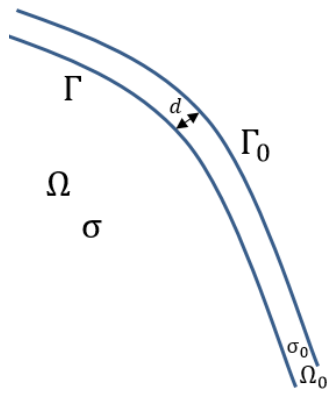

Figure 4: For the Robin boundary condition.

Rearranging this equality, we immediately obtain a Robin-type boundary condition along Γ:

$$\frac{\sigma_0}{d} \cdot u + \sigma \cdot \frac{\partial u}{\partial n} = \frac{\sigma_0}{d} \cdot u_{ext}$$

That is, the physical problem itself leads to the Robin boundary condition in a natural way.

In practice, the most frequently appearing situation is that the boundary is divided into several (disjoint) parts, and along the different parts, different types of boundary condition are prescribed (*mixed boundary condition*). Roughly speaking, this means that to each boundary point, exactly one boundary condition has to be assigned. Note also that the boundary condition should not contain second of higher order derivatives of the unknown function.

We demonstrate the very natural appearance of mixed boundary conditions through an example of seepage hydraulics, which also contains a non-familiar character. Consider a 2D seepage problem through a dam. Figure 5 shows the cross-section of a dam which is made of porous material. On the left-hand side of the dam, the water level is high (H_1), while on the right-hand side, it is low (H_2). The pressure difference induces a seepage motion through the dam, which is assumed to be homogeneous for simplicity, i.e. the hydraulic conductivity is constant everywhere.

In the steady state, the seepage is bounded by the curve Γ from above. This is the *free surface* of the seepage. The location of Γ is a priori not

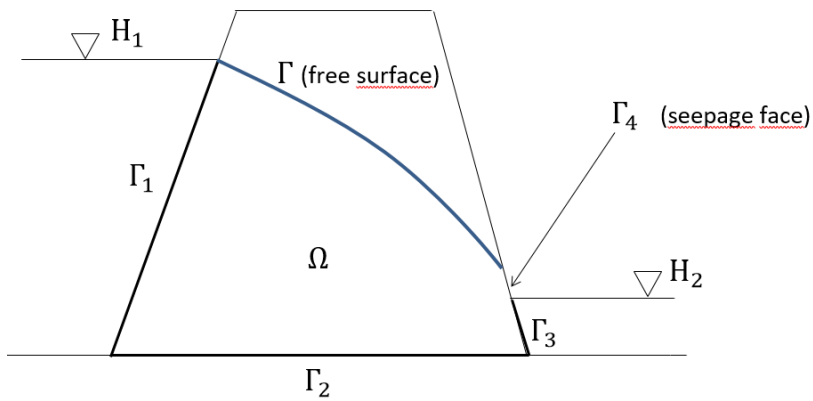

Figure 5: The classical dam problem (free boundary problem).

known, the determination of Γ is a part of the problem called *free boundary problem*. In the domain Ω, the velocity potential u satisfies the Laplace equation:

$$\Delta u = 0$$

Along the boundary, several types of boundary conditions are taken:

- At a point $(x, y) \in \Gamma_1$, the pressure can be calculated: $p = (H_1 - y) \cdot \gamma$, where γ is the specific weight of water. Substituting this into the expression of the velocity potential $u = \frac{p}{\gamma} + y$, we obtain that

  $$\boxed{u \equiv H_1 \qquad \text{along } \Gamma_1}$$

  (Dirichlet boundary condition).

- Along Γ_2, the velocity vector is tangential; consequently, its normal component is identically zero:

  $$\boxed{\frac{\partial u}{\partial n} \equiv 0 \qquad \text{along } \Gamma_2}$$

  (Neumann boundary condition).

- On Γ_3, the situation is similar to the case of Γ_1, i.e.:

  $$\boxed{u \equiv H_2 \qquad \text{along } \Gamma_3}$$

(Dirichlet boundary condition).

- At a point $(x, y) \in \Gamma_4$, the pressure is atmospheric, therefore:

$$\boxed{u(x, y) = y \qquad \text{along } \Gamma_4}$$

  (Dirichlet boundary condition). Note that, in contrast to the parts Γ_1 and Γ_3, the boundary value is not constant; it depends on the location of the point of boundary point. Γ_4 is called *seepage face*.

- At the points $(x, y) \in \Gamma$, where Γ is the (a priori unknown) free surface, two independent boundary conditions can be prescribed. The velocity vector is tangential, therefore:

$$\boxed{\frac{\partial u}{\partial n} \equiv 0 \qquad \text{along } \Gamma}$$

  (similarly to the case of Γ_2). On the other hand, the pressure is atmospheric, which implies:

$$\boxed{u(x, y) = y \qquad \text{along } \Gamma}$$

The presence of the free surface makes the whole problem nonlinear. However, it can be solved by a special iterative technique, by correcting the position of the approximate free surface in each step. Note that, in every iteration, a Laplace equation supplied with mixed boundary conditions has to be solved. Details are omitted.

5 The Finite Element Method

5.1 Theory in a nutshell

Here we briefly summarize the main theoretical aspects of the Finite Element Method. Most of the propositions, theorems are mentioned here without proofs.

First we recall the most important concepts.

5.1.1 A reminder

Normed spaces. A vector space X is called *normed space*, if a function $||.||$ is defined in X (called *norm*), which has the following properties:

- For every vector $x \in X$: $||x|| \geq 0$ and $||x|| = 0$ if and only if $x = \mathbf{0}$.
- For every vector $x \in X$ and every scalar $\alpha \in \mathbf{R}$: $||\alpha \cdot x|| = |\alpha| \cdot ||x||$.
- For arbitrary vectors $x, y \in X$: $||x + y|| \leq ||x|| + ||y||$ ('triangle inequality').

The *distance* of the vectors $x, y \in X$ is defined as the norm of their difference, i.e. $||x - y||$.

The vector sequence $(x_n) \subset X$ is said to be *convergent* and *to tend to the vector* $x \in X$, if $||x_n - x|| \to 0$, as $n \to \infty$.

The vector sequence $(x_n) \subset X$ is said to be a *Cauchy sequence* if for every $\varepsilon > 0$, there exists an index N such that for all indices $n, m \geq N$, the inequality $||x_n - x_m|| < \varepsilon$ holds.

In arbitrary normed space, if a vector sequence is convergent, then it is necessarily a Cauchy sequence. The converse statement is generally not true.

The normed space X is called *complete* or *Banach space*, if every Cauchy sequence is convergent in X. Every normed space can be made complete by adding extra vectors to the space and by properly expanding the operations as well as the norm to these extra vectors.

The most frequently appearing examples for normed spaces: $\mathbf{R}$ itself with the absolute value as a norm; $\mathbf{R}^N$ with the most popular norms:

- $||x||_{\max} := \max\limits_{1 \leq k \leq N} |x_k|$ (maximum norm)
- $||x||_1 := \sum\limits_{k=1}^{N} |x_k|$ (sum norm)

- $||x|| := \sqrt{\sum_{k=1}^{N} |x_k|^2}$ (Euclidean norm)

where $x = (x_1, x_2, ..., x_N)$. These spaces are finite dimensional spaces; every finite dimensional normed space is automatically complete, i.e. Banach space. For infinite dimensional spaces, this is not the case. The most important infinite dimensional spaces are certain *function spaces* as follows:

- The space of the continuous functions defined on a finite interval $[a, b]$ denoted by $C[a, b]$ with the maximum norm:

$$||f||_{\max} := \max_{a \le x \le b} |f(x)|$$

- The space of the m times continuously differentiable functions defined on a finite interval $[a, b]$ denoted by $C^m[a, b]$ with the maximum norm:

$$||f||_{C^m[a,b]} := \sum_{k=0}^{m} \max_{a \le x \le b} |f^{(k)}(x)|$$

- The space of the integrable functions defined on a finite or infinite interval (a, b) denoted by $L_1(a, b)$ with the L_1-norm:

$$||f||_{L_1(a,b)} := \int_a^b |f(x)|\, dx$$

- The space of the square integrable functions defined on a finite or infinite interval (a, b) denoted by $L_2(a, b)$ with the L_2-norm:

$$||f||_{L_2(a,b)} := \sqrt{\int_a^b |f(x)|^2\, dx}$$

The above spaces can be defined for multivariate functions in a straightforward way. The role of the interval (a, b) is played by an N-dimensional domain Ω.

All the above function spaces are complete i.e. Banach spaces. Note however, that the e.g. space $C[a, b]$ is a normed space with the $L_1(a, b)$-norm, but it is not complete: the completion of the space is exactly equal to the space $L_1(a, b)$.

Euclidean spaces. A vector space X is called *Euclidean space*, if a bivariate, real-valued function $\langle ., . \rangle$ is defined in X (called *scalar product* or *inner product*), which has the following properties:

- For every vector $x \in X$: $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ if and only if $x = \mathbf{0}$
- For arbitrary vectors $x, y \in X$: $\langle x, y \rangle = \langle y, x \rangle$
- For arbitrary vectors $x, y \in X$ and every scalar $\alpha \in \mathbf{R}$: $\langle \alpha x, y \rangle = \alpha \cdot \langle x, y \rangle$
- For arbitrary vectors $x, y, z \in X$: $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$

Note that these properties imply that for arbitrary vectors $x, y \in X$ and every scalar $\alpha \in \mathbf{R}$: $\langle x, \alpha y \rangle = \alpha \cdot \langle x, y \rangle$, and for arbitrary vectors $x, y, z \in X$: $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$.

A Euclidean space is always a normed space with the norm induced by the inner product:

$$||x|| := \sqrt{\langle x, x \rangle}$$

In this statement, only the triangle inequality is not obvious. It is a consequence of the following important inequality:

Theorem (Cauchy inequality): If X is a Euclidean space, then for arbitrary vectors $x, y \in X$:

$$|\langle x, y \rangle| \leq ||x|| \cdot ||y||,$$

and the equality is valid if and only if x and y are linearly dependent, i.e. one of them is equal to the other multiplied by a scalar constant.

Proof: For arbitrary scalar $\alpha \in \mathbf{R}$, obviously $||x - \alpha y||^2 \geq 0$, i.e.:

$$||x - \alpha y||^2 = \langle x - \alpha y, x - \alpha y \rangle =$$

$$= ||x||^2 - 2\alpha \langle x, y \rangle + \alpha^2 ||y||^2 \geq 0$$

Now define α by $\alpha := \frac{||x||}{||y||}$ (provided that $y \neq \mathbf{0}$; otherwise, the statement is simplified to the trivial equality $0 = 0$). We have:

$$||x||^2 - 2\frac{||x||}{||y||}\langle x, y \rangle + \frac{||x||^2}{||y||^2} \cdot ||y||^2 \geq 0.$$

This can be simplified to the inequality

$$\langle x, y \rangle \leq ||x|| \cdot ||y||$$

This is valid for arbitrary vectors $x, y \in X$. If y is substituted by $(-y)$, the inequality remains valid:

$$\langle x, -y \rangle \leq ||x|| \cdot ||-y|| = ||x|| \cdot ||y||,$$

whence $\langle x, y\rangle \geq -||x|| \cdot ||y||$. We have obtained that

$$-||x|| \cdot ||y|| \leq \langle x, y\rangle \leq ||x|| \cdot ||y||,$$

i.e. $|\langle x, y\rangle| \leq ||x|| \cdot ||y||$. Equality is valid only if $||x - \alpha y||^2 = 0$, i.e. $x = \alpha y$.

From the Cauchy inequality, the triangle inequality simply follows:

$||x+y||^2 = ||x||^2 + 2\langle x, y\rangle + ||y||^2 \leq ||x||^2 + 2||x|| \cdot ||y|| + ||y||^2 = (||x|| + ||y||)^2.$

Taking the square root of both sides, we have the triangle inequality.

It is worth mentioning the following simple but useful equalities as well:

$$||x + y||^2 = ||x||^2 + 2\langle x, y\rangle + ||y||^2$$
$$||x - y||^2 = ||x||^2 - 2\langle x, y\rangle + ||y||^2$$

for arbitrary $x, y \in X$.

A Euclidean space X is called *Hilbert space* if it is complete with respect to the norm induced by the scalar product.

One of the most important Hilbert spaces is the previously mentioned $L_2(a, b)$, the space of square integrable functions equipped with the scalar product

$$\langle f, g\rangle_{L_2(a,b)} := \int_a^b f(x) \cdot g(x)\, dx$$

and the corresponding multivariate space $L_2(\Omega)$ with the scalar product

$$\langle f, g\rangle_{L_2(\Omega)} := \int_\Omega f(x) \cdot g(x)\, dx$$

Orthogonality. In a Euclidean space X, the vectors $x, y \in X$ are called *orthogonal*, if their scalar product is zero: $\langle x, y\rangle = 0$. In this case, the theorem of Pythagoras is valid:

$$||x + y||^2 = ||x||^2 + ||y||^2,$$

which can be generalized as follows: if $x_1, x_2, ..., x_m \in X$ are pairwise orthogonal vectors, then:

$$||\sum_{j=1}^{m} x_j||^2 = \sum_{j=1}^{m} ||x_j||^2.$$

Bounded linear operators. Let X, Y be normed spaces and let $A : X \to Y$ be a linear mapping (operator). The operator A is called *bounded*, if there exists a constant K such that for every $x \in X$, the inequality

$$||Ax|| \leq K \cdot ||x||$$

is valid. K is called a *bound* of the operator A. The boundedness is equivalent to the continuity of A: A is bounded if and only if it preserves the limit i.e. for arbitrary convergent sequence $x_n \to x$, it follows that $Ax_n \to Ax$. The least bound of A is called the *norm* of the operator A (denoted by $||A||$). Therefore for arbitrary $x \in X$:

$$||Ax|| \leq ||A|| \cdot ||x||$$

is valid. The operator norm has the following extremal property:

$$||A|| = \max_{x \in X, ||x||=1} ||Ax||$$

If the linear operator A is real-valued i.e. $Y = \mathbf{R}$, then A is called linear *functional*.

The operator norm is really a norm, and makes the vector space of the bounded linear $X \to Y$ operators a normed space. (This space is complete, if Y is complete.) In particular, if $A, B : X \to Y$ are bounded linear operators, then:

$$||A + B|| \leq ||A|| + ||B||$$

Moreover, if Z is another normed space, $A : X \to Y$, $B : Y \to Z$ are bounded linear operators then so is their composition $BA : X \to Z$, and:

$$||BA|| \leq ||B|| \cdot ||A||$$

As an important special case, an $N \times M$ matrix can also be regarded as a bounded, linear $\mathbf{R}^M \to \mathbf{R}^N$ mapping. Its operator norm depends on the choice of the norms in the spaces $\mathbf{R}^M$ and $\mathbf{R}^N$. The most frequently used matrix norms (induced by vector norms) are as follows:

- *Row norm*: If both in $\mathbf{R}^N$ and in $\mathbf{R}^M$, the maximum norm is defined, then

$$||A|| = \max_{1 \leq k \leq N} \sum_{j=1}^{M} |A_{kj}|$$

- *Column norm*: If both in $\mathbf{R}^N$ and in $\mathbf{R}^M$, the sum norm is defined, then
$$||A|| = \max_{1\leq j\leq M} \sum_{k=1}^{N} |A_{kj}|$$

- *Matrix norm induced by the Euclidean vector norm*: If in $\mathbf{R}^N$, the Euclidean norm is defined, then for arbitrary square matrix $A \in \mathbf{M}_{N\times N}$:
$$||A|| = \max_{1\leq k\leq N} \sqrt{\lambda_k},$$
where λ_1,...,λ_N are the eigenvalues of the (self-adjoint, positive semidefinite) matrix A^*A.

As a consequence, if A is self-adjoint, the the matrix norm induced by the Euclidean norm is as follows:
$$||A|| = \max_{1\leq k\leq N} |\lambda_k|,$$
where now λ_1,...,λ_N are the eigenvalues of the matrix A. In addition to it, if A is positive definite, then
$$||A|| = \max_{1\leq k\leq N} \lambda_k, \quad \text{and} \quad ||A^{-1}|| = \frac{1}{\min\limits_{1\leq k\leq N} \lambda_k}.$$

5.1.2 Abstract variational problems

Let H be a real Hilbert space and let $a : H \times H \to \mathbf{R}$ be a *symmetric, bounded and coercive bilinear functional*, i.e.

- a is linear in its both variables
- $a(u, v) = a(v, u)$ for all $u, v \in H$ (symmetry)
- $|a(u, v)| \leq M \cdot ||u|| \cdot ||v||$ is valid for some $M \geq 0$ (boundedness)
- $|a(u, u)| \geq m \cdot ||u||^2$ is valid for some $m > 0$ (coercivity)

Note that in this case:
$$m \cdot ||u||^2 \leq a(u, u) \leq M \cdot ||u||^2$$
is valid for all $u \in H$.

Let $\ell : H \to \mathbf{R}$ be a given, *bounded linear functional*.

Variational problem: Find a vector $u \in H$ such that for every $v \in H$, the equality

$$\boxed{a(u, v) = \ell v}$$

is valid.

The fundamental theorem of the variational problems and methods is as follows (without proof):

Theorem (Lax-Milgram): Under the above conditions taken on a and ℓ, the variational problem has a unique solution.

In fact, the functional a is an *inner product* in the space H, which induces the norm

$$||u||_a := \sqrt{a(u, u)}$$

(*energetic norm*, or simply a-norm).

Remark: In practice, the bilinear functional a comes from the elliptic partial differential equation to be solved.

Denote by u^* the (unique) solution of the variational problem. Let $V_h \subset H$ be a finite dimensional subspace, the dimension N of which is characterized by the parameter h (in practice, this means a length). The functions belonging to V_h are often referred to as *trial functions*.

In order to approximate the exact solution u^*, restrict the variational problem to the subspace V_h.

Variational problem in the subspace V_h: Find a vector $u_h \in V_h$ such that the *variational equalities*

$$\boxed{a(u_h, v_h) = \ell v_h}$$

are satisfied for every vector $v_h \in V_h$.

By virtue of the Lax-Milgram theorem, this variational problem also has a unique solution $u_h \in V_h$. The vector u_h is interpreted as an approximate solution of the original variational problem. Two tasks arise immediately:

1. How to compute u_h ?

2. How to estimate the error $u^* - u_h$?

As far as the first task is concerned, denote by N the dimension of the subspace V_h. Define a basis φ_1, φ_2, ..., φ_N in V_h, and seek the solution u_h as a linear combination of the basis vectors:

$$u_h = \sum_{k=1}^{N} \alpha_j \varphi_j$$

It is sufficient to enforce the variational equalities to the vectors $v_h = \varphi_k$ $(k = 1, 2, ..., N)$. Thus, we obtain the **discrete variational problem**:

$$\boxed{\sum_{j=1}^{N} \alpha_j \cdot a(\varphi_j, \varphi_k) = \ell\varphi_k \qquad (k = 1, 2, ..., N)}$$

In a more compact form: $A\alpha = b$. The matrix A is called *stiffness matrix* with the entries: $A_{kj} := a(\varphi_j, \varphi_k)$. The components of the right-hand side are: $b_k := \ell\varphi_k$.

Due to the symmetry of the bilinear functional a, the stiffness matrix A is self-adjoint. Moreover:

Proposition: The matrix A is positive definite.
Proof: It is sufficient to investigate the quadratic form of A. Let $x = (x_1, x_2, ..., x_N) \in \mathbf{R}^N$ be an arbitrary nonzero vector, then:

$$\langle Ax, x\rangle = \sum_{k=1}^{N}\sum_{j=1}^{N} A_{kj}x_k x_j = \sum_{k=1}^{N}\sum_{j=1}^{N} a(\varphi_j, \varphi_k)x_k x_j =$$

$$= a\left(\sum_{k=1}^{N} x_k\varphi_k, \sum_{j=1}^{N} x_j\varphi_j\right) = a(w_h, w_h),$$

where $w_h := \sum_{j=1}^{N} x_j\varphi_j$. Due to the linear independence of the vectors φ_j, the vector w_h differs from zero. The proposition is now a consequence of the coercivity of a.

An important property of the discrete solution is a special *orthogonality property*:

Proposition: The error of the approximation i.e. the vector $(u^* - u_h)$ is a-orthogonal to the subspace V_h, i.e.

$$a(u^* - u_h, v_h) = 0 \qquad \text{for all } v_h \in V_h$$

Proof: Consider the original variational problem. In particular, for all $v_h \in V_h \subset H$:

$$a(u^*, v_h) = \ell v_h$$

For the discrete variational problem, by definition:

$$a(u_h, v_h) = \ell v_h$$

Subtracting the two equalities, we have the proposition.

Based on the orthogonality property, we obtain a special error estimation:

Proposition: For arbitrary vector $v_h \in V_h$

$$||u^* - u_h||_a \leq ||u^* - v_h||_a$$

Proof: Let $v_h \in V_h$ be arbitrary. Then:

$$||u^* - v_h||_a^2 = ||(u^* - u_h) + (u_h - v_h)||_a^2 =$$

$$= ||u^* - u_h||_a^2 + 2 \cdot a(u^* - u_h, u_h - v_h) + ||u_h - v_h||_a^2$$

Due to the othogonality property, $a(u^* - u_h, u_h - v_h) = 0$. On the other hand, obviously $||u_h - v_h||_a^2 \geq 0$. Thus, we obtain that:

$$||u^* - v_h||_a^2 \geq ||u^* - u_h||_a^2,$$

which implies the proposition.

The above proposition can be interpreted as an error estimation. The left-hand side is the *error of the approximate solution* measured in a-norm, while then right-hand side is an *approximation error* which shows, how precisely the vector u^* can be approximated by the elements of the subspace V_h. The proposition says that the vector u_h results in the minimal distance (measured in a-norm):

$$||u^* - u_h||_a = \min_{v_h \in V_h} ||u^* - v_h||_a$$

The above estimation can be performed with respect to the original norm of the Hilbert space H:

Theorem (Céa's lemma): For arbitrary vector $v_h \in V_h$

$$||u^* - u_h|| \leq \frac{M}{m} \cdot ||u^* - v_h||$$

Proof: Let $v_h \in V_h$ be arbitrary. Utilizing the inequalities $m \cdot ||u||^2 \leq a(u,u) \leq M \cdot ||u||^2$:

$$m \cdot ||u^* - u_h||^2 \leq a(u^* - u_h, u^* - u_h) = a(u^* - u_h, u^* - v_h + v_h - u_h) =$$

$$= a(u^* - u_h, u^* - v_h) + a(u^* - u_h, v_h - u_h)$$

Due to the orthogonality property, the last term in the right-hand side vanishes, therefore:

$$m \cdot ||u^* - u_h||^2 \leq a(u^* - u_h, u^* - v_h) \leq M \cdot ||u^* - u_h|| \cdot ||u^* - v_h||,$$

which implies the theorem.

The estimation of the right-hand side can be performed by more standard tools *without knowing the discrete variational solution* u_h.

Remarks:

1. Using the previous proposition

   $$||u^* - u_h||_a \leq ||u^* - v_h||_a,$$

   and the inequalities $m \cdot ||u||^2 \leq a(u,u) \leq M \cdot ||u||^2$, we immediately obtain a sharper estimation for Céa's lemma:

   $$m \cdot ||u^* - u_h||^2 \leq ||u^* - u_h||_a^2 \leq ||u^* - v_h||_a^2 \leq M \cdot ||u^* - v_h||^2,$$

   whence:

   $$||u^* - u_h|| \leq \sqrt{\frac{M}{m}} \cdot ||u^* - v_h||$$

   for all $v_h \in V_h$.

2. The above technique (when we use the same subspace V_h in the variational equalities) is known also as *Galerkin method*. However, it is also possible to define another finite-dimensional subspace W_h (the space of *test functions*) with the basis ψ_1, ψ_2, ..., ψ_M (the dimensions N and M may differ). This technique is called *Petrov-Galerkin method*. In this case, the variational equalities are enforced to the vectors ψ_k $(k = 1, 2, ..., M)$. Thus, we obtain another discrete variational problem:

$$\sum_{j=1}^{N} \alpha_j \cdot a(\varphi_j, \psi_k) = \ell\psi_k \qquad (k = 1, 2, ..., M)$$

However, the matrix of tis system does not remain self-adjoint in general, and the solvability is not guaranteed in advance.

5.2 Finite Element Method for 1D Poisson problems

Let $(a, b) \subset \mathbf{R}$ be a finite interval. Let $f : (a, b) \to \mathbf{R}$ be a given (square integrable) function, and consider the simplest 1D Poisson equation:

$$-u'' = f \qquad \text{in } (a, b)$$

supplied with homogeneous Dirichlet boundary condition:

$$u(a) = u(b) = 0.$$

First, a proper Hilbert space H should be defined.

Introduce the function space:

$$H := H_0^1(a, b) := \{w \in L_2(a, b) : \ w' \in L_2(a, b), \ w(a) = w(b) = 0\}$$

with the inner product:

$$\langle u, v \rangle_H := \int_a^b u'(x) \cdot v'(x)\, dx$$

For the sake of simplicity, the inner product and the norm in the space $L_2(a, b)$ will be denoted as follows:

$$\langle u, v \rangle_0 := \int_a^b u(x) \cdot v(x)\, dx, \qquad ||u||_0 := \sqrt{\int_a^b |u(x)|^2\, dx}$$

It can be easily checked that all the properties of the H-inner product are fulfilled, thus, H is a Euclidean space. The completeness is not trivial. It is stated here without proof:

Theorem: The space H is a Hilbert space.

The norm in the space H is characterized by the following important property:

Theorem (Poincaré's inequality): There exists a positive constant $c > 0$ such that for every $u \in H_0^1(a, b)$:

$$\int_a^b |u'(x)|^2\, dx \geq c \cdot \int_a^b |u(x)|^2\, dx$$

that is:

$$||u||_H^2 \geq c \cdot ||u||_0^2$$

Proof: Since $u(a) = 0$, for arbitrary $x \in (a, b)$, obviously:

$$u(x) = \int_a^x u'(t)\, dt$$

Applying the Cauchy inequality in the function space $L_2(a, b)$:

$$|u(x)|^2 = \left| \int_a^x 1 \cdot u'(t)\, dt \right|^2 \leq \left(\int_a^x 1^2\, dt \right) \cdot \left(\int_a^x |u'(t)|^2\, dt \right) \leq$$

$$\leq x \cdot \int_a^b |u'(t)|^2\, dt$$

Integrating both sides with respect to x, we have:

$$\int_a^b |u'(x)|^2\, dx \leq \left(\int_a^b x\, dx \right) \cdot \left(\int_a^b |u'(t)|^2\, dt \right) = \frac{b^2 - a^2}{2} \cdot \int_a^b |u'(t)|^2\, dt,$$

which completes the proof.

Remark: Observe that it is not exploited in the proof that u vanishes at *both* endpoints of the interval. The proof is essentially unchanged if the condition $u(a) = 0$ *or* the condition $u(b) = 0$ is assumed only.

Now the bilinear functional a and the linear functional ℓ should be defined. Multiplying the original differential equation by an arbitrary function $v \in H_0^1(a, b)$ and integrating over the interval:

$$-\int_a^b u''(x) \cdot v(x)\, dx = \int_a^b f(x) \cdot v(x)\, dx$$

Integrating by parts on the left-hand side, we obtain:

$$\int_a^b u'(x) \cdot v'(x)\, dx = \int_a^b f(x) \cdot v(x)\, dx$$

This is called the *weak form* of the original Poisson problem. Now let us define the desired functionals as follows:

$$a(u, v) := \langle u, v \rangle_H = \int_a^b u'(x) \cdot v'(x)\, dx$$

$$\ell v := \langle f, v \rangle_{L_2(a,b)} = \int_a^b f(x) \cdot v(x)\, dx$$

Now it should be checked whether these functionals have the desired properties. Indeed:

- a is linear in its both variables
- $a(u,v) = a(v,u)$ for all $u, v \in H$ i.e a is symmetric;
- $|a(u,v)| = |\langle u, v\rangle_H| \leq ||u||_H \cdot ||v||_H$, i.e. a is bounded (here the Cauchy inequality has been applied);
- $|a(u,u)| = ||u||_H^2$, i.e. a is coercive;
- $|\ell v| = |\langle f, v\rangle_0| \leq ||f||_0 \cdot ||v||_0 \leq \dfrac{1}{c} \cdot ||f||_0 \cdot ||v||_H$, i.e. ℓ is bounded (here the Cauchy and the Poincaré inequalities have been applied).

Thanks to the abstract results, we already know that the variational problem does have a unique solution in the Hilbert space H. It should be pointed out that in the variational problem, *no second-order derivatives occur*. This simplifies also the determination of the approximate solution.

5.2.1 Finite element subspaces

The next task is to properly define the finite dimensional subspace V_h. As a simple example, let V_h be the subspace of the *piecewise linear functions that satisfy the homogeneous Dirichlet boundary condition*. For this reason, define a subdivision of the interval (a, b) by the not necessarily equidistant point set $a = x_0 < x_1 < ... < x_N = b$. Denote by $h_k := x_{k+1} - x_k$ $(k = 0, 1, ...N - 1)$.

Now consider the piecewise linear functions defined at the gridpoints as follows:

$$\varphi_k(x_k) := 1, \qquad \text{and} \quad \varphi_k(x_j) := 0 \quad \text{if } j \neq k$$

$(k = 1, 2, ..., N - 1)$. Then

$$V_h = \text{span}\{\varphi_1, \varphi_2, ..., \varphi_{N-1}\}$$

It is clear that V_h is an $(N - 1)$-dimensional subspace, and the functions φ_1, φ_2, ..., φ_{N-1} form a basis in V_h. The functions φ_k are often called *hat functions* because of the shape of their graph (see Figure 6) The abstract results assure that the discrete variational problem also has a unique solution in the subspace V_h, and it is expressed in the form:

$$u_h = \sum_{j=1}^{N-1} \alpha_j \varphi_j,$$

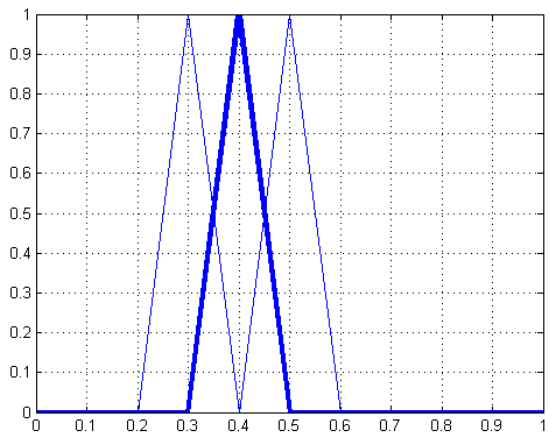

Figure 6: Some hat functions.

where the coefficients α_1, ..., α_{N-1} satisfy the discrete system

$$A\alpha = b,$$

where

$$A_{k,j} = a(\varphi_j, \varphi_k), \qquad b_k = \int_a^b f(x) \cdot \varphi_k(x)\, dx$$

Due to the definition of the basis functions φ_k, the stiffness matrix A is *tridiagonal*. Since the derivative of a hat function is a piecewise constant function, namely:

$$\varphi_k'(x) = \begin{cases} \frac{1}{h_{k-1}} & (x \in (x_{k-1}, x_k)) \\ \\ -\frac{1}{h_k} & (x \in (x_k, x_{k+1})) \end{cases}$$

therefore the calculation of the entries of the stiffness matrix is extremely simple:

$$A_{k,k} = \int_{x_{k-1}}^{x_{k+1}} (\varphi_k'(x))^2\, dx = \frac{1}{h_{k-1}} + \frac{1}{h_k}$$

$$A_{k,k-1} = \int_{x_{k-1}}^{x_{k+1}} \varphi_{k-1}'(x) \cdot \varphi_k'(x)\, dx = -\frac{1}{h_{k-1}}$$

$$A_{k,k+1} = \int_{x_{k-1}}^{x_{k+1}} \varphi'_{k+1}(x) \cdot \varphi'_k(x)\, dx = -\frac{1}{h_k}$$

All the remaining entries of A equal to zero.

The components of the right-hand side should be approximated in general (if the integral cannot be calculated exactly). For instance, if the integral is approximated by the trapezoidal rule, we obtain:

$$b_k = \int_a^b f(x) \cdot \varphi_k(x)\, dx = \int_{x_{k-1}}^{x_{k+1}} f(x) \cdot \varphi_k(x)\, dx \approx$$

$$\approx \frac{0 + f(x_k)}{2} \cdot (x_k - x_{k-1}) + \frac{f(x_k) + 0}{2} \cdot (x_{k+1} - x_k) =$$

$$= \frac{h_{k-1} + h_k}{2} \cdot f(x_k)$$

5.2.2 Error estimations

According to Céa's lemma, it is sufficient to examine how exactly the functions belonging to H can be approximated by piecewise linear functions.

For the sake of simplicity, assume that the set of points x_0, x_1, ..., x_N is equidistant: $x_k = a + k \cdot h$, where $h = \dfrac{b-a}{N}$ is the stepsize; N is a predefined number, which characterizes the 'resolution' of the subspace V_h.

For the time being, let $F \in C^2[a, b]$ be an arbitrary smooth function. Denote by S the first-degree spline (i.e. piecewise linear function), which takes the value $F_k := F(x_k)$ at the point x_k for $k = 0, 1, ..., N$. On the interval $[x_k, x_{k+1}]$, by definition:

$$S(x) = F_k + \frac{F_{k+1} - F_k}{h} \cdot (x - x_k)$$

Applying Lagrange's mean value theorem, we obtain that:

$$S(x) = F_k + F'(\xi) \cdot (x - x_k)$$

for some $\xi \in (x_k, x_{k+1})$. That is:

$$S'(x) \equiv F'(\xi) = const.$$

on the subinterval $[x_k, x_{k+1}]$. On the other hand, expand the function F' in finite Taylor series around x_k:

$$F'(x) = F'(x_k) + F''(\tau) \cdot (x - x_k)$$

for some $\tau \in (x_k, x_{k+1})$.

Applying Lagrange's mean value theorem once more (to the function F'), we have:

$$|F'(x) - S'(x)|^2 = |F'(x_k) - F'(\xi) + F''(\tau) \cdot (x - x_k)|^2 =$$

$$= |F''(\theta) \cdot (x_k - \xi) + F''(\tau) \cdot (x - x_k)|^2 \leq$$

$$\leq ||F''||^2_{\max} \cdot (h + |x - x_k|)^2 \leq ||F''||^2_{\max} \cdot 4h^2$$

This estimation is valid on each subinterval. Integrating over the interval (a, b) with respect to x, we obtain:

$$\int_a^b |F'(x) - S'(x)|^2 \, dx \leq 4 \cdot (b - a) \cdot ||F''||^2_{\max} \cdot h^2$$

that is:

$$||F - S||_H = ||F - S||_{H_0^1(a,b)} \leq 2 \cdot ||F''||_{\max} \cdot \sqrt{b - a} \cdot h = \mathcal{O}(h)$$

This result means that the approximation of the exact solution u^* of the original problem by the elements of the subspace V_h is also at least $\mathcal{O}(h)$. Thus, Céa's lemma gives us the following error estimation:

$$\boxed{||u^* - u_h||_{H_0^1(a,b)} \leq 2 \cdot ||(u^*)''||_{\max} \cdot \sqrt{b - a} \cdot h = \mathcal{O}(h)}$$

provided that the exact solution u^* is smooth enough (which is the case, when the function f is sufficiently smooth).

That is, *the error of the discrete solution is $\mathcal{O}(h)$ with respect to the $H_0^1(a, b)$-norm.*

Remark:

- If we want to analyse the error *with respect to the weaker $L_2(a, b)$-norm*, a sharper estimation can be obtained (also known as 'Nitsche's trick').

  First, summarize the estimations which are known from the previous considerations. There exists a constant $C \geq 0$ independent of h such that:

  1. $||u^*||_H \leq C \cdot ||f||_0$
  2. $||u^* - u_h||_H \leq C \cdot h \cdot ||(u^*)''||_0$

Recall also the orthogonality property:

$$\langle u^* - u_h, v_h \rangle_H = 0 \qquad \text{for all } v_h \in V_h$$

Now consider the auxiliary problem:

$$-w'' = u^* - u_h, \qquad w(a) = 0, \quad w(b) = 0$$

Denote by w^* the exact solution, and let $w_h \in V_h$ be the solution of the corresponding discrete variational problem.

Applying the variational equality $a(w^*, v) = \int_a^b (u^* - u_h) \cdot v \, dx$ to the function $v := u^* - u_h$, we have:

$$a(w^*, u^* - u_h) = ||u^* - u_h||_0^2$$

On the other hand, using the orthogonality property:

$$a(w_h, u^* - u_h) = a(u^* - u_h, w_h) = 0$$

Subtracting this equality from the previous one:

$$a(w^* - w_h, u^* - u_h) = ||u^* - u_h||_0^2,$$

that is:

$$||u^* - u_h||_0^2 = a(w^* - w_h, u^* - u_h) = \langle w^* - w_h, u^* - u_h \rangle_H \leq$$

$$\leq ||w^* - w_h||_H \cdot ||u^* - u_h||_H,$$

where we applied the Cauchy inequality. Now let us apply the estimation 2. to both factors of the right-hand side. We have:

$$||u^* - u_h||_0^2 \leq const.h^2 \cdot ||(w^*)''||_0 \cdot ||(u^*)''||_0$$

But $||(w^*)''||_0 = ||u^* - u_h||_0$, which implies that:

$$\boxed{||u^* - u_h||_0 \leq const.h^2 \cdot ||(u^*)''||_0 = \mathcal{O}(h^2)}$$

That is, *the error of the discrete solution is $\mathcal{O}(h^2)$ with respect to the $L_2(a, b)$-norm.*

5.2.3 Finite elements for more general 1D problems

Let $(a, b) \subset \mathbf{R}$ be a finite interval. Let $f : (a, b) \to \mathbf{R}$ be a given (square integrable) function, and consider the following 1D elliptic equation:

$$-(k \cdot u')' + d \cdot u = f \qquad \text{in } (a, b)$$

supplied with homogeneous Dirichlet boundary condition:

$$u(a) = u(b) = 0.$$

Here k is a given, positive, bounded function, d is a nonnegative function.

Assume that

$$0 < k_0 \leq k(x) \leq k_1, \qquad 0 \leq d(x) \leq d_1$$

Now the problem should be reformulated in a variational form. Again, introduce the Hilbert space:

$$H := H_0^1(a, b) := \{w \in L_2(a, b) : \ w' \in L_2(a, b), \ w(a) = w(b) = 0\}$$

with the inner product:

$$\langle u, v \rangle_H := \int_a^b u'(x) \cdot v'(x) \, dx$$

Now the bilinear functional a and the linear functional ℓ should be defined. Multiplying the original differential equation by an arbitrary function $v \in H_0^1(a, b)$ and integrating over the interval:

$$-\int_a^b ((k(x) \cdot u'(x))' + d \cdot u(x) \cdot v(x)) \, dx = \int_a^b f(x) \cdot v(x) \, dx$$

Integrating by parts on the left-hand side, we obtain:

$$\int_a^b (k(x) \cdot u'(x) \cdot v'(x) + d(x) \cdot u(x) \cdot v(x)) \, dx = \int_a^b f(x) \cdot v(x) \, dx$$

This is called the *weak form* of the original elliptic problem. Now let us define the desired functionals as follows:

$$a(u, v) := \int_a^b (k(x) \cdot u'(x) \cdot v'(x) + d(x) \cdot u(x) \cdot v(x)) \, dx$$

$$\ell v := \langle f, v \rangle_{L_2(a,b)} = \int_a^b f(x) \cdot v(x) \, dx$$

It should be checked whether these functionals have the desired properties. Straightforward calculations show that:

- a is linear in its both variables
- $a(u,v) = a(v,u)$ for all $u, v \in H$ i.e a is symmetric
- $|a(u,v)| \leq const. \cdot ||u||_H \cdot ||v||_H$, i.e. a is bounded (here the Cauchy inequality has been applied)
- $|a(u,u)| \geq k_0 \cdot ||u||_H^2$, i.e. a is coercive
- $|\ell v| = |\langle f, v\rangle_0| \leq ||f||_0 \cdot ||v||_0 \leq \frac{1}{c} \cdot ||f||_0 \cdot ||v||_H$, i.e. ℓ is bounded (here the Cauchy and the Poincaré inequalities have been applied).

Define the subspace V_h in such a way than earlier: let V_h be the subspace of the piecewise linear functions that satisfy the homogeneous Dirichlet boundary condition. Define a subdivision of the interval (a, b) by the not necessarily equidistant point set $a = x_0 < x_1 < ... < x_N = b$. Introduce the basis functions again:

$$\varphi_k(x_k) := 1, \qquad \text{and} \quad \varphi_k(x_j) := 0 \quad \text{if } j \neq k$$

$(k = 1, 2, ..., N-1)$.

The abstract results still assure that the discrete variational problem also has a unique solution in the subspace V_h, and it is expressed in the form:

$$u_h = \sum_{j=1}^{N-1} \alpha_j \varphi_j,$$

where the coefficients α_1, ..., α_{N-1} satisfy the discrete system

$$A\alpha = b,$$

where

$$A_{k,j} = a(\varphi_j, \varphi_k) = \int_a^b (k(x)\varphi_j'(x)\varphi_k'(x) + d\varphi_j(x)\varphi_k(x))\, dx$$

and

$$b_k = \int_a^b f(x) \cdot \varphi_k(x)\, dx$$

It is typical that the function k is piecewise constant. The entries of the stiffness matrix and the components of the right-hand side can now be calculated in a straightforward, but possibly a more difficult way than earlier.

5.2.4 1D problems, inhomogeneous Dirichlet boundary condition

Let $(a, b) \subset \mathbf{R}$ be a finite interval. Let $f : (a, b) \to \mathbf{R}$ be a given (square integrable) function, and consider the following 1D elliptic equation:

$$-(k \cdot u')' + d \cdot u = f \qquad \text{in } (a, b)$$

supplied with inhomogeneous Dirichlet boundary condition:

$$u(a) = A, \qquad u(b) = B.$$

Again, here k is a given, positive, bounded function, d is a nonnegative function. Assume again that

$$0 < k_0 \leq k(x) \leq k_1, \qquad 0 \leq d(x) \leq d_1$$

In principle, the exact solution can be expressed as follows. Let g be a sufficiently regular function that satisfies the inhomogeneous boundary conditions:

$$g(a) = A, \qquad g(b) = B,$$

and seek the solution in the form $u := w + g$, where w satisfies a modified partial differential equation:

$$-(k \cdot u')' + d \cdot u = f + (k \cdot g')' + d \cdot g \qquad \text{in } (a, b)$$

supplied with homogeneous boundary conditions:

$$w(a) = 0, \qquad w(b) = 0.$$

From here, the previous techniques can be applied.

In practice, the algorithm is much simpler. Introduce the 'one-sided' piecewise linear hat functions φ_0 and φ_N by defining $\varphi_0(x_0) = 1$ and $\varphi_N(x_N) = 1$ (see Figure 7). Now seek the discrete solution u_h in the following form:

$$u_h := \sum_{j=1}^{N-1} \alpha_j \varphi_j + A \cdot \varphi_0 + B \cdot \varphi_N,$$

and the discrete variational equations are as follows:

$$\sum_{j=1}^{N-1} \alpha_j a(\varphi_j, \varphi_k) + A \cdot a(\varphi_0, \varphi_k) + B \cdot a(\varphi_N, \varphi_k) = \ell \varphi_k$$

for $k = 1, 2, ..., N - 1$.

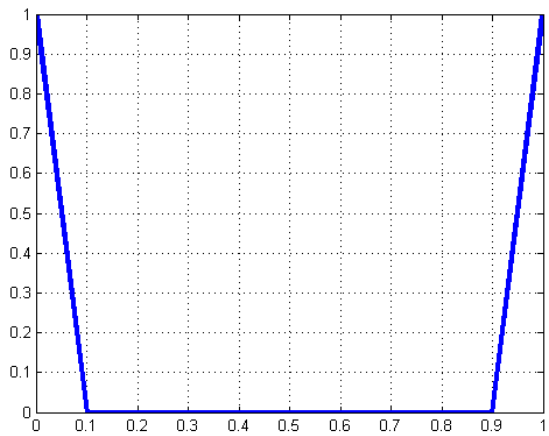

Figure 7: One-sided hat functions.

5.2.5 1D problems, mixed boundary condition

Let $(a, b) \subset \mathbf{R}$ be a finite interval. Let $f : (a, b) \to \mathbf{R}$ be a given (square integrable) function, and consider the following 1D elliptic equation:

$$-(k \cdot u')' + d \cdot u = f \qquad \text{in } (a, b)$$

supplied with mixed boundary condition:

$$u(a) = 0, \qquad k \cdot u'(b) = C$$

i.e. at the right-hand endpoint of the interval, a Neumann boundary condition is prescribed.

As earlier, k is a given, positive, bounded function, d is a nonnegative function. Assume again that

$$0 < k_0 \leq k(x) \leq k_1, \qquad 0 \leq d(x) \leq d_1$$

In contrast to the pure Dirichlet boundary condition, where the type of boundary condition affects the proper definition of the Hilbert space H, the Neumann type boundary condition affects the proper definition of the linear functional ℓ.

Introduce the Hilbert space:

$$H := \{w \in L_2(a, b) : \ w' \in L_2(a, b), \ w(a) = 0\}$$

with the inner product:

$$\langle u, v\rangle_H := \int_a^b u'(x) \cdot v'(x)\, dx$$

That is, the Dirichlet condition appears *in the definition of* H. It is not trivial that H is a Hilbert space again (this statement is not proved here), moreover, as observed earlier, the Poincaré inequality remains valid in this space; there exists a positive constant c such that the inequality

$$\int_a^b |u'(x)|^2\, dx \geq c \cdot \int_a^b |u(x)|^2\, dx$$

is valid for arbitrary $u \in H$.

In order that the bilinear functional a and the linear functional ℓ is defined properly, let us multiply the original differential equation by an arbitrary function $v \in H$ and integrate over the interval:

$$-\int_a^b ((k(x) \cdot u'(x))' + d \cdot u(x) \cdot v(x))\, dx = \int_a^b f(x) \cdot v(x)\, dx$$

Integrating by parts on the left-hand side, we obtain:

$$\left[-ku'v\right]_a^b + \int_a^b (k(x) \cdot u'(x) \cdot v'(x) + d(x) \cdot u(x) \cdot v(x))\, dx = \int_a^b f(x) \cdot v(x)\, dx$$

that is, since $v(a) = 0$ but $v(b) \neq 0$ in general:

$$\int_a^b (k(x) \cdot u'(x) \cdot v'(x) + d(x) \cdot u(x) \cdot v(x))\, dx = \int_a^b f(x) \cdot v(x)\, dx + (ku')(b) \cdot v(b)$$

But the Neumann boundary condition assures that $(k \cdot u'(b)) = C$. Thus, we have obtained the *weak form* of the original problem:

$$\int_a^b (k(x) \cdot u'(x) \cdot v'(x) + d(x) \cdot u(x) \cdot v(x))\, dx = \int_a^b f(x) \cdot v(x)\, dx + C \cdot v(b)$$

Now let us define the desired functionals as follows:

$$a(u, v) := \int_a^b (k(x) \cdot u'(x) \cdot v'(x) + d(x) \cdot u(x) \cdot v(x))\, dx$$

$$\ell v := \langle f, v\rangle_{L_2(a,b)} + C \cdot v(b) = \int_a^b f(x) \cdot v(x)\, dx + C \cdot v(b)$$

It should be checked whether these functionals have the desired properties. However, one should be careful since the definition of H and ℓ have been changed. We summarize the necessary statements:

- a is linear in its both variables
- $a(u,v) = a(v,u)$ for all $u, v \in H$ i.e a is symmetric
- $|a(u,v)| \leq const. \cdot ||u||_H \cdot ||v||_H$, i.e. a is bounded (here the Cauchy inequality has been applied)
- $|a(u,u)| \geq k_0 \cdot ||u||_H^2$, i.e. a is coercive
- $|\ell v| = |\langle f, v\rangle_0| \leq ||f||_0 \cdot ||v||_0 \leq \dfrac{1}{c} \cdot ||f||_0 \cdot ||v||_H$, i.e. ℓ is bounded (here the Cauchy and the Poincaré inequalities have been applied).

Remark: At the last item, we utilized the fact that the functional $v \to v(b)$ is bounded with respect to the norm of H. Indeed:

$$v(b) = \int_a^b 1 \cdot v'(x)\, dx \leq \sqrt{(b-a) \cdot \int_a^b |v'(x)|^2\, dx} = \sqrt{b-a} \cdot ||v||_H.$$

Now let V_h be the subspace of the piecewise linear functions that satisfy the homogeneous Dirichlet boundary condition at the point $x = a$ only. Define a subdivision of the interval (a, b) by the not necessarily equidistant point set $a = x_0 < x_1 < ... < x_N = b$. Introduce the basis functions again:

$$\varphi_k(x_k) := 1, \qquad \text{and} \quad \varphi_k(x_j) := 0 \quad \text{if } j \neq k$$

$(k = 1, 2, ..., N)$. Observe that the 'one-sided' function φ_N is chosen to the basis in V_h; thus, the dimension of V_h is N (in contrast to the case of the pure Dirichlet boundary condition).

The discrete variational problem still has a unique solution in the subspace V_h, and it is expressed in the form:

$$u_h = \sum_{j=1}^{N} \alpha_j \varphi_j,$$

where the coefficients α_1, ..., α_N satisfy the discrete system

$$A\alpha = b,$$

where

$$A_{k,j} = a(\varphi_j, \varphi_k) = \int_a^b (k(x)\varphi_j'(x)\varphi_k'(x) + d\varphi_j(x)\varphi_k(x))\, dx$$

and

$$b_k = \int_a^b f(x) \cdot \varphi_k(x)\, dx + C \cdot \varphi_k(b)$$

$(k, j = 1, 2, ..., N)$.

Remark: The case when the Dirichlet boundary condition is inhomogeneous can be treated in a similar way as earlier. If, for instance

$$u(a) = A$$

is prescribed, then the discrete variational solution u_h is expressed in the following form:

$$u_h = \sum_{j=1}^{N} \alpha_j \varphi_j + A \cdot \varphi_0,$$

which causes a change in the right-hand side of the discrete equations.

Summarizing the main tricks of the construction of a finite element technique, the essential steps are as follows:

- Define the partial differential equation together with the boundary conditions in a traditional form.
- Define a proper Hilbert space H, the element of which satisfy the corresponding *homogeneous* Dirichlet boundary condition.
- Multiply both sides of the partial differential equation by an arbitrary test function v and apply an integral transform theorem (integration by parts). Then define the bilinear functional a and the linear functional ℓ properly. Note that the Neumann type boundary condition (if exists) appears always in the definition of ℓ, while the Dirichlet type boundary condition affects the choice of the space H.
- Check the necessary properties of the above defined functionals (coercivity, boundedness).
- Define a finite dimensional space V_h of the trial functions.
- Assemble and solve the discrete system of variational equations.
- If possible, try to estimate the accuracy of the method.

5.3 Finite Element Method for 2D Poisson problems

Let $\Omega \subset \mathbf{R}^2$ be a bounded domain, denote by Γ the boundary of the domain. Let $f : \Omega \to \mathbf{R}$ be a given (square integrable) function, and consider the 2D Poisson equation:

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

supplied with homogeneous Dirichlet boundary condition:

$$u|_\Gamma = 0.$$

Introduce the function space:

$$H := H_0^1(\Omega) := \left\{ w \in L_2(\Omega) : \quad \frac{\partial w}{\partial x}, \frac{\partial w}{\partial y} \in L_2(\Omega),\ w|_\Gamma = 0 \right\}$$

with the inner product:

$$\langle u, v \rangle_H := \int_\Omega \langle \operatorname{grad} u, \operatorname{grad} v \rangle \, dxdy$$

For brevity, the inner product and the norm in the space $L_2(\Omega)$ will be denoted by:

$$\langle u, v \rangle_0 := \int_\Omega u(x,y) \cdot v(x,y) \, dxdy, \qquad ||u||_0 := \sqrt{\int_\Omega |u(x,y)|^2 \, dxdy}$$

The properties of the H-inner product are fulfilled, i.e. H is a Euclidean space. Moreover (without proof):

Theorem: The Euclidean space H is complete, i.e. it is a Hilbert space.

The Poincaré inequality is still valid. The essential ideas of the proof are unchanged:

Theorem (Poincaré's inequality): There exists a positive constant $c > 0$ such that for every $u \in H_0^1(\Omega)$:

$$\int_\Omega ||\operatorname{grad} u||^2 \, dxdy \geq c \cdot \int_\Omega |u(x,y)|^2 \, dxdy$$

that is:

$$||u||_H^2 \geq c \cdot ||u||_0^2$$

Proof: Since $u|_\Gamma = 0$, the function u can be extended to a larger rectangle $(0,a) \times (0,b)$ by zero outside of Ω. Let $y \in (0,b)$ be an arbitrary, fixed number. Since $u(0,y) = 0$, we have:

$$u(x,y) = \int_0^x \frac{\partial u}{\partial x}(t,y)\,dt$$

Hence:

$$|u(x,y)|^2 = \left(\int_0^x 1 \cdot \frac{\partial u}{\partial x}(t,y)\,dt\right)^2 \leq \left(\int_0^x 1^2\,dt\right) \cdot \left(\int_0^a \left(\frac{\partial u}{\partial x}(t,y)\right)^2 dt\right) =$$

$$= x \cdot \int_0^a \left(\frac{\partial u}{\partial x}(t,y)\right)^2 dt$$

Integrating both sides over the interval $(0,a)$ with respect to x:

$$\int_0^a |u(x,y)|^2\,dx \leq \frac{a^2}{2} \cdot \int_0^a \left(\frac{\partial u}{\partial x}(x,y)\right)^2 dx$$

Integrating both sides over the interval $(0,b)$ with respect to y:

$$\int_0^b \int_0^a |u(x,y)|^2\,dxdy = \int_\Omega |u(x,y)|^2\,dxdy \leq$$

$$\leq \frac{a^2}{2} \cdot \int_0^b \int_0^a \left(\frac{\partial u}{\partial x}(x,y)\right)^2 dxdy =$$

$$= \frac{a^2}{2} \cdot \int_\Omega \left(\frac{\partial u}{\partial x}(x,y)\right)^2 dxdy$$

Similarly:

$$\int_0^b \int_0^a |u(x,y)|^2\,dxdy = \int_\Omega |u(x,y)|^2\,dxdy \leq \frac{b^2}{2} \cdot \int_\Omega \left(\frac{\partial u}{\partial y}(x,y)\right)^2 dxdy$$

Adding the two inequalities:

$$\int_\Omega |u(x,y)|^2\,dxdy \leq \frac{a^2+b^2}{4} \cdot \int_\Omega \left(\left(\frac{\partial u}{\partial y}(x,y)\right)^2 + \left(\frac{\partial u}{\partial y}(x,y)\right)^2\right) dxdy$$

which implies the theorem.

Remark: Without going into details we note that the theorem remains valid,

if, in the definition of H, the property $u|_{\Gamma_0} = 0$ is required (instead of $u|_{\Gamma} = 0$), where Γ_0 is a nonempty part of the boundary. That is, it is sufficient to require that the function u vanishes along a part of the boundary only.

Now the bilinear functional a and the linear functional ℓ will be defined. Multiplying the original differential equation by an arbitrary function $v \in H_0^1(\Omega)$ and integrating over the interval:

$$-\int_\Omega \Delta u(x,y) \cdot v(x,y)\, dxdy = \int_\Omega f(x,y) \cdot v(x,y)\, dxdy$$

Using Green's first formula on the left-hand side, we obtain:

$$\int_\Omega \langle \operatorname{grad}(x,y), \operatorname{grad} v(x,y) \rangle\, dxdy = \int_\Omega f(x,y) \cdot v(x,y)\, dxdy$$

This is called the *weak form* of the original Poisson problem. Now let us define the desired functionals as follows:

$$a(u,v) := \langle u, v \rangle_H = \int_\Omega \langle \operatorname{grad} u(x,y), \operatorname{grad} v(x,y) \rangle\, dxdy$$

$$\ell v := \langle f, v \rangle_0 = \int_\Omega f(x,y) \cdot v(x,y)\, dxdy$$

Now it should be checked whether these functionals have the desired properties. Indeed:

- a is linear in its both variables
- $a(u,v) = a(v,u)$ for all $u, v \in H$ i.e a is symmetric
- $|a(u,v)| = |\langle u, v \rangle_H| \leq ||u||_H \cdot ||v||_H$, i.e. a is bounded (here the Cauchy inequality has been applied)
- $|a(u,u)| = ||u||_H^2$, i.e. a is coercive
- $|\ell v| = |\langle f, v \rangle_0| \leq ||f||_0 \cdot ||v||_0 \leq \dfrac{1}{c} \cdot ||f||_0 \cdot ||v||_H$, i.e. ℓ is bounded (here the Cauchy and the Poincaré inequalities have been applied).

Thanks to the abstract results, we already know that the variational problem does have a unique solution in the Hilbert space H. It should be pointed out that in the variational problem, *only first-order derivatives occur*. This simplifies also the determination of the approximate solution.

5.3.1 Finite element subspaces

The proper definition of the finite dimensional subspace V_h is a much more complicated task than in 1D problems, where an interval has to be subdivided only. In 2D problems, the domain Ω may have a complicated shape and the structure of the subspace V_h should fit the domain in a certain sense.

Perhaps the most straightforward way is to cover the domain Ω by a *set of triangles* ('mesh' in the following). The following restrictions are required:

- the interiors of the triangles have to be disjoint
- the union of the triangles should be identical to the domain as precisely as possible
- the triangles should fit the boundary of the domain as exactly as possible
- the triangles should join each others by whole sides
- the angles of the triangles should not be 'too small'.

The triangles of the systems are called *finite elements*. The corner points of the triangles are called *node points* or simply nodes. As an illustration, see Figure 8. In 2D (and especially in 3D), the definition of the basis func-

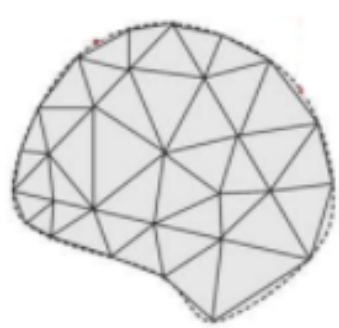

Figure 8: Triangular covering of a 2D domain.

tion (trial functions) is much more sophisticated task than in 1D. One of the simplest examples for trial functions is the system of *piecewise linear (bivariate)* functions. A typical basis in the subspace of these functions is the system of 'tent functions'. Such a function takes the value 1 in a specific node and zeros in all other nodes. This is a straightforward generalization of

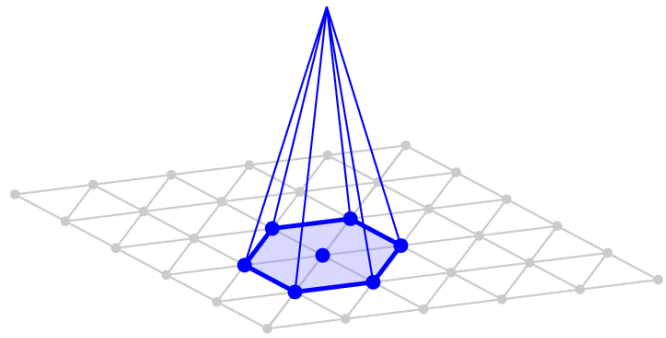

Figure 9: The graph of a typical tent function.

the one-dimensional 'hat functions' (see Figure 9 for illustration). Since the gradient of a bivariate piecewise linear function is piecewise constant vector functions, the assembly of the stiffness matrix and the right-hand vector is simple (in principle). However, in practice, one should create

- a list of nodes
- a list of elements (referred to by the indices of the nodes)
- a list of the elements that contain the node as a corner.

The stiffness matrix is sparse, because only the trial functions which belong to neighbouring nodes generate nonzero entries. Thus, to numerically solve the discrete variational equations, an iterative (preferably based on Krylov subspaces) method can be recommended.

As far as the Neumann type boundary conditions are concerned, we point out again that the proper definition of the Hilbert space H is affected by the Dirichlet boundary condition. The Neumann type boundary condition affects the concrete form of the linear functional ℓ.

As an illustrative example, consider a square domain covered by an equidistant grid with stepsize h in both principal directions. Let us divide each cell into two congruent triangles by drawing a diagonal between the lower left and the upper right corners. The support of the basis function φ_C corresponding to the central point C is shown in Figure 10. The gradient vector of φ_C is constant on each triangle belonging to the support of φ_C. Namely:

- $\operatorname{grad} \varphi_C = \frac{1}{h} \cdot (0, -1)$ on the triangle 1

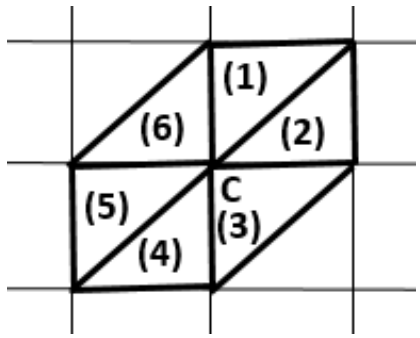

Figure 10: The support of the basis function φ_C.

- $\operatorname{grad}\varphi_C = \frac{1}{h}\cdot(-1,0)$ on the triangle 2
- $\operatorname{grad}\varphi_C = \frac{1}{h}\cdot(-1,1)$ on the triangle 3
- $\operatorname{grad}\varphi_C = \frac{1}{h}\cdot(0,1)$ on the triangle 4
- $\operatorname{grad}\varphi_C = \frac{1}{h}\cdot(1,0)$ on the triangle 5
- $\operatorname{grad}\varphi_C = \frac{1}{h}\cdot(1,-1)$ on the triangle 6

The area of each triangle is $\frac{h^2}{2}$. Therefore the diagonal element of the stiffness matrix belonging to φ_C is:

$$\int_\Omega ||\operatorname{grad}\varphi_C||^2\,dxdy = \frac{1}{2}+\frac{1}{2}+1+\frac{1}{2}+\frac{1}{2}+1 = 4.$$

As far as the off-diagonal entries located in the same row are concerned, only the basis function result in nonzero contributions that belong to a node which is neighbouring with C. The case of the node N is shown in Figure 11 As can be easily checked, the corresponding matrix entry is:

$$\int_\Omega \langle \operatorname{grad}\varphi_C, \operatorname{grad}\varphi_N\rangle\,dxdy = -1,$$

and the same results are valid for the other principal directions W, S and E:

$$\int_\Omega \langle \operatorname{grad}\varphi_C, \operatorname{grad}\varphi_W\rangle\,dxdy = -1,$$

$$\int_\Omega \langle \operatorname{grad}\varphi_C, \operatorname{grad}\varphi_S\rangle\,dxdy = -1,$$

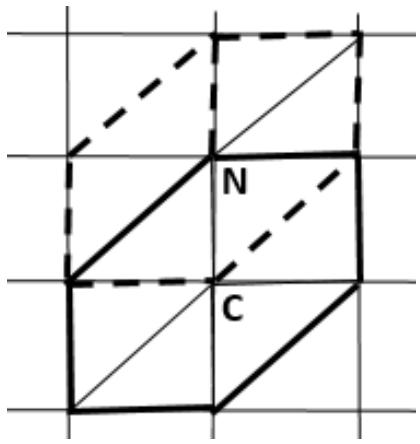

Figure 11: The supports of the basis function φ_C and φ_N.

$$\int_\Omega \langle \operatorname{grad} \varphi_C, \operatorname{grad} \varphi_E \rangle \, dxdy = -1.$$

There are two additional neighbouring basis functions in the direction NE and SW; the first one is shown in Figure 12. Here the contributions of the

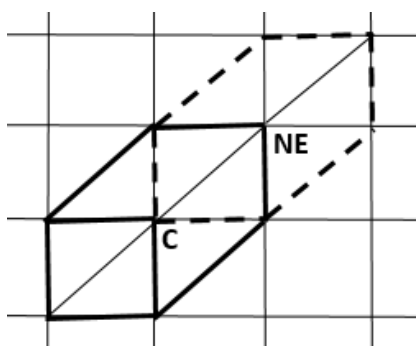

Figure 12: The supports of the basis function φ_C and φ_{NE}.

two triangles cancel out (check it!):

$$\int_\Omega \langle \operatorname{grad} \varphi_C, \operatorname{grad} \varphi_{NE} \rangle \, dxdy = \int_\Omega \langle \operatorname{grad} \varphi_C, \operatorname{grad} \varphi_{SW} \rangle \, dxdy = 0.$$

Thus, the discrete variational equation belonging to the node C is as follows:

$$4\alpha_C - \alpha_N - \alpha_W - \alpha_S - \alpha_E = b_C,$$

where

$$b_C := \int_\Omega f(x,y) \cdot \varphi_C(x,y)\, dxdy$$

It can be seen that the assembly of the discrete variational equations is rather complicated especially when the structure of the mesh is more complex. From a practical point of view, it is more comfortable to treat the trial functions *elementwise*, i.e. to compute the terms that the individual elements contribute to the entries of the stiffness matrix with (and to the components of the right-hand side, respectively).

The trial functions restricted to an element are called *shape functions*. To generate a discrete variational system of equations, it is sufficient to know the shape functions only (instead of the trial functions).

5.3.2 Some 2D finite elements

The description of the shape function is often simpler it the barycentric coordinates are used instead of the familiar x, y coordinates. First we briefly recall the definition:

Barycentric coordinates: Let the not collinear points $(x_1, y_1), (x_2, y_2), (x_3, y_3) \in \mathbf{R}^2$ be given. Denote by (x, y) an arbitrary points of the plane. Then the points (x, y) can be uniquely expressed as a *convex combination of* $(x_1, y_1), (x_2, y_2), (x_3, y_3)$, i.e. there exist some coefficients $\lambda_1, \lambda_2, \lambda_3 \in \mathbf{R}$ such that:

$$x_1\lambda_1 + x_2\lambda_2 + x_3\lambda_3 = x$$

$$y_1\lambda_1 + y_2\lambda_2 + y_3\lambda_3 = y$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

The numbers $\lambda_1, \lambda_2, \lambda_3$ are called the *barycentric coordinates* of the point (x, y). It is obvious that as functions of x and y, they are polynomials of first degree. It is also obvious that

$$\lambda_k(x_k, y_k) = 1, \qquad \lambda_k(x_j, y_j) = 0, \quad \text{if } j \neq k$$

$(k, j = 1, 2, 3)$.

Courant element (or T3 element): Consider the triangular element with the vertices (nodes) (x_1, y_1), (x_2, y_2), (x_3, y_3) (see Figure 13) supplied with the barycentric coordinate functions as shape functions:

$$\lambda_1, \quad \lambda_2, \quad \lambda_3$$

Then V_h is the elementwise linear functions (the 'tent' functions mentioned above).

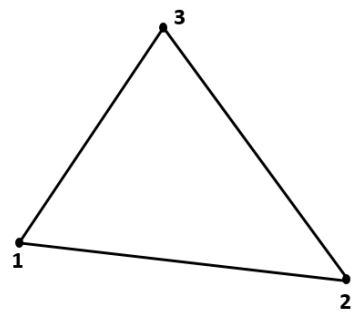

Figure 13: Courant (T_3) element

Second degree triangular element (or T6 element): Consider the triangular element with the vertices (nodes) (x_1, y_1), (x_2, y_2), (x_3, y_3) (see Figure 14). Denote by (x_4, y_4), (x_5, y_5), (x_6, y_6) the midpoints of the sides of the triangle.

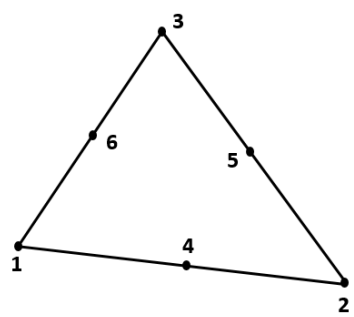

Figure 14: Second-order (T_6) triangular element

Define the following shape functions:

$$w_1 := \lambda_1 \cdot (2 \cdot \lambda_1 - 1), \quad w_2 := \lambda_2 \cdot (2 \cdot \lambda_2 - 1), \quad w_3 := \lambda_3 \cdot (2 \cdot \lambda_3 - 1)$$

$$w_4 := 4 \cdot \lambda_1 \cdot \lambda_2, \quad w_5 := 4 \cdot \lambda_2 \cdot \lambda_3, \quad w_6 := 4 \cdot \lambda_3 \cdot \lambda_1$$

Then

$$w_k(x_k, y_k) = 1, \qquad w_k(x_j, y_j) \ \text{ if } j \neq k$$

$(k, j = 1, 2, ..., 6)$. Moreover the subspace V_h is the elementwise quadratic functions which are continuously connected along the sides of the triangles.

The above elements can be generalized for 3D problems (tetrahedral meshes) in a natural way. Note, however, that a really good mesh generation, which can fit complicated domains (especially in 3D) is still a difficult task in practice and requires special tools and algorithms.

6 Other computational techniques - an outlook

6.1 Method of Fourier

The method is introduced through the example of the 2D Poisson equation. Throughout this subsection, we will use the more familiar notations x, y for the spatial variable (instead of the vector notations).

6.1.1 Fourier's method for 2D Poisson equation

Consider the Poisson equation

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

supplied with the homogeneous Dirichlet boundary condition:

$$u|_\Gamma = 0,$$

where the domain Ω is a rectangle: $\Omega = (0, A) \times (0, B)$.

The main idea of the method is to seek the solution in terms of sinusoidal Fourier series:

$$u(x, y) = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{k,j} \cdot \sin \frac{k\pi x}{A} \sin \frac{j\pi y}{B},$$

where the Fourier coefficients $a_{k,j}$ are unknown.

First, expand the function f in a sinusoidal Fourier series:

$$f(x, y) = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} c_{k,j} \cdot \sin \frac{k\pi x}{A} \sin \frac{j\pi y}{B},$$

where the Fourier coefficients can be computed by evaluating the following integral:

$$c_{k,j} = \frac{4}{AB} \int_\Omega f(x, y) \cdot \sin \frac{k\pi x}{A} \sin \frac{j\pi y}{B} \, dxdy$$

Proof reminder (with simplifications): Multiplying both sides of the equality

$$f(x, y) = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} c_{k,j} \cdot \sin \frac{k\pi x}{A} \sin \frac{j\pi y}{B},$$

by $\sin\frac{p\pi x}{A}\sin\frac{q\pi y}{B}$, and integrating over Ω, we have:

$$\int_\Omega f(x,y)\cdot\sin\frac{p\pi x}{A}\sin\frac{q\pi y}{B}\,dxdy =$$

$$= \sum_{k,j=1}^{\infty} c_{k,j}\cdot\left(\int_0^A \sin\frac{k\pi x}{A}\sin\frac{p\pi x}{A}\,dx\right)\cdot\left(\int_0^B \sin\frac{j\pi y}{B}\sin\frac{q\pi y}{B}\,dy\right)$$

Direct computations show that if $k \neq p$, then:

$$\int_0^A \sin\frac{k\pi x}{A}\sin\frac{p\pi x}{A}\,dx = 0,$$

and similarly, if $j \neq q$, then:

$$\int_0^A \sin\frac{j\pi x}{B}\sin\frac{q\pi x}{B}\,dx = 0.$$

Thus, the above equality is simplified to:

$$\int_\Omega f(x,y)\cdot\sin\frac{p\pi x}{A}\sin\frac{q\pi y}{B}\,dxdy =$$

$$= c_{p,q}\cdot\left(\int_0^A \sin^2\frac{p\pi x}{A}\,dx\right)\cdot\left(\int_0^B \sin^2\frac{q\pi y}{B}\,dy\right)$$

The integrals in the right-hand side can be calculated easily by substitution $t := \frac{\pi x}{A}$ yielding:

$$\int_0^A \sin^2\frac{p\pi x}{A}\,dx = \frac{A}{\pi}\cdot\int_0^\pi \sin^2 kt\,dt = \frac{A}{\pi}\cdot\int_0^\pi \frac{1-\cos 2kt}{2}\,dt = \frac{A}{2}$$

Similarly:

$$\int_0^B \sin^2\frac{q\pi x}{B}\,dy = \frac{B}{2}.$$

This implies that:

$$,\int_\Omega f(x,y)\cdot\sin\frac{p\pi x}{A}\sin\frac{q\pi y}{B}\,dxdy = c_{p,q}\cdot\frac{AB}{4},$$

from which the proposition follows.

Now *define* the function u in the form:

$$u(x,y) = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} a_{k,j}\cdot\sin\frac{k\pi x}{A}\sin\frac{j\pi y}{B},$$

with the Fourier coefficient:

$$a_{k,j} := -\frac{c_{k,j}}{\frac{k^2\pi^2}{A^2} + \frac{j^2\pi^2}{B}} \qquad (k, j = 1, 2, ...)$$

Then the function u vanishes along the boundary i.e. along the lines $y = 0$, $x = A$, $y = B$, $x = 0$ (check it!). Moreover, the function u satisfies the Poisson equation, since:

$$\Delta u(x, y) = \sum_{k,j=1}^{\infty} a_{k,j} \cdot \left(\left(\frac{\partial^2}{\partial x^2} \sin \frac{k\pi x}{A} \right) \sin \frac{j\pi y}{B} + \sin \frac{k\pi x}{A} \left(\frac{\partial^2}{\partial y^2} \sin \frac{j\pi y}{B} \right) \right) =$$

$$= \sum_{k,j=1}^{\infty} \frac{c_{k,j}}{\frac{k^2\pi^2}{A^2} + \frac{j^2\pi^2}{B^2}} \cdot \left(\frac{k^2\pi^2}{A^2} + \frac{j^2\pi^2}{B^2} \right) \cdot \sin \frac{k\pi x}{A} \sin \frac{j\pi y}{B} =$$

$$= \sum_{k,j=1}^{\infty} c_{k,j} \cdot \sin \frac{k\pi x}{A} \sin \frac{j\pi y}{B} = f(x, y)$$

Summarizing the above calculations, the algorithm is as follows:

- Calculate the Fourier coefficients:

$$\boxed{c_{k,j} := \frac{4}{AB} \int_{\Omega} f(x, y) \cdot \sin \frac{k\pi x}{A} \sin \frac{j\pi y}{B} \, dxdy \qquad (k, j = 1, 2, ...)}$$

- Calculate the Fourier coefficients:

$$\boxed{a_{k,j} := -\frac{c_{k,j}}{\frac{k^2\pi^2}{A^2} + \frac{j^2\pi^2}{B}} \qquad (k, j = 1, 2, ...)}$$

- Evaluate the Fourier series

$$\boxed{u(x, y) := \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{k,j} \cdot \sin \frac{k\pi x}{A} \sin \frac{j\pi y}{B}}$$

The method can be applied very rarely, since it requires that:

- the domain Ω is a rectangle
- the problem to be solved is a Poisson equation.

Generalizations are hard. However, if the method is applicable, then it is economic from computation point of view, provided the Fourier series are truncated by taking into account the first N terms and by performing both the expansions and the evaluations by the Fast Fourier Transform algorithm. In this case, the necessary number of arithmetic operations is $\mathcal{O}(N^2 \log N)$.

The generalization of the algorithm is not trivial even to the case of the Laplace equation. In the following, this technique will be outlined.

6.1.2 Fourier's method for 2D Laplace equation

Consider the Laplace equation

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

supplied with the nonhomogeneous Dirichlet boundary condition:

$$u|_\Gamma = u_0,$$

where the domain Ω is again a rectangle: $\Omega = (0, A) \times (0, B)$.

Without loss of generality, we can assume that *the boundary condition function u_0 vanishes at the corner points of the rectangle.* Otherwise, seek the solution u in the form: $u(x, y) = w(x, y) + w_0(x, y)$, where $w_0(x, y) = a + bx + cy + dxy$ and the numbers a, b, c, d are the solutions of the simple system of equations:

$$\begin{aligned}
a &= u_0(0,0)\\
a + b \cdot A &= u_0(A,0)\\
a + b \cdot A + c \cdot B &= u_0(A,B)\\
a + c \cdot B &= u_0(0,B)
\end{aligned}$$

Then, by definition, the values of u_0 and w_0 coincide at the corner points. The function w_0 is obviously harmonic, thus, the function w satisfies the Dirichlet problem:

$$\Delta w = 0 \quad \text{in } \Omega, \qquad w|_\Gamma = u_0 - w_0|_\Gamma,$$

where the boundary condition function vanishes at the corner points.

From now on, assume that the boundary condition function u_0 itself has the above property i.e. it vanishes at the corner points of the domain Ω. Denote by Γ_1, Γ_2, Γ_3 and Γ_4 the sides of the rectangle Ω:

$$\Gamma_1 := \{(x, 0) : \ 0 \le x \le A\}$$

$$\Gamma_2 := \{(A, y) : \ 0 \le y \le B\}$$
$$\Gamma_3 := \{(x, B) : \ 0 \le x \le A\}$$
$$\Gamma_4 := \{(0, B) : \ 0 \le y \le B\}$$

Expand the function u_0 in sinusoidal Fourier series along the four sides of the rectangle:

- Along Γ_1:
$$u_0(x, 0) = \sum_{k=1}^{\infty} a_k^{(1)} \cdot \sin \frac{k\pi x}{A}$$
- Along Γ_2:
$$u_0(A, y) = \sum_{k=1}^{\infty} a_k^{(2)} \cdot \sin \frac{j\pi y}{B}$$
- Along Γ_3:
$$u_0(x, B) = \sum_{k=1}^{\infty} a_k^{(3)} \cdot \sin \frac{k\pi x}{A}$$
- Along Γ_4:
$$u_0(0, y) = \sum_{k=1}^{\infty} a_k^{(4)} \cdot \sin \frac{k\pi y}{B}$$

Since the function u_0 is assumed to be known, the Fourier coefficients $a^{(1)}$, $a^{(2)}$, $a^{(3)}$, $a^{(4)}$ can be computed without difficulty.

Now define the following four bivariate functions:

$$u^{(1)}(x, y) := \sum_{k=1}^{\infty} a_k^{(1)} \cdot \frac{1}{\sinh \frac{k\pi B}{A}} \cdot \sin \frac{k\pi x}{A} \cdot \sinh \frac{k\pi(B - y)}{A}$$

$$u^{(2)}(x, y) := \sum_{k=1}^{\infty} a_k^{(2)} \cdot \frac{1}{\sinh \frac{k\pi A}{B}} \cdot \sinh \frac{k\pi x}{B} \cdot \sin \frac{k\pi y}{B}$$

$$u^{(3)}(x, y) := \sum_{k=1}^{\infty} a_k^{(3)} \cdot \frac{1}{\sinh \frac{k\pi B}{A}} \cdot \sin \frac{k\pi x}{A} \cdot \sinh \frac{k\pi y}{A}$$

$$u^{(4)}(x, y) := \sum_{k=1}^{\infty} a_k^{(4)} \cdot \frac{1}{\sinh \frac{k\pi A}{B}} \cdot \sinh \frac{k\pi(A - x)}{B} \cdot \sin \frac{k\pi y}{B}$$

Elementary calculations show that $\Delta u^{(j)} = 0$ for $j = 1, 2, 3, 4$. Indeed, for any k:

$$\Delta \left(\sin \frac{k\pi x}{A} \sinh \frac{k\pi y}{A} \right) =$$

$$= -\frac{k^2\pi^2}{A^2} \cdot \sin \frac{k\pi x}{A} \sinh \frac{k\pi y}{A} + \frac{k^2\pi^2}{A^2} \cdot \sin \frac{k\pi x}{A} \sinh \frac{k\pi y}{A} \equiv 0$$

(the other three functions appearing in the above Fourier series can be treated similarly). Note that, if the function u_0 is smooth enough, the Laplace operator can be applied to the Fourier series term-by-term.

It can be checked by direct computations that for each $j = 1, 2, 3, 4$:

$$u^{(j)}|_{\Gamma_j} = u_0|_{\Gamma_j},$$

i.e. along the side Γ_j, $u^{(j)}$ is identical to the boundary function u_0, *and vanishes on the remaining three sides*:

$$u^{(j)}|_{\Gamma_k} \equiv 0$$

for $k \neq j$.

Thus, the function defined by

$$u := u^{(1)} + u^{(2)} + u^{(3)} + u^{(4)}$$

is harmonic in Ω, and coincides with the boundary condition on Γ.

In short, the algorithm is as follows:

- Calculate the Fourier coefficients of u_0 on each side of the rectangle Ω. Thus, we obtain the numbers $a_k^{(1)}$, $a_k^{(2)}$, $a_k^{(3)}$, $a_k^{(4)}$ $(k = 1, 2, ...)$.
- Define the functions $u^{(1)}$, $u^{(2)}$, $u^{(3)}$, $u^{(4)}$ by the above sinusoidal Fourier series.
- The solution of the Dirichlet problem is expressed as $u := u^{(1)} + u^{(2)} + u^{(3)} + u^{(4)}$.

The applicability of the method is restricted to the 2D Laplace equations (or, at most, to a bit more general elliptic equations) defined on a rectangle. In practice, the Fourier series expansions and evaluations are recommended to be performed by the Fast Fourier Transform algorithm. In this case, the necessary number of arithmetic operations is $\mathcal{O}(N \log N)$ for computing the Fourier coefficients $a_k^{(j)}$ $(k = 1, 2, ..., N,\ j = 1, 2, 3, 4)$. However, the

computational cost of the evaluation of the solution is also $\mathcal{O}(N \log N)$ *in each point* (x, y). If the evaluation points are located on a grid with stepsize $h := \frac{const.}{N}$, the total number of necessary operations is $\mathcal{O}(N^3 \log N)$. However, if the solution has to be evaluated in the vicinity of the boundary only (which is often the case), then the computational complexity is reduced, typically to $\mathcal{O}(N^2 \log N)$.

Remark:

- In the seemingly more general case, when an inhomogeneous Poisson equation is to be solved:

$$\Delta u = f \quad \text{in } \Omega, \qquad u|_\Gamma = u_0,$$

  the solution can be expressed as a sum of the solutions of the following two problems:

$$\Delta v = f \quad \text{in } \Omega, \qquad v|_\Gamma = 0,$$

$$\Delta w = 0 \quad \text{in } \Omega, \qquad w|_\Gamma = u_0.$$

  It is obvious that the function $u := v + w$ satisfies both the Poisson equation and also the boundary condition.

6.2 The Finite Difference Method

The most traditional computational tool for solving partial differential equations is the finite difference method. The basic idea of the method is to approximate the derivatives appearing in the differential equation by difference schemes. Therefore the first task is to define a *computational grid* i.e. a finite set of points. All the values of functions are evaluated and computed at the gridpoints only. To compute the (approximate) solution between the gridpoints is a completely different task (interpolation problem in general), which does not belong to the finite difference method.

6.2.1 Finite difference method for 1D elliptic problems

Let Ω be an 1D domain i.e. an interval: $\Omega := (0, A)$. Consider the 1D Poisson equation:

$$u'' = f \qquad \text{in } \Omega$$

supplied with Dirichlet-type boundary condition:

$$u(0) = a, \qquad u(A) = b$$

where a, b are predefined numbers, f is a given function which is defined on the interval Ω.

Define a computational grid on the closed interval $[0, A]$ by

$$x_k := k \cdot h \qquad (k = 0, 1, ..., N)$$

where h denotes the *stepsize* of the grid: $h := \frac{A}{N}$. The given number N indicates how many parts the original interval is subdivided into. The numbers x_0, x_1, ... , x_N are called *gridpoints*. The above defined grid is *equidistant* i.e. the distances between the consecutive gridpoints is constant. A part of the grid is shown in Figure 15.

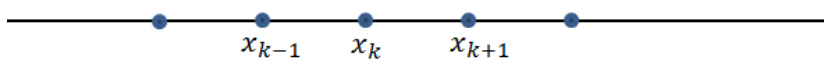

Figure 15: A part of a 1D equidistant grid.

A natural tool to create difference schemes is the well-known Taylor series expansion. Recall that if u is a sufficiently smooth function (more precisely, four times continuously differentiable), then it can be expanded in terms of finite Taylor series around x_k as:

$$u(x_{k+1}) = u(x_k + h) = u(x_k) + \frac{u'(x_k)}{1!} \cdot h + \frac{u''(x_k)}{2!} \cdot h^2 + \frac{u'''(x_k)}{3!} \cdot h^3 + \mathcal{O}(h^4)$$

where in the right-hand side, only the order of magnitude of the remainder term is indicated ($\mathcal{O}(h^4)$) Similarly:

$$u(x_{k-1}) = u(x_k - h) = u(x_k) - \frac{u'(x_k)}{1!} \cdot h + \frac{u''(x_k)}{2!} \cdot h^2 - \frac{u'''(x_k)}{3!} \cdot h^3 + \mathcal{O}(h^4)$$

Straightforward calculations show that

$$u'(x_k) = \frac{u(x_{k+1}) - u(x_k)}{h} + \mathcal{O}(h) \qquad \text{(forward scheme)}$$

$$u'(x_k) = \frac{u(x_k) - u(x_{k-1})}{h} + \mathcal{O}(h) \qquad \text{(backward scheme)}$$

$$u'(x_k) = \frac{u(x_{k+1}) - u(x_{k-1})}{2h} + \mathcal{O}(h^2) \qquad \text{(central scheme)}$$

$$u''(x_k) = \frac{u(x_{k+1}) - 2u(x_k) + u(x_{k-1})}{h^2} + \mathcal{O}(h^2) \qquad \text{(central scheme)}$$

The terms $\mathcal{O}(h)$ and $\mathcal{O}(h^2)$ characterize the errors of the schemes. We say that a scheme is of order p, if the error term of the scheme (the difference between the exact derivative and the difference scheme) is $\mathcal{O}(h^p)$. We see that the forward and the backward schemes are of first order, while the last two central schemes are of second order.

In practice, the stepsize is predefined. A smaller stepsize results in more precise schemes especially when the scheme is of high order. On the other hand, a smaller stepsize increases the number of gridpoints, thus, the amount of computational work is also increased. It should be pointed out, that the exactness of the applied schemes does *not* mean the exactness of the approximate solution of the differential equation; the error analysis is a much more complicated task and requires special mathematical tools.

Returning to the model Poisson problem:

$$u'' = f \qquad u(0) = a, \quad u(A) = b,$$

denote by u_0, u_1, ..., u_N the values of the *approximate* solution. At the gridpoint x_k the second order derivative $u''(x_k)$ is approximated by the above three-point central scheme. This results in the *discrete problem*:

$$\frac{u_{k-1} - 2u_k + u_{k+1}}{h^2} = f_k := f(x_k) \qquad (k = 1, 2, ..., N-1)$$

The values u_0 and u_N are known form the boundary condition: $u_0 = a$, $u_N = b$. We have obtained a linear system of algebraic equations with $(N-1)$ unknowns.

To analyse the exactness, introduce the following error terms. Denote by u the exact solution and:

$$g_k := \frac{u(x_{k-1}) - 2u(x_k) + u(x_{k+1})}{h^2} - f_k \qquad (k = 1, 2, ..., N-1).$$

The numbers g_k are called *local error terms*. Introduce also the *global error terms* as the differences of the exact and approximate solution at the gridpoints:

$$e_k := u(x_k) - u_k \qquad (k = 0, 2, ..., N).$$

Of course, $e_0 = e_N = 0$.

The local error can be estimated easily, based on the Taylor expansion of u. However, the error which has a direct importance in practice is the global error. In the following, an error estimation will be deduced.

Observe that for the local error terms, the following equalities are valid:

$$\frac{u(x_{k-1}) - 2u(x_k) + u(x_{k+1})}{h^2} = f_k + g_k \qquad (k = 1, 2, ..., N-1).$$

Consider the discrete system of equations again:

$$\frac{u_{k-1} - 2u_k + u_{k+1}}{h^2} = f_k \qquad (k = 1, 2, ..., N-1)$$

Subtracting the two equations, we obtain that the global error terms satisfy a similar system of equations, and the local error terms appear on the right-hand side:

$$\frac{e_{k-1} - 2e_k + e_{k+1}}{h^2} = g_k \qquad (k = 1, 2, ..., N-1)$$

and $e_0 = e_N = 0$.

Denote by $A_h \in \mathbf{M}_{(N-1)\times(N-1)}$ the matrix of the system. Obviously, for every $w \in \mathbf{R}^{N-1}$, the equality

$$(A_h w)_k = \frac{w_{k-1} - 2w_k + w_{k+1}}{h^2} \qquad (k = 1, 2, ..., N-1)$$

is valid (where, by definition, $w_0 := w_N := 0$.

Thus, the connection between the local and global error terms can be written in the following compact form:

$$A_h e = g$$

where $e := (e_1, ..., e_{N-1})$, and $g := (g_1, ..., g_{N-1})$. From this system, an error estimation immediately follows:

$$||e|| \leq ||A_h^{-1}|| \cdot ||g||,$$

where $||.||$ denotes the Euclidean norm in $\mathbf{R}^{N-1}$.

This estimation is not suitable for practical purposes yet, since the Euclidean norm directly depend on h (through N). A much better choice for measuring the errors is the use of the *root mean square*:

$$||e||_{RMS} := \sqrt{\frac{e_1^2 + e_2^2 + ... + e_{N-1}^2}{N-1}}$$

Note that the root mean square of a vector which consists of ones is always equal to 1, independently of the dimension of the vector. The above estimation immediately implies that:

$$||e||_{RMS} \leq ||A_h^{-1}|| \cdot ||g||_{RMS}$$

We will show that the norm of the inverse matrix A_h^{-1} is uniformly bounded i.e. it is independent of h. More precisely:

$$||A_h^{-1}|| \leq C$$

for some constant C, which is independent of h. This is often called the *stability* of the scheme, and results in the estimation

$$||e||_{RMS} \leq C \cdot ||g||_{RMS},$$

which means that the global error can be estimated by the local error from above. The local error terms can be estimated easily (using Taylor series expansions), thus, this will result in the desired global error estimation.

The stability result is based on the following theorem:

Theorem: The matrix A_h is self-adjoint and negative definite. The eigenvalues of A_h are λ_1, λ_2, ..., λ_{N-1}, and the corresponding eigenvectors are $s^{(1)}$, $s^{(2)}$, ..., $s^{(N-1)}$, where:

$$\lambda_k = -\frac{4}{h^2} \cdot \sin^2 \frac{k\pi}{2N}, \qquad s_j^{(k)} = \sin \frac{kj\pi}{N} \qquad (k = 1, 2, ..., N-1)$$

Proof: Based on the definition of A_h, it is obvious that A_h is self-adjoint. The only thing that should be proved is that $A_h s^{(k)} = \lambda_k \cdot s^{(k)}$. By definition:

$$\left(A_h s^{(k)}\right)_j = \frac{1}{h^2} \cdot \left(\sin \frac{k(j-1)\pi}{N} - 2\sin \frac{kj\pi}{N} + \sin \frac{k(j+1)\pi}{N}\right) =$$

$$= \frac{1}{h^2} \cdot \left(\sin \frac{kj\pi}{N} \cos \frac{k\pi}{N} - \cos \frac{kj\pi}{N} \sin \frac{k\pi}{N} - 2\sin \frac{kj\pi}{N} + \right.$$

$$\left. \sin \frac{kj\pi}{N} \cos \frac{k\pi}{N} + \cos \frac{kj\pi}{N} \sin \frac{k\pi}{N}\right) =$$

$$= -\frac{2}{h^2} \cdot \sin \frac{kj\pi}{N} \cdot \left(1 - \cos \frac{k\pi}{N}\right)$$

Utilizing the elementary trigonometric identity $\sin^2 \alpha = \dfrac{1 - \cos 2\alpha}{2}$, we obtain:

$$\left(A_h s^{(k)}\right)_j = -\frac{4}{h^2} \cdot \sin \frac{kj\pi}{N} \cdot \sin^2 \frac{k\pi}{2N},$$

which completes the proof.

Due to the self-adjoint property of the matrix A_h, the norm of A_h as well as the norm of A^{-1} can be expressed with the eigenvalues of A_h:

Corollary:

$$||A_h|| = |\lambda|_{\max} = |\lambda_{N-1}| = \frac{4}{h^2} \cdot \sin^2 \frac{(N-1)\pi}{2N} \leq \frac{4}{h^2}$$

and

$$||A_h^{-1}|| = \frac{1}{|\lambda|_{\min}} = \frac{1}{|\lambda_1|} = \frac{h^2}{4\sin^2 \frac{\pi}{2N}} \sim \frac{h^2}{4\frac{\pi^2}{4N^2}} = \frac{A^2}{\pi^2}$$

independently of h (where $\sim$ means the asymptotic equality when $N \to \infty$ i.e. $h \to 0$).

Thus, the stability of the scheme is proved, and the estimation

$$||e||_{RMS} \leq C \cdot ||g||_{RMS},$$

is valid.

The estimation of the right-hand side is much simpler. Expanding the exact solution u in finite Taylor series around x_k, we have:

$$u(x_{k+1}) = u(x_k + h) = u(x_k) + \frac{u'(x_k)}{1!} \cdot h + \frac{u''(x_k)}{2!} \cdot h^2 + \frac{u'''(x_k)}{3!} \cdot h^3 + \mathcal{O}(h^4)$$

and

$$u(x_{k-1}) = u(x_k - h) = u(x_k) - \frac{u'(x_k)}{1!} \cdot h + \frac{u''(x_k)}{2!} \cdot h^2 - \frac{u'''(x_k)}{3!} \cdot h^3 + \mathcal{O}(h^4)$$

By adding these equalities, we obtain:

$$u(x_{k-1}) + u(x_{k+1}) = 2u(x_k) + u''(x_k) \cdot h^2 + \mathcal{O}(h^4),$$

whence

$$g_k = \frac{u(x_{k-1}) - 2u(x_k) + u(x_{k+1})}{h^2} - f_k =$$

$$= \frac{u(x_{k-1}) - 2u(x_k) + u(x_{k+1})}{h^2} - u''(x_k) = \mathcal{O}(h^2)$$

i.e.

$$|g_k| \leq C_0 \cdot h^2$$

$(k = 1, 2, ..., N-1)$ for some constant C_0 which is independent of h and k. Therefore:

$$||g||_{RMS} \leq C_0 \cdot h^2$$

which implies that the same type of estimation is valid for the global error:

$$||e||_{RMS} \leq C \cdot C_0 \cdot h^2$$

Summarizing the above results we have the following theorem:

Theorem: The accuracy of the above defined finite difference method based on the central scheme is of second order, i.e. the global error (with respect to the root mean square) is $\mathcal{O}(h^2)$ (provided that the exact solution is smooth enough).

Remarks:

1. The above 3-point central scheme can be generalized to the more general problem

$$\frac{d}{dx}\left(\sigma \cdot \frac{du}{dx}\right) = f,$$

which is the 1D equivalent of the differential equation $\operatorname{div}(\sigma \cdot \operatorname{grad} u) = f$. The most usual scheme is:

$$\frac{1}{h^2} \cdot \left(\frac{\sigma_{k-1} + \sigma_k}{2} u_{k-1} - \frac{\sigma_{k-1} + 2\sigma_k + \sigma_{k+1}}{2} u_k + \frac{\sigma_k + \sigma_{k+1}}{2} u_{k+1}\right)$$

2. If, for instance, at the point x_0, a Neumann type boundary condition $u'(x_0) = a$ is given, then the first derivative of u should be approximated at x_0. The naive solution is to use the one-sided scheme:

$$u'(x_0) \approx \frac{u_1 - u_0}{h} = a$$

However, this scheme is of first order only, which decreases the accuracy on the whole interval $(0, A)$. A possible remedy is the use of a higher order scheme for the derivative $u'(x_0)$. A more elegant way is as follows. Let us expand the function u' in a finite Taylor series around $x_0 + \frac{h}{2}$:

$$u'(x_0) = u'\left(x_0 + \frac{h}{2}\cdot\right) - \frac{1}{2}u''\left(x_0 + \frac{h}{2}\right) + \mathcal{O}(h^2) =$$

$$= \frac{u_1 - u_0}{h} - \frac{1}{2} \cdot f\left(x_0 + \frac{h}{2}\right) + \mathcal{O}(h^2)$$

And since $f\left(x_0 + \frac{h}{2}\right) = f(x_0) + \mathcal{O}(h)$, therefore we have:

$$u'(x_0) = \frac{u_1 - u_0}{h} - \frac{h}{2} \cdot f(x_0) + \mathcal{O}(h^2)$$

Thus, the Neumann boundary condition at x_0 can be approximated by the *second-order* scheme:

$$\frac{u_1 - u_0}{h} = a + \frac{h}{2} \cdot f(x_0)$$

6.2.2 Finite difference method for 2D elliptic problems

Let Ω be an 2D rectangular domain $\Omega := (0, A) \times (0, B)$. Consider the 2D Poisson equation:

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

supplied with Dirichlet-type boundary condition:

$$u|_\Gamma = u_0$$

where Γ denotes the boundary of Ω, f is a given function which is defined on the domain Ω.

Define a computational grid on the closure of Ω by

$$(x_{k,j}, y_{k,j}) := (k \cdot h_x, j \cdot h_y) \qquad (k = 0, 1, ..., N; \quad j = 0, 1, ..., M)$$

where h_x, h_y denote the *stepsizes* of the grid: $h_x := \frac{A}{N}$, $h_y := \frac{B}{M}$. The given numbers N, M indicate how many parts the sides of the original rectangle are subdivided into. The points $(x_{k,j}, y_{k,j})$, are called *gridpoints*. The above defined grid is equidistant. The stepsizes h_x, h_y need not be equal. A part of the grid is shown in Figure 16.

The 2D scheme comes from the 1D central scheme, by approximating the derivatives $\frac{\partial^2 u}{\partial x^2}$ and $\frac{\partial^2 u}{\partial y^2}$ separately. This results in the following discrete system:

$$\frac{u_{k-1,j} - 2u_{k,j} + u_{k+1,j}}{h_x^2} + \frac{u_{k,j-1} - 2u_{k,j} + u_{k,j+1}}{h_y^2} = f_{k,j} := f(x_k, y_j)$$

where $u_{k,j}$ denotes the approximate value of the solution at the gridpoint (x_k, y_j), and $k = 0, 1, ..., N$, $j = 0, 1, ..., M$.

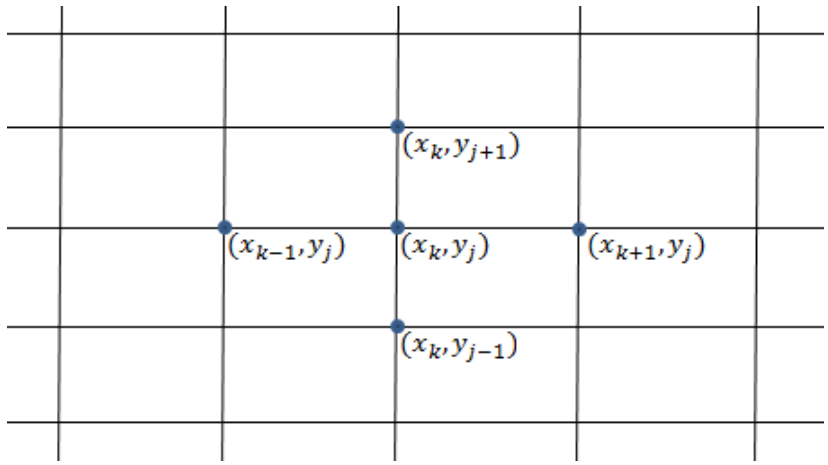

Figure 16: A part of a 2D equidistant grid.

Remark:

- If the stepsizes in the different directions coincide: $h_x = h_y =: h$, the above 5-point scheme becomes even simpler:

$$\frac{u_N + u_W + u_S + e_E - 4u_C}{h^2} = f_C$$

  where the index C refers to a central gridpoint, and the neighbours of C taken in the main cardinal directions are denoted by N, W, S and E.

Let u be the exact solution of the problem. Similarly to the 1D case, introduce the local error terms:

$$g_{k,j} := \frac{u(x_{k-1}, y_j) - 2u(x_k, y_j) + u(x_{k+1}, y_j)}{h_x^2} +$$

$$+ \frac{u(x_k, y_{j-1}) - 2u(x_k, y_j) + u(x_k, y_{j+1})}{h_y^2} - f_{k,j},$$

and also the global error terms:

$$e_{k,j} := u(x_k, y_j) - u_{k,j}$$

Again, by definition, the global error terms satisfy the following system of equations:

$$\frac{e_{k-1,j} - 2e_{k,j} + e_{k+1,j}}{h_x^2} + \frac{e_{k,j-1} - 2e_{k,j} + e_{k,j+1}}{h_y^2} = g_{k,j},$$

and $e_{k,j} = 0$ at the boundary gridpoints $x_{k,j}$ (i.e. when $k = 0$ or $k = N$ or $j = 0$ or $j = M$).

Denote by A_{h_x,h_y} the discrete Laplace operator, i.e. for every grid function w which vanishes at the boundary gridpoints:

$$(A_{h_x,h_y} w)_{k,j} = \frac{w_{k-1,j} - 2w_{k,j} + w_{k+1,j}}{h_x^2} + \frac{w_{k,j-1} - 2w_{k,j} + w_{k,j+1}}{h_y^2}$$

Theorem: The mapping A_{h_x,h_y} is self-adjoint and negative definite. The eigenvalues of A_{h_x,h_y} are $\lambda_{k,j}$, and the corresponding eigenvectors are $s^{(k,j)}$, where:

$$\lambda_{k,j} = -\frac{4}{h_x^2} \cdot \sin^2 \frac{k\pi}{2N} - \frac{4}{h_y^2} \cdot \sin^2 \frac{j\pi}{2M}, \qquad s_{p,q}^{(k,j)} = \sin \frac{kp\pi}{N} \cdot \sin \frac{jq\pi}{M}$$

where $k, p = 1, ..., N-1$ and $j, q = 1, ..., M-1$.

Proof: The mapping A_{h_x,h_y} is self-adjoint (check it!). The only thing that should be proved is that $A_{h_x,h_y} s^{(k,j)} = \lambda_{k,j} \cdot s^{(k,j)}$. By definition:

$$\left(A_{h_x,h_y} s^{(k,j)}\right)_{p,q} =$$

$$= \frac{1}{h_x^2} \cdot \left(\sin \frac{k(p-1)\pi}{N} - 2\sin \frac{kp\pi}{N} + \sin \frac{k(p+1)\pi}{N}\right) \cdot \sin \frac{jq\pi}{M} +$$

$$+ \frac{1}{h_x^2} \cdot \left(\sin \frac{j(q-1)\pi}{M} - 2\sin \frac{jq\pi}{M} + \sin \frac{j(q+1)\pi}{M}\right) \cdot \sin \frac{kp\pi}{N}$$

Using the same trigonometric calculations as in the 1D case, we have:

$$\left(A_{h_x,h_y} s^{(k,j)}\right)_{p,q} =$$

$$= -\sin \frac{kp\pi}{N} \cdot \frac{4}{h_x^2} \cdot \sin^2 \frac{k\pi}{2N} \cdot \sin \frac{jq\pi}{M} - \sin \frac{jq\pi}{M} \cdot \frac{4}{h_y^2} \cdot \sin^2 \frac{j\pi}{2M} \cdot \sin \frac{kp\pi}{N} =$$

$$= -\sin \frac{kp\pi}{N} \sin \frac{jq\pi}{M} \cdot \left(\frac{4}{h_x^2} \sin^2 \frac{k\pi}{2N} + \frac{4}{h_y^2} \sin^2 \frac{j\pi}{2M}\right)$$

which completes the proof.

Corollary:

$$||A^{-1}_{h_x,h_y}|| = \frac{1}{|\lambda|_{\min}} = \frac{1}{|\lambda_{1,1}|} = \frac{1}{\frac{4}{h_x^2}\sin^2\frac{\pi}{2N} + \frac{4}{h_y^2}\sin^2\frac{\pi}{2M}} \sim$$

$$\sim \frac{1}{\frac{4}{h_x^2}\cdot\frac{\pi^2}{4N^2} + \frac{4}{h_y^2}\cdot\frac{\pi^2}{4M^2}} = \frac{1}{\pi^2}\cdot\frac{A^2B^2}{A^2+B^2}$$

independently of h_x and h_y ($\sim$ means again the asymptotic equality when $N, M \to \infty$ i.e. $h_x, h_y \to 0$).

This is the stability result for the 2D scheme. Similarly to the 1D case, this implies that the estimation

$$||e||_{RMS} \leq C \cdot ||g||_{RMS},$$

is valid also in the 2D case. The local error terms can be estimated easily, based on Taylor series expansion around x_k and y_j, respectively. Utilizing the 1D results, we have:

$$u(x_{k-1}, y_j) - 2u(x_k, y_j) + u(x_{k+1}, y_j) = \frac{\partial^2 u}{\partial x^2}(x_k, y_j)\cdot h_x^2 + \mathcal{O}(h_x^4),$$

and also

$$u(x_k, y_{j-1}) - 2u(x_k, y_j) + u(x_k, y_{j+1}) = \frac{\partial^2 u}{\partial y^2}(x_k, y_j)\cdot h_y^2 + \mathcal{O}(h_y^4),$$

Hence

$$g_{k,j} = \frac{u(x_{k-1}, y_j) - 2u(x_k, y_j) + u(x_{k+1}, y_j)}{h_x^2} +$$

$$+\frac{u(x_k, y_{j-1}) - 2u(x_k, y_j) + u(x_k, y_{j+1})}{h_y^2} - f_{k,j} =$$

$$= \frac{u(x_{k-1}, y_j) - 2u(x_k, y_j) + u(x_{k+1}, y_j)}{h_x^2} +$$

$$+\frac{u(x_k, y_{j-1}) - 2u(x_k, y_j) + u(x_k, y_{j+1})}{h_y^2} - \Delta u(x_k, y_j) =$$

$$= \mathcal{O}(h_x^2 + h_y^2)$$

Consequently:

$$|g_{k,j}| \leq C_0 \cdot (h_x^2 + h_y^2)$$

for some constant C_0 which is independent of h_x, h_y and also of k, j. Therefore:

$$||g||_{RMS} \leq C_0 \cdot (h_x^2 + h_y^2)$$

which implies that the same type of estimation is valid for the global error:

$$||e||_{RMS} \leq C \cdot C_0 \cdot (h_x^2 + h_y^2)$$

Summarizing the above results we have the following theorem:

Theorem: The accuracy of the above defined finite difference method based on the central scheme is of second order, i.e. the global error (with respect to the root mean square) is $\mathcal{O}(h_x^2 + h_y^2)$ (provided that the exact solution is smooth enough).

Remark:

- Assume that a Neumann type boundary condition

  $$\frac{\partial u}{\partial n} = a$$

  is given along a part of the boundary, for instance, along an eastern part (see Figure 17). Here $\frac{\partial u}{\partial n} = \frac{\partial u}{\partial x}$, so that the first derivative of u

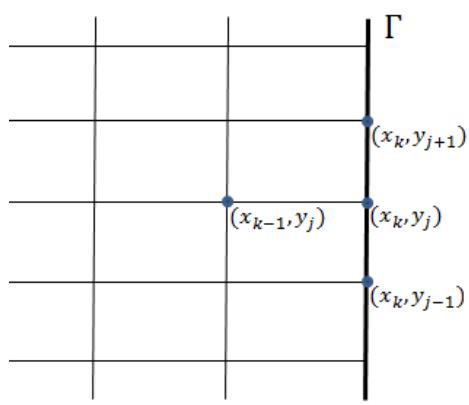

Figure 17: A part of a 2D equidistant grid, Neumann boundary condition.

should be approximated at a point (x_k, y_j). The naive solution is to use the one-sided scheme:

$$\frac{\partial u}{\partial n}(x_k, y_j) \approx \frac{u_{k,j} - u_{k-1,j}}{h_x}$$

However, this scheme is of first order only, which decreases the accuracy on the whole domain Ω. A possible remedy is the use of a higher order scheme for the derivative $\frac{\partial u}{\partial n}(x_k, y_j)$. A more elegant way is as follows. Let us expand the function u in a finite Taylor series around (x_k, y_j):

$$u(x_{k-1}, y_j) = u(x_k, y_j) - \frac{\partial u}{\partial x}(x_k, y_j) \cdot h_x + \frac{1}{2} \cdot \frac{\partial^2 u}{\partial x^2}(x_k, y_j) \cdot h_x^2 + \mathcal{O}(h_x^3)$$

However:

$$\frac{\partial^2 u}{\partial x^2}(x_k, y_j) = -\frac{\partial^2 u}{\partial y^2}(x_k, y_j) + f(x_k, y_j),$$

and the second order derivative of u with respect to y can be approximated by the 3-point central difference scheme at an accuracy of $\mathcal{O}(h_y^2)$. Thus, we have obtained that:

$$u(x_{k-1}, y_j) = u(x_k, y_j) - \frac{\partial u}{\partial x}(x_k, y_j) \cdot h_x +$$

$$+\frac{1}{2} \cdot \left(-\frac{u(x_k, y_{j-1}) - 2u(x_k, y_j) + u(x_k, y_{j+1})}{h_y^2} + f(x_k, y_j) + \mathcal{O}(h_y^2) \right) \cdot h_x^2 +$$

$$+\mathcal{O}(h_x^3),$$

which implies:

$$\frac{\partial u}{\partial x}(x_k, y_j) = \frac{u(x_k, y_j) - u(x_{k-1}, y_j)}{h_x} -$$

$$-\frac{1}{2} \cdot \frac{u(x_k, y_{j+1}) - 2u(x_k, y_j) + u(x_k, y_{j-1}))}{h_y^2} \cdot h_x + \frac{1}{2} \cdot f(x_k, y_j) \cdot h_x +$$

$$+\mathcal{O}(h_x^2 + h_y^2)$$

This results in the following *second order* scheme at Neumann boundary:

$$\frac{u_{k,j} - u_{k-1,j}}{h_x} - \frac{1}{2} \cdot \frac{u_{k,j+1} - 2u_{k,j} + u_{k,j-1}}{h_y^2} \cdot h_x + \frac{1}{2} \cdot f_{k,j} \cdot h_x = a_{k,j}$$

Finally, it should be pointed out that the finite difference method still works on non-rectangular domains as well (but the error analysis is more difficult). One has to register the inner and the boundary gridpoints; at each inner gridpoint, a discrete scheme is defined, while at the boundary gridpoints, the boundary condition is to be approximated. This results in a large linear system of equations but the matrix of the system is extremely sparse, which makes it possible to apply efficient solution techniques e.g. Krylov subspace methods. Note, however, that the approximation of the boundary is rough, which may decrease the accuracy. A possible solution technique is to define gridpoints which are located on the exact boundary and to define non-equidistant schemes in the vicinity of the boundary. It should be emphasize that the structure of the finite difference grid cannot follow the characteristic properties of the solution. In a lot of cases, the solution is much 'smoother' in the middle of the domain than in the vicinity of the boundary, so that much coarser grid would be sufficient. The above disadvantages and difficulties reflect to the limitations of the classical finite difference method.

6.3 The Method of Fundamental Solutions

The Method of Fundamental Solution (MFS) is a relatively new technique for solving *homogeneous* partial differential equations. We introduce the method through the example of the 2D Laplace equation. It should be pointed out in advance that the method is a *meshless method*, i.e. in contrast to the finite difference method or the finite element method, the MFS requires neither grid nor element structure in the domain or on the boundary. Only some finite sets of points are needed without eny grid of element structure.

Consider the model problem

$$\Delta u = 0$$

in a bounded 2D domain Ω supplied with the mixed boundary condition:

$$u|_{\Gamma_D} = u_0, \qquad \frac{\partial u}{\partial n}|_{\Gamma_N} = v_0$$

where Γ_D and Γ_N form a disjoint decomposition of the boundary Γ.

The Method of Fundamental Solutions gives an approximate solution of the problem in the form:

$$u(x) := \sum_{j=1}^{N} \alpha_j \cdot \Phi(x - \tilde{x}_j),$$

where α_1,..., α_N are a priori unknown coefficients and Φ denotes the *fundamental solution* of the 2D Laplace operator:

$$\Phi(x) := \log ||x||$$

The points $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_N$ are predefined external points (*source points*), where N is a given integer number.

The method is based on the fact that Φ is a harmonic function (except for the origin):

$$\Delta\Phi(x) = 0 \qquad (x \neq \mathbf{0})$$

Indeed, a simple vector calculus shows that if $x \neq \mathbf{0}$, then:

$$\Delta\Phi(x) = \operatorname{div}\operatorname{grad}\left(\frac{1}{2}\log ||x||^2\right) = \operatorname{div}\left(\frac{1}{||x||^2}\cdot x\right) =$$

$$= \left\langle \operatorname{grad}\frac{1}{||x||^2}, x\right\rangle + \frac{1}{||x||^2}\cdot(\operatorname{div} x) =$$

$$= \left\langle -\frac{2x}{||x||^4}, x\right\rangle + \frac{2}{||x||^2} = 0.$$

The origin is clearly a singular point of Φ.

Consequently, the function

$$u(x) := \sum_{j=1}^{N} \alpha_j \cdot \Phi(x - \tilde{x}_j),$$

satisfies the Laplace equation exactly in all points of Ω. The coefficients α_1, ..., α_N can be calculated by enforcing the boundary conditions in some predefined *boundary collocation points* $x_1, x_2, ..., x_N \in \Gamma$ (cf Figure 18).

This results in the following system of equations:

$$\sum_{j=1}^{N} \alpha_j \Phi(x_k - \tilde{x}_j) = u_0(x_k) \qquad \text{if } x_k \in \Gamma_D$$

$$\sum_{j=1}^{N} \alpha_j \frac{\partial\Phi}{\partial n_k}(x_k - \tilde{x}_j) = v_0(x_k) \qquad \text{if } x_k \in \Gamma_N$$

where n_k denotes the outward normal unit vector at the boundary collocation point x_k. The normal derivative of Φ can be calculated easily, yielding:

$$\frac{\partial\Phi}{\partial n_k}(x) = \frac{\langle x, n_k\rangle}{||x||^2}$$

The basic algorithm of the method is extremely simple:

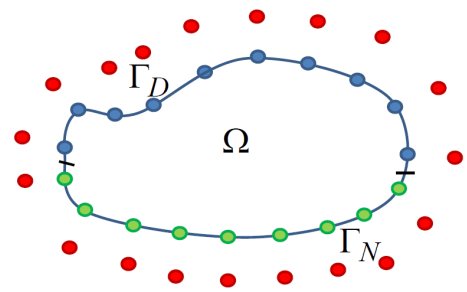

Figure 18: Source and boundary collocation points.

- Define a set of external source points $\tilde{x}_1$, $\tilde{x}_2$, ... , $\tilde{x}_N$.
- Define a set of boundary collocation points x_1, x_2, ... , x_N.
- Generate and solve the system of equations

$$\sum_{j=1}^{N} \alpha_j \Phi(x_k - \tilde{x}_j) = u_0(x_k) \qquad \text{if } x_k \in \Gamma_D$$

$$\sum_{j=1}^{N} \alpha_j \frac{\partial \Phi}{\partial n_k}(x_k - \tilde{x}_j) = v_0(x_k) \qquad \text{if } x_k \in \Gamma_N$$

- The approximate solution is:

$$u(x) := \sum_{j=1}^{N} \alpha_j \cdot \Phi(x - \tilde{x}_j)$$

The algorithm can be programmed in a very simple way. However, the matrix of the resulting system is generally fully populated and non-symmetric. In many cases, the matrix of the system is extremely ill-conditioned. The greater the distance of the source points from the boundary, the higher the condition number of the matrix. It should be also pointed out that the proper location of the source points is far from being evident, and no 'optimal' arrangement of source points exist in general.

Remark:

- The numbers of source and collocation points need not be equal. If they are different, the linear system has a nonsquare matrix. In this case, a least squares approach or the Singular Value Decomposition can be used.

The method can be generalized to Poisson equations as well, using a scattered data interpolation technique based on radial basis functions (RBF). Consider the Poisson equation

$$\Delta u = f$$

in the domain Ω supplied with the mixed boundary condition:

$$u|_{\Gamma_D} = u_0, \qquad \frac{\partial u}{\partial n}|_{\Gamma_N} = v_0$$

The solution is sought as a sum of a *particular solution* and a *homogeneous solution*:

$$u = u_P + u_H,$$

where the function u_P is assumed to satisfy the Poisson equation

$$\Delta u_P = f$$

without requiring any boundary condition. Once a particular solution u_P has been determined, the homogeneous solution can be obtained by solving a Laplace equation supplied with modified boundary conditions:

$$u_H|_{\Gamma_D} = u_0 - u_P|_{\Gamma_D}, \qquad \frac{\partial u_H}{\partial n}|_{\Gamma_N} = v_0 - \frac{\partial u_P}{\partial n}|_{\Gamma_N}$$

Obviously, the function $u = u_P + u_H$ satisfies the Poisson equation and the original boundary conditions. This technique is known as the *method of particular solutions.*

The homogeneous equation can be solved by the Method of Fundamental Solutions, so that the only problem is to find a particular solution.

Define additional points scattered in the *interior* of the domain: w_1, w_2, ... , $w_M \in \Omega$. Let Ψ be a radial basis function, and approximate the function f by a scattered data interpolation based on the radial basis function Ψ (e.g. a multiquadric function or the thin plate spline function etc.). Recall that the interpolation function has the form:

$$f(x) = \sum_{j=1}^{M} \beta_j \cdot \Psi(x - w_j),$$

where the coefficients β_1, ..., β_M can be calculated by enforcing the interpolation conditions:

$$\sum_{j=1}^{M} \beta_j \cdot \Psi(w_k - w_j) = f(w_k) \qquad (k = 1, ..., M)$$

The crucial idea is *to find another radial basis function* Θ *such that*

$$\Delta\Theta = \Psi$$

is satisfied. If Θ is such a radial basis function, then the function

$$u_P(x) := \sum_{j=1}^{M} \beta_j \cdot \Theta(x - w_j)$$

is an (approximate) particular solution, since at the interior interpolation points:

$$\Delta u_P(w_k) = \sum_{j=1}^{M} \beta_j \cdot \Delta\Theta(w_k - w_j) =$$

$$= \sum_{j=1}^{M} \beta_j \cdot \Psi(w_k - w_j) = f(w_k) \qquad (k = 1, ..., M)$$

The overall algorithm can be summarized as follows:

- Define some interpolation points w_1, w_2, ..., w_M scattered in the interior of the domain Ω.
- Choose a radial basis function Ψ, and perform an interpolation based on the radial basis function Ψ, the interpolation points w_1, w_2, ..., w_M and the corresponding values $f(w_1)$, $f(w_2)$, ..., $f(w_M)$. Calculate the coefficients β_1, β_2, ..., β_M by solving the interpolation equations.
- Using the same coefficients β_j computed in the previous step, define the particular solution by

  $$u_P(x) := \sum_{j=1}^{M} \beta_j \cdot \Theta(x - w_j)$$

  where the radial basis function Θ satisfies the equality $\Delta\Theta = \Psi$.

- Solve the Dirichlet problem

$$\Delta u_H = 0$$

supplied with the modified boundary conditions

$$u_H|_{\Gamma_D} = u_0 - u_P|_{\Gamma_D}, \qquad \frac{\partial u_H}{\partial n}|_{\Gamma_N} = v_0 - \frac{\partial u_P}{\partial n}|_{\Gamma_N}$$

using the Method of Fundamental Solutions.

- The (approximate) solution of the original mixed problem is:

$$u = u_H + u_P$$

The problem which has still to be solved is the proper definition of the radial basis function Θ provided that the function Ψ is known. Here three concrete choices are presented. For the sake of simplicity, polar coordinates are used. The formulas can be verified by straightforward but more or less lengthy calculations; they are left as an exercise.

1. Define Ψ by

$$\boxed{\Psi(r) := 1 + r}$$

Then

$$\boxed{\Theta(r) = \frac{1}{4}r^2 + \frac{1}{9}r^3}$$

2. (Thin plate spline.) Define Ψ by

$$\boxed{\Psi(r) := r^2 \cdot \log r}$$

Then

$$\boxed{\Theta(r) = \frac{1}{16}r^4 \log r - \frac{1}{32}r^4}$$

3. (Multiquadric function.) Define Ψ by

$$\boxed{\Psi(r) := \sqrt{r^2 + c^2}}$$

Then

$$\boxed{\Theta(r) = \frac{1}{9} \cdot (4c^2 + r^2) \cdot \sqrt{r^2 + c^2} - \frac{c^3}{3} \cdot \log\left(c + \sqrt{r^2 + c^2}\right)}$$