# The Finite Element Method

Theory and Application to Heat Conduction Problems

2026

**UDC** 519.63
**LBC** 22.19

**Elyukov A. S.**

**The Finite Element Method: Theory and Application to Heat Conduction Problems.**

The book is devoted to the numerical solution of heat conduction equations by the finite element method (FEM). It presents analytical solutions of one-, two- and three-dimensional boundary value problems by the Fourier method of separation of variables, the variational formulation of the problem and its discretisation by the Bubnov–Galerkin method. Stiffness and damping matrices and load vectors for simplex finite elements are described, together with the global system assembly algorithm and the treatment of boundary conditions.

For students, postgraduates and engineers familiar with the basics of mathematical analysis and linear algebra.

# Contents

# Chapter 1

# Introduction

## Overview

The goal of this guide is to present the theory of the finite element method using simple examples, without excessive mathematical rigor. The material is aimed at an engineer with a knowledge of the basics of higher mathematics.

Partial differential equations have long been successfully solved both by analytical methods — for simple geometries — and by numerical ones — for complex geometries. In practice geometries are usually complex, so one has to solve numerically. Among numerical methods, the most widespread are the finite difference method (FDM) and the finite element method (FEM). The finite element method is preferable because it allows a nonuniform mesh of simplices, which makes it possible to concentrate computational resources where they are really needed.

This guide deals with one of the simplest equations — the nonstationary heat conduction equation. In practice you will most likely face more complex problems, but once you master the method presented here, more general cases will become much clearer. You are strongly encouraged to verify all formulas yourself.

## The equation

This guide is devoted to thermal fields, which in general are nonstationary. Let us consider a small volume $dV$ inside a body with boundary $S$ and outward unit normal $\overrightarrow{n}$ (figure 1.1). The idea is simple: the amount of heat accumulated in the volume per unit time equals the heat inflow through its boundary plus the heat released inside by internal sources. This equation is called the energy balance equation and follows from the law of conservation of energy. Through the boundary there can be not only an inflow of heat but also an outflow, which is mathematically described as a negative inflow; similarly with internal sources, which can be consumers.

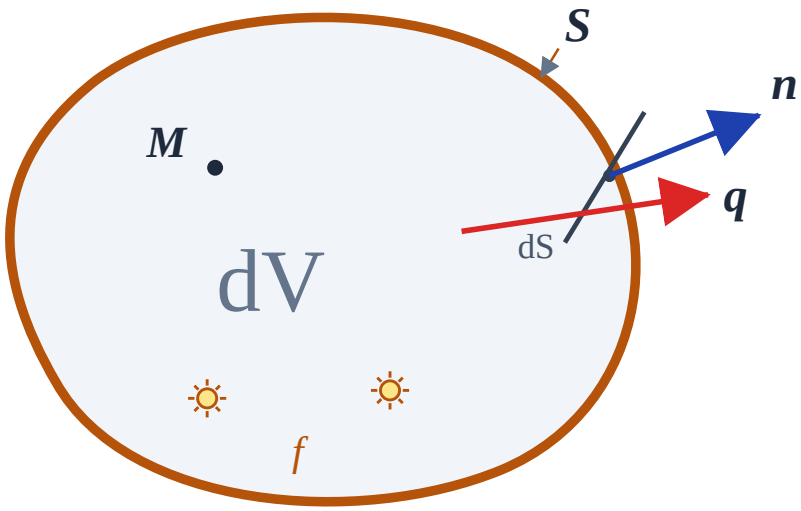


**Fig. 1.1.** The energy balance in the small volume $dV$: the heat flux $\overrightarrow{q}$ through the boundary $S$ and internal heat sources.

The formula for thermal energy is known $dQ = c \cdot dm \cdot T$, where $dQ$ is the thermal energy of the small volume [J], $c$ is the specific heat capacity [J/(kg · K)], $dm$ is the mass of the small volume [kg], $T$ is the temperature [K]. The mass of the small volume $dV$ can be written in terms of the density $\rho$ [kg/m$^3$] and the volume itself $dV$ [m$^3$]: $dm = \rho \cdot dV$. Then the energy stored in this small volume equals $dQ = c \cdot \rho \cdot T \cdot dV$. The rate of change of the thermal energy of the volume $V$ equals

$$\frac{\partial}{\partial t} \int_V c \cdot \rho \cdot T \, dV = \int_V c \cdot \rho \cdot \frac{\partial T}{\partial t} \, dV. \tag{1.1}$$

Now let us describe the heat transfer. Fourier's law states that the heat flux density $\overrightarrow{q}$ [W/m$^2$] is proportional to the temperature gradient and directed against it — heat flows from hot regions to cold ones:

$$\overrightarrow{q} = -k \, \nabla T, \tag{1.2}$$

where $k$ is the thermal conductivity [W/(m · K)]. The minus sign precisely reflects the fact that the flux is directed toward decreasing temperature.

Let us assemble the energy balance. The rate of change of the energy of the volume equals the heat inflow through the boundary (the inward flux is $-\overrightarrow{q} \cdot \overrightarrow{n}$, since $\overrightarrow{n}$ is the outward one) plus the power of the internal sources with volume density $F$ [W/m$^3$]:

$$\int_V c \cdot \rho \cdot \frac{\partial T}{\partial t} \, dV = -\oint_S \overrightarrow{q} \cdot \overrightarrow{n} \, dS + \int_V F \, dV. \tag{1.3}$$

The surface integral is inconvenient — we get rid of it by the Ostrogradsky–Gauss theorem, turning the flux through the boundary into a volume integral of the divergence:

$$\oint_S \overrightarrow{q} \cdot \overrightarrow{n} \, dS = \int_V \operatorname{div} \overrightarrow{q} \, dV. \tag{1.4}$$

Substituting (1.4) into (1.3) and collecting everything under one integral, we obtain

$$\int_V \left( c \cdot \rho \cdot \frac{\partial T}{\partial t} + \operatorname{div} \overrightarrow{q} - F \right) dV = 0. \tag{1.5}$$

The volume $V$ was chosen arbitrarily, and the integral over any such volume equals zero — hence the integrand itself equals zero. Substituting Fourier's law (1.2), we arrive at the equation in differential form:

$$c \cdot \rho \cdot \frac{\partial T}{\partial t} = \operatorname{div}(k \cdot \nabla T) + F. \tag{1.6}$$

It remains to use the assumption made above. The medium is homogeneous and isotropic, so the thermal conductivity $k$ is the same at all points and in all directions — it can be taken outside the divergence sign, and $\operatorname{div} \nabla \equiv \Delta$ is the Laplace operator:

$$c \cdot \rho \cdot \frac{\partial T}{\partial t} = k \cdot \Delta T + F. \tag{1.7}$$

Dividing both sides by $c \cdot \rho$ and denoting $a^2 = \dfrac{k}{c \cdot \rho}$ — the thermal diffusivity [$\mathrm{m}^2/\mathrm{s}$] — and $f = \dfrac{F}{c \cdot \rho}$ the reduced source density [K/s], we finally obtain the heat conduction equation:

$$\frac{\partial T(M,t)}{\partial t} = a^2 \cdot \Delta T(M,t) + f(M,t), \tag{1.8}$$

where $a^2$ is the thermal diffusivity, $\Delta$ is the Laplace operator, $M$ is the geometry, $t$ is time [s], $T(M,t)$ is the sought solution — the temperature field [K], $f(M,t)$ is the heat source density function inside the body [K/s].

Equation (1.8) is an inhomogeneous equation, since the right-hand side contains the heat source density function $f(M,t)$. The corresponding homogeneous equation has the form

$$\frac{\partial T(M,t)}{\partial t} = a^2 \cdot \Delta T(M,t). \tag{1.9}$$

The importance of the homogeneous equation will become apparent later, when we obtain the general solution by the Fourier method of separation of variables.

For the test problems we will consider the following coordinate systems and Laplace operators.

In the one-dimensional case the Cartesian coordinate system is used $M \equiv x$:

$$\Delta \equiv \frac{\partial^2}{\partial x^2}. \tag{1.10}$$

In the two-dimensional case the polar coordinate system is used $M \equiv (r, \theta)$:

$$\Delta \equiv \frac{\partial^2}{\partial r^2} + \frac{1}{r} \cdot \frac{\partial}{\partial r} + \frac{1}{r^2} \cdot \frac{\partial^2}{\partial \theta^2}. \tag{1.11}$$

In the three-dimensional case the spherical coordinate system is used $M \equiv (r, \theta, \phi)$:

$$\Delta \equiv \frac{\partial^2}{\partial r^2} + \frac{2}{r} \cdot \frac{\partial}{\partial r} + \frac{1}{r^2 \sin\theta} \cdot \frac{\partial}{\partial \theta}\left(\sin\theta \cdot \frac{\partial}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta} \cdot \frac{\partial^2}{\partial \phi^2}. \tag{1.12}$$

## Boundary conditions

The equation alone is not enough: it only describes the relations between the particles of the object itself, but the object does not exist like a spherical horse in a vacuum, so the object under study must somehow be connected to the surrounding world. The boundary value conditions for our example are the combination of the boundary conditions and the initial condition $T_0(M)$. The initial condition is clear: it describes the temperature in the object $M$ at all interior points and on the boundary $S$. There may be a case where the boundary is not explicitly specified: for example, we compute the electromagnetic field above the ground; clearly the domain under consideration is the whole Universe, but we artificially bound it within some reasonable limits and adapt the size of the finite elements in proportion to the distance from the field source. This is an example of the advantage of FEM over FDM, where we would have to compute the field where it is of little interest to us. I will also note that we will not specify the thermal conductivity coefficients or the density function — all of this will be in the most general form that FEM can accept; the type of boundary condition, for example, directly affects the construction of the method and must not be overlooked.

The initial condition is clear, so let us move on to the boundary conditions. Imagine that we heat a steel ball in a blast furnace at 1000 K; we can assume the ball is placed in an environment on which it has no influence, and the sphere bounding it will stay at 1000 K throughout the heating cycle. Such conditions are described by equation (1.13), which is called a boundary condition of the first kind, or a Dirichlet boundary condition:

$$T(M,t) = \Phi(M,t), \quad M \in S. \tag{1.13}$$

That is, the temperature on the surface is prescribed by some function, which in general may depend on time; a particular case is a constant (in our example, 1000 K).

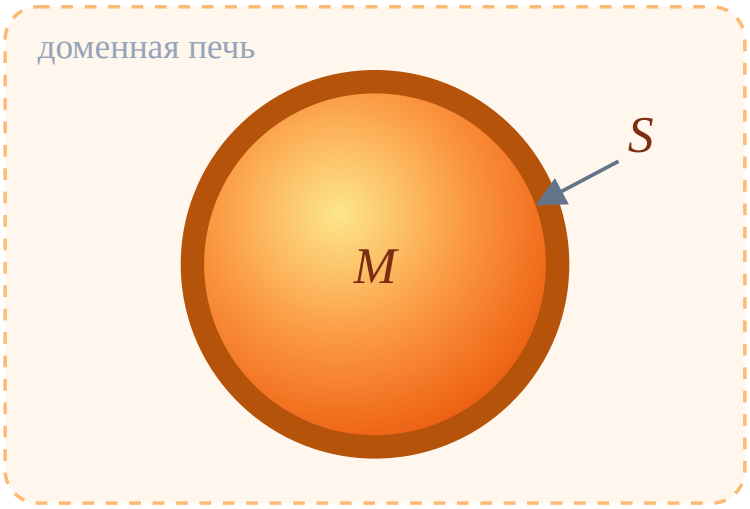


**Fig. 1.2.** Boundary condition of the first kind (Dirichlet): the surface temperature is prescribed.

Now let us take the ball out of the furnace heated to 1000 K and wrap it in thermal insulation. The ball will cool at roughly one rate, that is, the cooling pace will be more or

less constant. This is a boundary condition of the second kind (Neumann); it prescribes the normal heat flux through the surface (1.14):

$$\lambda \cdot \frac{\partial T(M,t)}{\partial \overrightarrow{n}} = \Phi(M,t), \quad M \in S, \tag{1.14}$$

where $\overrightarrow{n}$ is the outward normal to the surface $S$ at the point $M$, $\lambda$ is the thermal conductivity. This condition applies when the surface is being heated or cooled: if $\Phi(M,t) \equiv 0$, the surface is insulated; if $\Phi(M,t) < 0$, cooling takes place; if $\Phi(M,t) > 0$, heating.

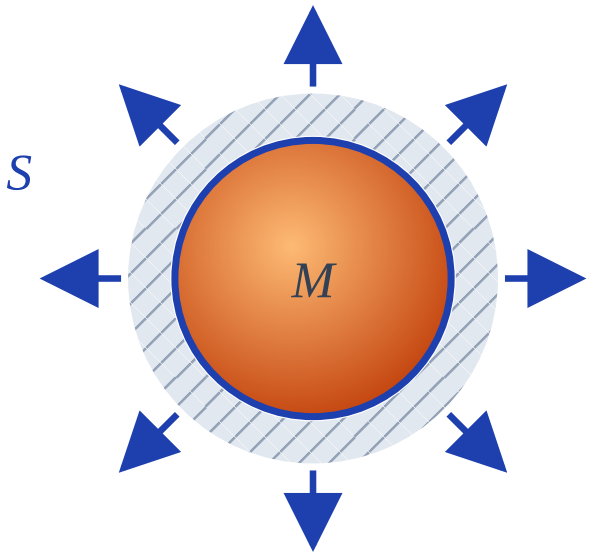


**Fig. 1.3.** Boundary condition of the second kind (Neumann): the flux through the surface is prescribed.

There is also a condition of the third kind (Robin). Let the same ball cool in the open air with no wind: heat is gradually removed from the surface into the environment, the more intensely the hotter the ball is relative to the air. Such heat exchange is described by equation (1.15):

$$-\lambda \cdot \frac{\partial T(M,t)}{\partial \overrightarrow{n}} = \alpha \left(T(M,t) - T_c\right), \quad M \in S, \tag{1.15}$$

where $\alpha$ is the heat transfer coefficient, $T_c$ is the ambient temperature. The flux through the surface is proportional to the temperature difference — this is Newton's law of cooling.

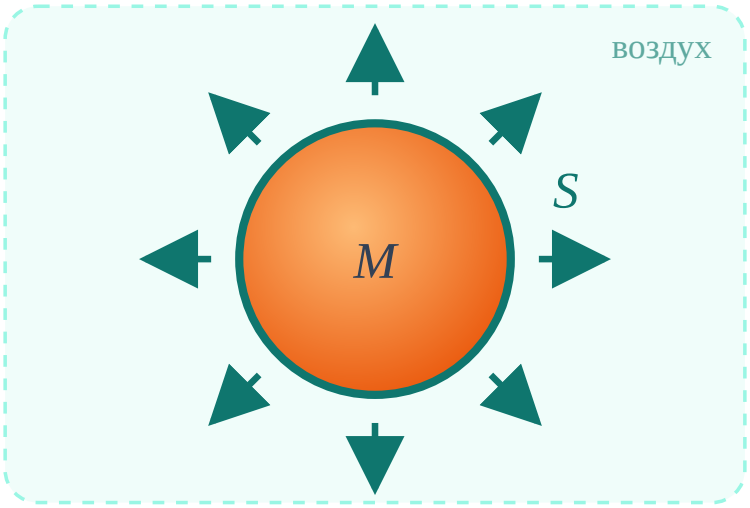


**Fig. 1.4.** Boundary condition of the third kind (Robin): convective heat exchange with the environment.

In this guide we restrict ourselves to the first two kinds of boundary conditions. The corresponding homogeneous boundary conditions have the form

$$T(M,t) = 0, \quad M \in S. \tag{1.16}$$

$$\lambda \cdot \frac{\partial T(M,t)}{\partial \overrightarrow{n}} = 0, \quad M \in S. \tag{1.17}$$

The importance of homogeneous boundary conditions will become apparent later, when we obtain the general solution by the Fourier method of separation of variables.

Equation (1.8) with the initial condition $T_0(M)$ and the boundary condition (1.13) is called the Dirichlet boundary value problem, while the same equation with the same initial condition but with the boundary condition (1.14) is called the Neumann boundary value problem. We will consider both problems in this guide.

# Chapter 2

# Analytical solution of boundary value problems

## General analytical solution of the boundary value problem

In this chapter we will construct the general solution of the heat conduction boundary value problem by the method of separation of variables (the Fourier method). But first, a few introductory remarks. We want to solve the inhomogeneous equation (1.8) with the initial condition $T_0(M)$ and inhomogeneous boundary conditions (1.13) and (1.14). Solving it head-on can be very difficult, so first we have to get rid of the inhomogeneity of the boundary conditions.

Let us represent the solution as the sum $T(M,t) = \widehat{T}(M,t)+U(M,t)$, where $\widehat{T}(M,t)$ is the solution of the problem with homogeneous boundary conditions, and $U(M,t)$ is a function that will absorb the inhomogeneity of the boundary conditions. Let us substitute the sum into equation (1.8) and obtain

$$\frac{\partial \widehat{T}(M,t)}{\partial t} = a^2 \cdot \Delta \widehat{T}(M,t) + f(M,t) + a^2 \cdot \Delta U(M,t) - \frac{\partial U(M,t)}{\partial t}.$$

The form of the function $U(M,t)$ we will determine later, when considering specific problems; for now let us note that it can be chosen as a polynomial defined up to coefficients, which are found from the boundary conditions. The new heat source density function has the form

$$\widehat{f}(M,t) = f(M,t) + a^2 \cdot \Delta U(M,t) - \frac{\partial U(M,t)}{\partial t}. \tag{2.1}$$

The coefficients $U(M,t)$ for the Dirichlet boundary conditions are found by substituting the solution into the boundary conditions (1.13). The solution $\widehat{T}(M,t)$ satisfies the homogeneous Dirichlet condition, that is, vanishes on the boundary $S$, whence

$$\widehat{T}(M,t) + U(M,t) = \Phi(M,t), \quad \widehat{T}(M,t) = 0 \;\Rightarrow\; U(M,t) = \Phi(M,t), \quad M \in S. \tag{2.2}$$

Similarly, the coefficients $U(M,t)$ for the Neumann boundary conditions are found by substituting the solution into the boundary conditions (1.14). The solution $\widehat{T}(M,t)$ satis-

fies the homogeneous Neumann condition, that is, its normal derivative on the boundary $S$ vanishes, whence

$$\frac{\partial \widehat{T}(M,t)}{\partial \overrightarrow{n}} + \frac{\partial U(M,t)}{\partial \overrightarrow{n}} = \Phi(M,t), \quad \frac{\partial \widehat{T}(M,t)}{\partial \overrightarrow{n}} = 0 \;\Rightarrow\; \frac{\partial U(M,t)}{\partial \overrightarrow{n}} = \Phi(M,t), \quad M \in S. \tag{2.3}$$

Let us not forget the initial condition, which does not compute any coefficients but is simply rewritten as

$$\widehat{T}(M,0) = T_0(M) - U(M,0). \tag{2.4}$$

After computing the coefficients of the polynomial function $U(M,t)$ the inhomogeneous boundary value problem with inhomogeneous boundary conditions reduces to an inhomogeneous boundary value problem with homogeneous boundary conditions. Further, to simplify the notation somewhat, we will not write the hat over the solution $\widehat{T}(M,t)$, the initial condition $\widehat{T}(M,0)$ and the source function $\widehat{f}(M,t)$, denoting them simply $T(M,t)$, $T(M,0)$ and $f(M,t)$.

Now we have to get rid of the inhomogeneity in the equation itself, but unlike the procedure for removing the inhomogeneity in the boundary conditions, here we go from the other side: first we solve the homogeneous equation, and then, based on it, we explain how to solve the inhomogeneous one. To see that such a transition is possible, we will rely on *Steklov's expansion theorem*, since it is precisely what gives the basis in which everything is subsequently expanded. Having solved the homogeneous problem, we obtain the set of its eigenfunctions. Steklov's theorem states that this system is complete: every function with a continuous first and piecewise continuous second derivative, satisfying the same homogeneous boundary conditions, for example (1.16) or (1.17), expands into a uniformly convergent series in the eigenfunctions of the homogeneous problem (1.9). This means that both the sought solution of the inhomogeneous problem and the source function can be represented in the same basis: projecting the inhomogeneous equation onto each eigenfunction and using their orthogonality, we reduce it to independent ordinary equations for the coefficients — that is, we build the solution of the inhomogeneous problem on the solution of the homogeneous one.

We use the method of separation of variables (the Fourier method). We will look for the solution of equation (1.9) as a sum of products $T(M,t) = \sum_{n=1}^{\infty} \Psi_n(M) \cdot \phi_n(t)$, where each term is a product of a function of geometry only and a function of time only. The equation is linear and homogeneous, so it suffices to require that each term satisfy it separately; let us consider one product $\Psi(M) \cdot \phi(t)$ and substitute it into equation (1.9) — we obtain

$$\Psi(M) \cdot \frac{\partial \phi(t)}{\partial t} = a^2 \cdot \Delta\Psi(M) \cdot \phi(t), \tag{2.5}$$

which after division by $\Psi(M) \cdot \phi(t)$ leads to

$$\frac{\partial \phi(t)}{\partial t} \cdot \frac{1}{a^2 \cdot \phi(t)} = -\gamma^2 = \Delta\Psi(M) \cdot \frac{1}{\Psi(M)}, \tag{2.6}$$

where $\gamma^2$ is a constant. Indeed, the left-hand side of the equality depends only on time, while the right-hand side depends only on the geometry; a function of time can

coincide with a function of geometry at all moments of time and at all points of the domain only when both are equal to one and the same constant — this is the essence of the separation of variables. The minus sign in front of it is chosen so that the solution converges, which is easy to verify by examining the equation for time. The specific values of this constant we will determine later — from the equation for the geometry.

Let us write the equation for time and its solution

$$\frac{\partial\phi(t)}{\partial t}\cdot\frac{1}{a^2\cdot\phi(t)}=-\gamma^2$$

$$\phi(t)=A\cdot e^{-a^2\cdot\gamma^2\cdot t}, \tag{2.7}$$

where $A$ is a constant.

The equation for the geometry has the form

$$\Delta\Psi(M)+\gamma^2\cdot\Psi(M)=0. \tag{2.8}$$

This is the homogeneous Helmholtz equation. To solve it, boundary conditions must be specified. Let us substitute the product $\Psi(M)\cdot\phi(t)$ into the boundary conditions (1.16) and (1.17) respectively and obtain for the Dirichlet conditions

$$\Psi(M)\cdot\phi(t)=0\Rightarrow\Psi(M)=0,\quad M\in S, \tag{2.9}$$

and for the Neumann conditions

$$\frac{\partial\Psi(M)}{\partial\overrightarrow{n}}\cdot\phi(t)=0\Rightarrow\frac{\partial\Psi(M)}{\partial\overrightarrow{n}}=0,\quad M\in S. \tag{2.10}$$

Clearly, the case when $\phi(t)=0$ is of no interest, since we are constructing conditions for the geometry, not for time. Equation (2.8) with the boundary conditions (2.9) and (2.10) is called the Sturm–Liouville problem. It is well studied, and below we will use its known properties without proof (they can be found in the literature). This problem has not one but a countable set of eigenvalues, and each of them corresponds to its own eigenfunction; eigenfunctions corresponding to different eigenvalues are mutually orthogonal and form a complete system.

So, the Sturm–Liouville problem yielded a countable set of eigenvalues $\gamma_n^2$ and the corresponding eigenfunctions $\Psi_n(M)$. For the mode with index $n$ the equation for time gives the factor $e^{-a^2\cdot\gamma_n^2\cdot t}$, so that $\Psi_n(M)\cdot e^{-a^2\cdot\gamma_n^2\cdot t}$ is a particular solution of equation (1.9). Returning to the sum we started with and using linearity, we obtain the general solution

$$T(M,t)=\sum_{n=1}^{\infty}c_n\cdot\Psi_n(M)\cdot e^{-a^2\cdot\gamma_n^2\cdot t}, \tag{2.11}$$

where $c_n$ are arbitrary constants that have absorbed the constant factors of both parts of each term. The completeness of the system of eigenfunctions (Steklov's theorem) guarantees that any admissible solution can be represented by such a sum.

Now we have to determine the coefficients $c_n$; to do so, note that the solution must satisfy the initial condition (2.4), and hence

$$T(M,0)=\sum_{n=1}^{\infty}c_n\cdot\Psi_n(M)=T_0(M).$$

It is easy to see that the coefficients $c_n$ are determined uniquely thanks to the mutual orthogonality of the eigenfunctions — as the expansion coefficients of the function $T_0(M)$ into a Fourier series in the eigenfunctions $\Psi_n(M)$ — by the general formula (A.4) from the appendix "Fourier series"

$$c_n = \frac{1}{\|\Psi_n(M)\|^2} \cdot \iiint_G \rho(M) \cdot T_0(M) \cdot \Psi_n(M)\, dG, \tag{2.12}$$

where $\|\Psi_n(M)\|^2$ is the squared norm of the eigenfunction, defined by formula (A.5).

Thus, we have obtained the solution of equation (1.9) with the initial condition $T_0(M)$ and homogeneous boundary conditions

$$T(M,t) = \sum_{n=1}^{\infty} \frac{\Psi_n(M)}{\|\Psi_n(M)\|^2} \cdot e^{-a^2 \cdot \gamma_n^2 \cdot t} \cdot \iiint_G \rho(M) \cdot T_0(M) \cdot \Psi_n(M)\, dG. \tag{2.13}$$

Now we can proceed to solving the inhomogeneous equation, which, as we said, in accordance with Steklov's expansion theorem, can be built on the solution of the homogeneous problem. To do so, we expand the heat source density function $f(M,t)$ into a Fourier series in the eigenfunctions $\Psi_n(M)$

$$f(M,t) = \sum_{n=1}^{\infty} \mu_n(t) \cdot \Psi_n(M), \tag{2.14}$$

where $\mu_n(t)$ are the expansion coefficients of the function $f(M,t)$ into a Fourier series in the eigenfunctions $\Psi_n(M)$

$$\mu_n(t) = \frac{1}{\|\Psi_n(M)\|^2} \cdot \iiint_G \rho(M) \cdot f(M,t) \cdot \Psi_n(M)\, dG.$$

We will look for the solution of the inhomogeneous equation in the same form — as a sum of products $\sum_{n=1}^{\infty} \Psi_n(M) \cdot \phi_n(t)$. Let us substitute it together with the expansion (2.14) into the inhomogeneous equation (1.8); since $\Delta\Psi_n(M) = -\gamma_n^2 \cdot \Psi_n(M)$, we obtain

$$\sum_{n=1}^{\infty} \Psi_n(M) \cdot \frac{\partial \phi_n(t)}{\partial t} = -a^2 \cdot \sum_{n=1}^{\infty} \gamma_n^2 \cdot \Psi_n(M) \cdot \phi_n(t) + \sum_{n=1}^{\infty} \mu_n(t) \cdot \Psi_n(M).$$

The geometric part remains the same — these are the same eigenfunctions $\Psi_n(M)$. Equating the coefficients of each $\Psi_n(M)$ (this is legitimate due to orthogonality), we see that only the equation for time has changed — it has acquired the term $\mu_n(t)$:

$$\frac{\partial \phi_n(t)}{\partial t} = -a^2 \cdot \gamma_n^2 \cdot \phi_n(t) + \mu_n(t). \tag{2.15}$$

The solution of this equation has the form

$$\phi_n(t) = c_n \cdot e^{-a^2 \cdot \gamma_n^2 \cdot t} + \int_0^t \mu_n(\tau) \cdot e^{-a^2 \cdot \gamma_n^2 \cdot (t-\tau)}\, d\tau, \tag{2.16}$$

where $c_n$ are arbitrary coefficients equal to the coefficients determined in the solution of the homogeneous problem, since they are computed from the initial condition, and at $t = 0$ the integral equals zero.

The solution of the inhomogeneous equation with homogeneous boundary conditions takes the form

$$\begin{aligned} &T(M,t) = \\ &\sum_{n=1}^{\infty} \frac{\Psi_n(M)}{\|\Psi_n(M)\|^2} \cdot e^{-a^2 \cdot \gamma_n^2 \cdot t} \cdot \iiint_G \rho(M) \cdot T_0(M) \cdot \Psi_n(M)\, dG + \\ &\sum_{n=1}^{\infty} \frac{\Psi_n(M)}{\|\Psi_n(M)\|^2} \cdot \int_0^t e^{-a^2 \cdot \gamma_n^2 \cdot (t-\tau)} \cdot \iiint_G \rho(M) \cdot f(M,\tau) \cdot \Psi_n(M)\, dG\, d\tau. \end{aligned} \tag{2.17}$$

One could also obtain the solution of the inhomogeneous equation with inhomogeneous boundary conditions, but this is rather cumbersome — we will construct the final solution later using specific examples.

When solving stationary equations — and the electromagnetic field is generally stationary unless relativistic effects are considered — we need to obtain the general solution without time, and for that we take the limit $t \to \infty$ in the solution of the nonstationary equation. Note that the first term of the solution vanishes immediately, since the exponential tends to zero and the initial conditions disappear from the solution, which is more than obvious. The heat source density function stops depending on time a priori. It only remains to clarify the integral

$$\begin{aligned} \int_0^t e^{-a^2 \cdot \gamma_n^2 \cdot (t-\tau)}\, d\tau &= \frac{1}{a^2 \cdot \gamma_n^2} \cdot e^{-a^2 \cdot \gamma_n^2 \cdot (t-\tau)} \Big|_0^t \\ &= \frac{1}{a^2 \cdot \gamma_n^2} \cdot (1 - e^{-a^2 \cdot \gamma_n^2 \cdot t}) \to \frac{1}{a^2 \cdot \gamma_n^2} \quad \text{as} \quad t \to \infty, \quad \gamma_n \\ &\neq 0. \end{aligned}$$

Let us write the general solution for the stationary equation

$$\begin{aligned} &T(M) \\ &\quad = \sum_{n=1}^{\infty} \frac{\Psi_n(M)}{\|\Psi_n(M)\|^2} \cdot \frac{1}{a^2 \cdot \gamma_n^2} \cdot \iiint_G \rho(M) \cdot f(M) \cdot \Psi_n(M)\, dG, \quad \gamma_n \\ &\qquad \neq 0. \end{aligned} \tag{2.18}$$

And what if $\gamma_n = 0$? In fact this case is possible, but it is better considered separately for each equation, which is what we will do in the corresponding chapters.

# Errors in posing a boundary value problem

It is not enough to solve a boundary value problem — it must be solved correctly, and for that one must at least avoid errors at the problem formulation stage. We will consider a couple of typical inaccuracies.

**Compatibility of initial and boundary conditions**

The first typical error is to prescribe initial and boundary conditions that do not agree with each other at the junction. In a nonstationary problem the solution is sought for $t \geq 0$ and $M \in G$; on the boundary $S$ for $t = 0$ the initial and boundary conditions describe the same point. If they give different values there, a discontinuity arises at that point.

Let us consider the Dirichlet problem with the initial condition $T(M, 0) = T_0(M)$ and the boundary condition $T(M, t) = \Phi(M, t)$, $M \in S$. The zeroth-order compatibility condition requires the data to coincide

$$T_0(M) = \Phi(M, 0), \quad M \in S. \tag{2.19}$$

If condition (2.19) is violated, the Fourier series of the solution approaches the boundary with different limits in $t$ and in $S$: the convergence is nonuniform, the Gibbs phenomenon appears, and no classical (smooth) solution exists.

But even the fulfillment of (2.19) is not enough. The equation itself $\frac{\partial T}{\partial t} = a^2 \cdot \Delta T$ relates the derivatives of the data: differentiating the boundary condition with respect to time and taking $t = 0$, we obtain the first-order compatibility condition

$$\frac{\partial \Phi(M, 0)}{\partial t} = a^2 \cdot \Delta T_0(M), \quad M \in S. \tag{2.20}$$

Higher-order conditions are constructed similarly; for a solution of class $C^k$ compatibility conditions are required up to roughly order $k$. Violating them does not make the problem unsolvable, but it limits the smoothness of the solution — something easy to forget when formally writing down "arbitrary" $T_0(M)$ and $\Phi(M, t)$. For the Neumann problem the compatibility conditions are the same but are written for the normal derivative:

$$\frac{\partial T_0}{\partial \overrightarrow{n}} = \Phi(M, 0), \quad M \in S.$$

**Solvability condition of the stationary Neumann problem**

And here let us also discuss the solvability condition of the stationary Neumann problem. First, it should be noted that the solution of the Neumann problem is defined up to a constant, which can be fixed at any point of the geometry. But the main issue is that for the Neumann problem it is essential that the sum of the energy fluxes — incoming, outgoing, sources and sinks inside the geometry — equals zero. If this condition is not met, the solution will have an incorrect physical meaning. For example, what happens if a heat flux enters a rod while its other end is insulated? Then the temperature will grow indefinitely in the steady state, which essentially will not exist. In practice, of course, this cannot happen, because the rod is not one-dimensional and heat can leak through the lateral surface, and pure Neumann conditions do not exist anyway; nevertheless, the model must be correct. Let us consider an example of a stationary Neumann problem. The time derivative in equation (1.8) then vanishes, and it takes the form

$$a^2 \cdot \Delta T(M) + f(M) = 0. \tag{2.21}$$

The Neumann boundary conditions for it are given by the boundary condition (1.14).

Green's second formula is known

$$\iiint_G (U(M)\cdot\Delta T(M) - T(M)\cdot\Delta U(M))\,dM = \iint_S \left(U(M)\cdot\frac{\partial T(M)}{\partial\overrightarrow{n}} - T(M)\cdot\frac{\partial U(M)}{\partial\overrightarrow{n}}\right) dS.$$

If we set $U(M) = -1$, we obtain

$$-\iiint_G \Delta T(M)\,dM = -\iint_S \frac{\partial T(M)}{\partial\overrightarrow{n}}\,dS.$$

Now let us substitute equation (2.21) and the boundary conditions (1.14)

$$\iiint_G f(M)\,dM = -a^2\cdot\iint_S \Phi(M)\,dS, \tag{2.22}$$

that is, the amount of heat produced by the sources inside the geometry equals the amount of heat leaving through the boundary. If the condition is violated, the solution will have an incorrect physical meaning.

## One-dimensional heat conduction boundary value problem

Let us consider the one-dimensional heat conduction problem on the interval from $\alpha$ to $\beta$. The heat conduction equation (1.8) in this case takes the form

$$\frac{\partial T(x,t)}{\partial t} = a^2\cdot\frac{\partial^2 T(x,t)}{\partial x^2} + f(x,t). \tag{2.23}$$

The Dirichlet boundary conditions (1.13) in the one-dimensional case have the form

$$\begin{cases} T(x,0) = T_0(x), \\ T(\alpha,t) = \Phi_\alpha(t), \\ T(\beta,t) = \Phi_\beta(t) \end{cases} \tag{2.24}$$

To get rid of the inhomogeneity, we introduce $T(x,t) = \widehat{T}(x,t) + U(x,t)$, let us choose the function $U(x,t) = a_1(t) + a_2(t)\cdot x$, and compute the coefficients, in accordance with (2.2)

$$\begin{cases} \Phi_\alpha(t) - a_1(t) - a_2(t)\cdot\alpha = 0 \\ \Phi_\beta(t) - a_1(t) - a_2(t)\cdot\beta = 0 \end{cases} \tag{2.25}$$

$$\begin{cases} a_1(t) = \dfrac{\Phi_\alpha(t)\cdot\beta - \Phi_\beta(t)\cdot\alpha}{\beta-\alpha} \\ a_2(t) = \dfrac{\Phi_\beta(t) - \Phi_\alpha(t)}{\beta-\alpha} \end{cases} \tag{2.26}$$

Then equation (2.23) and the boundary conditions (2.24) take the form

$$\begin{cases} \dfrac{\partial \widehat{T}(x,t)}{\partial t} = a^2 \cdot \dfrac{\partial^2 \widehat{T}(x,t)}{\partial x^2} + \widehat{f}(x,t), \\ \widehat{T}(x,0) = T_0(x) - a_1(0) - a_2(0) \cdot x, \\ \widehat{T}(\alpha,t) = 0, \widehat{T}(\beta,t) = 0, \end{cases} \tag{2.27}$$

where $\widehat{f}(x,t) = f(x,t) - \dfrac{da_1(t)}{dt} - \dfrac{da_2(t)}{dt} \cdot x$.
The Neumann boundary conditions (1.14) in the one-dimensional case have the form

$$\begin{cases} T(x,0) = T_0(x), \\ \left.\dfrac{\partial T(x,t)}{\partial x}\right|_{x=\alpha} = \Phi_\alpha(t), \\ \left.\dfrac{\partial T(x,t)}{\partial x}\right|_{x=\beta} = \Phi_\beta(t) \end{cases} \tag{2.28}$$

To get rid of the inhomogeneity, we introduce $T(x,t) = \widehat{T}(x,t) + U(x,t)$, let us choose the function $U(x,t) = a_1(t) \cdot x + a_2(t) \cdot x^2$, and compute the coefficients, in accordance with (2.3)

$$\begin{cases} \Phi_\alpha(t) - a_1(t) - 2 \cdot a_2(t) \cdot \alpha = 0 \\ \Phi_\beta(t) - a_1(t) - 2 \cdot a_2(t) \cdot \beta = 0 \end{cases} \tag{2.29}$$

$$\begin{cases} a_1(t) = \dfrac{\Phi_\alpha(t) \cdot \beta - \Phi_\beta(t) \cdot \alpha}{\beta - \alpha} \\ a_2(t) = \dfrac{\Phi_\beta(t) - \Phi_\alpha(t)}{2 \cdot (\beta - \alpha)} \end{cases} \tag{2.30}$$

Then equation (2.23) and the boundary conditions (2.28) take the form

$$\begin{cases} \dfrac{\partial \widehat{T}(x,t)}{\partial t} = a^2 \cdot \dfrac{\partial^2 \widehat{T}(x,t)}{\partial x^2} + \widehat{f}(x,t), \\ \widehat{T}(x,0) = T_0(x) - a_1(0) \cdot x - a_2(0) \cdot x^2, \\ \left.\dfrac{\partial \widehat{T}(x,t)}{\partial x}\right|_{x=\alpha} = 0, \left.\dfrac{\partial \widehat{T}(x,t)}{\partial x}\right|_{x=\beta} = 0, \end{cases} \tag{2.31}$$

where $\widehat{f}(x,t) = f(x,t) - \dfrac{da_1(t)}{dt} \cdot x - \dfrac{da_2(t)}{dt} \cdot x^2 + 2 \cdot a^2 \cdot a_2(t)$.
Equation (2.8) takes the form

$$\frac{d^2\Psi(x)}{dx^2} + \gamma^2 \cdot \Psi(x) = 0. \tag{2.32}$$

The general solution of this equation has the form

$$\Psi(x) = c_1 \cdot \cos[\gamma \cdot (x - \alpha)] + c_2 \cdot \sin[\gamma \cdot (x - \alpha)], \tag{2.33}$$

where $c_1$ and $c_2$ are coefficients determined from the boundary conditions, and the weight ensuring the orthogonality of the eigenfunctions is $\rho(x) = 1$.
Let us substitute the general solution into the boundary conditions (2.27) and obtain

$$\begin{cases} c_1 \cdot \cos(\gamma \cdot 0) + c_2 \cdot \sin(\gamma \cdot 0) = 0, \\ c_1 \cdot \cos[\gamma \cdot (\beta - \alpha)] + c_2 \cdot \sin[\gamma \cdot (\beta - \alpha)] = 0. \end{cases} \tag{2.34}$$

Clearly, from the first equation $c_1 = 0$ and the solution makes sense only when $c_2 \neq 0$, which is possible only when $\sin[\gamma \cdot (\beta - \alpha)] = 0$

$$\gamma \cdot (\beta - \alpha) = \pi \cdot n, n \in (1..\infty). \tag{2.35}$$

Zero and negative values of $n$ are no good: for $n = 0$ we have $\gamma = 0$, and $\Psi(x) = c_2 \cdot \sin[0 \cdot (x - \alpha)] \equiv 0$ — identically zero, which is not an eigenfunction; to negative $n$ corresponds the same eigenvalue $\gamma^2$ (since $\gamma_{-n} = -\gamma_n$), and $\sin[-\gamma_n \cdot (x - \alpha)] = -\sin[\gamma_n \cdot (x - \alpha)]$ is linearly dependent with the solution for positive $n$ and gives no new contribution. Therefore $n \in (1..\infty)$.

Thus, the eigenvalues and eigenfunctions have the form

$$\begin{cases} \gamma_n = \dfrac{\pi \cdot n}{\beta - \alpha}, n \in (1..\infty), \\ \Psi_n(x) = c_{2n} \cdot \sin\left[\dfrac{\pi \cdot n}{\beta - \alpha} \cdot (x - \alpha)\right], \end{cases} \tag{2.36}$$

where $c_{2n}$ is an arbitrary constant factor: the eigenfunction is defined up to it. Let us set $c_{2n} = 1$ — this factor is absorbed into the expansion coefficients anyway $T_{0n}$ and $f_n$, as in the general solution.

To expand functions into a Fourier series in $\Psi_n(x)$ we need to compute the norm $\|\Psi_n(x)\|^2$; the weight for Cartesian coordinates is $\rho(x) = 1$.

$$\|\Psi_n(x)\|^2 = \int_\alpha^\beta \Psi_n(x)^2\, dx = \int_\alpha^\beta \sin\left[\frac{\pi \cdot n}{\beta - \alpha} \cdot (x - \alpha)\right]^2 dx = \frac{\beta - \alpha}{2}. \tag{2.37}$$

Now let us substitute the solution into the boundary conditions (2.31) and obtain

$$\begin{cases} -c_1 \cdot \sin(\gamma_n \cdot 0) + c_2 \cdot \cos(\gamma_n \cdot 0) = 0, \\ -c_1 \cdot \sin[\gamma_n \cdot (\beta - \alpha)] + c_2 \cdot \cos[\gamma_n \cdot (\beta - \alpha)] = 0. \end{cases} \tag{2.38}$$

Clearly, from the first equation $c_2 = 0$ and the solution makes sense only when $\sin[\gamma_n \cdot (\beta - \alpha)] = 0$

$$\gamma_n \cdot (\beta - \alpha) = \pi \cdot n, n \in (0..\infty). \tag{2.39}$$

Here, unlike the Dirichlet problem, the mode $n = 0$ survives: for $\gamma_0 = 0$ the eigenfunction $\Psi_0(x) = c_{10} \cdot \cos[0 \cdot (x - \alpha)] = c_{10}$ — a nonzero constant, that is, a fully fledged eigenfunction (the constant mode). Hence the indexing starts from zero.

Thus, the eigenvalues and eigenfunctions have the form

$$\begin{cases} \gamma_n = \dfrac{\pi \cdot n}{\beta - \alpha}, n \in (0..\infty), \\ \Psi_n(x) = c_{1n} \cdot \cos\left[\dfrac{\pi \cdot n}{\beta - \alpha} \cdot (x - \alpha)\right], \end{cases} \tag{2.40}$$

where $c_{1n}$ is an arbitrary constant factor: the eigenfunction is defined up to it. Let us set $c_{1n} = 1$ — this factor is absorbed into the expansion coefficients anyway $T_{0n}$ and $f_n$, as in the general solution.

We compute the norm of the eigenfunctions in the same way as in the sine case, reducing the power of the cosine by the half-angle formula. For $n \geq 1$

$$\|\Psi_n(x)\|^2 = \int_\alpha^\beta \Psi_n(x)^2\,dx = \int_\alpha^\beta \cos\left[\gamma_n\cdot(x-\alpha)\right]^2\,dx$$
$$= \int_\alpha^\beta \frac{1+\cos\left[2\cdot\gamma_n\cdot(x-\alpha)\right]}{2}\,dx = \frac{\beta-\alpha}{2}, \quad n\in(1..\infty), \quad (2.41)$$

since the integral of the cosine over an integer number of half-periods equals zero: $\sin\left[2\cdot\gamma_n\cdot(\beta-\alpha)\right] = \sin(2\cdot\pi\cdot n) = 0$. The special case $n=0$ (for $\gamma_0 = 0$ the eigenfunction $\Psi_0(x) = 1$) we compute separately

$$\|\Psi_0(x)\|^2 = \int_\alpha^\beta \Psi_0(x)^2\,dx = \int_\alpha^\beta 1\,dx = \beta-\alpha. \quad (2.42)$$

The final solution of the one-dimensional Dirichlet boundary value problem with inhomogeneous boundary conditions takes the form

$$\begin{aligned}
&a_1(t) = \frac{\Phi_\alpha(t)\cdot\beta - \Phi_\beta(t)\cdot\alpha}{\beta-\alpha}, \quad a_2(t) = \frac{\Phi_\beta(t)-\Phi_\alpha(t)}{\beta-\alpha},\\
&U(x,t) = a_1(t) + a_2(t)\cdot x,\\
&\widehat{T}(x,0) = T_0(x) - U(x,0),\\
&\widehat{f}(x,t) = f(x,t) - \frac{da_1(t)}{dt} - \frac{da_2(t)}{dt}\cdot x,\\
&\gamma_n = \frac{\pi\cdot n}{\beta-\alpha}, \quad \Psi_n(x) = \sin\left[\gamma_n\cdot(x-\alpha)\right], \quad \|\Psi_n(x)\|^2 = \frac{\beta-\alpha}{2}, \quad n\in(1..\infty),\\
&T_{0n} = \int_\alpha^\beta \widehat{T}(\omega,0)\cdot\Psi_n(\omega)\,d\omega, \quad f_n(\tau) = \int_\alpha^\beta \widehat{f}(\omega,\tau)\cdot\Psi_n(\omega)\,d\omega,\\
&\widehat{T}(x,t) = \frac{2}{\beta-\alpha}\cdot\sum_{n=1}^{\infty}\Psi_n(x)\cdot e^{-a^2\cdot\gamma_n^2\cdot t}\cdot\left[T_{0n} + \int_0^t e^{a^2\cdot\gamma_n^2\cdot\tau}\cdot f_n(\tau)\,d\tau\right],\\
&T(x,t) = U(x,t) + \widehat{T}(x,t).
\end{aligned} \quad (2.43)$$

In the Neumann problem, unlike the Dirichlet one, the spectrum contains the zero mode ($\gamma_0 = 0$, $\Psi_0(x) = 1$) — the case $\gamma_n = 0$, which in the general solution (2.18) was set aside for separate consideration; it is precisely what gives the first (non-decaying) term. The final solution of the one-dimensional Neumann boundary value problem with inhomogeneous boundary conditions takes the form

$$
\begin{aligned}
&a_1(t) = \frac{\Phi_\alpha(t)\cdot\beta - \Phi_\beta(t)\cdot\alpha}{\beta-\alpha}, \quad a_2(t) = \frac{\Phi_\beta(t) - \Phi_\alpha(t)}{2\cdot(\beta-\alpha)},\\
&U(x,t) = a_1(t)\cdot x + a_2(t)\cdot x^2,\\
&\widehat{T}(x,0) = T_0(x) - U(x,0),\\
&\widehat{f}(x,t) = f(x,t) - \frac{da_1(t)}{dt}\cdot x - \frac{da_2(t)}{dt}\cdot x^2 + 2\cdot a^2\cdot a_2(t),\\
&\gamma_n = \frac{\pi\cdot n}{\beta-\alpha}, \quad \Psi_n(x) = \cos\left[\gamma_n\cdot(x-\alpha)\right], \quad n\in(0..\infty),\\
&\|\Psi_n(x)\|^2 = \frac{\beta-\alpha}{2} \quad (n\ge 1), \qquad \|\Psi_0(x)\|^2 = \beta-\alpha,\\
&T_{0n} = \int_\alpha^\beta \widehat{T}(\omega,0)\cdot\Psi_n(\omega)\,d\omega, \quad f_n(\tau) = \int_\alpha^\beta \widehat{f}(\omega,\tau)\cdot\Psi_n(\omega)\,d\omega,\\
&\widehat{T}(x,t) = \frac{1}{\beta-\alpha}\cdot\left[T_{00} + \int_0^t f_0(\tau)\,d\tau\right] + \frac{2}{\beta-\alpha}\cdot\sum_{n=1}^\infty \Psi_n(x)\cdot e^{-a^2\cdot\gamma_n^2\cdot t}\cdot\left[T_{0n} + \int_0^t e^{a^2\cdot\gamma_n^2\cdot\tau}\cdot f_n(\tau)\,d\tau\right],\\
&T(x,t) = U(x,t) + \widehat{T}(x,t).
\end{aligned}
\tag{2.44}
$$

To obtain the stationary solutions, we let time tend to infinity in the solutions found ($t\to\infty$), as was done for the general case in (2.18). The term with the initial condition vanishes, since $e^{-a^2\cdot\gamma_n^2\cdot t}\to 0$; the sources and boundary conditions stop depending on time, and the time integral for $\gamma_n\neq 0$ gives the factor $\frac{1}{a^2\cdot\gamma_n^2}$:

$$
\begin{aligned}
\int_0^t e^{-a^2\cdot\gamma_n^2\cdot(t-\tau)}\,d\tau &= \frac{1}{a^2\cdot\gamma_n^2}\cdot e^{-a^2\cdot\gamma_n^2\cdot(t-\tau)}\Big|_0^t\\
&= \frac{1}{a^2\cdot\gamma_n^2}\cdot(1-e^{-a^2\cdot\gamma_n^2\cdot t}) \to \frac{1}{a^2\cdot\gamma_n^2} \quad \text{as} \quad t\to\infty, \quad \gamma_n \neq 0.
\end{aligned}
$$

The stationary solution of the one-dimensional Dirichlet boundary value problem takes the form

$$
\begin{aligned}
&a_1 = \frac{\Phi_\alpha\cdot\beta - \Phi_\beta\cdot\alpha}{\beta-\alpha}, \quad a_2 = \frac{\Phi_\beta - \Phi_\alpha}{\beta-\alpha}, \quad U(x) = a_1 + a_2\cdot x,\\
&\gamma_n = \frac{\pi\cdot n}{\beta-\alpha}, \quad \Psi_n(x) = \sin\left[\gamma_n\cdot(x-\alpha)\right], \quad n\in(1..\infty),\\
&f_n = \int_\alpha^\beta f(\omega)\cdot\Psi_n(\omega)\,d\omega,\\
&T(x) = U(x) + \frac{2}{\beta-\alpha}\cdot\sum_{n=1}^\infty \frac{\Psi_n(x)}{a^2\cdot\gamma_n^2}\cdot f_n.
\end{aligned}
\tag{2.45}
$$

The stationary solution of the one-dimensional Neumann boundary value problem is not so simple. Unlike the Dirichlet problem, the purely Neumann stationary problem is not always solvable. The reduced function $\widehat{T}$ satisfies the homogeneous Neumann conditions $\widehat{T}'(\alpha) = \widehat{T}'(\beta) = 0$ and the stationary equation $a^2\cdot\widehat{T}''(x) + \widehat{f}(x) = 0$. Let us integrate it over the interval $[\alpha,\beta]$:

$$
\begin{aligned}
0 = \int_\alpha^\beta \left(a^2\cdot\widehat{T}''(x) + \widehat{f}(x)\right)dx &= a^2\cdot\left[\widehat{T}'(\beta) - \widehat{T}'(\alpha)\right]\\
&\quad + \int_\alpha^\beta \widehat{f}(x)\,dx = \int_\alpha^\beta \widehat{f}(x)\,dx.
\end{aligned}
$$

The boundary term vanished due to the homogeneous Neumann conditions, and what remains is the solvability condition (2.22): $\int_\alpha^\beta \widehat{f}(\omega)\,d\omega = 0$ — the total reduced source must vanish (otherwise heat accumulates, the mean grows without bound and there is no steady state). When this condition holds, the solution is defined only up to an arbitrary additive constant $C$: it corresponds to the zero mode ($\gamma_0 = 0$, $\Psi_0(x) = 1$), whose amplitude is not fixed by the stationary equation. With this caveat the solution takes the form

$$\begin{aligned}
&a_1 = \frac{\Phi_\alpha \cdot \beta - \Phi_\beta \cdot \alpha}{\beta - \alpha}, \quad a_2 = \frac{\Phi_\beta - \Phi_\alpha}{2 \cdot (\beta - \alpha)}, \quad U(x) = a_1 \cdot x + a_2 \cdot x^2, \\
&\widehat{f}(x) = f(x) + 2 \cdot a^2 \cdot a_2, \\
&\gamma_n = \frac{\pi \cdot n}{\beta - \alpha}, \quad \Psi_n(x) = \cos\left[\gamma_n \cdot (x - \alpha)\right], \quad n \in (1..\infty), \\
&f_n = \int_\alpha^\beta \widehat{f}(\omega) \cdot \Psi_n(\omega)\,d\omega, \\
&T(x) = U(x) + \frac{2}{\beta - \alpha} \cdot \sum_{n=1}^{\infty} \frac{\Psi_n(x)}{a^2 \cdot \gamma_n^2} \cdot f_n + C, \quad C = \text{const}.
\end{aligned} \tag{2.46}$$

## Two-dimensional heat conduction boundary value problem

Let us consider the two-dimensional nonsymmetric heat conduction problem in a disk of radius $R$. The heat conduction equation (1.8) in polar coordinates takes the form

$$\frac{\partial T(r,\theta,t)}{\partial t} = a^2 \cdot \frac{\partial^2 T(r,\theta,t)}{\partial r^2} + a^2 \cdot \frac{1}{r} \cdot \frac{\partial T(r,\theta,t)}{\partial r} + a^2 \cdot \frac{1}{r^2} \cdot \frac{\partial^2 T(r,\theta,t)}{\partial \theta^2} + f(r,\theta,t). \tag{2.47}$$

The Dirichlet boundary conditions (1.13) in the two-dimensional case have the form

$$\begin{cases} T(r,\theta,0) = T_0(r,\theta), \\ T(R,\theta,t) = \Phi(\theta,t), \\ T(r,\theta,t) = T(r,\theta+2\pi,t), \end{cases} \tag{2.48}$$

where $T_0(r,\theta)$ is the initial condition, $\Phi(\theta,t)$ is the boundary condition, which is, of course, periodic in $\theta$.

To get rid of the inhomogeneity, we introduce $T(r,\theta,t) = \widehat{T}(r,\theta,t) + U(\theta,t)$, let us choose the function $U(\theta,t) = \Phi(\theta,t)$, in accordance with (2.2). Then equation (2.47) and the boundary conditions (2.48) take the form

$$\begin{cases} \dfrac{\partial \widehat{T}(r,\theta,t)}{\partial t} = a^2 \cdot \dfrac{\partial^2 \widehat{T}(r,\theta,t)}{\partial r^2} + a^2 \cdot \dfrac{1}{r} \cdot \dfrac{\partial \widehat{T}(r,\theta,t)}{\partial r} + a^2 \cdot \dfrac{1}{r^2} \cdot \dfrac{\partial^2 \widehat{T}(r,\theta,t)}{\partial \theta^2} + \widehat{f}(r,\theta,t), \\ \widehat{T}(r,\theta,0) = T_0(r,\theta) - U(\theta,0), \\ \widehat{T}(R,\theta,t) = 0, \\ \widehat{T}(r,\theta,t) = \widehat{T}(r,\theta+2\pi,t), \end{cases} \tag{2.49}$$

where $\widehat{f}(r,\theta,t) = f(r,\theta,t) - \dfrac{\partial U(\theta,t)}{\partial t} + a^2 \cdot \dfrac{1}{r^2} \cdot \dfrac{\partial^2 U(\theta,t)}{\partial \theta^2}$.

The Neumann boundary conditions (1.14) in the two-dimensional case have the form

$$\begin{cases} T(r,\theta,0) = T_0(r,\theta), \\ \dfrac{\partial T(r,\theta,t)}{\partial \mathbf{n}}\Big|_{r=R} = \Phi(\theta,t), \\ T(r,\theta,t) = T(r,\theta+2\pi,t), \end{cases} \tag{2.50}$$

where $\dfrac{\partial T(r,\theta,t)}{\partial \mathbf{n}}\Big|_{r=R}$ is the normal derivative of the temperature on the circle of radius $R$; the direction of the normal coincides with the direction of the coordinate $r$.

To get rid of the inhomogeneity, we introduce $T(r,\theta,t) = \widehat{T}(r,\theta,t) + U(r,\theta,t)$, let us choose the function $U(r,\theta,t) = \Phi(\theta,t) \cdot r$ — it satisfies condition (2.3): $\dfrac{\partial U(r,\theta,t)}{\partial \mathbf{n}}\Big|_{r=R} = \Phi(\theta,t)$. Then equation (2.47) and the boundary conditions (2.50) take the form

$$\begin{cases} \dfrac{\partial \widehat{T}(r,\theta,t)}{\partial t} = a^2 \cdot \dfrac{\partial^2 \widehat{T}(r,\theta,t)}{\partial r^2} + a^2 \cdot \dfrac{1}{r} \cdot \dfrac{\partial \widehat{T}(r,\theta,t)}{\partial r} + a^2 \cdot \dfrac{1}{r^2} \cdot \dfrac{\partial^2 \widehat{T}(r,\theta,t)}{\partial \theta^2} + \widehat{f}(r,\theta,t), \\ \widehat{T}(r,\theta,0) = T_0(r,\theta) - U(r,\theta,0), \\ \dfrac{\partial \widehat{T}(r,\theta,t)}{\partial \mathbf{n}}\Big|_{r=R} = 0, \\ \widehat{T}(r,\theta,t) = \widehat{T}(r,\theta+2\pi,t), \end{cases} \tag{2.51}$$

where $\widehat{f}(r,\theta,t) = f(r,\theta,t) - \dfrac{\partial U(r,\theta,t)}{\partial t} + a^2 \cdot \dfrac{1}{r} \cdot \dfrac{\partial U(r,\theta,t)}{\partial r} + a^2 \cdot \dfrac{1}{r^2} \cdot \dfrac{\partial^2 U(r,\theta,t)}{\partial \theta^2}$.

Equation (2.8) takes the form

$$\frac{\partial^2 \Psi(r,\theta)}{\partial r^2} + \frac{1}{r} \cdot \frac{\partial \Psi(r,\theta)}{\partial r} + \frac{1}{r^2} \cdot \frac{\partial^2 \Psi(r,\theta)}{\partial \theta^2} + \gamma^2 \cdot \Psi(r,\theta) = 0. \tag{2.52}$$

Let us represent the function $\Psi(r,\theta)$ as a product of two functions $\Psi(r,\theta) = \Lambda(r) \cdot \Upsilon(\theta)$, substitute it into equation (2.52) and carry out the transformations

$$\frac{d^2\Lambda(r)}{dr^2} \cdot \Upsilon(\theta) + \frac{1}{r} \cdot \frac{d\Lambda(r)}{dr} \cdot \Upsilon(\theta) + \frac{1}{r^2} \cdot \frac{d^2\Upsilon(\theta)}{d\theta^2} \cdot \Lambda(r) + \gamma^2 \cdot \Lambda(r) \cdot \Upsilon(\theta) = 0,$$

$$r^2 \cdot \frac{d^2\Lambda(r)}{dr^2} \cdot \Upsilon(\theta) + r \cdot \frac{d\Lambda(r)}{dr} \cdot \Upsilon(\theta) + \frac{d^2\Upsilon(\theta)}{d\theta^2} \cdot \Lambda(r) + \gamma^2 \cdot r^2 \cdot \Lambda(r) \cdot \Upsilon(\theta) = 0,$$

$$\left[r^2 \cdot \frac{d^2\Lambda(r)}{dr^2} + r \cdot \frac{d\Lambda(r)}{dr} + \gamma^2 \cdot r^2 \cdot \Lambda(r)\right] \cdot \Upsilon(\theta) = -\frac{d^2\Upsilon(\theta)}{d\theta^2} \cdot \Lambda(r),$$

$$\frac{1}{\Lambda(r)} \cdot \left[r^2 \cdot \frac{d^2\Lambda(r)}{dr^2} + r \cdot \frac{d\Lambda(r)}{dr} + \gamma^2 \cdot r^2 \cdot \Lambda(r)\right] = m^2 = -\frac{1}{\Upsilon(\theta)} \cdot \frac{d^2\Upsilon(\theta)}{d\theta^2}.$$

We introduced the constant $m^2$ by analogy with the separation of variables for time and geometry. Now let us rewrite it as a system of two equations: one in the angle $\theta$, the other in the radius $r$.

$$\begin{cases} \frac{d^2\Upsilon(\theta)}{d\theta^2} + m^2 \cdot \Upsilon(\theta) = 0, \\ \frac{d^2\Lambda(r)}{dr^2} + \frac{1}{r} \cdot \frac{d\Lambda(r)}{dr} + \left(\gamma^2 - \frac{m^2}{r^2}\right) \cdot \Lambda(r) = 0. \end{cases} \tag{2.53}$$

The solution of the equation for the angle $\theta$ can be represented by a sine or a cosine, since both of these functions are periodic and satisfy the boundary condition $\Upsilon(\theta) = \Upsilon(\theta + 2\pi)$, which imposes a restriction on $m$ — it must be an integer. We get two families of solutions; later we will take their superposition.

$$\begin{cases} \Upsilon_{1m}(\theta) = C_{1m} \cdot \sin(m \cdot \theta), m \in (1..\infty), \\ \Upsilon_{2m}(\theta) = C_{2m} \cdot \cos(m \cdot \theta), m \in (0..\infty). \end{cases} \tag{2.54}$$

Let us take the coefficients $C_{1m}$ and $C_{2m}$ in front of the sine and cosine equal to one — these factors are absorbed into the expansion coefficients anyway. The solution of the equation for the radius $r$ has the form

$$\Lambda_m(r) = A_{1m} \cdot J_m(\gamma \cdot r) + A_{2m} \cdot Y_m(\gamma \cdot r), \tag{2.55}$$

where $J_m(\gamma \cdot r)$ and $Y_m(\gamma \cdot r)$ are the Bessel functions of the first and second kind respectively, $A_{1m}$ and $A_{2m}$ are coefficients determined from the boundary conditions, and the weight ensuring the orthogonality of the eigenfunctions is $\rho(r, \theta) = r$. Moreover, for the function $\Lambda_m(r)$ to be bounded as $r \to 0$, it is necessary that $A_{2m} = 0$: the Bessel function of the second kind is infinite at zero, as seen in figure (B.2).

Thus, the solutions of equation (2.52) have the form

$$\begin{cases} \Psi_{1m}(r, \theta) = \Lambda_m(r) \cdot \Upsilon_{1m}(\theta) = A_{1m} \cdot J_m(\gamma \cdot r) \cdot \sin(m \cdot \theta), m \in (1..\infty), \\ \Psi_{2m}(r, \theta) = \Lambda_m(r) \cdot \Upsilon_{2m}(\theta) = A_{1m} \cdot J_m(\gamma \cdot r) \cdot \cos(m \cdot \theta), m \in (0..\infty). \end{cases} \tag{2.56}$$

Let us substitute the solutions (2.56) into the boundary conditions (2.49) and obtain

$$\begin{cases} A_{1m} \cdot J_m(\gamma \cdot R) \cdot \sin(m \cdot \theta) = 0, m \in (1..\infty), \\ A_{1m} \cdot J_m(\gamma \cdot R) \cdot \cos(m \cdot \theta) = 0, m \in (0..\infty). \end{cases}$$

Clearly, the solution makes sense only when $A_{1m} \neq 0$, which is possible only when $J_m(\gamma \cdot R) = 0$. Let us denote the roots of the equation $J_m(\mu) = 0$ by $\mu_{mk}$ — then the eigenvalues take the form

$$J_m(\gamma \cdot R) = 0 \;\Rightarrow\; \gamma_{mk} = \frac{\mu_{mk}}{R}, \quad k \in (1..\infty). \tag{2.57}$$

Zero and negative roots are no good: for $\mu = 0$ we have $\gamma = 0$, and for $m \geq 1$ the eigenfunction $J_m(0 \cdot r) \equiv 0$ — identically zero, which is not an eigenfunction, while for $m = 0$ zero is not a root at all: $J_0(0) = 1$. Negative roots give no new solutions: $J_m(-\mu) = (-1)^m \cdot J_m(\mu)$ — the eigenfunction is the same up to sign. Therefore we take only the positive roots: $\mu_{mk}, k \in (1..\infty)$.

Thus, the eigenvalues and eigenfunctions have the form

$$\begin{cases} \gamma_{mk} = \frac{\mu_{mk}}{R}, k \in (1..\infty), \\ \Psi_{1mk}(r,\theta) = A_{1mk} \cdot J_m(\gamma_{mk} \cdot r) \cdot \sin(m \cdot \theta), m \in (1..\infty), \\ \Psi_{2mk}(r,\theta) = A_{1mk} \cdot J_m(\gamma_{mk} \cdot r) \cdot \cos(m \cdot \theta), m \in (0..\infty), \end{cases} \tag{2.58}$$

where $A_{1mk}$ is an arbitrary constant factor: the eigenfunction is defined up to it. Let us set $A_{1mk} = 1$ — this factor is absorbed into the expansion coefficients anyway, as in the general solution.

To expand functions into a Fourier series in $\Psi_{1mk}(r,\theta)$ and $\Psi_{2mk}(r,\theta)$ we need to compute the norms $\|\Psi_{1mk}(r,\theta)\|^2$ and $\|\Psi_{2mk}(r,\theta)\|^2$; the weight for polar coordinates is $\rho(r,\theta) = r$.

$$\begin{cases} \|\Psi_{1mk}(r,\theta)\|^2 = \int_0^R r \cdot J_m^2(\gamma_{mk} \cdot r)\, dr \int_0^{2\pi} \sin^2(m \cdot \theta)\, d\theta, m \in (1..\infty), k \in (1..\infty), \\ \|\Psi_{2mk}(r,\theta)\|^2 = \int_0^R r \cdot J_m^2(\gamma_{mk} \cdot r)\, dr \int_0^{2\pi} \cos^2(m \cdot \theta)\, d\theta, m \in (0..\infty), k \in (1..\infty). \end{cases}$$

$$\begin{cases} \int_0^{2\pi} \sin^2(m \cdot \theta)\, d\theta = \pi, m \in (1..\infty), \\ \int_0^{2\pi} \cos^2(m \cdot \theta)\, d\theta = \begin{cases} 2 \cdot \pi, & m = 0, \\ \pi, & m \geq 1 \end{cases}, m \in (0..\infty). \end{cases} \tag{2.59}$$

Let us make the substitution $x = \gamma_{mk} \cdot r = \frac{\mu_{mk}}{R} \cdot r$.

$$\int_0^R r \cdot J_m^2(\gamma_{mk} \cdot r)\, dr = \int_0^R r \cdot J_m^2\left(\frac{\mu_{mk}}{R} \cdot r\right) dr = \frac{R^2}{\mu_{mk}^2} \cdot \int_0^{\mu_{mk}} x \cdot J_m^2(x)\, dx.$$

We use formula (C.6) to compute the norm

$$\int_0^{\mu_{mk}} x \cdot J_m^2(x)\, dx = \frac{\mu_{mk}^2}{2} \cdot J_{m+1}^2(\mu_{mk}).$$

Finally we obtain

$$\int_0^R r \cdot J_m^2(\gamma_{mk} \cdot r)\, dr = \frac{R^2}{\mu_{mk}^2} \cdot \frac{\mu_{mk}^2}{2} \cdot J_{m+1}^2(\mu_{mk}) = \frac{R^2}{2} \cdot J_{m+1}^2(\mu_{mk}). \quad (2.60)$$

The norms $\|\Psi_{1mk}(r,\theta)\|^2$ and $\|\Psi_{2mk}(r,\theta)\|^2$ have the form

$$\begin{cases} \|\Psi_{1mk}(r,\theta)\|^2 = \dfrac{\pi \cdot R^2}{2} \cdot J_{m+1}^2(\mu_{mk}), \\ \|\Psi_{2mk}(r,\theta)\|^2 = \begin{cases} \pi \cdot R^2 \cdot J_{m+1}^2(\mu_{mk}), & m = 0, \\ \dfrac{\pi \cdot R^2}{2} \cdot J_{m+1}^2(\mu_{mk}), & m \geq 1 \end{cases} \end{cases}. \quad (2.61)$$

Now let us substitute the solution (2.56) into the boundary conditions (2.51) and obtain

$$\begin{cases} A_{1m} \cdot \gamma \cdot J_m'(\gamma \cdot R) \cdot \sin(m \cdot \theta) = 0, m \in (1..\infty), \\ A_{1m} \cdot \gamma \cdot J_m'(\gamma \cdot R) \cdot \cos(m \cdot \theta) = 0, m \in (0..\infty). \end{cases}$$

Clearly, the solution makes sense only when $A_{1m} \neq 0$, which is possible only when $J_m'(\gamma \cdot R) = 0$. Let us denote the roots of the equation $J_m'(\mu) = 0$ by $\mu_{mk}$ — then the eigenvalues take the form

$$J_m'(\gamma \cdot R) = 0 \;\Rightarrow\; \gamma_{mk} = \frac{\mu_{mk}}{R}, \quad k \in (1..\infty). \quad (2.62)$$

Here, unlike the Dirichlet problem, the zero mode survives: for $m = 0$ zero is a root of the equation $J_0'(\mu) = 0$, and it corresponds to $\gamma = 0$ and the eigenfunction $\Psi(r,\theta) = J_0(0) = 1$ — a nonzero constant, that is, a fully fledged eigenfunction (the constant mode). We will count it as the first root: $\mu_{01} = 0$. For $m \geq 1$ the zero root gives no eigenfunction ($J_m(0 \cdot r) \equiv 0$), so there we still take only the positive roots.

Thus, the eigenvalues and eigenfunctions have the form

$$\begin{cases} \gamma_{mk} = \dfrac{\mu_{mk}}{R}, k \in (1..\infty), \quad \mu_{01} = 0, \\ \Psi_{1mk}(r,\theta) = A_{1mk} \cdot J_m(\gamma_{mk} \cdot r) \cdot \sin(m \cdot \theta), m \in (1..\infty), \\ \Psi_{2mk}(r,\theta) = A_{1mk} \cdot J_m(\gamma_{mk} \cdot r) \cdot \cos(m \cdot \theta), m \in (0..\infty), \end{cases} \quad (2.63)$$

where $A_{1mk}$ is an arbitrary constant factor; as before, we set it equal to one.

We compute the norms of the eigenfunctions in the same way as in the Dirichlet problem: the integral over the angle remains the same (2.59), while the integral over the radius takes a different value — we use formula (D.2) to compute the norm

$$\int_0^{\mu_{mk}} x \cdot J_m^2(x)\, dx = \left(\mu_{mk}^2 - m^2\right) \cdot \frac{J_m^2(\mu_{mk})}{2}.$$

Finally we obtain

$$\int_0^R r \cdot J_m^2(\gamma_{mk} \cdot r)\, dr = \frac{R^2}{\mu_{mk}^2} \cdot \left(\mu_{mk}^2 - m^2\right) \cdot \frac{J_m^2(\mu_{mk})}{2} = R^2 \cdot \left(\mu_{mk}^2 - m^2\right) \cdot \frac{J_m^2(\mu_{mk})}{2 \cdot \mu_{mk}^2}. \quad (2.64)$$

The special case $\mu_{01} = 0$ (for $\gamma_{01} = 0$ the eigenfunction $\Psi_{201}(r, \theta) = 1$) we compute separately

$$\|\Psi_{201}(r, \theta)\|^2 = \int_0^R \int_0^{2\pi} 1 \cdot r \, dr \, d\theta = \pi \cdot R^2. \tag{2.65}$$

The norms $\|\Psi_{1mk}(r, \theta)\|^2$ and $\|\Psi_{2mk}(r, \theta)\|^2$ have the form

$$\begin{cases} \|\Psi_{1mk}(r, \theta)\|^2 = \dfrac{\pi \cdot R^2 \cdot (\mu_{mk}^2 - m^2) \cdot J_m^2(\mu_{mk})}{2 \cdot \mu_{mk}^2}, \\ \|\Psi_{2mk}(r, \theta)\|^2 = \begin{cases} \pi \cdot R^2 \cdot J_m^2(\mu_{mk}), & m = 0, \\ \dfrac{\pi \cdot R^2 \cdot (\mu_{mk}^2 - m^2) \cdot J_m^2(\mu_{mk})}{2 \cdot \mu_{mk}^2}, & m \geq 1 \end{cases} . \end{cases} \tag{2.66}$$

The final solution of the two-dimensional Dirichlet boundary value problem with inhomogeneous boundary conditions takes the form

$$
\begin{aligned}
&U(\theta,t)=\Phi(\theta,t),\\
&\widehat{T}(r,\theta,0)=T_0(r,\theta)-U(\theta,0),\\
&\widehat{f}(r,\theta,t)=f(r,\theta,t)-\frac{\partial\Phi(\theta,t)}{\partial t}+\frac{a^2}{r^2}\cdot\frac{\partial^2\Phi(\theta,t)}{\partial\theta^2},\\
&\gamma_{mk}=\frac{\mu_{mk}}{R},\quad J_m(\mu_{mk})=0,\quad m\in(0..\infty),\quad k\in(1..\infty),\\
&\Psi_{1mk}(r,\theta)=J_m(\gamma_{mk}\cdot r)\cdot\sin(m\cdot\theta),\quad m\in(1..\infty),\quad k\in(1..\infty),\\
&\Psi_{2mk}(r,\theta)=J_m(\gamma_{mk}\cdot r)\cdot\cos(m\cdot\theta),\quad m\in(0..\infty),\quad k\in(1..\infty),\\
&\|\Psi_{1mk}(r,\theta)\|^2=\frac{\pi\cdot R^2}{2}\cdot J_{m+1}^2(\mu_{mk}),\\
&\|\Psi_{2mk}(r,\theta)\|^2=\begin{cases}\pi\cdot R^2\cdot J_{m+1}^2(\mu_{mk}), & m=0,\\ \dfrac{\pi\cdot R^2}{2}\cdot J_{m+1}^2(\mu_{mk}), & m\geq 1\end{cases},\\
&T_{01mk}=\int_0^R\int_0^{2\pi}\widehat{T}(r,\theta,0)\cdot\Psi_{1mk}(r,\theta)\cdot r\,dr\,d\theta,\\
&T_{02mk}=\int_0^R\int_0^{2\pi}\widehat{T}(r,\theta,0)\cdot\Psi_{2mk}(r,\theta)\cdot r\,dr\,d\theta,\\
&f_{1mk}(t)=\int_0^R\int_0^{2\pi}\widehat{f}(r,\theta,t)\cdot\Psi_{1mk}(r,\theta)\cdot r\,dr\,d\theta,\\
&f_{2mk}(t)=\int_0^R\int_0^{2\pi}\widehat{f}(r,\theta,t)\cdot\Psi_{2mk}(r,\theta)\cdot r\,dr\,d\theta,\\
&\widehat{T}(r,\theta,t)=\\
&\sum_{m=1}^{\infty}\sum_{k=1}^{\infty}\frac{\Psi_{1mk}(r,\theta)}{\|\Psi_{1mk}(r,\theta)\|^2}\cdot e^{-a^2\cdot\gamma_{mk}^2\cdot t}\cdot\left[T_{01mk}+\int_0^t e^{a^2\cdot\gamma_{mk}^2\cdot\tau}\cdot f_{1mk}(\tau)\,d\tau\right]+\\
&\sum_{m=0}^{\infty}\sum_{k=1}^{\infty}\frac{\Psi_{2mk}(r,\theta)}{\|\Psi_{2mk}(r,\theta)\|^2}\cdot e^{-a^2\cdot\gamma_{mk}^2\cdot t}\cdot\left[T_{02mk}+\int_0^t e^{a^2\cdot\gamma_{mk}^2\cdot\tau}\cdot f_{2mk}(\tau)\,d\tau\right],\\
&T(r,\theta,t)=U(\theta,t)+\widehat{T}(r,\theta,t).
\end{aligned}
\tag{2.67}
$$

In the Neumann problem, unlike the Dirichlet one, the spectrum contains the zero mode ($\gamma_{01}=0$, $\Psi_{201}(r,\theta)=1$) — the case $\gamma=0$, which in the general solution (2.18) was set aside for separate consideration; it is precisely what gives the non-decaying term — the term with $m=0, k=1$ in the second sum. The final solution of the two-dimensional Neumann boundary value problem with inhomogeneous boundary conditions takes the form

$$
\begin{aligned}
&U(r,\theta,t) = \Phi(\theta,t)\cdot r,\\
&\widehat{T}(r,\theta,0) = T_0(r,\theta) - U(r,\theta,0),\\
&\widehat{f}(r,\theta,t) = f(r,\theta,t) - r\cdot\frac{\partial\Phi(\theta,t)}{\partial t} + \frac{a^2\cdot\Phi(\theta,t)}{r} + \frac{a^2}{r}\cdot\frac{\partial^2\Phi(\theta,t)}{\partial\theta^2},\\
&\gamma_{mk} = \frac{\mu_{mk}}{R},\quad J'_m(\mu_{mk}) = 0,\quad \mu_{01} = 0,\quad m\in(0..\infty),\quad k\in(1..\infty),\\
&\Psi_{1mk}(r,\theta) = J_m(\gamma_{mk}\cdot r)\cdot\sin(m\cdot\theta),\quad m\in(1..\infty),\quad k\in(1..\infty),\\
&\Psi_{2mk}(r,\theta) = J_m(\gamma_{mk}\cdot r)\cdot\cos(m\cdot\theta),\quad m\in(0..\infty),\quad k\in(1..\infty),\\
&\|\Psi_{1mk}(r,\theta)\|^2 = \frac{\pi\cdot R^2\cdot(\mu_{mk}^2 - m^2)\cdot J_m^2(\mu_{mk})}{2\cdot\mu_{mk}^2},\\
&\|\Psi_{2mk}(r,\theta)\|^2 = \begin{cases}\pi\cdot R^2\cdot J_m^2(\mu_{mk}), & m = 0,\\ \dfrac{\pi\cdot R^2\cdot(\mu_{mk}^2 - m^2)\cdot J_m^2(\mu_{mk})}{2\cdot\mu_{mk}^2}, & m\geq 1\end{cases},\\
&T_{01mk} = \int_0^R\int_0^{2\pi}\widehat{T}(r,\theta,0)\cdot\Psi_{1mk}(r,\theta)\cdot r\,dr\,d\theta,\\
&T_{02mk} = \int_0^R\int_0^{2\pi}\widehat{T}(r,\theta,0)\cdot\Psi_{2mk}(r,\theta)\cdot r\,dr\,d\theta,\\
&f_{1mk}(t) = \int_0^R\int_0^{2\pi}\widehat{f}(r,\theta,t)\cdot\Psi_{1mk}(r,\theta)\cdot r\,dr\,d\theta,\\
&f_{2mk}(t) = \int_0^R\int_0^{2\pi}\widehat{f}(r,\theta,t)\cdot\Psi_{2mk}(r,\theta)\cdot r\,dr\,d\theta,\\
&\widehat{T}(r,\theta,t) =\\
&\sum_{m=1}^{\infty}\sum_{k=1}^{\infty}\frac{\Psi_{1mk}(r,\theta)}{\|\Psi_{1mk}(r,\theta)\|^2}\cdot e^{-a^2\cdot\gamma_{mk}^2\cdot t}\cdot\left[T_{01mk} + \int_0^t e^{a^2\cdot\gamma_{mk}^2\cdot\tau}\cdot f_{1mk}(\tau)\,d\tau\right] +\\
&\sum_{m=0}^{\infty}\sum_{k=1}^{\infty}\frac{\Psi_{2mk}(r,\theta)}{\|\Psi_{2mk}(r,\theta)\|^2}\cdot e^{-a^2\cdot\gamma_{mk}^2\cdot t}\cdot\left[T_{02mk} + \int_0^t e^{a^2\cdot\gamma_{mk}^2\cdot\tau}\cdot f_{2mk}(\tau)\,d\tau\right],\\
&T(r,\theta,t) = U(r,\theta,t) + \widehat{T}(r,\theta,t).
\end{aligned}
\tag{2.68}
$$

To obtain the stationary solutions, we let time tend to infinity in the solutions found ($t\to\infty$), as was done for the general case in (2.18). The term with the initial condition vanishes, since $e^{-a^2\cdot\gamma_{mk}^2\cdot t}\to 0$; the sources and boundary conditions stop depending on time, and the time integral for $\gamma_{mk}\neq 0$ gives the factor $\frac{1}{a^2\cdot\gamma_{mk}^2}$:

$$
\begin{aligned}
\int_0^t e^{-a^2\cdot\gamma_{mk}^2\cdot(t-\tau)}\,d\tau &= \frac{1}{a^2\cdot\gamma_{mk}^2}\cdot(1 - e^{-a^2\cdot\gamma_{mk}^2\cdot t})\\
&\to \frac{1}{a^2\cdot\gamma_{mk}^2}\quad\text{as}\quad t\to\infty,\quad \gamma_{mk}\neq 0.
\end{aligned}
$$

The stationary solution of the two-dimensional Dirichlet boundary value problem takes the form

$$
\begin{aligned}
&U(\theta) = \Phi(\theta),\\
&\widehat{f}(r,\theta) = f(r,\theta) + \frac{a^2}{r^2}\cdot\frac{d^2\Phi(\theta)}{d\theta^2},\\
&\gamma_{mk} = \frac{\mu_{mk}}{R}, \quad J_m(\mu_{mk}) = 0, \quad m \in (0..\infty), \quad k \in (1..\infty),\\
&\Psi_{1mk}(r,\theta) = J_m(\gamma_{mk}\cdot r)\cdot\sin(m\cdot\theta), \quad m \in (1..\infty),\\
&\Psi_{2mk}(r,\theta) = J_m(\gamma_{mk}\cdot r)\cdot\cos(m\cdot\theta), \quad m \in (0..\infty),\\
&\|\Psi_{1mk}(r,\theta)\|^2 = \frac{\pi\cdot R^2}{2}\cdot J_{m+1}^2(\mu_{mk}),\\
&\|\Psi_{2mk}(r,\theta)\|^2 = \pi\cdot R^2\cdot J_{m+1}^2(\mu_{mk}), \quad m = 0,\\
&\|\Psi_{2mk}(r,\theta)\|^2 = \frac{\pi\cdot R^2}{2}\cdot J_{m+1}^2(\mu_{mk}), \quad m \geq 1,\\
&f_{1mk} = \int_0^R\int_0^{2\pi}\widehat{f}(r,\theta)\cdot\Psi_{1mk}(r,\theta)\cdot r\,dr\,d\theta,\\
&f_{2mk} = \int_0^R\int_0^{2\pi}\widehat{f}(r,\theta)\cdot\Psi_{2mk}(r,\theta)\cdot r\,dr\,d\theta,\\
&T(r,\theta) = U(\theta) + \sum_{m=1}^{\infty}\sum_{k=1}^{\infty}\frac{\Psi_{1mk}(r,\theta)}{\|\Psi_{1mk}(r,\theta)\|^2}\cdot\frac{f_{1mk}}{a^2\cdot\gamma_{mk}^2} + \sum_{m=0}^{\infty}\sum_{k=1}^{\infty}\frac{\Psi_{2mk}(r,\theta)}{\|\Psi_{2mk}(r,\theta)\|^2}\cdot\frac{f_{2mk}}{a^2\cdot\gamma_{mk}^2}.
\end{aligned}
\tag{2.69}
$$

The stationary solution of the two-dimensional Neumann boundary value problem is not so simple. Unlike the Dirichlet problem, the purely Neumann stationary problem is not always solvable. The reduced function $\widehat{T}(r,\theta)$ satisfies the homogeneous Neumann conditions and the stationary equation $a^2\cdot\Delta\widehat{T}(r,\theta) + \widehat{f}(r,\theta) = 0$. Let us integrate it over the domain:

$$
\begin{aligned}
0 = \int_0^R\int_0^{2\pi}\left(a^2\cdot\Delta\widehat{T} + \widehat{f}\right)r\,dr\,d\theta = a^2\oint_{r=R}\frac{\partial\widehat{T}}{\partial\mathbf{n}}\,dl\\
+\int_0^R\int_0^{2\pi}\widehat{f}\cdot r\,dr\,d\theta = \int_0^R\int_0^{2\pi}\widehat{f}(r,\theta)\cdot r\,dr\,d\theta.
\end{aligned}
$$

The boundary integral vanished due to the homogeneous Neumann conditions, and what remains is the solvability condition (2.22): $\int_0^R\int_0^{2\pi}\widehat{f}(r,\theta)\cdot r\,dr\,d\theta = 0$ — the total reduced source over the domain must vanish (otherwise heat accumulates and there is no steady state). When it holds, the solution is defined only up to an arbitrary additive constant $C$: it corresponds to the zero mode ($\gamma_{01} = 0$, $\Psi_{201}(r,\theta) = 1$), whose amplitude is not fixed by the stationary equation. With this caveat the solution takes the form

$$
\begin{aligned}
&U(r,\theta) = \Phi(\theta)\cdot r,\\
&\widehat{f}(r,\theta) = f(r,\theta) + \frac{a^2\cdot\Phi(\theta)}{r} + \frac{a^2}{r}\cdot\frac{d^2\Phi(\theta)}{d\theta^2},\\
&\gamma_{mk} = \frac{\mu_{mk}}{R},\quad J'_m(\mu_{mk}) = 0,\quad \mu_{mk} > 0,\quad m\in(0..\infty),\quad k\in(1..\infty),\\
&\Psi_{1mk}(r,\theta) = J_m(\gamma_{mk}\cdot r)\cdot\sin(m\cdot\theta),\quad m\in(1..\infty),\\
&\Psi_{2mk}(r,\theta) = J_m(\gamma_{mk}\cdot r)\cdot\cos(m\cdot\theta),\quad m\in(0..\infty),\\
&\|\Psi_{1mk}(r,\theta)\|^2 = \frac{\pi\cdot R^2\cdot(\mu_{mk}^2 - m^2)\cdot J_m^2(\mu_{mk})}{2\cdot\mu_{mk}^2},\\
&\|\Psi_{2mk}(r,\theta)\|^2 = \pi\cdot R^2\cdot J_m^2(\mu_{mk}),\quad m = 0,\\
&\|\Psi_{2mk}(r,\theta)\|^2 = \frac{\pi\cdot R^2\cdot(\mu_{mk}^2 - m^2)\cdot J_m^2(\mu_{mk})}{2\cdot\mu_{mk}^2},\quad m\geq 1,\\
&f_{1mk} = \int_0^R\int_0^{2\pi}\widehat{f}(r,\theta)\cdot\Psi_{1mk}(r,\theta)\cdot r\,dr\,d\theta,\\
&f_{2mk} = \int_0^R\int_0^{2\pi}\widehat{f}(r,\theta)\cdot\Psi_{2mk}(r,\theta)\cdot r\,dr\,d\theta,\\
&T(r,\theta) = U(r,\theta) + \sum_{m=1}^{\infty}\sum_{k=1}^{\infty}\frac{\Psi_{1mk}(r,\theta)}{\|\Psi_{1mk}(r,\theta)\|^2}\cdot\frac{f_{1mk}}{a^2\cdot\gamma_{mk}^2} + \sum_{m=0}^{\infty}\sum_{k=1}^{\infty}\frac{\Psi_{2mk}(r,\theta)}{\|\Psi_{2mk}(r,\theta)\|^2}\cdot\frac{f_{2mk}}{a^2\cdot\gamma_{mk}^2} + C,\quad C = \text{const}.
\end{aligned}
\tag{2.70}
$$

# Three-dimensional heat conduction boundary value problem

Let us consider the three-dimensional nonsymmetric heat conduction problem in a ball of radius $R$. The heat conduction equation (1.8) in spherical coordinates takes the form

$$
\begin{aligned}
\frac{\partial T(r,\theta,\phi,t)}{\partial t} = a^2\cdot\frac{\partial^2 T(r,\theta,\phi,t)}{\partial r^2} + \frac{2\cdot a^2}{r}\cdot\frac{\partial T(r,\theta,\phi,t)}{\partial r}\\
+ \frac{a^2}{r^2\cdot\sin(\theta)}\cdot\frac{\partial}{\partial\theta}\left(\sin(\theta)\cdot\frac{\partial T(r,\theta,\phi,t)}{\partial\theta}\right)\\
+ \frac{a^2}{r^2\cdot\sin^2\theta}\cdot\frac{\partial^2 T(r,\theta,\phi,t)}{\partial\phi^2} + f(r,\theta,\phi,t),
\end{aligned}
\tag{2.71}
$$

where $r\in(0,R)$, $\theta\in(0,\pi)$, $\phi\in(0,2\pi)$, $R$ is the radius, $T(r,\theta,\phi,t)$ is the temperature, $a^2$ is the thermal diffusivity, $f(r,\theta,\phi,t)$ is the heat source density function.

The Dirichlet boundary conditions (1.13) in the three-dimensional case have the form

$$
\begin{cases}
T(r,\theta,\phi,0) = T_0(r,\theta,\phi),\\
T(R,\theta,\phi,t) = \Phi(\theta,\phi,t),\\
|T(r,0,\phi,t)| < \infty,\quad |T(r,\pi,\phi,t)| < \infty,\\
T(r,\theta,\phi,t) = T(r,\theta,\phi+2\pi,t),
\end{cases}
\tag{2.72}
$$

where $T_0(r,\theta,\phi)$ is the initial condition, $\Phi(\theta,\phi,t)$ is the boundary condition, which is, of course, periodic in $\phi$; at the poles ($\theta = 0$ and $\theta = \pi$) the boundary conditions are replaced by the requirement that the solution be bounded.

To get rid of the inhomogeneity, we introduce $T(r,\theta,\phi,t) = \widehat{T}(r,\theta,\phi,t) + U(\theta,\phi,t)$, let us choose the function $U(\theta,\phi,t) = \Phi(\theta,\phi,t)$, in accordance with (2.2). Then equation (2.71) and the boundary conditions (2.72) take the form

$$\begin{cases} \frac{\partial \widehat{T}(r,\theta,\phi,t)}{\partial t} = a^2 \cdot \frac{\partial^2 \widehat{T}(r,\theta,\phi,t)}{\partial r^2} + \frac{2\cdot a^2}{r} \cdot \frac{\partial \widehat{T}(r,\theta,\phi,t)}{\partial r} + \frac{a^2}{r^2 \cdot \sin(\theta)} \cdot \frac{\partial}{\partial \theta}\left(\sin(\theta) \cdot \frac{\partial \widehat{T}(r,\theta,\phi,t)}{\partial \theta}\right) + \frac{a^2}{r^2 \cdot \sin^2\theta} \cdot \frac{\partial^2 \widehat{T}(r,\theta,\phi,t)}{\partial \phi^2} + \widehat{f}(r,\theta,\phi,t), \\ \widehat{T}(r,\theta,\phi,0) = T_0(r,\theta,\phi) - U(\theta,\phi,0), \\ \widehat{T}(R,\theta,\phi,t) = 0, \\ |\widehat{T}(r,0,\phi,t)| < \infty, \quad |\widehat{T}(r,\pi,\phi,t)| < \infty, \\ \widehat{T}(r,\theta,\phi,t) = \widehat{T}(r,\theta,\phi+2\pi,t), \end{cases} \tag{2.73}$$

where

$$\begin{aligned} \widehat{f}(r,\theta,\phi,t) = f(r,\theta,\phi,t) &- \frac{\partial U(\theta,\phi,t)}{\partial t} \\ &+ \frac{a^2}{r^2 \cdot \sin(\theta)} \cdot \frac{\partial}{\partial \theta}\left(\sin(\theta) \cdot \frac{\partial U(\theta,\phi,t)}{\partial \theta}\right) \\ &+ \frac{a^2}{r^2 \cdot \sin^2\theta} \cdot \frac{\partial^2 U(\theta,\phi,t)}{\partial \phi^2}. \end{aligned}$$

The Neumann boundary conditions (1.14) in the three-dimensional case have the form

$$\begin{cases} T(r,\theta,\phi,0) = T_0(r,\theta,\phi), \\ \left.\dfrac{\partial T(r,\theta,\phi,t)}{\partial \mathbf{n}}\right|_{r=R} = \Phi(\theta,\phi,t), \\ |T(r,0,\phi,t)| < \infty, \quad |T(r,\pi,\phi,t)| < \infty, \\ T(r,\theta,\phi,t) = T(r,\theta,\phi+2\pi,t), \end{cases} \tag{2.74}$$

where $\left.\dfrac{\partial T(r,\theta,\phi,t)}{\partial \mathbf{n}}\right|_{r=R}$ is the normal derivative of the temperature on the sphere of radius $R$; the direction of the normal coincides with the direction of the coordinate $r$.

To get rid of the inhomogeneity, we introduce $T(r,\theta,\phi,t) = \widehat{T}(r,\theta,\phi,t) + U(r,\theta,\phi,t)$, let us choose the function $U(r,\theta,\phi,t) = \Phi(\theta,\phi,t)\cdot r$ — it satisfies condition (2.3): $\left.\dfrac{\partial U(r,\theta,\phi,t)}{\partial \mathbf{n}}\right|_{r=R} = \Phi(\theta,\phi,t)$. Then equation (2.71) and the boundary conditions (2.74) take the form

$$\begin{cases} \frac{\partial \widehat{T}(r,\theta,\phi,t)}{\partial t} = a^2 \cdot \frac{\partial^2 \widehat{T}(r,\theta,\phi,t)}{\partial r^2} + \frac{2\cdot a^2}{r} \cdot \frac{\partial \widehat{T}(r,\theta,\phi,t)}{\partial r} + \frac{a^2}{r^2 \cdot \sin(\theta)} \cdot \frac{\partial}{\partial \theta}\left(\sin(\theta) \cdot \frac{\partial \widehat{T}(r,\theta,\phi,t)}{\partial \theta}\right) + \frac{a^2}{r^2 \cdot \sin^2\theta} \cdot \frac{\partial^2 \widehat{T}(r,\theta,\phi,t)}{\partial \phi^2} + \widehat{f}(r,\theta,\phi,t), \\ \widehat{T}(r,\theta,\phi,0) = T_0(r,\theta,\phi) - U(r,\theta,\phi,0), \\ \left.\frac{\partial \widehat{T}(r,\theta,\phi,t)}{\partial \mathbf{n}}\right|_{r=R} = 0, \\ |\widehat{T}(r,0,\phi,t)| < \infty, \quad |\widehat{T}(r,\pi,\phi,t)| < \infty, \\ \widehat{T}(r,\theta,\phi,t) = \widehat{T}(r,\theta,\phi+2\pi,t), \end{cases} \tag{2.75}$$

where

$$\begin{aligned} \widehat{f}(r,\theta,\phi,t) = f(r,\theta,\phi,t) &- \frac{\partial U(r,\theta,\phi,t)}{\partial t} + a^2 \cdot \frac{\partial^2 U(r,\theta,\phi,t)}{\partial r^2} \\ &+ \frac{2\cdot a^2}{r} \cdot \frac{\partial U(r,\theta,\phi,t)}{\partial r} \\ &+ \frac{a^2}{r^2 \cdot \sin(\theta)} \cdot \frac{\partial}{\partial \theta}\left(\sin(\theta) \cdot \frac{\partial U(r,\theta,\phi,t)}{\partial \theta}\right) \\ &+ \frac{a^2}{r^2 \cdot \sin^2\theta} \cdot \frac{\partial^2 U(r,\theta,\phi,t)}{\partial \phi^2}. \end{aligned}$$

Equation (2.8) takes the form

$$\frac{\partial^2\Psi(r,\theta,\phi)}{\partial r^2}+\frac{2}{r}\cdot\frac{\partial\Psi(r,\theta,\phi)}{\partial r}+\frac{1}{r^2\cdot\sin(\theta)}\cdot\frac{\partial}{\partial\theta}\left(\sin(\theta)\cdot\frac{\partial\Psi(r,\theta,\phi)}{\partial\theta}\right)+\frac{1}{r^2\cdot\sin^2\theta}\cdot\frac{\partial^2\Psi(r,\theta,\phi)}{\partial\phi^2}+\gamma^2\cdot\Psi(r,\theta,\phi)=0. \quad (2.76)$$

Let us represent the function $\Psi(r,\theta,\phi)$ as a product of three functions $\Psi(r,\theta,\phi)=R(r)\cdot\Theta(\theta)\cdot\Upsilon(\phi)$, substitute it into equation (2.76) and carry out the transformations

$$\frac{\partial^2(R(r)\cdot\Theta(\theta)\cdot\Upsilon(\phi))}{\partial r^2}+\frac{2}{r}\cdot\frac{\partial(R(r)\cdot\Theta(\theta)\cdot\Upsilon(\phi))}{\partial r}+\frac{1}{r^2\cdot\sin(\theta)}\cdot\frac{\partial}{\partial\theta}\left(\sin(\theta)\cdot\frac{\partial(R(r)\cdot\Theta(\theta)\cdot\Upsilon(\phi))}{\partial\theta}\right)+\frac{1}{r^2\cdot\sin^2\theta}\cdot\frac{\partial^2(R(r)\cdot\Theta(\theta)\cdot\Upsilon(\phi))}{\partial\phi^2}+\gamma^2\cdot R(r)\cdot\Theta(\theta)\cdot\Upsilon(\phi)=0,$$

$$\Theta(\theta)\cdot\Upsilon(\phi)\cdot\left[\frac{d^2R(r)}{dr^2}+\frac{2}{r}\cdot\frac{dR(r)}{dr}\right]+\frac{R(r)}{r^2}\cdot\left[\frac{1}{\sin(\theta)}\cdot\frac{\partial}{\partial\theta}\left(\sin(\theta)\cdot\frac{\partial(\Theta(\theta)\cdot\Upsilon(\phi))}{\partial\theta}\right)+\frac{1}{\sin^2(\theta)}\cdot\frac{\partial^2(\Theta(\theta)\cdot\Upsilon(\phi))}{\partial\phi^2}\right]+\gamma^2\cdot R(r)\cdot\Theta(\theta)\cdot\Upsilon(\phi)=0,$$

$$\Theta(\theta)\cdot\Upsilon(\phi)\cdot\left[r^2\cdot\frac{d^2R(r)}{dr^2}+2\cdot r\cdot\frac{dR(r)}{dr}+r^2\cdot\gamma^2\cdot R(r)\right]+R(r)\cdot\left[\frac{1}{\sin(\theta)}\cdot\frac{\partial}{\partial\theta}\left(\sin(\theta)\cdot\frac{\partial(\Theta(\theta)\cdot\Upsilon(\phi))}{\partial\theta}\right)+\frac{1}{\sin^2(\theta)}\cdot\frac{\partial^2(\Theta(\theta)\cdot\Upsilon(\phi))}{\partial\phi^2}\right]=0,$$

The expression in the first bracket depends only on $r$, the one in the second bracket only on the angles; therefore, dividing by $R(r)\cdot\Theta(\theta)\cdot\Upsilon(\phi)$, both sides can be set equal to a constant $\gamma_1^2$ — by analogy with the separation of variables for time and geometry. The equation for the radius takes the form

$$r^2\cdot\frac{d^2R(r)}{dr^2}+2\cdot r\cdot\frac{dR(r)}{dr}+\left(r^2\cdot\gamma^2-\gamma_1^2\right)\cdot R(r)=0, \quad (2.77)$$

and continue the transformations

$$\frac{1}{\sin(\theta)}\cdot\frac{\partial}{\partial\theta}\left(\sin(\theta)\cdot\frac{\partial(\Theta(\theta)\cdot\Upsilon(\phi))}{\partial\theta}\right)+\frac{1}{\sin^2(\theta)}\cdot\frac{\partial^2(\Theta(\theta)\cdot\Upsilon(\phi))}{\partial\phi^2}+\gamma_1^2\cdot\Theta(\theta)\cdot\Upsilon(\phi)=0,$$

$$\Upsilon(\phi)\cdot\left[\frac{1}{\sin(\theta)}\cdot\frac{\partial}{\partial\theta}\left(\sin(\theta)\cdot\frac{\partial\Theta(\theta)}{\partial\theta}\right)+\gamma_1^2\cdot\Theta(\theta)\right]+\Theta(\theta)\cdot\frac{1}{\sin^2(\theta)}\cdot\frac{\partial^2(\Upsilon(\phi))}{\partial\phi^2}=0.$$

Here the expression in brackets depends only on $\theta$, while the last term depends only on $\phi$; therefore, multiplying by $\frac{\sin^2(\theta)}{\Theta(\theta)\cdot\Upsilon(\phi)}$, we separate the variables again with the constant $m^2$ and obtain the equations for the polar and azimuthal angles

$$\frac{1}{\sin(\theta)}\cdot\frac{d}{d\theta}\left(\sin(\theta)\cdot\frac{d\Theta(\theta)}{d\theta}\right)+\left[\gamma_1^2-\frac{m^2}{\sin^2(\theta)}\right]\cdot\Theta(\theta) = 0, \quad (2.78)$$

$$\frac{d^2\Upsilon(\phi)}{d\phi^2}+m^2\cdot\Upsilon(\phi)=0. \quad (2.79)$$

The solution of equation (2.79) we already found in the two-dimensional case, so we refer to (2.54). And to solve equation (2.78) let us introduce the substitution $z=\cos(\theta)$: for $\theta\in(0,\pi)$ we have $z\in(-1,1)$ and $dz=-\sin(\theta)\cdot d\theta$. Then the equation takes the form

$$\frac{d}{dz}\left((1-z^2)\cdot\frac{d\Theta(z)}{dz}\right)+\left[\gamma_1^2-\frac{m^2}{1-z^2}\right]\cdot\Theta(z)=0, \quad -1<z<1. \quad (2.80)$$

This equation is analyzed in the appendix "The Legendre equation"; its solution has the form (E.6)

$$\Theta_{km}(z)=P_k^{(m)}(z)=(1-z^2)^{\frac{m}{2}}\cdot\frac{d^m}{dz^m}P_k(z),$$

where $\gamma_{1k}^2=k\cdot(k+1)$ are the eigenvalues, $P_k^{(m)}(z)$ are the associated Legendre polynomials, and nontrivial solutions exist only for $m\le k$. The norm of the associated Legendre polynomial has the form (F.6). Now let us return to the angle $\theta$, making the inverse substitution

$$\Theta_{km}(\theta)=\sin^m(\theta)\cdot\frac{d^m}{d(\cos(\theta))^m}P_k(\cos(\theta)). \quad (2.81)$$

Generally speaking, this solution is defined up to a constant factor — let us set it equal to one: it is absorbed into the expansion coefficients anyway. Now let us return to equation (2.77) and rewrite it taking into account $\gamma_1^2=k\cdot(k+1)$

$$r^2\cdot\frac{d^2R(r)}{dr^2}+2\cdot r\cdot\frac{dR(r)}{dr}+\left(r^2\cdot\gamma^2-k\cdot(k+1)\right)\cdot R(r)=0.$$

This equation is analyzed in the appendix "The Bessel equation in spherical coordinates": the substitution $R(r)=\frac{\widehat{R}(r)}{\sqrt{r}}$ reduces it to the Bessel equation with half-integer index, and the solution has the form (G.2). It should be noted that the solution is

bounded at the point $r = 0$: the Bessel function $J_{k+1/2}(\gamma \cdot r)$ tends to zero no slower than $\sqrt{r}$.

Thus, the solutions of equation (2.76) have the form

$$\begin{cases} \Psi_{1km}(r,\theta,\phi) = A_{km} \cdot \dfrac{1}{\sqrt{r}} \cdot J_{k+1/2}(\gamma \cdot r) \cdot P_k^{(m)}(\cos(\theta)) \cdot \sin(m \cdot \phi), \\ 1 \le m \le k, \quad k \ge 1, \\ \Psi_{2km}(r,\theta,\phi) = A_{km} \cdot \dfrac{1}{\sqrt{r}} \cdot J_{k+1/2}(\gamma \cdot r) \cdot P_k^{(m)}(\cos(\theta)) \cdot \cos(m \cdot \phi), \\ 0 \le m \le k, \quad k \ge 0, \end{cases} \tag{2.82}$$

Let us substitute the solutions (2.82) into the boundary conditions (2.73) and obtain

$$\begin{cases} A_{km} \cdot \dfrac{1}{\sqrt{R}} \cdot J_{k+1/2}(\gamma \cdot R) \cdot P_k^{(m)}(\cos(\theta)) \cdot \sin(m \cdot \phi) = 0, \quad 1 \le m \le k, \quad k \ge 1, \\ A_{km} \cdot \dfrac{1}{\sqrt{R}} \cdot J_{k+1/2}(\gamma \cdot R) \cdot P_k^{(m)}(\cos(\theta)) \cdot \cos(m \cdot \phi) = 0, \quad 0 \le m \le k, \quad k \ge 0. \end{cases}$$

Clearly, the solution makes sense only when $A_{km} \neq 0$, which is possible only when $J_{k+1/2}(\gamma \cdot R) = 0$. Let us denote the roots of the equation $J_{k+1/2}(\mu) = 0$ by $\mu_{ks}$ — then the eigenvalues take the form

$$J_{k+1/2}(\gamma \cdot R) = 0 \;\Rightarrow\; \gamma_{ks} = \frac{\mu_{ks}}{R}, \quad s \in (1..\infty). \tag{2.83}$$

Zero and negative roots are no good: for $\mu = 0$ the eigenfunction $J_{k+1/2}(0 \cdot r) \equiv 0$ — identically zero, which is not an eigenfunction (the index $k + 1/2 > 0$), and negative roots give no new solutions. Therefore we take only the positive roots: $\mu_{ks}, s \in (1..\infty)$.

Thus, the eigenvalues and eigenfunctions have the form

$$\begin{cases} \gamma_{ks} = \dfrac{\mu_{ks}}{R}, s \in (1..\infty), \\ \Psi_{1kms}(r,\theta,\phi) = A_{kms} \cdot \dfrac{1}{\sqrt{r}} \cdot J_{k+1/2}(\gamma_{ks} \cdot r) \cdot P_k^{(m)}(\cos(\theta)) \cdot \sin(m \cdot \phi), \\ 1 \le m \le k, \quad k \ge 1, \\ \Psi_{2kms}(r,\theta,\phi) = A_{kms} \cdot \dfrac{1}{\sqrt{r}} \cdot J_{k+1/2}(\gamma_{ks} \cdot r) \cdot P_k^{(m)}(\cos(\theta)) \cdot \cos(m \cdot \phi), \\ 0 \le m \le k, \quad k \ge 0, \end{cases} \tag{2.84}$$

where $A_{kms}$ is an arbitrary constant factor: the eigenfunction is defined up to it. Let us set $A_{kms} = 1$ — this factor is absorbed into the expansion coefficients anyway, as in the general solution.

To expand functions into a Fourier series in $\Psi_{1kms}(r,\theta,\phi)$ and $\Psi_{2kms}(r,\theta,\phi)$ we need to compute the norms $\|\Psi_{1kms}(r,\theta,\phi)\|^2$ and $\|\Psi_{2kms}(r,\theta,\phi)\|^2$; the weight for spherical coordinates is $\rho(r,\theta,\phi) = r^2 \cdot \sin(\theta)$.

$$\|\Psi_{1kms}\|^2 = \int_0^R r \cdot J_{k+1/2}^2(\gamma_{ks} \cdot r)\, dr \int_0^\pi \left[P_k^{(m)}(\cos\theta)\right]^2 \sin(\theta)\, d\theta \cdot \cdot \int_0^{2\pi} \sin^2(m\phi)\, d\phi, \quad 1 \le m \le k, \quad k \ge 1, \quad s \ge 1,$$

$$\|\Psi_{2kms}\|^2 = \int_0^R r \cdot J_{k+1/2}^2(\gamma_{ks} \cdot r)\, dr \int_0^\pi \left[P_k^{(m)}(\cos\theta)\right]^2 \sin(\theta)\, d\theta \cdot \cdot \int_0^{2\pi} \cos^2(m\phi)\, d\phi, \quad 0 \le m \le k, \quad k \ge 0, \quad s \ge 1.$$

The integral over the azimuthal angle $\phi$ was computed in (2.59), the integral over the radius is the same as in the Dirichlet problem for the disk: after the substitution $x = \gamma_{ks} \cdot r$ it is evaluated by formula (C.6), and the integral over the angle $\theta$ after the substitution $z = \cos(\theta)$ is exactly the norm of the associated Legendre polynomial (F.6). Let us put it all together

$$\begin{cases} \|\Psi_{1kms}(r,\theta,\phi)\|^2 = \dfrac{\pi \cdot R^2}{2 \cdot k + 1} \cdot \dfrac{(k+m)!}{(k-m)!} \cdot J_{k+3/2}^2(\mu_{ks}), \\ \|\Psi_{2kms}(r,\theta,\phi)\|^2 = \begin{cases} \dfrac{2 \cdot \pi \cdot R^2}{2 \cdot k + 1} \cdot \dfrac{(k+m)!}{(k-m)!} \cdot J_{k+3/2}^2(\mu_{ks}), & m = 0, \\ \dfrac{\pi \cdot R^2}{2 \cdot k + 1} \cdot \dfrac{(k+m)!}{(k-m)!} \cdot J_{k+3/2}^2(\mu_{ks}), & m \ge 1 \end{cases} \end{cases}. \tag{2.85}$$

Now let us substitute the solution (2.82) into the boundary conditions (2.75); since the direction of the normal coincides with the direction of the coordinate $r$, we obtain

$$\begin{cases} A_{km} \cdot \left[ -\dfrac{1}{2 \cdot R \cdot \sqrt{R}} \cdot J_{k+1/2}(\gamma \cdot R) + \dfrac{\gamma}{\sqrt{R}} \cdot J'_{k+1/2}(\gamma \cdot R) \right] \cdot P_k^{(m)}(\cos(\theta)) \cdot \sin(m \cdot \phi) = 0, \\ 1 \le m \le k, \quad k \ge 1, \\ A_{km} \cdot \left[ -\dfrac{1}{2 \cdot R \cdot \sqrt{R}} \cdot J_{k+1/2}(\gamma \cdot R) + \dfrac{\gamma}{\sqrt{R}} \cdot J'_{k+1/2}(\gamma \cdot R) \right] \cdot P_k^{(m)}(\cos(\theta)) \cdot \cos(m \cdot \phi) = 0, \\ 0 \le m \le k, \quad k \ge 0, \end{cases}$$

Clearly, the solution makes sense only when $A_{km} \neq 0$, which is possible only when $\gamma \cdot J'_{k+1/2}(\gamma \cdot R) = \dfrac{1}{2 \cdot R} \cdot J_{k+1/2}(\gamma \cdot R)$. Let us denote the roots of the equation $J'_{k+1/2}(\mu) = \dfrac{1}{2 \cdot \mu} \cdot J_{k+1/2}(\mu)$ by $\mu_{ks}$ — then the eigenvalues take the form

$$J'_{k+1/2}(\mu_{ks}) = \frac{1}{2 \cdot \mu_{ks}} \cdot J_{k+1/2}(\mu_{ks}) \;\Rightarrow\; \gamma_{ks} = \frac{\mu_{ks}}{R}, \quad s \in (1..\infty). \tag{2.86}$$

Here, unlike the Dirichlet problem, the zero mode survives: for $k = 0$ the radial solution is the spherical Bessel function $j_0(\gamma \cdot r) = \dfrac{\sin(\gamma \cdot r)}{\gamma \cdot r}$, and zero is a root of the Neumann condition ($j_0'(0) = 0$); it corresponds to $\gamma = 0$ and the eigenfunction $\Psi(r,\theta,\phi) = 1$ — a nonzero constant, that is, a fully fledged eigenfunction (the constant mode). We will count it as the first root: $\mu_{01} = 0$. For $k \ge 1$ the zero root gives no eigenfunction ($j_k(0) = 0$), so there we still take only the positive roots.

Thus, the eigenvalues and eigenfunctions have the form

$$\begin{cases} \gamma_{ks} = \dfrac{\mu_{ks}}{R}, s \in (1..\infty), \quad \mu_{01} = 0, \\ \Psi_{1kms}(r,\theta,\phi) = A_{kms} \cdot \dfrac{1}{\sqrt{r}} \cdot J_{k+1/2}(\gamma_{ks} \cdot r) \cdot P_k^{(m)}(\cos(\theta)) \cdot \sin(m \cdot \phi), \\ 1 \le m \le k, \quad k \ge 1, \\ \Psi_{2kms}(r,\theta,\phi) = A_{kms} \cdot \dfrac{1}{\sqrt{r}} \cdot J_{k+1/2}(\gamma_{ks} \cdot r) \cdot P_k^{(m)}(\cos(\theta)) \cdot \cos(m \cdot \phi), \\ 0 \le m \le k, \quad k \ge 0, \end{cases} \tag{2.87}$$

where $A_{kms}$ is an arbitrary constant factor; as before, we set it equal to one.

We compute the norms of the eigenfunctions in the same way as in the Dirichlet problem: the integrals over the angles remain the same, while the integral over the radius takes a different value — we use formula (H.2) to compute the norm

$$\int_0^R r \cdot J_{k+1/2}^2(\gamma_{ks} \cdot r)\, dr = \frac{R^2}{2} \cdot \left[1 - \frac{k \cdot (k+1)}{\mu_{ks}^2}\right] \cdot J_{k+1/2}^2(\mu_{ks}).$$

The special case $\mu_{01} = 0$ (for $\gamma_{01} = 0$ the eigenfunction $\Psi_{2001}(r,\theta,\phi) = 1$) we compute separately

$$\|\Psi_{2001}(r,\theta,\phi)\|^2 = \int_0^R \int_0^\pi \int_0^{2\pi} 1 \cdot r^2 \cdot \sin(\theta)\, dr\, d\theta\, d\phi = \frac{4 \cdot \pi \cdot R^3}{3}. \tag{2.88}$$

The norms $\|\Psi_{1kms}(r,\theta,\phi)\|^2$ and $\|\Psi_{2kms}(r,\theta,\phi)\|^2$ have the form

$$\begin{aligned} \|\Psi_{1kms}(r,\theta,\phi)\|^2 &= \frac{\pi \cdot R^2}{2 \cdot k + 1} \cdot \frac{(k+m)!}{(k-m)!} \cdot \left[1 - \frac{k \cdot (k+1)}{\mu_{ks}^2}\right] \cdot J_{k+1/2}^2(\mu_{ks}), \\ \|\Psi_{2kms}(r,\theta,\phi)\|^2 &= \frac{2 \cdot \pi \cdot R^2}{2 \cdot k + 1} \cdot \frac{(k+m)!}{(k-m)!} \cdot \left[1 - \frac{k \cdot (k+1)}{\mu_{ks}^2}\right] \cdot J_{k+1/2}^2(\mu_{ks}), \quad m = 0, \\ \|\Psi_{2kms}(r,\theta,\phi)\|^2 &= \frac{\pi \cdot R^2}{2 \cdot k + 1} \cdot \frac{(k+m)!}{(k-m)!} \cdot \left[1 - \frac{k \cdot (k+1)}{\mu_{ks}^2}\right] \cdot J_{k+1/2}^2(\mu_{ks}), \quad m \ge 1. \end{aligned} \tag{2.89}$$

The final solution of the three-dimensional Dirichlet boundary value problem with inhomogeneous boundary conditions takes the form

$$
\begin{aligned}
&U(\theta,\phi,t)=\Phi(\theta,\phi,t),\\
&\widehat{T}(r,\theta,\phi,0)=T_0(r,\theta,\phi)-U(\theta,\phi,0),\\
&\widehat{f}(r,\theta,\phi,t)=f(r,\theta,\phi,t)-\frac{\partial\Phi(\theta,\phi,t)}{\partial t}+\frac{a^2}{r^2\cdot\sin(\theta)}\cdot\frac{\partial}{\partial\theta}\left(\sin(\theta)\cdot\frac{\partial\Phi(\theta,\phi,t)}{\partial\theta}\right)+\\
&\frac{a^2}{r^2\cdot\sin^2(\theta)}\cdot\frac{\partial^2\Phi(\theta,\phi,t)}{\partial\phi^2},\\
&\gamma_{ks}=\frac{\mu_{ks}}{R},\quad J_{k+1/2}(\mu_{ks})=0,\quad k\in(0..\infty),\quad s\in(1..\infty),\\
&\Psi_{1kms}(r,\theta,\phi)=\frac{1}{\sqrt{r}}\cdot J_{k+1/2}(\gamma_{ks}\cdot r)\cdot P_k^{(m)}(\cos(\theta))\cdot\sin(m\cdot\phi),\quad 1\le m\le k,\\
&\Psi_{2kms}(r,\theta,\phi)=\frac{1}{\sqrt{r}}\cdot J_{k+1/2}(\gamma_{ks}\cdot r)\cdot P_k^{(m)}(\cos(\theta))\cdot\cos(m\cdot\phi),\quad 0\le m\le k,\\
&\|\Psi_{1kms}(r,\theta,\phi)\|^2=\frac{\pi\cdot R^2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!}\cdot J_{k+3/2}^2(\mu_{ks}),\\
&\|\Psi_{2kms}(r,\theta,\phi)\|^2=\frac{2\cdot\pi\cdot R^2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!}\cdot J_{k+3/2}^2(\mu_{ks}),\quad m=0,\\
&\|\Psi_{2kms}(r,\theta,\phi)\|^2=\frac{\pi\cdot R^2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!}\cdot J_{k+3/2}^2(\mu_{ks}),\quad m\ge 1,\\
&T_{01kms}=\int_0^R\int_0^\pi\int_0^{2\pi}\widehat{T}(r,\theta,\phi,0)\cdot\Psi_{1kms}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi,\\
&T_{02kms}=\int_0^R\int_0^\pi\int_0^{2\pi}\widehat{T}(r,\theta,\phi,0)\cdot\Psi_{2kms}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi,\\
&f_{1kms}(t)=\int_0^R\int_0^\pi\int_0^{2\pi}\widehat{f}(r,\theta,\phi,t)\cdot\Psi_{1kms}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi,\\
&f_{2kms}(t)=\int_0^R\int_0^\pi\int_0^{2\pi}\widehat{f}(r,\theta,\phi,t)\cdot\Psi_{2kms}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi,\\
&\widehat{T}(r,\theta,\phi,t)=\\
&\sum_{k=1}^{\infty}\sum_{m=1}^{k}\sum_{s=1}^{\infty}\frac{\Psi_{1kms}(r,\theta,\phi)}{\|\Psi_{1kms}(r,\theta,\phi)\|^2}\cdot e^{-a^2\cdot\gamma_{ks}^2\cdot t}\cdot\left[T_{01kms}+\int_0^t e^{a^2\cdot\gamma_{ks}^2\cdot\tau}\cdot f_{1kms}(\tau)\,d\tau\right]+\\
&\sum_{k=0}^{\infty}\sum_{m=0}^{k}\sum_{s=1}^{\infty}\frac{\Psi_{2kms}(r,\theta,\phi)}{\|\Psi_{2kms}(r,\theta,\phi)\|^2}\cdot e^{-a^2\cdot\gamma_{ks}^2\cdot t}\cdot\left[T_{02kms}+\int_0^t e^{a^2\cdot\gamma_{ks}^2\cdot\tau}\cdot f_{2kms}(\tau)\,d\tau\right],\\
&T(r,\theta,\phi,t)=U(\theta,\phi,t)+\widehat{T}(r,\theta,\phi,t).
\end{aligned}
\tag{2.90}
$$

In the Neumann problem, unlike the Dirichlet one, the spectrum contains the zero mode ($\gamma_{01}=0$, $\Psi_{2001}(r,\theta,\phi)=1$) — the case $\gamma=0$, which in the general solution (2.18) was set aside for separate consideration; it is precisely what gives the non-decaying term — the term with $k=0, m=0, s=1$ in the second sum. The final solution of the three-dimensional Neumann boundary value problem with inhomogeneous boundary conditions takes the form

$$
\begin{aligned}
&U(r,\theta,\phi,t) = \Phi(\theta,\phi,t)\cdot r,\\
&\widehat{T}(r,\theta,\phi,0) = T_0(r,\theta,\phi) - U(r,\theta,\phi,0),\\
&\widehat{f}(r,\theta,\phi,t) = f(r,\theta,\phi,t) - r\cdot\frac{\partial\Phi(\theta,\phi,t)}{\partial t} + \frac{2\cdot a^2\cdot\Phi(\theta,\phi,t)}{r} +\\
&\frac{a^2}{r\cdot\sin(\theta)}\cdot\frac{\partial}{\partial\theta}\left(\sin(\theta)\cdot\frac{\partial\Phi(\theta,\phi,t)}{\partial\theta}\right) +\\
&\frac{a^2}{r\cdot\sin^2(\theta)}\cdot\frac{\partial^2\Phi(\theta,\phi,t)}{\partial\phi^2},\\
&\gamma_{ks} = \frac{\mu_{ks}}{R},\quad J'_{k+1/2}(\mu_{ks}) = \frac{1}{2\cdot\mu_{ks}}\cdot J_{k+1/2}(\mu_{ks}),\quad \mu_{01}=0,\\
&k\in(0..\infty),\quad s\in(1..\infty),\\
&\Psi_{1kms}(r,\theta,\phi) = \frac{1}{\sqrt{r}}\cdot J_{k+1/2}(\gamma_{ks}\cdot r)\cdot P_k^{(m)}(\cos(\theta))\cdot\sin(m\cdot\phi),\quad 1\le m\le k,\\
&\Psi_{2kms}(r,\theta,\phi) = \frac{1}{\sqrt{r}}\cdot J_{k+1/2}(\gamma_{ks}\cdot r)\cdot P_k^{(m)}(\cos(\theta))\cdot\cos(m\cdot\phi),\quad 0\le m\le k,\\
&\|\Psi_{1kms}(r,\theta,\phi)\|^2 = \frac{\pi\cdot R^2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!}\cdot\left[1-\frac{k\cdot(k+1)}{\mu_{ks}^2}\right]\cdot J_{k+1/2}^2(\mu_{ks}),\\
&\|\Psi_{2kms}(r,\theta,\phi)\|^2 = \frac{2\cdot\pi\cdot R^2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!}\cdot\left[1-\frac{k\cdot(k+1)}{\mu_{ks}^2}\right]\cdot J_{k+1/2}^2(\mu_{ks}),\quad m=0,\\
&\|\Psi_{2kms}(r,\theta,\phi)\|^2 = \frac{\pi\cdot R^2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!}\cdot\left[1-\frac{k\cdot(k+1)}{\mu_{ks}^2}\right]\cdot J_{k+1/2}^2(\mu_{ks}),\quad m\ge 1,\\
&\|\Psi_{2001}(r,\theta,\phi)\|^2 = \frac{4\cdot\pi\cdot R^3}{3},\\
&T_{01kms} = \int_0^R\int_0^\pi\int_0^{2\pi}\widehat{T}(r,\theta,\phi,0)\cdot\Psi_{1kms}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi,\\
&T_{02kms} = \int_0^R\int_0^\pi\int_0^{2\pi}\widehat{T}(r,\theta,\phi,0)\cdot\Psi_{2kms}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi,\\
&f_{1kms}(t) = \int_0^R\int_0^\pi\int_0^{2\pi}\widehat{f}(r,\theta,\phi,t)\cdot\Psi_{1kms}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi,\\
&f_{2kms}(t) = \int_0^R\int_0^\pi\int_0^{2\pi}\widehat{f}(r,\theta,\phi,t)\cdot\Psi_{2kms}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi,\\
&\widehat{T}(r,\theta,\phi,t) =\\
&\sum_{k=1}^{\infty}\sum_{m=1}^{k}\sum_{s=1}^{\infty}\frac{\Psi_{1kms}(r,\theta,\phi)}{\|\Psi_{1kms}(r,\theta,\phi)\|^2}\cdot e^{-a^2\cdot\gamma_{ks}^2\cdot t}\cdot\left[T_{01kms}+\int_0^t e^{a^2\cdot\gamma_{ks}^2\cdot\tau}\cdot f_{1kms}(\tau)\,d\tau\right] +\\
&\sum_{k=0}^{\infty}\sum_{m=0}^{k}\sum_{s=1}^{\infty}\frac{\Psi_{2kms}(r,\theta,\phi)}{\|\Psi_{2kms}(r,\theta,\phi)\|^2}\cdot e^{-a^2\cdot\gamma_{ks}^2\cdot t}\cdot\left[T_{02kms}+\int_0^t e^{a^2\cdot\gamma_{ks}^2\cdot\tau}\cdot f_{2kms}(\tau)\,d\tau\right],\\
&T(r,\theta,\phi,t) = U(r,\theta,\phi,t) + \widehat{T}(r,\theta,\phi,t).
\end{aligned}
\tag{2.91}
$$

To obtain the stationary solutions, we let time tend to infinity in the solutions found ($t\to\infty$), as was done for the general case in (2.18). The term with the initial condition vanishes, since $e^{-a^2\cdot\gamma_{ks}^2\cdot t}\to 0$; the sources and boundary conditions stop depending on time, and the time integral for $\gamma_{ks}\neq 0$ gives the factor $\frac{1}{a^2\cdot\gamma_{ks}^2}$:

$$\int_0^t e^{-a^2\cdot\gamma_{ks}^2\cdot(t-\tau)}\,d\tau = \frac{1}{a^2\cdot\gamma_{ks}^2}\cdot(1-e^{-a^2\cdot\gamma_{ks}^2\cdot t})$$

$$\to \frac{1}{a^2\cdot\gamma_{ks}^2} \quad \text{as} \quad t\to\infty, \quad \gamma_{ks}\neq 0.$$

The stationary solution of the three-dimensional Dirichlet boundary value problem takes the form

$$\begin{aligned}
&U(\theta,\phi)=\Phi(\theta,\phi),\\
&\widehat{f}(r,\theta,\phi)=f(r,\theta,\phi)+\frac{a^2}{r^2\cdot\sin(\theta)}\cdot\frac{\partial}{\partial\theta}\left(\sin(\theta)\cdot\frac{\partial\Phi(\theta,\phi)}{\partial\theta}\right)+\frac{a^2}{r^2\cdot\sin^2(\theta)}\cdot\frac{\partial^2\Phi(\theta,\phi)}{\partial\phi^2},\\
&\gamma_{ks}=\frac{\mu_{ks}}{R}, \quad J_{k+1/2}(\mu_{ks})=0, \quad k\in(0..\infty), \quad s\in(1..\infty),\\
&\Psi_{1kms}(r,\theta,\phi)=\frac{1}{\sqrt{r}}\cdot J_{k+1/2}(\gamma_{ks}\cdot r)\cdot P_k^{(m)}(\cos(\theta))\cdot\sin(m\cdot\phi), \quad 1\le m\le k,\\
&\Psi_{2kms}(r,\theta,\phi)=\frac{1}{\sqrt{r}}\cdot J_{k+1/2}(\gamma_{ks}\cdot r)\cdot P_k^{(m)}(\cos(\theta))\cdot\cos(m\cdot\phi), \quad 0\le m\le k,\\
&\|\Psi_{1kms}(r,\theta,\phi)\|^2=\frac{\pi\cdot R^2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!}\cdot J_{k+3/2}^2(\mu_{ks}),\\
&\|\Psi_{2kms}(r,\theta,\phi)\|^2=\frac{2\cdot\pi\cdot R^2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!}\cdot J_{k+3/2}^2(\mu_{ks}), \quad m=0,\\
&\|\Psi_{2kms}(r,\theta,\phi)\|^2=\frac{\pi\cdot R^2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!}\cdot J_{k+3/2}^2(\mu_{ks}), \quad m\ge 1,\\
&f_{1kms}=\int_0^R\int_0^\pi\int_0^{2\pi}\widehat{f}(r,\theta,\phi)\cdot\Psi_{1kms}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi,\\
&f_{2kms}=\int_0^R\int_0^\pi\int_0^{2\pi}\widehat{f}(r,\theta,\phi)\cdot\Psi_{2kms}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi,\\
&T(r,\theta,\phi)=U(\theta,\phi)+\sum_{k=1}^{\infty}\sum_{m=1}^{k}\sum_{s=1}^{\infty}\frac{\Psi_{1kms}(r,\theta,\phi)}{\|\Psi_{1kms}(r,\theta,\phi)\|^2}\cdot\frac{f_{1kms}}{a^2\cdot\gamma_{ks}^2}+\\
&\sum_{k=0}^{\infty}\sum_{m=0}^{k}\sum_{s=1}^{\infty}\frac{\Psi_{2kms}(r,\theta,\phi)}{\|\Psi_{2kms}(r,\theta,\phi)\|^2}\cdot\frac{f_{2kms}}{a^2\cdot\gamma_{ks}^2}.
\end{aligned} \tag{2.92}$$

The stationary solution of the three-dimensional Neumann boundary value problem is not so simple. Unlike the Dirichlet problem, the purely Neumann stationary problem is not always solvable. The reduced function $\widehat{T}(r,\theta,\phi)$ satisfies the homogeneous Neumann conditions and the stationary equation $a^2\cdot\Delta\widehat{T}(r,\theta,\phi)+\widehat{f}(r,\theta,\phi)=0$. Let us integrate it over the domain:

$$0=\int_0^R\int_0^\pi\int_0^{2\pi}\left(a^2\cdot\Delta\widehat{T}+\widehat{f}\right)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi=$$

$$a^2\cdot\oint_{r=R}\frac{\partial\widehat{T}}{\partial\mathbf{n}}\,dS+\int_0^R\int_0^\pi\int_0^{2\pi}\widehat{f}\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi=\int_0^R\int_0^\pi\int_0^{2\pi}\widehat{f}\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi.$$

The boundary integral vanished due to the homogeneous Neumann conditions, and what remains is the solvability condition (2.22): $\int_0^R\int_0^\pi\int_0^{2\pi}\widehat{f}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi=0$

— the total reduced source over the domain must vanish (otherwise heat accumulates and there is no steady state). When it holds, the solution is defined only up to an arbitrary additive constant $C$: it corresponds to the zero mode ($\gamma_{01} = 0$, $\Psi_{2001}(r,\theta,\phi) = 1$), whose amplitude is not fixed by the stationary equation. With this caveat the solution takes the form

$$
\begin{aligned}
&U(r,\theta,\phi) = \Phi(\theta,\phi)\cdot r,\\
&\widehat{f}(r,\theta,\phi) = f(r,\theta,\phi) + \frac{2\cdot a^2\cdot\Phi(\theta,\phi)}{r} +\\
&\frac{a^2}{r\cdot\sin(\theta)}\cdot\frac{\partial}{\partial\theta}\left(\sin(\theta)\cdot\frac{\partial\Phi(\theta,\phi)}{\partial\theta}\right) + \frac{a^2}{r\cdot\sin^2(\theta)}\cdot\frac{\partial^2\Phi(\theta,\phi)}{\partial\phi^2},\\
&\gamma_{ks} = \frac{\mu_{ks}}{R},\quad J'_{k+1/2}(\mu_{ks}) = \frac{1}{2\cdot\mu_{ks}}\cdot J_{k+1/2}(\mu_{ks}),\quad \mu_{ks} > 0,\\
&k\in(0..\infty),\quad s\in(1..\infty),\\
&\Psi_{1kms}(r,\theta,\phi) = \frac{1}{\sqrt{r}}\cdot J_{k+1/2}(\gamma_{ks}\cdot r)\cdot P_k^{(m)}(\cos(\theta))\cdot\sin(m\cdot\phi),\quad 1\le m\le k,\\
&\Psi_{2kms}(r,\theta,\phi) = \frac{1}{\sqrt{r}}\cdot J_{k+1/2}(\gamma_{ks}\cdot r)\cdot P_k^{(m)}(\cos(\theta))\cdot\cos(m\cdot\phi),\quad 0\le m\le k,\\
&\|\Psi_{1kms}(r,\theta,\phi)\|^2 = \frac{\pi\cdot R^2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!}\cdot\left[1-\frac{k\cdot(k+1)}{\mu_{ks}^2}\right]\cdot J_{k+1/2}^2(\mu_{ks}),\\
&\|\Psi_{2kms}(r,\theta,\phi)\|^2 = \frac{2\cdot\pi\cdot R^2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!}\cdot\left[1-\frac{k\cdot(k+1)}{\mu_{ks}^2}\right]\cdot J_{k+1/2}^2(\mu_{ks}),\quad m=0,\\
&\|\Psi_{2kms}(r,\theta,\phi)\|^2 = \frac{\pi\cdot R^2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!}\cdot\left[1-\frac{k\cdot(k+1)}{\mu_{ks}^2}\right]\cdot J_{k+1/2}^2(\mu_{ks}),\quad m\ge 1,\\
&f_{1kms} = \int_0^R\int_0^\pi\int_0^{2\pi}\widehat{f}(r,\theta,\phi)\cdot\Psi_{1kms}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi,\\
&f_{2kms} = \int_0^R\int_0^\pi\int_0^{2\pi}\widehat{f}(r,\theta,\phi)\cdot\Psi_{2kms}(r,\theta,\phi)\cdot r^2\cdot\sin(\theta)\,dr\,d\theta\,d\phi,\\
&T(r,\theta,\phi) = U(r,\theta,\phi) + \sum_{k=1}^{\infty}\sum_{m=1}^{k}\sum_{s=1}^{\infty}\frac{\Psi_{1kms}(r,\theta,\phi)}{\|\Psi_{1kms}(r,\theta,\phi)\|^2}\cdot\frac{f_{1kms}}{a^2\cdot\gamma_{ks}^2} +\\
&\sum_{k=0}^{\infty}\sum_{m=0}^{k}\sum_{s=1}^{\infty}\frac{\Psi_{2kms}(r,\theta,\phi)}{\|\Psi_{2kms}(r,\theta,\phi)\|^2}\cdot\frac{f_{2kms}}{a^2\cdot\gamma_{ks}^2} + C,\quad C = \text{const}.
\end{aligned}
\tag{2.93}
$$

# Chapter 3

# Solving large sparse matrices

## Storage formats

After applying the finite element method, the problem reduces to a system of linear algebraic equations

$$A \cdot x = b. \tag{3.1}$$

The matrix $A$ in such a system is usually large but sparse: almost all of its elements are zero. This is natural for FEM — each simplex is connected only to a small number of neighbouring nodes, so a single row of the global matrix contains nonzero elements only at the positions of the local neighbourhood of a node. Storing such a matrix as a dense table is wasteful: most of the memory is spent on zeros, and matrix-vector multiplication performs many useless operations.

In a FEM implementation it is convenient to use three representations of one and the same sparse matrix.

1. **COO** (coordinate list) stores the matrix as a list of triples $(i,\ j,\ a_{ij})$, where $i$ is the row index, $j$ is the column index, and $a_{ij}$ is the element value. The format is convenient when assembling the global matrix: the local matrices of the simplices are simply appended to the common list, and duplicate indices are summed.

2. **CSR** (compressed sparse row) stores the matrix row by row in three arrays: the values of the nonzero elements, their column indices, and the array of offsets of the beginning of each row. The format is inconvenient for adding new elements, but optimal for the operation $A \cdot x$, which is performed at every iteration.

3. **CSC** (compressed sparse column) is the column analogue of CSR. It is not needed for solving the system, but it is useful when multiplying two sparse matrices, when fast access to the elements of the right matrix by columns is required.

All three formats for one and the same matrix $A$ are shown in figure (3.1).

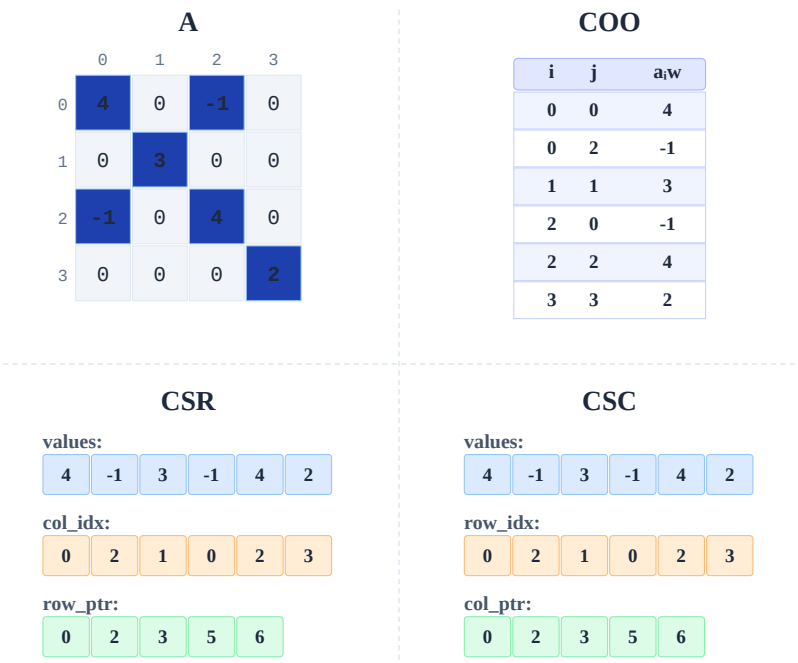


**Fig. 3.1.** Storage of the sparse matrix $A$ in the COO, CSR and CSC formats.

Next, let us talk about the solution method.

# Solution method

For large sparse matrices, direct methods that build a full decomposition of the matrix are usually unprofitable: during the decomposition, new nonzero elements appear in the positions of the original zeros — the so-called *fill-in*. As a result, the main advantage of a sparse matrix is lost — a small memory footprint and a small number of arithmetic operations.

We will solve system (3.1) by the iterative *preconditioned conjugate gradient method* (PCG). The method applies to symmetric positive definite matrices — exactly those that arise in FEM problems with elliptic operators and in implicit heat conduction schemes. Let us write out the algorithm step by step.

Let $x_0 = 0$ be the initial guess. The initial residual takes the form

$$r_0 = b - A \cdot x_0. \tag{3.2}$$

The initial search direction and the preconditioning vector have the form

$$P \cdot z_0 = r_0, \qquad p_0 = z_0, \tag{3.3}$$

where $P$ is the preconditioner matrix: it approximates $A$, but systems with it are much cheaper to solve. Then at every iteration $k = 0, 1, 2, \ldots$ we compute

$$\alpha_k = \frac{r_k^T \cdot z_k}{p_k^T \cdot A \cdot p_k}, \tag{3.4}$$

$$x_{k+1} = x_k + \alpha_k \cdot p_k, \tag{3.5}$$

$$r_{k+1} = r_k - \alpha_k \cdot A \cdot p_k, \tag{3.6}$$

$$P \cdot z_{k+1} = r_{k+1}, \tag{3.7}$$

$$\beta_k = \frac{r_{k+1}^T \cdot z_{k+1}}{r_k^T \cdot z_k}, \tag{3.8}$$

$$p_{k+1} = z_{k+1} + \beta_k \cdot p_k. \tag{3.9}$$

The iterations continue until the squared norm of the residual drops below a given threshold

$$r_{k+1}^T \cdot r_{k+1} < \varepsilon. \tag{3.10}$$

The main computational cost of each iteration is one multiplication of the sparse matrix by a vector, $A \cdot p_k$. This is precisely why the matrix is converted from COO to CSR before the iterations start: in this format the multiplication of a row by the vector runs only over the nonzero elements.

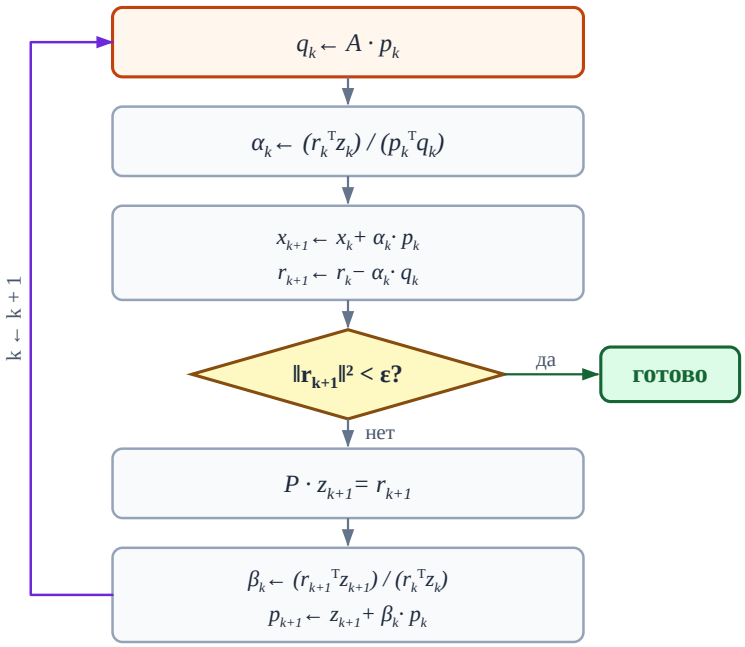


**Fig. 3.2.** One PCG iteration: the step $A \cdot p_k$ is the most expensive one.

Next, let us talk about preconditioners.

# Preconditioners

The convergence rate of the conjugate gradient method is determined by the condition number of the matrix $A$. The larger it is, the slower the method converges — many iterations are required to reach acceptable accuracy. The preconditioner $P$ is meant to

improve this number by replacing the original system with an equivalent one with the matrix $P^{-1} \cdot A$, whose eigenvalues should cluster in a narrow neighbourhood of one.

In practice there is no need to build $P^{-1} \cdot A$ explicitly: at every iteration it suffices to be able to solve the auxiliary problem

$$P \cdot z_k = r_k, \tag{3.11}$$

using $z_k$ instead of the original residual $r_k$ — see formulas (3.7)–(3.8). Let us consider three variants.

**1. No preconditioning ($P = I$).** In this case $z_k = r_k$, and the method coincides with the classical conjugate gradient method. This variant is useful for debugging, but for ill-conditioned systems it converges slowly.

**2. Jacobi preconditioner.** As $P$ we take the diagonal part of the matrix $A$

$$P = \operatorname{diag}(A) = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn}). \tag{3.12}$$

Its application reduces to element-wise division

$$z_{k,i} = \frac{r_{k,i}}{a_{ii}}. \tag{3.13}$$

This preconditioner is practically free in memory and time, but it takes into account only the diagonal, ignoring the couplings between the variables.

**3. Incomplete Cholesky factorization (IC).** For a symmetric positive definite matrix an approximate factorization is built

$$A \approx L \cdot L^T, \tag{3.14}$$

where $L$ is a lower triangular matrix. Unlike the full Cholesky factorization, the fill-in elements in the positions of the original zeros are discarded — the nonzero structure of $L$ coincides with the structure of the lower triangle of $A$. Once $L$ is built, applying the preconditioner consists of two triangular substitutions

$$L \cdot y = r_k, \tag{3.15}$$

$$L^T \cdot z_k = y. \tag{3.16}$$


**Fig. 3.3.** Preconditioners: identity ($P = I$), Jacobi ($P = \operatorname{diag}(A)$) and incomplete Cholesky ($A \approx L \cdot L^T$).

In FEM problems the incomplete Cholesky factorization usually gives the best trade-off: it is more expensive than the diagonal preconditioner (building $L$ and two triangular

sweeps per iteration), but it noticeably reduces the number of iterations. For nonsymmetric or indefinite systems, which fall outside standard heat conduction problems, one should turn to the GMRES or BiCGSTAB methods with appropriate preconditioners.

This concludes the solution of sparse systems — next, let us talk about partitioning the geometry into simplices.

# Chapter 4

# Partitioning the geometry into simplices

## Segmentation

In the one-dimensional case, partitioning the geometry into simplices reduces to building segments between neighbouring points. We will call this process segmentation. The input is a set of points on a line, the output is a set of one-dimensional simplices.

The algorithm in this case is very simple.

1. Take all the points of the geometry and sort them by the coordinate $x$.
2. After sorting, connect each point to the next one: if the ordered points have coordinates $x_0 < x_1 < \ldots < x_n$, the simplices have the form $s_i = [x_i, x_{i+1}]$, $i = 0, 1, \ldots, n-1$.
3. Assign a new index to each resulting segment. These indices refer no longer to the original points but to the simplices.

The figure "Segmentation" shows both key actions: first the points are ordered by coordinate, then each neighbouring pair turns into a separate simplex.

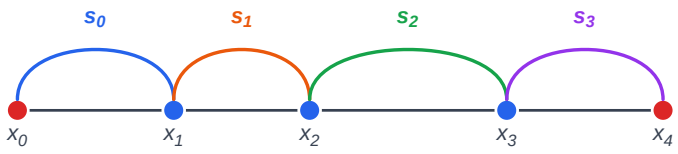


**Fig. 4.1.** Segmentation: the ordered points and the simplex segments.

In the one-dimensional case the whole geometry is already given by the arrangement of the points on the line, so it suffices to order them. If two points have the same coordinate, this case must be handled before segmentation: the algorithm itself assumes that the sorted sequence of points defines a correct chain of segments.

# Triangulation

In the two-dimensional case the simplex is a triangle, so for a given set of points we must build a family of triangles that covers the domain completely and has no internal intersections. Additionally we require the resulting mesh to satisfy the *Delaunay condition*: the circumscribed circle of each triangle must contain no other vertices of the triangulation. For a non-degenerate point set without prescribed edges, such a triangulation maximizes the minimum angle. Thus the Delaunay condition helps reduce the number of badly elongated triangles, while the stability of the subsequent finite-element computations also depends on the input geometry quality, point density and correct handling of constraints.

Building a Delaunay mesh by hand is trickier than it seems: the computational domain is often a square or an annulus, whose boundaries produce degenerate configurations on which naive floating-point computations fail (see below). The low-level construction is therefore entrusted to the mature **CGAL** library with exact geometric predicates, with a data-preparation and validation pipeline implemented on top. The project's own implementation of the classical divide-and-conquer scheme, with which it all started, has been moved to the appendix "Triangulation by divide and conquer".

## Geometric predicates

Almost every geometric decision in the algorithm reduces to two sign predicates.

**Position of a point relative to an edge.** For a directed edge with endpoints $p_1 = (x_1, y_1)$, $p_2 = (x_2, y_2)$ and a point $p_3 = (x_3, y_3)$, the sign of the expression

$$\operatorname{orient}(p_1, p_2, p_3) = (x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1) \tag{4.1}$$

determines whether $p_3$ lies to the left, to the right, or exactly on the line $p_1p_2$. The same predicate is used when building the convex hull and when controlling triangle orientation.

**Testing the Delaunay condition.** For a triangle with vertices $p_1, p_2, p_3$ listed counter-clockwise and a query point $p_0 = (x_0, y_0)$, the condition is conveniently written via the sign of the determinant

$$\det \begin{pmatrix} x_1 & y_1 & x_1^2 + y_1^2 & 1 \\ x_2 & y_2 & x_2^2 + y_2^2 & 1 \\ x_3 & y_3 & x_3^2 + y_3^2 & 1 \\ x_0 & y_0 & x_0^2 + y_0^2 & 1 \end{pmatrix} \le 0 \tag{4.2}$$

If condition (4.2) holds, the point $p_0$ lies outside the circumscribed circle or on its boundary; otherwise it is inside, and the pair of adjacent triangles must be restructured. With clockwise vertex order the sign of the determinant flips, so triangle orientation is controlled by predicate (4.1).

## Robustness on degenerate input

Predicates (4.1) and (4.2) are signs of determinants. On inputs in general position their value differs noticeably from zero, and ordinary double-precision arithmetic yields the correct sign. But a square computational domain produces two unpleasant cases:

- *collinearity* — many boundary points lie exactly on one vertical or horizontal line, and predicate (4.1) evaluates to exact zero;
- *cocircularity* — the four corners of any cell of a regular grid lie on one circle, and determinant (4.2) is also exactly zero.

The trouble is that when such a "zero" determinant is computed in floating-point arithmetic, rounding errors accumulate, and instead of zero we get a tiny number of *arbitrary* sign — noise. Different parts of the algorithm, having received contradictory answers for the same triple of points, start building an inconsistent mesh: flat (zero-area) triangles appear along the boundary and the Delaunay condition is violated.

The solution is to compute the sign of the predicate *exactly*. The library we use evaluates the determinants with filtered arithmetic: a fast preliminary computation with a guaranteed error bound, and in doubtful (near-zero) cases an exact re-evaluation. As a result the sign is never wrong, and a true zero is recognized as zero. Importantly, only the predicates need exactness: the algorithm constructs no new points — the mesh vertices are exactly the input points — so exact construction arithmetic is not required and the coordinates remain plain double-precision numbers. This compromise costs almost nothing in speed.

## Constrained triangulation

In practice, besides the points themselves we must honour user-specified edges — the boundaries of the computational domain and the contours of holes. Such edges must be present in the final mesh and must not be destroyed by local restructurings. This problem is known as *constrained Delaunay triangulation*: near a constraint edge the Delaunay condition may legitimately be violated, but in return the mesh is guaranteed to follow the prescribed contours. The edge-insertion operation itself is described below in the implementation section, and in full detail in the appendix.

## Meshing the computational domain

The full pipeline that meshes a domain with a boundary and holes consists of six steps.

1. The contours — the outer one and the hole contours — are subdivided into segments no longer than the target step $\sqrt{S/n}$, where $S$ is the domain area and $n$ is the number of interior points; this matches the boundary point density to the interior density and reduces the risk of elongated triangles along the boundary.
2. $n$ random points are generated inside the domain (outside the holes) by rejection sampling, optionally with a prescribed minimum distance between them.
3. The *interior* points are inserted one by one into an empty triangulation, followed by the *boundary* points of all contours.
4. Every segment of every contour is forcibly inserted into the mesh as a *constraint edge*.
5. The resulting triangulation — it still covers the whole convex hull — is validated; on any defect the construction aborts with an error.

6. Triangles whose centroid lies outside the domain (inside a hole or beyond the outer contour) are discarded — what remains is the mesh of the computational domain proper.

The validation in step five is a safety net comprising four checks:

- *vertex completeness* — every input point is present in the mesh as a vertex, and the mesh contains no extraneous vertices;
- *triangle count* — it must equal the value $T = 2N - 2 - h$ that follows from Euler's formula, where $N$ is the number of vertices and $h$ is the number of vertices on the convex hull;
- *non-intersection* — no two triangles intersect across edge interiors;
- *Delaunay condition* — checked for pairs of triangles whose shared side is not a constraint edge; near constraints the Delaunay condition may legitimately be violated.

## Storing the triangulation

The finished triangulation is stored not as a mere list of triangles but as a structure that maps every edge to its adjacent triangles. Through a shared edge one can always quickly find the pair of triangles potentially violating the Delaunay condition — this is what flip operations use. The same structure yields the mesh boundary (edges belonging to a single triangle): before the holes are cut out, that is the convex hull. The output is a set of triangular simplices from which the local stiffness matrices are later built and the global finite-element system is assembled.

## Implementation: the CGAL library

**CGAL** (Computational Geometry Algorithms Library) is an open-source C++ computational-geometry library and the de facto standard in the field. We use its classes *Delaunay_triangulation_2* (plain Delaunay triangulation) and *Constrained_Delaunay_triangulation_2* (constrained triangulation) with the *Exact_predicates_inexact_constructions_kernel*. The staged pipeline described above, the constraint insertion and the step-five checks are implemented on top of it.

CGAL builds the mesh *incrementally* — points are inserted one by one into an existing triangulation.

- Instead of an auxiliary "super-triangle", which would have to be sized and later deleted, it uses an *infinite vertex*: the convex hull of the current mesh is closed off by fictitious triangles whose third vertex is a formal "point at infinity".
- Each next point is first *located* by walking the mesh: starting from some triangle, the algorithm steps from neighbour to neighbour towards the point, guided by orientation predicate (4.1).
- The located triangle is split by the inserted point into three (a point on an edge splits the two adjacent triangles into four), after which the Delaunay condition is restored by a cascade of flips spreading from the new vertex: if by predicate (4.2)

the fourth vertex of the neighbouring triangle falls inside the circumscribed circle, the common diagonal of the quadrilateral is replaced by the opposite one (figure Diagonal swap).

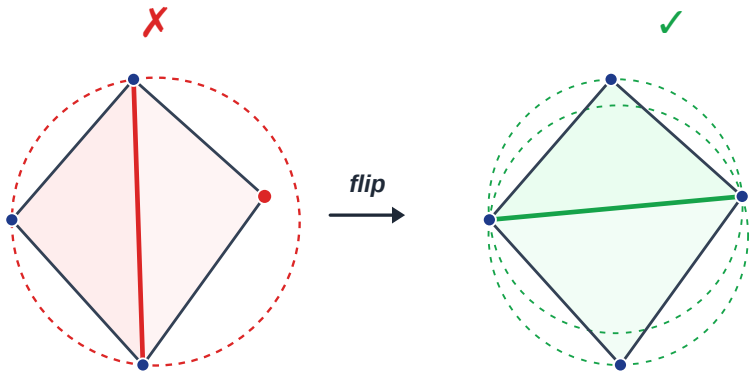


**Fig. 4.2.** Local restructuring of a pair of triangles (flip) to restore the Delaunay condition.

A constraint edge is inserted into the finished mesh by a separate operation: all triangles crossed by the segment are removed, the resulting cavity is split by the segment into two chains of vertices, and each chain is re-triangulated; in subsequent restructurings, flipping across the constraint edge is forbidden. Since by insertion time both endpoints of every segment are already mesh vertices and the contours do not intersect, no new (Steiner) points appear. A detailed walkthrough of this operation is given in the appendix.

The chosen kernel combines *exact predicates* with *inexact constructions*. Predicates are evaluated with filtering: interval arithmetic with a guaranteed error bound first, and only when the interval straddles zero — an exact re-evaluation in multiprecision arithmetic. This is what ensures that the collinear and cocircular configurations of the square are handled correctly. Constructions, in turn, stay plain doubles: the algorithm creates no new points, so every vertex of the final mesh maps back losslessly to an input point. With a random insertion order — which our random point generator provides — the expected construction cost is $O(n \log n)$.

## Examples

All meshes below are real outputs of the described pipeline (the *triangulate* utility of the computational core), drawn without any retouching.

The figure "Square" shows the triangulation of the unit square with $n = 200$ interior points: after the boundary is subdivided with step $\sqrt{S/n}$, the mesh contains 260 vertices and 458 triangles. The square's boundary consists of dozens of exactly collinear points — the very degenerate input on which floating-point arithmetic without exact predicates falls apart; here the mesh is correct along the entire boundary, with no flat triangles.

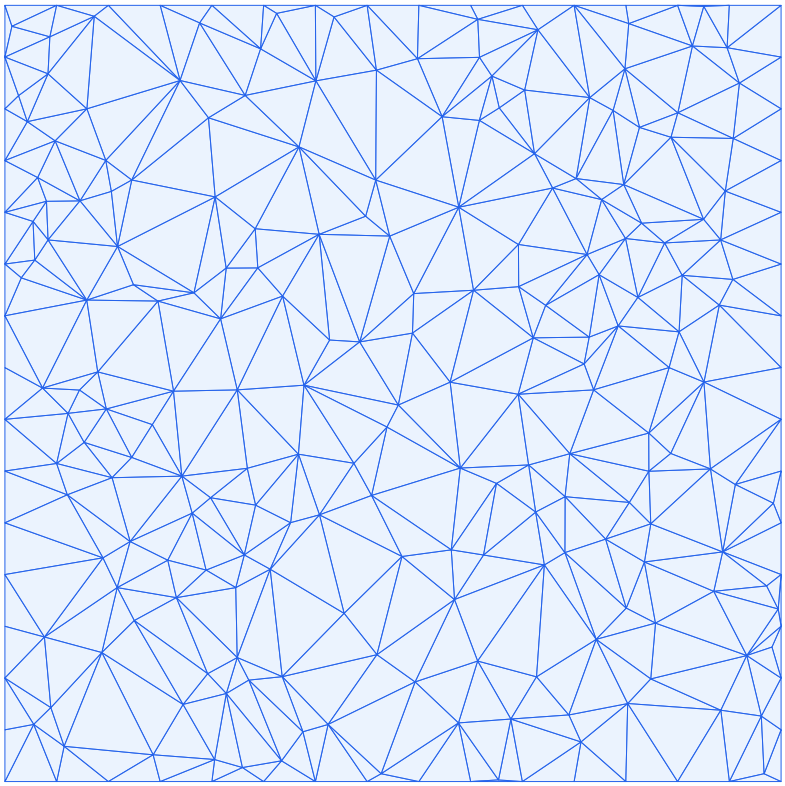

**Fig. 4.3.** Triangulation of the unit square ($n = 200$; 260 vertices, 458 triangles): collinear boundary points are handled correctly.

The figure "Square with a hole" demonstrates constraint edges at work: a square hole is cut out of the square, and both contours are inserted as constraints ($n = 300$; 396 vertices, 696 triangles). The mesh follows the hole contour exactly and does not enter it — triangles whose centroid landed inside the hole were discarded at the final pipeline step.

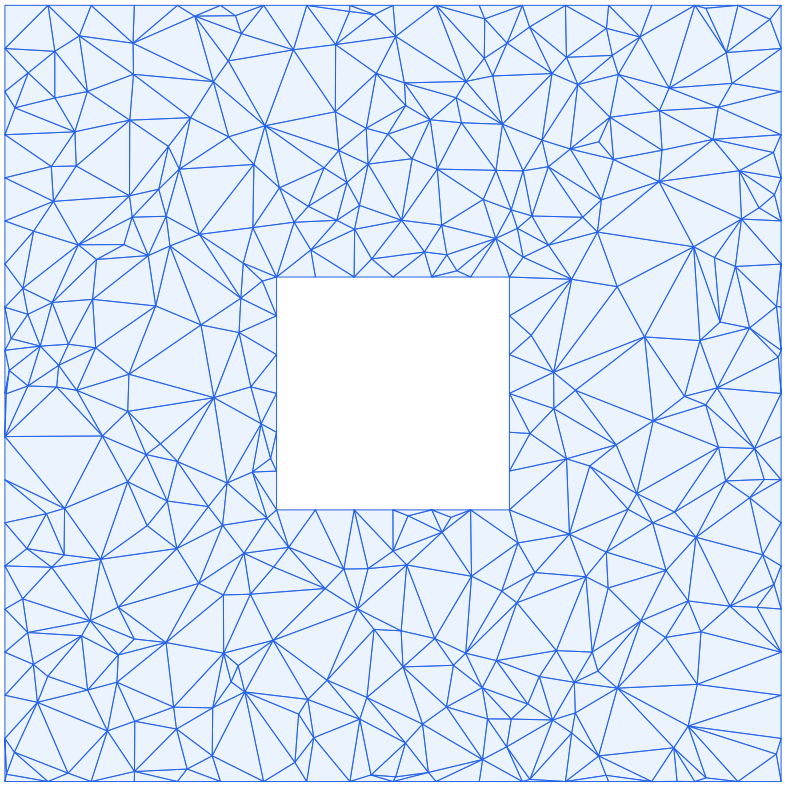

**Fig. 4.4.** Triangulation of a square with a hole ($n = 300$; 396 vertices, 696 triangles): the contours are inserted as constraint edges.

Finally, the figure "Annulus" shows the triangulation of an annulus (outer radius 1, inner radius 0.4) with $n = 1000$: the final mesh contains 1192 vertices and 2192 triangles. Both contours are approximated by polygons (128 and 64 vertices) and inserted as constraint edges, so the mesh follows the prescribed polygonal boundary of the domain, while the interior points produce a nearly uniform Delaunay mesh.

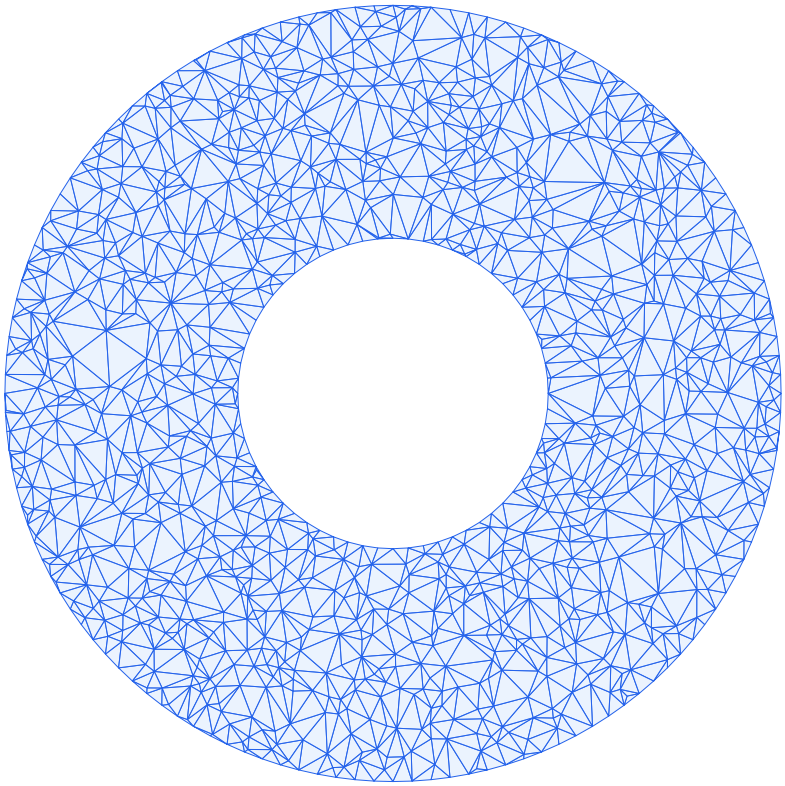

**Fig. 4.5.** Triangulation of an annulus ($R = 1$, $r = 0.4$, $n = 1000$; 1192 vertices, 2192 triangles): the outer contour and the hole boundary are inserted as constraint edges.

# Tetrahedralization

In the three-dimensional case the simplex is a tetrahedron, so the analogue of triangulation is *tetrahedralization*: the volumetric domain must be partitioned into tetrahedra so that they do not overlap, cover the domain completely and meet pairwise only along shared faces, edges or vertices. As in the two-dimensional case, we require the *Delaunay condition* — now for the circumscribed *sphere* of a tetrahedron: it must contain no other mesh vertices. This condition helps avoid some badly elongated elements, but by itself it does not guarantee a good shape for every tetrahedron; this is why we also prescribe quality criteria and run mesh optimization.

The minimal three-dimensional simplex is built on four points:

$$T = (p_1, p_2, p_3, p_4). \tag{4.3}$$

## Why the three-dimensional case is harder

The classical divide-and-conquer scheme, described for triangulation in the appendix, becomes much more involved in three dimensions. When stitching two parts in the plane, the gap between them is a pair of edge chains, and filling it is easy. In space the gap is formed by a whole *set of faces*, and it has to be closed with tetrahedra correctly. Moreover, the local Delaunay restructuring in the plane is a simple diagonal swap of a quadrilateral (the 2–2 flip), while in space the elementary restructurings (the 2–3 and 3–2 flips) change the very number of tetrahedra and require careful admissibility checks. Finally, unlike

the plane, not every three-dimensional domain can be partitioned into tetrahedra *without adding new points* — the classical example is the Schönhardt polyhedron.

For these reasons we do not implement the low-level tetrahedralization by hand but delegate it to the same CGAL library (see below), focusing on the *problem statement*: describing the domain, its boundary and the mesh quality requirements.

## The domain as a piecewise linear complex

The three-dimensional computational domain is specified by its *boundary* — a closed, oriented, triangulated surface. Such a description is called a piecewise linear complex (PLC): a set of vertices, edges and planar faces consistently covering the body's surface. It is the direct three-dimensional analogue of the boundary contours of the two-dimensional triangulation: there the boundary consisted of closed polylines, here — of closed triangular meshes.

For three characteristic bodies the surface is built explicitly:

- **cube** — eight vertices and twelve triangular faces (two per each of the six sides);
- **ball** — an icosphere: an icosahedron whose faces are repeatedly subdivided into four and projected onto the sphere;
- **hollow cylinder** — the outer and inner lateral surfaces plus two annular end caps; the result is a closed surface of genus 1 enclosing a body with a through hole.

## Constraints: edge and face recovery

The faces and edges of the boundary surface play the role of *constraints* — just like the non-crossable edges in the two-dimensional constrained Delaunay triangulation. They must be present in the final mesh: tetrahedra must not cross the body's boundary through their interiors, and sharp edges (the cube's edges, the rims of the cylinder caps) must be preserved as mesh edges rather than be "shaved off". Sharp edges are detected automatically by the dihedral angle between adjacent faces and are recovered separately, after which the surface remains embedded in the volumetric mesh.

## Quality criteria

Unlike the planar case, where the density was set merely by the number and placement of points, the volumetric mesh is driven by *sizing fields* and *shape criteria*:

- a bound on the tetrahedron size controls the mesh density: in CGAL it is set by the *cell_size* parameter, an upper bound on the circumradius of a mesh tetrahedron. In the demonstration utility the target scale is estimated as $\sqrt[3]{V/n}$, where $V$ is the body volume and $n$ is the prescribed number of points;
- a bound on the ratio of the circumscribed sphere radius to the shortest edge (the radius–edge ratio) prevents many elongated "needle" tetrahedra; the flat "slivers" that this criterion does not catch are removed by the final optimization;

- boundary-approximation criteria (the angle, size and distance error of boundary facets) control the accuracy of the surface mesh, while for the ball and the cylinder the accuracy also depends on how well the initial polyhedral surface approximates the intended curved surface.

The algorithm inserts interior points (Delaunay insertion) and refines the mesh until all criteria are met. A final optimization step then improves element shapes through vertex perturbation and sliver exudation, that is, vertex reweighting used to remove nearly flat tetrahedra.

## Implementation

The low-level construction is delegated to the *Mesh_3* module of the **CGAL** library — the same one used for the two-dimensional triangulation. The boundary surface of each body is represented using *Mesh_polyhedron_3<K>::type*, the domain by *Polyhedral_mesh_domain_with_features_3* (with automatic sharp-edge detection), and the tetrahedralization itself is performed by calling *make_mesh_3* with the criteria described above (*Mesh_criteria_3*). The exact-predicates kernel *Exact_predicates_inexact_constructions_kernel* is used, so the in-sphere and orientation checks are evaluated robustly. The output is a set of tetrahedral simplices of the form (4.3), from which the local matrices are later built and the global finite-element system is assembled.

## Examples

All meshes below are real outputs of the *tetrahedralize* utility of the computational core; the text gives the parameters of the full tetrahedral mesh. The interactive models show its surface shell: the surface can be rotated with the mouse or a finger, and the lighting-based shading emphasizes boundary facets of the tetrahedra.

The figure "Cube" shows the tetrahedralization of a unit cube with $n = 600$: 674 vertices, 2809 tetrahedra. The cube faces are flat, so all the mesh "curvature" goes inside, while the surface shows the triangular faces of the near-boundary tetrahedra; the cube edges are preserved as sharp features.

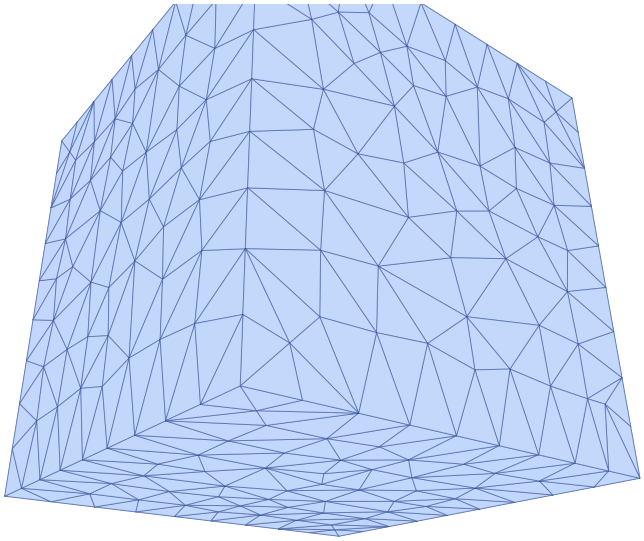

**Fig. 4.6.** Tetrahedralization of a cube (side 1, $n = 600$; 674 vertices, 2809 tetrahedra).
(interactive version on the website)

The figure "Ball" demonstrates surface approximation: the icosphere boundary of a ball ($R = 1$, $n = 800$; 737 vertices, 3417 tetrahedra) approximates the sphere with the prescribed accuracy, with no sharp edges.

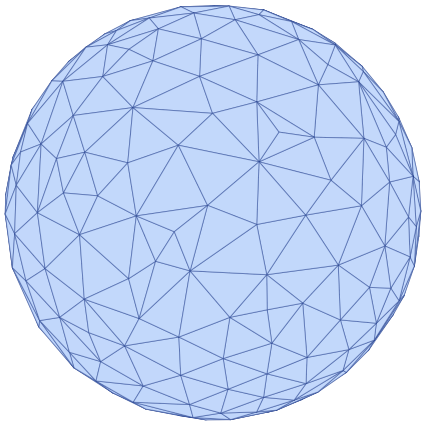

**Fig. 4.7.** Tetrahedralization of a ball ($R = 1$, $n = 800$; 737 vertices, 3417 tetrahedra): the boundary approximates the sphere with the prescribed accuracy.
(interactive version on the website)

Finally, the figure "Hollow Cylinder" shows a hollow cylinder ($R = 1$, $r = 0.4$, height 1.5) with $n = 1000$: 1118 vertices, 4588 tetrahedra. The through hole and the cap rims are preserved by the constraints: the sharp edges are detected by the dihedral angle and recovered in the mesh.

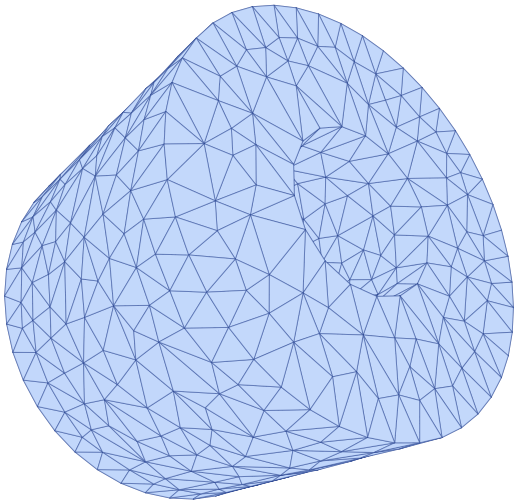

**Fig. 4.8.** Tetrahedralization of a hollow cylinder ($R = 1$, $r = 0.4$, $h = 1.5$, $n = 1000$; 1118 vertices, 4588 tetrahedra): the through hole and the cap rims are preserved.
(interactive version on the website)

# Chapter 5

# Formulation of the FEM problem

## Variational formulation of the problem

This chapter is the key one: here we lay the foundation of the finite element method. Having mastered it, you will be able to apply the same approach to equations of other types containing a linear operator.

We have considered the analytical approach to solving the heat conduction equation; however, for more complex geometries it quickly becomes cumbersome or altogether inapplicable, so now we proceed to our main topic — the numerical solution of the heat conduction equation. Let us write equation (1.8) for convenience, replacing $T$ by $u$

$$\frac{\partial u(M,t)}{\partial t} = a^2 \cdot \Delta u(M,t) + f(M,t).$$

We need to find a function $u(M,t)$ satisfying the equation and the boundary conditions; we first homogenize the boundary conditions, as was done in the analytical part: the inhomogeneity goes into the heat source density function and the initial condition, and is added back to the solution at the end. The problem itself can be represented as the Euler equation

$$L[u] = f, \tag{5.1}$$

where $L = \frac{\partial}{\partial t} - a^2 \cdot \Delta$ is the linear operator of the differential equation, $u$ is the unknown function, $f$ is the load function. As I said above, our equation also fits this form, which means we can apply the well-known approach to solving the Euler equation. That approach is the Bubnov–Galerkin method: in the weak form it is required that for any trial function $v(M,t)$ at any moment of time $t$ the following equality holds

$$\int_M L[u(M,t)] \cdot v(M,t)\, dM = \int_M f(M,t) \cdot v(M,t)\, dM. \tag{5.2}$$

Why is all this needed, and what are these trial functions? Above, when constructing the analytical solution, we saw how the solution is expanded into a Fourier series — the same thing happens here: if the functions are chosen orthogonal, the problem splits into several solutions independent of each other, which finally add up to the sought one. The freedom of choice is very large, so the functions can be taken as simple and convenient for computation as possible. One glance at the equation above makes it clear that the solution can be represented by the sum $u(M,t) = \sum_{i=1}^{n} a_i \cdot v_i(M,t)$, where $a_i$ are the coefficients to be found, and $v_i(M,t)$ are the basis functions; in the Bubnov–Galerkin method the

trial functions are taken from the same family. The approach to the numerical solution is analogous to the analytical one and essentially also reduces to finding the coefficients of known basis functions. I would say more: there is no fundamental distinction between analytical and numerical methods — they are essentially the same thing.

Writing integrals is inconvenient, so let us write (5.2) in the form of a scalar product

$$(L[u], v) = (f, v). \tag{5.3}$$

It is known that the Euler equation is related to the functional as follows

$$I(v) = (L[v], v) - 2 \cdot (f, v). \tag{5.4}$$

Minimizing this functional brings us back to the Euler equation, and the function $v$ on which the minimum is attained is precisely the sought solution $u$. Let us show this.

If on the function $u$ the functional attains its minimum, let us give it the increment $u + \epsilon \cdot v$, where $\epsilon$ is a small number, and $v$ is an arbitrary function, and carry out the transformations taking into account the obvious relation $I(u) \leq I(u + \epsilon \cdot v)$

$$I(u + \epsilon \cdot v) = (L[u + \epsilon \cdot v], u + \epsilon \cdot v) - 2 \cdot (f, u + \epsilon \cdot v),$$

$$I(u + \epsilon \cdot v) = I(u) + 2 \cdot \epsilon \cdot [(L[u], v) - (f, v)] + \epsilon^2 \cdot (L[v], v).$$

When expanding the brackets we used the self-adjointness of the operator: $(L[v], u) = (L[u], v)$. For the time derivative this is, strictly speaking, not the case — a careful treatment of this point is given in the appendix "The Bubnov–Galerkin functional". Since $\epsilon$ can be negative, for the minimum condition to hold it is necessary that the term linear in $\epsilon$ vanish, that is, that (5.3) holds. Thus, to solve the Euler equation by the Bubnov–Galerkin method, one needs to minimize the functional (5.4).

Solving equations analytically showed us the approach of separation of variables, so let us try to apply that knowledge here. Logically, nonstationarity adds no complexity related to the geometry: geometry separately, time separately. Ideally, solving the nonstationary equation should reduce to successively solving $N$ stationary problems. In the appendix "The Bubnov–Galerkin functional" we obtained the functionals whose minimization yields the solution of the nonstationary (I.6) and the stationary (I.5) problems.

## Form of the solution

An analytical solution for simple geometries can be obtained by many methods; one of them — the so-called Fourier method — we have already used in the section "General analytical solution of the boundary value problem". We assume in advance that the solution, even before it is obtained, is already expanded into a Fourier series — the expansions in the individual coordinates are considered mutually independent — and all that remains is to determine the specific values of the general solution (2.17). The difference is that in the Fourier method we partition the solution but not the geometry — which is why an analytical solution is possible only for a limited set of geometries — while in FEM we partition the geometry itself into simple elements, usually called simplices. The solution on these elements is approximated by so-called trial functions.

So, let us partition the space $M$ into simplex subspaces $\sigma_h$, which completely cover the original space. In each subspace there will exist its own solution $v_h = T(\sigma_h)$ — the restriction of the sought solution to the simplex, and the collection of such solutions will be the solution of the boundary value problem. Next, let us talk about simplices.

## Simplices

Before describing the solution, the domain on which it is sought must be partitioned into the simplest non-intersecting elements — simplices. The ways of building such partitions are discussed in detail in the chapter "Partitioning the geometry into simplices", so here we only briefly recall the result we will rely on.

In the one-dimensional case the simplex is a segment. The interval $l$ on which the solution is sought is split into $n$ simplex segments $s_i,\ i \in (1..n)$; unlike the finite difference method, their lengths may differ. The construction of these segments is described in the section "Segmentation". An example is shown in figure Segment simplices: the length of simplex $s_2$ equals $x_2 - x_1$.

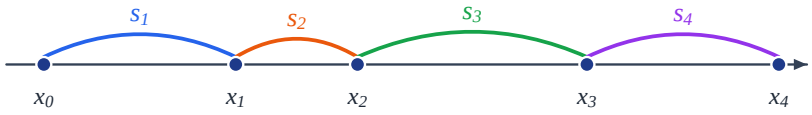


**Fig. 5.1.** One-dimensional segment simplices $s_1, \ldots, s_n$; their lengths may differ.

In the two-dimensional case the simplex is a triangle. The domain $S$ is partitioned into $n$ non-intersecting triangles $t_i,\ i \in (1..n)$ by the constrained Delaunay triangulation — see the section "Triangulation". An example is shown in figure Triangle simplices.

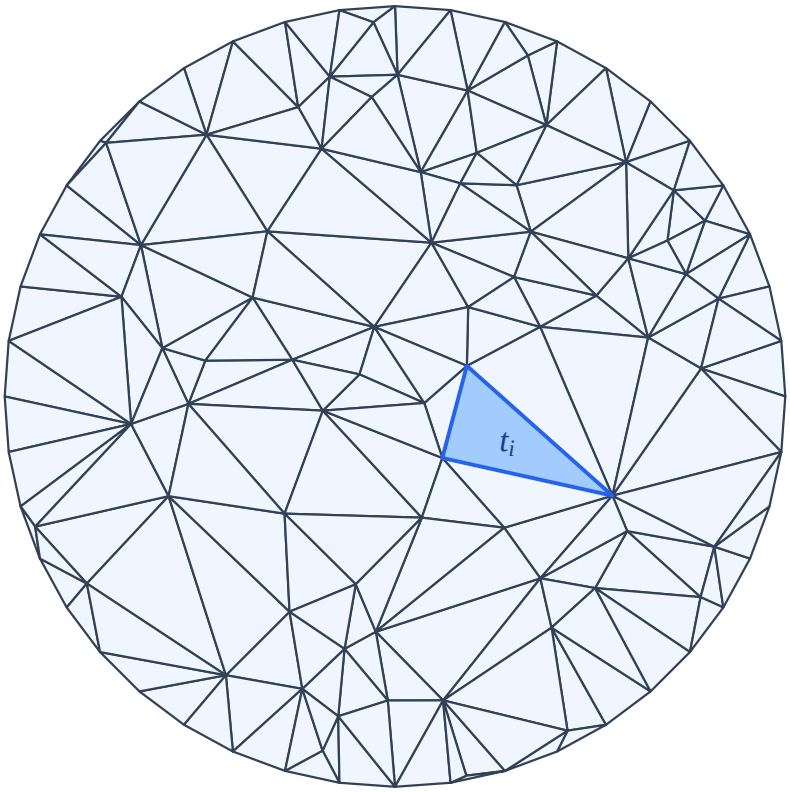


**Fig. 5.2.** A convex domain covered by non-intersecting triangle simplices (a real triangulation); one simplex is highlighted.

In the three-dimensional case the simplex is a tetrahedron. The domain $V$ is partitioned into $n$ non-intersecting tetrahedra $t_i,\ i \in (1..n)$; the corresponding procedure is discussed in the section "Tetrahedralization". The minimal simplex and its face-sharing neighbour are shown in figure Tetrahedron simplices.

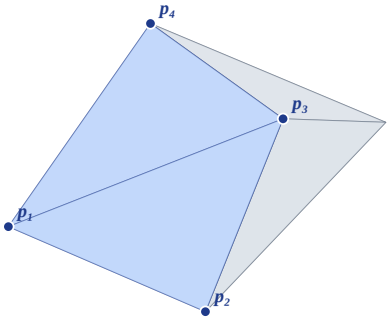


**Fig. 5.3.** A tetrahedral simplex on vertices $p_1, \ldots, p_4$ and its neighbour across the shared face (grey).

(interactive version on the website)

Once the domain is partitioned into simplices, the solution $v_h$ must be described on each of them.

## Test functions

The function that describes the behaviour on a simplex is called a test function — probably because in the early days of the method, given how weak computing hardware was, one had to try out and tune functions that would approximate the solution. There is a vast number of test functions — one can even use trigonometric ones — but in practice polynomial functions became the common choice, which hints at expanding the solution in a Taylor series, by analogy with the expansion in a Fourier series, with the sole difference that we carry out the Taylor expansion not over the whole geometry, but only on the simplices, which lets us use the simplest possible polynomials. As a rule, polynomials of degree 3 were chosen, with 9–10 coefficients, and this required adding extra points to the simplex in order to determine those coefficients, which led to 9–10 systems of linear equations on each simplex; but they had to be solved only once, which, computationally, is not a problem even by the standards of that time. This is, of course, a sound approach, but I decided to "try" building the solution with the simplest possible polynomials, namely polynomials of the form:

$$v_h(x, y, z) = \sum_{i=1}^{n} q_i \cdot \phi_i(x, y, z), \tag{5.5}$$

where $n$ is the number of nodes, $q_i$ are the coefficients, $\phi_i(x, y, z)$ are the so-called "hat" functions, which equal 1 at a node of the simplex and 0 at the remaining nodes of the geometry; and it is precisely because the "hat" functions are polynomials that the test functions, too, will be polynomials.

$$\phi_i(x, y, z) = a_i + b_i \cdot x + c_i \cdot y + d_i \cdot z, \tag{5.6}$$

where $i$ is the node index, not a power, $a_i, b_i, c_i, d_i$ are the coefficients that depend on the geometry of the particular node $i$ on the simplex. Once the geometry has been partitioned into simplices, they will need to be computed once for each node of the simplex. I wrote the "hat" function for a three-dimensional coordinate system, but further on we will work with the two-dimensional or one-dimensional case in the same way.

Let us discuss the "hat" functions in more detail and start with the one-dimensional case.

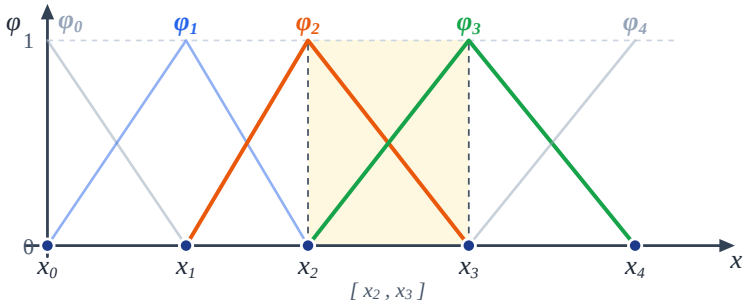


**Fig. 5.4.** "Hat" functions for segment simplices.

In figure "Hat" functions for segment simplices five "hat" functions are shown — $\phi_0, \ldots, \phi_4$ and four simplices: in total, for one-dimensional space, over $n$ points there are $n$ "hat" functions and $n-1$ simplices, respectively. At the ends, at the nodes $x_0$ and $x_4$, the functions take the form of half-"hats". Let us take $i = 3$; then the simplex between the points $x_{i-1}$ and $x_i$ (in the figure it is highlighted with a fill and bounded by dashed lines) is defined as follows: $\upsilon_{i-1}(x) = q_{i-1} \cdot \phi_{i-1}(x) + q_i \cdot \phi_i(x)$, since the remaining "hat" functions have no effect on it because they take the value zero on the interval $(x_{i-1}, x_i)$. The function $\phi_{i-1}(x)$ passes through the points $(x_{i-1}, 1)$ and $(x_i, 0)$, and the function $\phi_i(x)$ through the points $(x_{i-1}, 0)$ and $(x_i, 1)$. Then the following holds

$$\begin{cases} a_{i-1} + b_{i-1} \cdot x_{i-1} = 1 \\ a_{i-1} + b_{i-1} \cdot x_i = 0 \end{cases} \quad \begin{cases} a_i + b_i \cdot x_{i-1} = 0 \\ a_i + b_i \cdot x_i = 1 \end{cases} \tag{5.7}$$

The solutions of the systems are, respectively

$$\begin{cases} a_{i-1} = \dfrac{-x_i}{x_{i-1} - x_i} \\ b_{i-1} = \dfrac{1}{x_{i-1} - x_i} \end{cases} \quad \begin{cases} a_i = \dfrac{x_{i-1}}{x_{i-1} - x_i} \\ b_i = \dfrac{-1}{x_{i-1} - x_i} \end{cases} \tag{5.8}$$

If for one-dimensional simplices the "hat" functions look like lines in the plane, then for two-dimensional simplices they become surfaces in three-dimensional space.

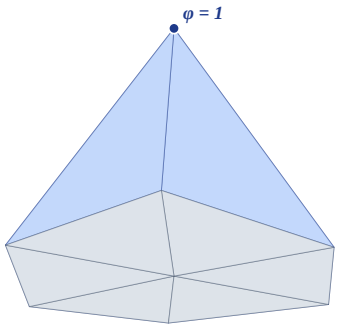


**Fig. 5.5.** "Hat" functions for triangle simplices.
(interactive version on the website)

We took all three vertices of the triangle $t_2$ from figure Triangle simplices and built three pyramids whose apexes are the vertices of the original triangle "raised" by 1, while the bases of the pyramids are the polygons from the same figure that contain the triangle's vertices inside them. Since the triangle under study is $t_2$ and we intend to write the test function for it, let us discard the extra parts and redraw the figure, keeping only the planes that project onto the triangle.

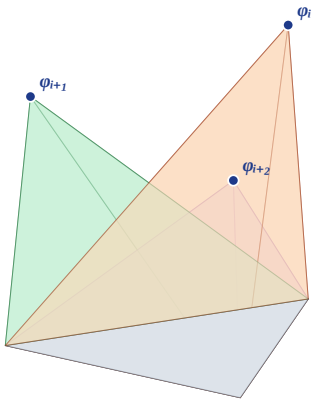


**Fig. 5.6.** Parts of the "hat" functions for a triangle simplex.
(interactive version on the website)

In figure Parts of the "hat" functions for a triangle simplex for two-dimensional space, just as for the one-dimensional case, to each of the $n$ mesh nodes there corresponds exactly one "hat" function, so there are $n$ of them in total; the number of triangle simplices, however, depends on the particular triangulation. Note also that within a single triangle exactly three "hat" functions are nonzero — one for each of its vertices. Let us denote the vertices of the simplex by $(x_i, y_i), (x_{i+1}, y_{i+1}), (x_{i+2}, y_{i+2})$. As a result, the test function takes the form $v_i(x, y) = q_i \cdot \phi_i(x, y) + q_{i+1} \cdot \phi_{i+1}(x, y) + q_{i+2} \cdot \phi_{i+2}(x, y)$. To determine the constants $a_i, b_i, c_i,\ a_{i+1}, b_{i+1}, c_{i+1},\ a_{i+2}, b_{i+2}, c_{i+2}$ let us take into account that the coordinate $z$ is nonzero only at the apexes of the pyramids; as a result, we obtain three systems of equations

$$\begin{cases} a_i + b_i \cdot x_i + c_i \cdot y_i = 1 \\ a_i + b_i \cdot x_{i+1} + c_i \cdot y_{i+1} = 0 \\ a_i + b_i \cdot x_{i+2} + c_i \cdot y_{i+2} = 0 \end{cases} \quad \begin{cases} a_{i+1} + b_{i+1} \cdot x_i + c_{i+1} \cdot y_i = 0 \\ a_{i+1} + b_{i+1} \cdot x_{i+1} + c_{i+1} \cdot y_{i+1} = 1 \\ a_{i+1} + b_{i+1} \cdot x_{i+2} + c_{i+1} \cdot y_{i+2} = 0 \end{cases} \quad \begin{cases} a_{i+2} + b_{i+2} \cdot x_i + c_{i+2} \cdot y_i = 0 \\ a_{i+2} + b_{i+2} \cdot x_{i+1} + c_{i+2} \cdot y_{i+1} = 0 \\ a_{i+2} + b_{i+2} \cdot x_{i+2} + c_{i+2} \cdot y_{i+2} = 1 \end{cases} \tag{5.9}$$

The solutions of the equations are, respectively

$$\begin{cases} a_i = (x_{i+1} \cdot y_{i+2} - x_{i+2} \cdot y_{i+1}) / \det \\ b_i = (y_{i+1} - y_{i+2}) / \det \\ c_i = (-x_{i+1} + x_{i+2}) / \det \end{cases} \quad \begin{cases} a_{i+1} = (-x_i \cdot y_{i+2} + x_{i+2} \cdot y_i) / \det \\ b_{i+1} = (-y_i + y_{i+2}) / \det \\ c_{i+1} = (x_i - x_{i+2}) / \det \end{cases} \quad \begin{cases} a_{i+2} = (x_i \cdot y_{i+1} - x_{i+1} \cdot y_i) / \det \\ b_{i+2} = (y_i - y_{i+1}) / \det \\ c_{i+2} = (-x_i + x_{i+1}) / \det \end{cases} \tag{5.10}$$

where $\det = x_i \cdot y_{i+1} - x_i \cdot y_{i+2} - x_{i+1} \cdot y_i + x_{i+1} \cdot y_{i+2} + x_{i+2} \cdot y_i - x_{i+2} \cdot y_{i+1}$ is the determinant of the matrix.

The difficulty of working with a three-dimensional coordinate system is that the solution, like the "hat" functions, is hard to visualise, so we will make do with formulas alone. In general, this guide places its main emphasis on two-dimensional boundary-value problems, since, unlike the one-dimensional ones, they are not trivial and at the same time visualise well. So, in three-dimensional space a tetrahedron serves as the base of a four-dimensional pentachoron-"hat". By analogy with the formulas above, we venture to assume that the "hat" function will take the value 1 at one of the vertices of the tetrahedron — raised into four-dimensional space perpendicular to the base — and be zero

at the remaining nodes of the tetrahedralization; the test function then takes the form $\upsilon_h(x,y,z) = q_1 \cdot \phi_1(x,y,z) + q_2 \cdot \phi_2(x,y,z) + q_3 \cdot \phi_3(x,y,z) + q_4 \cdot \phi_4(x,y,z)$. To determine the constants $a_i, b_i, c_i, d_i,\ a_{i+1}, b_{i+1}, c_{i+1}, d_{i+1},\ a_{i+2}, b_{i+2}, c_{i+2}, d_{i+2},\ a_{i+3}, b_{i+3}, c_{i+3}, d_{i+3}$ let us take into account that the coordinate $\mu$ of the fourth dimension is nonzero only at the apexes of the pentachora; as a result, we obtain four systems of equations

$$\begin{cases} a_i + b_i \cdot x_i + c_i \cdot y_i + d_i \cdot z_i = 1 \\ a_i + b_i \cdot x_{i+1} + c_i \cdot y_{i+1} + d_i \cdot z_{i+1} = 0 \\ a_i + b_i \cdot x_{i+2} + c_i \cdot y_{i+2} + d_i \cdot z_{i+2} = 0 \\ a_i + b_i \cdot x_{i+3} + c_i \cdot y_{i+3} + d_i \cdot z_{i+3} = 0 \end{cases} \quad \begin{cases} a_{i+1} + b_{i+1} \cdot x_i + c_{i+1} \cdot y_i + d_{i+1} \cdot z_i = 0 \\ a_{i+1} + b_{i+1} \cdot x_{i+1} + c_{i+1} \cdot y_{i+1} + d_{i+1} \cdot z_{i+1} = 1 \\ a_{i+1} + b_{i+1} \cdot x_{i+2} + c_{i+1} \cdot y_{i+2} + d_{i+1} \cdot z_{i+2} = 0 \\ a_{i+1} + b_{i+1} \cdot x_{i+3} + c_{i+1} \cdot y_{i+3} + d_{i+1} \cdot z_{i+3} = 0 \end{cases} \quad \begin{cases} a_{i+2} + b_{i+2} \cdot x_i + c_{i+2} \cdot y_i + d_{i+2} \cdot z_i = 0 \\ a_{i+2} + b_{i+2} \cdot x_{i+1} + c_{i+2} \cdot y_{i+1} + d_{i+2} \cdot z_{i+1} = 0 \\ a_{i+2} + b_{i+2} \cdot x_{i+2} + c_{i+2} \cdot y_{i+2} + d_{i+2} \cdot z_{i+2} = 1 \\ a_{i+2} + b_{i+2} \cdot x_{i+3} + c_{i+2} \cdot y_{i+3} + d_{i+2} \cdot z_{i+3} = 0 \end{cases} \quad \begin{cases} a_{i+3} + b_{i+3} \cdot x_i + c_{i+3} \cdot y_i + d_{i+3} \cdot z_i = 0 \\ a_{i+3} + b_{i+3} \cdot x_{i+1} + c_{i+3} \cdot y_{i+1} + d_{i+3} \cdot z_{i+1} = 0 \\ a_{i+3} + b_{i+3} \cdot x_{i+2} + c_{i+3} \cdot y_{i+2} + d_{i+3} \cdot z_{i+2} = 0 \\ a_{i+3} + b_{i+3} \cdot x_{i+3} + c_{i+3} \cdot y_{i+3} + d_{i+3} \cdot z_{i+3} = 1 \end{cases} \tag{5.11}$$

The solutions of the equations are, respectively

$$\begin{cases} a_i = (-x_{i+1} \cdot y_{i+2} \cdot z_{i+3} + x_{i+1} \cdot y_{i+3} \cdot z_{i+2} + x_{i+2} \cdot y_{i+1} \cdot z_{i+3} - x_{i+2} \cdot y_{i+3} \cdot z_{i+1} \\ \quad - x_{i+3} \cdot y_{i+1} \cdot z_{i+2} + x_{i+3} \cdot y_{i+2} \cdot z_{i+1})/\det \\ b_i = (y_{i+1} z_{i+2} - y_{i+2} z_{i+1} + y_{i+2} z_{i+3} + y_{i+3} z_{i+1} - y_{i+3} z_{i+2})/\det \\ c_i = (-x_{i+1} z_{i+2} + x_{i+1} z_{i+3} + x_{i+2} z_{i+1} - x_{i+2} z_{i+3} - x_{i+3} z_{i+1} + x_{i+3} z_{i+2})/\det \\ d_i = (x_{i+1} y_{i+2} - x_{i+1} y_{i+3} - x_{i+2} y_{i+1} + x_{i+2} y_{i+3} + x_{i+3} y_{i+1} - x_{i+3} y_{i+2})/\det \end{cases} \quad \begin{cases} a_{i+1} = (x_i \cdot y_{i+2} \cdot z_{i+3} - x_i \cdot y_{i+3} \cdot z_{i+2} - x_{i+2} \cdot y_i \cdot z_{i+3} + x_{i+2} \cdot y_{i+3} \cdot z_i \\ \quad + x_{i+3} \cdot y_i \cdot z_{i+2} - x_{i+3} \cdot y_{i+2} \cdot z_i)/\det \\ b_{i+1} = (-y_i z_{i+2} + y_i z_{i+3} + y_{i+2} z_i - y_{i+2} z_{i+3} - y_{i+3} z_i + y_{i+3} z_{i+2})/\det \\ c_{i+1} = (x_i z_{i+2} - x_i z_{i+3} - x_{i+2} z_i + x_{i+2} z_{i+3} + x_{i+3} z_i - x_{i+3} z_{i+2})/\det \\ d_{i+1} = (-x_i y_{i+2} + x_i y_{i+3} + x_{i+2} y_i - x_{i+2} y_{i+3} - x_{i+3} y_i + x_{i+3} y_{i+2})/\det \end{cases} \quad \begin{cases} a_{i+2} = (-x_i \cdot y_{i+1} \cdot z_{i+3} + x_i \cdot y_{i+3} \cdot z_{i+1} + x_{i+1} \cdot y_i \cdot z_{i+3} - x_{i+1} \cdot y_{i+3} \cdot z_i \\ \quad - x_{i+3} \cdot y_i \cdot z_{i+1} + x_{i+3} \cdot y_{i+1} \cdot z_i)/\det \\ b_{i+2} = (y_i z_{i+1} - y_i z_{i+3} - y_{i+1} z_i + y_{i+1} z_{i+3} + y_{i+3} z_i - y_{i+3} z_{i+1})/\det \\ c_{i+2} = (-x_i z_{i+1} + x_i z_{i+3} + x_{i+1} z_i - x_{i+1} z_{i+3} - x_{i+3} z_i + x_{i+3} z_{i+1})/\det \\ d_{i+2} = (x_i y_{i+1} - x_i y_{i+3} - x_{i+1} y_i + x_{i+1} y_{i+3} + x_{i+3} y_i - x_{i+3} y_{i+1})/\det \end{cases} \quad \begin{cases} a_{i+3} = (x_i \cdot y_{i+1} \cdot z_{i+2} - x_i \cdot y_{i+2} \cdot z_{i+1} - x_{i+1} \cdot y_i \cdot z_{i+2} + x_{i+1} \cdot y_{i+2} \cdot z_i \\ \quad + x_{i+2} \cdot y_i \cdot z_{i+1} - x_{i+2} \cdot y_{i+1} \cdot z_i)/\det \\ b_{i+3} = (-y_i z_{i+1} + y_i z_{i+2} + y_{i+1} z_i - y_{i+1} z_{i+2} - y_{i+2} z_i + y_{i+2} z_{i+1})/\det \\ c_{i+3} = (x_i z_{i+1} - x_i z_{i+2} - x_{i+1} z_i + x_{i+1} z_{i+2} + x_{i+2} z_i - x_{i+2} z_{i+1})/\det \\ d_{i+3} = (-x_i y_{i+1} + x_i y_{i+2} + x_{i+1} y_i - x_{i+1} y_{i+2} - x_{i+2} y_i + x_{i+2} y_{i+1})/\det \end{cases} \tag{5.12}$$

where the determinant of the matrix is:

$$\begin{aligned} \det = {} & x_i \cdot y_{i+1} \cdot z_{i+2} - x_i \cdot y_{i+1} \cdot z_{i+3} - x_i \cdot y_{i+2} \cdot z_{i+1} \\ & + x_i \cdot y_{i+2} \cdot z_{i+3} + x_i \cdot y_{i+3} \cdot z_{i+1} - x_i \cdot y_{i+3} \cdot z_{i+2} \\ & - x_{i+1} \cdot y_i \cdot z_{i+2} + x_{i+1} \cdot y_i \cdot z_{i+3} + x_{i+1} \cdot y_{i+2} \cdot z_i \\ & - x_{i+1} \cdot y_{i+2} \cdot z_{i+3} - x_{i+1} \cdot y_{i+3} \cdot z_i \\ & + x_{i+1} \cdot y_{i+3} \cdot z_{i+2} + x_{i+2} \cdot y_i \cdot z_{i+1} - x_{i+2} \cdot y_i \cdot z_{i+3} \\ & - x_{i+2} \cdot y_{i+1} \cdot z_i + x_{i+2} \cdot y_{i+1} \cdot z_{i+3} + x_{i+2} \cdot y_{i+3} \cdot z_i \\ & - x_{i+2} \cdot y_{i+3} \cdot z_{i+1} - x_{i+3} \cdot y_i \cdot z_{i+1} + x_{i+3} \cdot y_i \cdot z_{i+2} \\ & + x_{i+3} \cdot y_{i+1} \cdot z_i - x_{i+3} \cdot y_{i+1} \cdot z_{i+2} - x_{i+3} \cdot y_{i+2} \cdot z_i \\ & + x_{i+3} \cdot y_{i+2} \cdot z_{i+1} \end{aligned}$$

# Chapter 6

# Computing the matrices

## Stiffness matrix 1D

Let us compute the stiffness matrix for the one-dimensional case. In one-dimensional space the gradient has the form $\nabla v = \frac{\partial v}{\partial x}$, and the dot product of the gradient with itself is $\nabla v \cdot \nabla v = \left(\frac{\partial v}{\partial x}\right)^2$. Since the one-dimensional computational domain $M$ is partitioned into segment simplices, the part of the functional under study for a single segment $(x_i, x_{i+1})$ can be written as

$$\int_{x_i}^{x_{i+1}} \left(\frac{\partial v}{\partial x}\right)^2 dx. \tag{6.1}$$

In the general case the trial function $v(x) = \sum_{i=1}^{N} v_i(x)$, whereas on a simplex it has the form $v_{(i)(i+1)}(x) = q_i \cdot \phi_i(x) + q_{i+1} \cdot \phi_{i+1}(x)$. Taking into account (5.7) and (5.8) from the section on hat functions, we write the relations for the hat functions

$$\begin{cases} \phi_i(x) = a_i + b_i \cdot x \\ \phi_{i+1}(x) = a_{i+1} + b_{i+1} \cdot x \end{cases} \tag{6.2}$$

Compute the derivative of the trial function

$$\frac{\partial v_{(i)(i+1)}(x)}{\partial x} = q_i \cdot b_i + q_{i+1} \cdot b_{i+1} \tag{6.3}$$

Note that the derivative does not depend on $x$ and is constant on the segment. Substituting (6.3) into (6.1)

$$\int_{x_i}^{x_{i+1}} \left(\frac{\partial v_{(i)(i+1)}}{\partial x}\right)^2 dx = l_{(i)(i+1)} \cdot (q_i \cdot b_i + q_{i+1} \cdot b_{i+1})^2,$$

where $l_{(i)(i+1)} = x_{i+1} - x_i$ is the length of the segment.
Expand the square and regroup the terms

$$\int_{x_i}^{x_{i+1}} \left(\frac{\partial v_{(i)(i+1)}}{\partial x}\right)^2 dx = l_{(i)(i+1)} \cdot \left[q_i^2 \cdot b_i^2 + q_{i+1}^2 \cdot b_{i+1}^2 + 2 \cdot q_i \cdot q_{i+1} \cdot b_i \cdot b_{i+1}\right].$$

Let us take into account the formulas from (5.8) for the coefficients $b$ and introduce the notation for the elements of the local stiffness matrix of the segment

$$
\begin{aligned}
k_{(i)(i)} &= l_{(i)(i+1)} \cdot b_i^2 = \frac{1}{x_{i+1} - x_i} \\
k_{(i+1)(i+1)} &= l_{(i)(i+1)} \cdot b_{i+1}^2 = \frac{1}{x_{i+1} - x_i} \\
k_{(i)(i+1)} &= k_{(i+1)(i)} = l_{(i)(i+1)} \cdot b_i \cdot b_{i+1} = -\frac{1}{x_{i+1} - x_i}
\end{aligned} \tag{6.4}
$$

Thus, the local stiffness matrix for a one-dimensional element has the form

$$
\mathbf{K} = \begin{bmatrix} k_{(i)(i)} & k_{(i)(i+1)} \\ k_{(i+1)(i)} & k_{(i+1)(i+1)} \end{bmatrix} = \frac{1}{x_{i+1} - x_i} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{6.5}
$$

The global stiffness matrix $\mathbf{K}$ is obtained by summing the contributions from all segments of the mesh by the assembly method: the elements of the local matrices are added to the corresponding elements of the global matrix according to the global node numbering. The dimension of the global stiffness matrix is $N \times N$, where $N$ is the total number of mesh nodes.

For example, consider a mesh with nodes $0, 1, \ldots, i, i{+}1, \ldots, N$. During assembly each segment $(j)(j+1)$ contributes $\pm\frac{1}{l_{(j)(j+1)}}$, and an interior node $i$ receives a contribution from the two adjacent segments — $(i-1)(i)$ and $(i)(i+1)$ — so its main-diagonal entry is the sum $\frac{1}{l_{(i-1)(i)}} + \frac{1}{l_{(i)(i+1)}}$. Since the length $l_{(j)(j+1)}$ depends on the segment index, it cannot be taken out as a common factor — each entry of the global matrix keeps the length of its own segment. As a result the matrix is tridiagonal

$$
\mathbf{K} = \begin{bmatrix}
\frac{1}{l_{(0)(1)}} & -\frac{1}{l_{(0)(1)}} & 0 & \cdots & & & 0 \\
-\frac{1}{l_{(0)(1)}} & \frac{1}{l_{(0)(1)}} + \frac{1}{l_{(1)(2)}} & -\frac{1}{l_{(1)(2)}} & & & & \\
0 & -\frac{1}{l_{(1)(2)}} & \ddots & \ddots & & & \\
\vdots & & \ddots & \frac{1}{l_{(i-1)(i)}} + \frac{1}{l_{(i)(i+1)}} & -\frac{1}{l_{(i)(i+1)}} & & \vdots \\
 & & & -\frac{1}{l_{(i)(i+1)}} & \frac{1}{l_{(i)(i+1)}} + \frac{1}{l_{(i+1)(i+2)}} & \ddots & \\
 & & & & \ddots & \ddots & -\frac{1}{l_{(N-1)(N)}} \\
0 & & & \cdots & & -\frac{1}{l_{(N-1)(N)}} & \frac{1}{l_{(N-1)(N)}}
\end{bmatrix}.
$$

# Stiffness matrix 2D

Now let us compute the stiffness matrix for the two-dimensional case. In two-dimensional space the gradient has the form $\nabla\upsilon = \left(\frac{\partial\upsilon}{\partial x}, \frac{\partial\upsilon}{\partial y}\right)$, and the dot product of the gradient with itself is $\nabla\upsilon \cdot \nabla\upsilon = \left(\frac{\partial\upsilon}{\partial x}\right)^2 + \left(\frac{\partial\upsilon}{\partial y}\right)^2$. Since the two-dimensional computational domain $M$ is partitioned into triangular simplices, the part of the functional under study for a single triangle with vertices $(x_i, y_i), (x_{i+1}, y_{i+1}), (x_{i+2}, y_{i+2})$ can be written as

$$
\int_{\triangle} \left[ \left(\frac{\partial\upsilon}{\partial x}\right)^2 + \left(\frac{\partial\upsilon}{\partial y}\right)^2 \right] dS. \tag{6.6}
$$

The function $\upsilon(x,y) = \sum_{i=1}^{N} \upsilon_i(x,y)$. The trial function on the triangle has the form $\upsilon_{(i)(i+2)}(x,y) = q_i \cdot \phi_i(x,y) + q_{i+1} \cdot \phi_{i+1}(x,y) + q_{i+2} \cdot \phi_{i+2}(x,y)$. Taking into account (5.9) and (5.10) from the section on hat functions, we write the relations for the hat functions

$$\begin{cases} \phi_i(x,y) = a_i + b_i \cdot x + c_i \cdot y \\ \phi_{i+1}(x,y) = a_{i+1} + b_{i+1} \cdot x + c_{i+1} \cdot y \\ \phi_{i+2}(x,y) = a_{i+2} + b_{i+2} \cdot x + c_{i+2} \cdot y \end{cases} \tag{6.7}$$

Let us compute the partial derivatives of the trial function

$$\begin{cases} \dfrac{\partial \upsilon_{(i)(i+2)}(x,y)}{\partial x} = q_i \cdot b_i + q_{i+1} \cdot b_{i+1} + q_{i+2} \cdot b_{i+2} \\ \dfrac{\partial \upsilon_{(i)(i+2)}(x,y)}{\partial y} = q_i \cdot c_i + q_{i+1} \cdot c_{i+1} + q_{i+2} \cdot c_{i+2} \end{cases} \tag{6.8}$$

Note that the derivatives do not depend on $x$ and $y$ and are constants on the triangle. Substituting (6.8) into (6.6)

$$\int_{\triangle} \left[ \left( \frac{\partial \upsilon_{(i)(i+2)}}{\partial x} \right)^2 + \left( \frac{\partial \upsilon_{(i)(i+2)}}{\partial y} \right)^2 \right] dS = \quad S_{\triangle} \cdot \left[ (q_i \cdot b_i + q_{i+1} \cdot b_{i+1} + q_{i+2} \cdot b_{i+2})^2 + (q_i \cdot c_i + q_{i+1} \cdot c_{i+1} + q_{i+2} \cdot c_{i+2})^2 \right],$$

where $S_{\triangle}$ is the area of the triangle, which is computed by the formula

$$S_{\triangle} = \frac{|d|}{2}, \tag{6.9}$$

where $d = x_i \cdot y_{i+1} - x_i \cdot y_{i+2} - x_{i+1} \cdot y_i + x_{i+1} \cdot y_{i+2} + x_{i+2} \cdot y_i - x_{i+2} \cdot y_{i+1}$ is the determinant from (5.10).

Expand the squares and regroup the terms

$$\begin{aligned} \int_{\triangle} (\nabla \upsilon_{(i)(i+2)})^2 \, dS = S_{\triangle} \cdot \Big[ & q_i^2 \cdot [b_i^2 + c_i^2] \\ & + q_{i+1}^2 \cdot [b_{i+1}^2 + c_{i+1}^2] + q_{i+2}^2 \cdot [b_{i+2}^2 + c_{i+2}^2] \\ & + 2 \cdot q_i \cdot q_{i+1} \cdot (b_i \cdot b_{i+1} + c_i \cdot c_{i+1}) \\ & + 2 \cdot q_i \cdot q_{i+2} \cdot (b_i \cdot b_{i+2} + c_i \cdot c_{i+2}) \\ & + 2 \cdot q_{i+1} \cdot q_{i+2} \cdot (b_{i+1} \cdot b_{i+2} + c_{i+1} \cdot c_{i+2}) \Big] \end{aligned}$$

Taking into account the formulas from (5.10) for the coefficients $b$ and $c$ and introducing the notation for the elements of the local stiffness matrix of the triangle

$$
\begin{aligned}
k_{(i)(i)} &= S_{\triangle} \cdot [b_i^2 + c_i^2] = \frac{1}{2 \cdot |d|} \cdot [(y_{i+1} - y_{i+2})^2 + (-x_{i+1} + x_{i+2})^2] \\
k_{(i+1)(i+1)} &= S_{\triangle} \cdot [b_{i+1}^2 + c_{i+1}^2] = \frac{1}{2 \cdot |d|} \cdot [(-y_i + y_{i+2})^2 + (x_i - x_{i+2})^2] \\
k_{(i+2)(i+2)} &= S_{\triangle} \cdot [b_{i+2}^2 + c_{i+2}^2] = \frac{1}{2 \cdot |d|} \cdot [(y_i - y_{i+1})^2 + (-x_i + x_{i+1})^2] \\
k_{(i)(i+1)} &= k_{(i+1)(i)} = S_{\triangle} \cdot (b_i \cdot b_{i+1} + c_i \cdot c_{i+1}) \\
&= \frac{1}{2 \cdot |d|} \cdot [(y_{i+1} - y_{i+2}) \cdot (-y_i + y_{i+2}) + (-x_{i+1} + x_{i+2}) \cdot (x_i - x_{i+2})] \\
k_{(i)(i+2)} &= k_{(i+2)(i)} = S_{\triangle} \cdot (b_i \cdot b_{i+2} + c_i \cdot c_{i+2}) \\
&= \frac{1}{2 \cdot |d|} \cdot [(y_{i+1} - y_{i+2}) \cdot (y_i - y_{i+1}) + (-x_{i+1} + x_{i+2}) \cdot (-x_i + x_{i+1})] \\
k_{(i+1)(i+2)} &= k_{(i+2)(i+1)} = S_{\triangle} \cdot (b_{i+1} \cdot b_{i+2} + c_{i+1} \cdot c_{i+2}) \\
&= \frac{1}{2 \cdot |d|} \cdot [(-y_i + y_{i+2}) \cdot (y_i - y_{i+1}) + (x_i - x_{i+2}) \cdot (-x_i + x_{i+1})]
\end{aligned}
\tag{6.10}
$$

Thus, the local stiffness matrix for a triangular element has the form

$$
\mathbf{K}_{\triangle} = \begin{bmatrix} k_{(i)(i)} & k_{(i)(i+1)} & k_{(i)(i+2)} \\ k_{(i+1)(i)} & k_{(i+1)(i+1)} & k_{(i+1)(i+2)} \\ k_{(i+2)(i)} & k_{(i+2)(i+1)} & k_{(i+2)(i+2)} \end{bmatrix} = S_{\triangle} \begin{bmatrix} b_i^2 + c_i^2 & b_i \cdot b_{i+1} + c_i \cdot c_{i+1} & b_i \cdot b_{i+2} + c_i \cdot c_{i+2} \\ b_{i+1} \cdot b_i + c_{i+1} \cdot c_i & b_{i+1}^2 + c_{i+1}^2 & b_{i+1} \cdot b_{i+2} + c_{i+1} \cdot c_{i+2} \\ b_{i+2} \cdot b_i + c_{i+2} \cdot c_i & b_{i+2} \cdot b_{i+1} + c_{i+2} \cdot c_{i+1} & b_{i+2}^2 + c_{i+2}^2 \end{bmatrix}. \tag{6.11}
$$

The global stiffness matrix $\mathbf{K}$ is obtained by summing the contributions from all triangular elements of the mesh using the assembly method: the elements of the local matrices are added to the corresponding elements of the global matrix according to the global node numbering. The dimension of the global stiffness matrix is equal to $N \times N$, where $N$ is the total number of mesh nodes.

# Stiffness matrix 3D

Let us proceed to computing the stiffness matrix for the three-dimensional case. In three-dimensional space the gradient has the form $\nabla \upsilon = \left(\frac{\partial \upsilon}{\partial x}, \frac{\partial \upsilon}{\partial y}, \frac{\partial \upsilon}{\partial z}\right)$, and the dot product of the gradient with itself is $\nabla \upsilon \cdot \nabla \upsilon = \left(\frac{\partial \upsilon}{\partial x}\right)^2 + \left(\frac{\partial \upsilon}{\partial y}\right)^2 + \left(\frac{\partial \upsilon}{\partial z}\right)^2$. Since the three-dimensional computational domain $M$ is partitioned into tetrahedral simplices, the part of the functional under study for a single tetrahedron with vertices

$$
(x_i, y_i, z_i), \quad (x_{i+1}, y_{i+1}, z_{i+1}), \quad (x_{i+2}, y_{i+2}, z_{i+2}), \quad (x_{i+3}, y_{i+3}, z_{i+3})
$$

can be written as

$$
\int_{\text{tet}} \left[ \left(\frac{\partial \upsilon}{\partial x}\right)^2 + \left(\frac{\partial \upsilon}{\partial y}\right)^2 + \left(\frac{\partial \upsilon}{\partial z}\right)^2 \right] dV. \tag{6.12}
$$

The function $\upsilon(x,y,z) = \sum_{i=1}^{N} \upsilon_i(x,y,z)$. The trial function on the tetrahedron has the form $\upsilon_{(i)(i+3)}(x,y,z) = q_i \cdot \phi_i(x,y,z) + q_{i+1} \cdot \phi_{i+1}(x,y,z) + q_{i+2} \cdot \phi_{i+2}(x,y,z) + q_{i+3} \cdot \phi_{i+3}(x,y,z)$. Analogously to the two-dimensional case, the hat functions for the tetrahedron are linear

$$\begin{cases} \phi_i(x,y,z) = a_i + b_i \cdot x + c_i \cdot y + d_i \cdot z \\ \phi_{i+1}(x,y,z) = a_{i+1} + b_{i+1} \cdot x + c_{i+1} \cdot y + d_{i+1} \cdot z \\ \phi_{i+2}(x,y,z) = a_{i+2} + b_{i+2} \cdot x + c_{i+2} \cdot y + d_{i+2} \cdot z \\ \phi_{i+3}(x,y,z) = a_{i+3} + b_{i+3} \cdot x + c_{i+3} \cdot y + d_{i+3} \cdot z \end{cases} \tag{6.13}$$

Let us compute the partial derivatives of the trial function

$$\begin{cases} \dfrac{\partial \upsilon_{(i)(i+3)}(x,y,z)}{\partial x} = q_i \cdot b_i + q_{i+1} \cdot b_{i+1} + q_{i+2} \cdot b_{i+2} + q_{i+3} \cdot b_{i+3} \\ \dfrac{\partial \upsilon_{(i)(i+3)}(x,y,z)}{\partial y} = q_i \cdot c_i + q_{i+1} \cdot c_{i+1} + q_{i+2} \cdot c_{i+2} + q_{i+3} \cdot c_{i+3} \\ \dfrac{\partial \upsilon_{(i)(i+3)}(x,y,z)}{\partial z} = q_i \cdot d_i + q_{i+1} \cdot d_{i+1} + q_{i+2} \cdot d_{i+2} + q_{i+3} \cdot d_{i+3} \end{cases} \tag{6.14}$$

Note that the derivatives do not depend on $x$, $y$ and $z$ and are constants over the tetrahedron. Substituting (6.14) into (6.12)

$$\begin{aligned} \int_{\text{tet}} \left[ \left( \frac{\partial \upsilon_{(i)(i+3)}}{\partial x} \right)^2 + \left( \frac{\partial \upsilon_{(i)(i+3)}}{\partial y} \right)^2 + \left( \frac{\partial \upsilon_{(i)(i+3)}}{\partial z} \right)^2 \right] dV \\ = \quad V_{\text{tet}} \cdot \Big[ (q_i \cdot b_i + q_{i+1} \cdot b_{i+1} + q_{i+2} \cdot b_{i+2} \\ + q_{i+3} \cdot b_{i+3})^2 \\ + (q_i \cdot c_i + q_{i+1} \cdot c_{i+1} + q_{i+2} \cdot c_{i+2} + q_{i+3} \cdot c_{i+3})^2 \\ + (q_i \cdot d_i + q_{i+1} \cdot d_{i+1} + q_{i+2} \cdot d_{i+2} + q_{i+3} \cdot d_{i+3})^2 \Big], \end{aligned}$$

where $V_{\text{tet}}$ is the volume of the tetrahedron, which is computed by the formula

$$V_{\text{tet}} = \frac{|\Delta|}{6}, \tag{6.15}$$

where $\Delta$ is the determinant of the matrix

$$\Delta = \begin{vmatrix} x_i & y_i & z_i & 1 \\ x_{i+1} & y_{i+1} & z_{i+1} & 1 \\ x_{i+2} & y_{i+2} & z_{i+2} & 1 \\ x_{i+3} & y_{i+3} & z_{i+3} & 1 \end{vmatrix}. \tag{6.16}$$

Let us expand the squares and regroup the terms

$$\begin{aligned}\int_{\text{tet}} (\nabla \upsilon_{(i)(i+3)})^2 \, dV = V_{\text{tet}} \cdot \Big[ & q_i^2 \cdot [b_i^2 + c_i^2 + d_i^2] \\ & + q_{i+1}^2 \cdot [b_{i+1}^2 + c_{i+1}^2 + d_{i+1}^2] \\ & + q_{i+2}^2 \cdot [b_{i+2}^2 + c_{i+2}^2 + d_{i+2}^2] \\ & + q_{i+3}^2 \cdot [b_{i+3}^2 + c_{i+3}^2 + d_{i+3}^2] \\ & + 2 \cdot q_i \cdot q_{i+1} \cdot (b_i \cdot b_{i+1} + c_i \cdot c_{i+1} + d_i \cdot d_{i+1}) \\ & + 2 \cdot q_i \cdot q_{i+2} \cdot (b_i \cdot b_{i+2} + c_i \cdot c_{i+2} + d_i \cdot d_{i+2}) \\ & + 2 \cdot q_i \cdot q_{i+3} \cdot (b_i \cdot b_{i+3} + c_i \cdot c_{i+3} + d_i \cdot d_{i+3}) \\ & + 2 \cdot q_{i+1} \cdot q_{i+2} \cdot (b_{i+1} \cdot b_{i+2} + c_{i+1} \cdot c_{i+2} \\ & + d_{i+1} \cdot d_{i+2}) + 2 \cdot q_{i+1} \cdot q_{i+3} \cdot (b_{i+1} \cdot b_{i+3} \\ & + c_{i+1} \cdot c_{i+3} + d_{i+1} \cdot d_{i+3}) \\ & + 2 \cdot q_{i+2} \cdot q_{i+3} \cdot (b_{i+2} \cdot b_{i+3} + c_{i+2} \cdot c_{i+3} \\ & + d_{i+2} \cdot d_{i+3}) \Big]\end{aligned}$$

The coefficients $b_k$, $c_k$ and $d_k$ for each vertex $k$ of the tetrahedron are computed through the minors of the determinant $\Delta$. For the vertex with index $k$ the coefficients have the form

$$\begin{aligned} b_k &= (-1)^{k+1} \cdot \frac{\Delta_k^{(x)}}{\Delta} \\ c_k &= (-1)^{k+2} \cdot \frac{\Delta_k^{(y)}}{\Delta} \\ d_k &= (-1)^{k+3} \cdot \frac{\Delta_k^{(z)}}{\Delta} \end{aligned} \tag{6.17}$$

where $\Delta_k^{(x)}$, $\Delta_k^{(y)}$ and $\Delta_k^{(z)}$ are the minors obtained by deleting the $k$-th row and the corresponding column ($x$, $y$ or $z$) from the matrix (6.16).

We introduce the notation for the elements of the local stiffness matrix of the tetrahedron

$$\begin{aligned}k_{mn} = V_{\text{tet}} \cdot (b_m \cdot b_n + c_m \cdot c_n + d_m \cdot d_n), \quad m, n \in \{i, i+1, i \\ + 2, i+3\}.\end{aligned} \tag{6.18}$$

Thus, the local stiffness matrix for a tetrahedral element has the form

$$\mathbf{K}_{\text{tet}} = \begin{bmatrix} k_{(i)(i)} & k_{(i)(i+1)} & k_{(i)(i+2)} & k_{(i)(i+3)} \\ k_{(i+1)(i)} & k_{(i+1)(i+1)} & k_{(i+1)(i+2)} & k_{(i+1)(i+3)} \\ k_{(i+2)(i)} & k_{(i+2)(i+1)} & k_{(i+2)(i+2)} & k_{(i+2)(i+3)} \\ k_{(i+3)(i)} & k_{(i+3)(i+1)} & k_{(i+3)(i+2)} & k_{(i+3)(i+3)} \end{bmatrix}. \tag{6.19}$$

The local stiffness matrix is symmetric, that is $k_{mn}^{\text{loc}} = k_{nm}^{\text{loc}}$. The global stiffness matrix $\mathbf{K}$ is obtained by summing the contributions from all tetrahedral elements of the mesh by the assembly method: the elements of the local matrices are added to the corresponding elements of the global matrix according to the global node numbering. The dimension of the global stiffness matrix is $N \times N$, where $N$ is the total number of mesh nodes.

# Damping matrix 1D

Now let us compute the damping matrix for the one-dimensional case. The damping matrix is related to the integral of the square of the trial function. Consider the integral for a single segment with vertices $x_i, x_{i+1}$

$$\int_{x_i}^{x_{i+1}} \upsilon^2 \, dx. \tag{6.20}$$

The trial function on the segment has the form $\upsilon_{(i)(i+1)}(x) = q_i \cdot \phi_i(x) + q_{i+1} \cdot \phi_{i+1}(x)$. Take into account (5.7) and (5.8) from the section on hat functions and write down the relations for the hat functions

$$\begin{cases} \phi_i(x) = a_i + b_i \cdot x \\ \phi_{i+1}(x) = a_{i+1} + b_{i+1} \cdot x \end{cases} \tag{6.21}$$

Substitute the trial function into (6.20)

$$\int_{x_i}^{x_{i+1}} \upsilon_{(i)(i+1)}^2 \, dx = \int_{x_i}^{x_{i+1}} \left[ q_i \cdot (a_i + b_i \cdot x) + q_{i+1} \cdot (a_{i+1} + b_{i+1} \cdot x) \right]^2 \, dx.$$

Taking into account (5.8), we can write

$$\int_{x_i}^{x_{i+1}} \upsilon_{(i)(i+1)}^2 \, dx = \frac{1}{(x_i - x_{i+1})^2} \cdot \int_{x_i}^{x_{i+1}} \left[ (-q_i \cdot x_{i+1} + q_{i+1} \cdot x_i) + (q_i - q_{i+1}) \cdot x \right]^2 \, dx.$$

Expand the square and split the integral into three parts

$$\begin{aligned} \int_{x_i}^{x_{i+1}} \upsilon_{(i)(i+1)}^2 \, dx &= \frac{1}{(x_i - x_{i+1})^2} \cdot \Big[ \int_{x_i}^{x_{i+1}} (-q_i \cdot x_{i+1} + q_{i+1} \cdot x_i)^2 \, dx \\ &+ 2 \cdot \int_{x_i}^{x_{i+1}} (-q_i \cdot x_{i+1} + q_{i+1} \cdot x_i) \cdot (q_i - q_{i+1}) \cdot x \, dx \\ &+ \int_{x_i}^{x_{i+1}} (q_i - q_{i+1})^2 \cdot x^2 \, dx \Big] \end{aligned}$$

Let us compute each of the three integrals

$$\begin{aligned} \int_{x_i}^{x_{i+1}} \upsilon_{(i)(i+1)}^2 \, dx &= \frac{(-q_i \cdot x_{i+1} + q_{i+1} \cdot x_i)^2 \cdot x}{(x_i - x_{i+1})^2} \Big|_{x_i}^{x_{i+1}} \\ &+ \frac{(-q_i \cdot x_{i+1} + q_{i+1} \cdot x_i) \cdot (q_i - q_{i+1}) \cdot x^2}{(x_i - x_{i+1})^2} \Big|_{x_i}^{x_{i+1}} \\ &+ \frac{(q_i - q_{i+1})^2 \cdot x^3}{3 \cdot (x_i - x_{i+1})^2} \Big|_{x_i}^{x_{i+1}}. \end{aligned}$$

After simplification we obtain

$$\int_{x_i}^{x_{i+1}} \upsilon_{(i)(i+1)}^2 \, dx = -\frac{(q_i^2 + q_i \cdot q_{i+1} + q_{i+1}^2) \cdot (x_i - x_{i+1})^2}{3 \cdot (x_i - x_{i+1})} = \frac{x_{i+1} - x_i}{3} \cdot (q_i^2 + q_i \cdot q_{i+1} + q_{i+1}^2).$$

We introduce the notation for the segment length

$$\int_{x_i}^{x_{i+1}} \upsilon_{(i)(i+1)}^2 \, dx = \frac{l_{(i)(i+1)}}{3} \cdot \left[q_i^2 + q_i \cdot q_{i+1} + q_{i+1}^2\right],$$

where $l_{(i)(i+1)} = x_{i+1} - x_i$ is the length of the segment.
We introduce the notation for the elements of the local damping matrix of the segment

$$\begin{aligned} c_{(i)(i)} &= \frac{l_{(i)(i+1)}}{3} \\ c_{(i+1)(i+1)} &= \frac{l_{(i)(i+1)}}{3} \\ c_{(i)(i+1)} &= c_{(i+1)(i)} = \frac{l_{(i)(i+1)}}{6} \end{aligned} \tag{6.22}$$

Thus, the local damping matrix for a one-dimensional element has the form

$$\mathbf{C} = \begin{bmatrix} c_{(i)(i)} & c_{(i)(i+1)} \\ c_{(i+1)(i)} & c_{(i+1)(i+1)} \end{bmatrix} = \frac{l_{(i)(i+1)}}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \tag{6.23}$$

The global damping matrix $\mathbf{C}$ is obtained by summing the contributions from all segments of the mesh using the assembly method: the elements of the local matrices are added to the corresponding elements of the global matrix according to the global node numbering. The dimension of the global damping matrix is $N \times N$, where $N$ is the total number of nodes in the mesh.

For example, consider a mesh with nodes $0, 1, \ldots, i, i+1, \ldots, N$. During assembly each segment $(j)(j+1)$ adds $\frac{l_{(j)(j+1)}}{3}$ to the diagonal and $\frac{l_{(j)(j+1)}}{6}$ to the adjacent off-diagonal entries; an interior node $i$ receives a diagonal contribution from the two adjacent segments — $(i-1)(i)$ and $(i)(i+1)$ — so its main-diagonal entry is the sum $\frac{l_{(i-1)(i)}}{3} + \frac{l_{(i)(i+1)}}{3}$. Since the length $l_{(j)(j+1)}$ depends on the segment index, it cannot be taken out as a common factor — each entry of the global matrix keeps the length of its own segment. As a result the matrix is tridiagonal

$$\mathbf{C} = \begin{bmatrix} \frac{l_{(0)(1)}}{3} & \frac{l_{(0)(1)}}{6} & 0 & \cdots & & & 0 \\ \frac{l_{(0)(1)}}{6} & \frac{l_{(0)(1)}}{3} + \frac{l_{(1)(2)}}{3} & \frac{l_{(1)(2)}}{6} & & & & \\ 0 & \frac{l_{(1)(2)}}{6} & \ddots & \ddots & & & \\ \vdots & & \ddots & \frac{l_{(i-1)(i)}}{3} + \frac{l_{(i)(i+1)}}{3} & \frac{l_{(i)(i+1)}}{6} & & \vdots \\ & & & \frac{l_{(i)(i+1)}}{6} & \frac{l_{(i)(i+1)}}{3} + \frac{l_{(i+1)(i+2)}}{3} & \ddots & \\ & & & & \ddots & \ddots & \frac{l_{(N-1)(N)}}{6} \\ 0 & & & \cdots & & \frac{l_{(N-1)(N)}}{6} & \frac{l_{(N-1)(N)}}{3} \end{bmatrix}.$$

# Damping matrix 2D

Now we compute the damping matrix for the two-dimensional case. The damping matrix is related to the integral of the square of the trial function. Consider the integral over a single triangle with vertices $(x_i, y_i), (x_{i+1}, y_{i+1}), (x_{i+2}, y_{i+2})$

$$\int_{\triangle} v^2 \, dS. \tag{6.24}$$

The trial function on the triangle has the form $v_{(i)(i+2)}(x, y) = q_i \cdot \phi_i(x, y) + q_{i+1} \cdot \phi_{i+1}(x, y) + q_{i+2} \cdot \phi_{i+2}(x, y)$. We take into account (5.9) and (5.10) from the section on hat functions and write the relations for the hat functions

$$\begin{cases} \phi_i(x, y) = a_i + b_i \cdot x + c_i \cdot y \\ \phi_{i+1}(x, y) = a_{i+1} + b_{i+1} \cdot x + c_{i+1} \cdot y \\ \phi_{i+2}(x, y) = a_{i+2} + b_{i+2} \cdot x + c_{i+2} \cdot y \end{cases} \tag{6.25}$$

Substitute the trial function into (6.24)

$$\int_{\triangle} v^2_{(i)(i+2)} \, dS = \int_{\triangle} \Big[ q_i \cdot (a_i + b_i \cdot x + c_i \cdot y) + q_{i+1} \cdot (a_{i+1} + b_{i+1} \cdot x + c_{i+1} \cdot y) + q_{i+2} \cdot (a_{i+2} + b_{i+2} \cdot x + c_{i+2} \cdot y) \Big]^2 \, dS$$

Expand the square. To simplify the computations, we use the fact that for linear hat functions on the triangle the following relations hold

$$\int_{\triangle} \phi_m \cdot \phi_n \, dS = \begin{cases} \frac{S_{\triangle}}{6}, & m = n \\ \frac{S_{\triangle}}{12}, & m \neq n \end{cases} \tag{6.26}$$

where $S_{\triangle}$ is the area of the triangle, which is computed by formula (6.9).
Using these relations and expanding the square of the trial function, we obtain

$$\int_{\triangle} v^2_{(i)(i+2)}\,dS = \int_{\triangle} \Big[q_i^2\cdot\phi_i^2 + q_{i+1}^2\cdot\phi_{i+1}^2 + q_{i+2}^2\cdot\phi_{i+2}^2 + 2\cdot q_i\cdot q_{i+1}\cdot\phi_i\cdot\phi_{i+1} + 2\cdot q_i\cdot q_{i+2}\cdot\phi_i\cdot\phi_{i+2} + 2\cdot q_{i+1}\cdot q_{i+2}\cdot\phi_{i+1}\cdot\phi_{i+2}\Big]\,dS$$

Apply the formulas (6.26)

$$\int_{\triangle} v^2_{(i)(i+2)}\,dS = q_i^2\cdot\frac{S_\triangle}{6} + q_{i+1}^2\cdot\frac{S_\triangle}{6} + q_{i+2}^2\cdot\frac{S_\triangle}{6} + 2\cdot q_i\cdot q_{i+1}\cdot\frac{S_\triangle}{12} + 2\cdot q_i\cdot q_{i+2}\cdot\frac{S_\triangle}{12} + 2\cdot q_{i+1}\cdot q_{i+2}\cdot\frac{S_\triangle}{12}$$

Simplify the expression

$$\int_{\triangle} v^2_{(i)(i+2)}\,dS = \frac{S_\triangle}{12}\left[2q_i^2 + 2q_{i+1}^2 + 2q_{i+2}^2 + 2q_iq_{i+1} + 2q_iq_{i+2} + 2q_{i+1}q_{i+2}\right].$$

We introduce the notation for the elements of the local damping matrix of the triangle

$$\begin{aligned} c_{(i)(i)} &= \frac{S_\triangle}{6} \\ c_{(i+1)(i+1)} &= \frac{S_\triangle}{6} \\ c_{(i+2)(i+2)} &= \frac{S_\triangle}{6} \\ c_{(i)(i+1)} = c_{(i+1)(i)} &= \frac{S_\triangle}{12} \\ c_{(i)(i+2)} = c_{(i+2)(i)} &= \frac{S_\triangle}{12} \\ c_{(i+1)(i+2)} = c_{(i+2)(i+1)} &= \frac{S_\triangle}{12} \end{aligned} \tag{6.27}$$

Thus, the local damping matrix for the triangular element has the form

$$\mathbf{C}_\triangle = \begin{bmatrix} c_{(i)(i)} & c_{(i)(i+1)} & c_{(i)(i+2)} \\ c_{(i+1)(i)} & c_{(i+1)(i+1)} & c_{(i+1)(i+2)} \\ c_{(i+2)(i)} & c_{(i+2)(i+1)} & c_{(i+2)(i+2)} \end{bmatrix} = \frac{S_\triangle}{12}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}. \tag{6.28}$$

The global damping matrix $\mathbf{C}$ is obtained by summing the contributions from all triangular elements of the mesh using the assembly method: the elements of the local matrices are added to the corresponding elements of the global matrix according to the global node numbering. The dimension of the global damping matrix is equal to $N \times N$, where $N$ is the total number of nodes in the mesh.

## Damping matrix 3D

Let us proceed to computing the damping matrix for the three-dimensional case. The damping matrix is related to the integral of the square of the trial function. Consider the integral for a single tetrahedron with vertices

$$(x_i, y_i, z_i), \quad (x_{i+1}, y_{i+1}, z_{i+1}), \quad (x_{i+2}, y_{i+2}, z_{i+2}), \quad (x_{i+3}, y_{i+3}, z_{i+3})$$

$$\int_{\text{tet}} v^2 \, dV. \tag{6.29}$$

The trial function on the tetrahedron has the form $v_{(i)(i+3)}(x, y, z) = q_i \cdot \phi_i + q_{i+1} \cdot \phi_{i+1} + q_{i+2} \cdot \phi_{i+2} + q_{i+3} \cdot \phi_{i+3}$. Analogously to the two-dimensional case, the hat functions for the tetrahedron are linear according to (6.13)

$$\begin{cases} \phi_i(x, y, z) = a_i + b_i \cdot x + c_i \cdot y + d_i \cdot z \\ \phi_{i+1}(x, y, z) = a_{i+1} + b_{i+1} \cdot x + c_{i+1} \cdot y + d_{i+1} \cdot z \\ \phi_{i+2}(x, y, z) = a_{i+2} + b_{i+2} \cdot x + c_{i+2} \cdot y + d_{i+2} \cdot z \\ \phi_{i+3}(x, y, z) = a_{i+3} + b_{i+3} \cdot x + c_{i+3} \cdot y + d_{i+3} \cdot z \end{cases} \tag{6.30}$$

Substitute the trial function into (6.29)

$$\int_{\text{tet}} v^2_{(i)(i+3)} \, dV = \int_{\text{tet}} \Big[ q_i \cdot \phi_i(x, y, z) + q_{i+1} \cdot \phi_{i+1}(x, y, z) + q_{i+2} \cdot \phi_{i+2}(x, y, z) + q_{i+3} \cdot \phi_{i+3}(x, y, z) \Big]^2 \, dV$$

For linear hat functions on the tetrahedron the following relations hold

$$\int_{\text{tet}} \phi_m \cdot \phi_n \, dV = \begin{cases} \frac{V_{\text{tet}}}{10}, & m = n \\ \frac{V_{\text{tet}}}{20}, & m \neq n \end{cases} \tag{6.31}$$

where $V_{\text{tet}}$ is the volume of the tetrahedron, which is computed by formula (6.15). Expand the square of the trial function

$$\begin{aligned} \int_{\text{tet}} v^2_{(i)(i+3)} \, dV = \int_{\text{tet}} \Big[ & q_i^2 \cdot \phi_i^2 + q_{i+1}^2 \cdot \phi_{i+1}^2 \\ & + q_{i+2}^2 \cdot \phi_{i+2}^2 + q_{i+3}^2 \cdot \phi_{i+3}^2 + 2 \cdot q_i \cdot q_{i+1} \cdot \phi_i \cdot \phi_{i+1} \\ & + 2 \cdot q_i \cdot q_{i+2} \cdot \phi_i \cdot \phi_{i+2} + 2 \cdot q_i \cdot q_{i+3} \cdot \phi_i \cdot \phi_{i+3} \\ & + 2 \cdot q_{i+1} \cdot q_{i+2} \cdot \phi_{i+1} \cdot \phi_{i+2} \\ & + 2 \cdot q_{i+1} \cdot q_{i+3} \cdot \phi_{i+1} \cdot \phi_{i+3} \\ & + 2 \cdot q_{i+2} \cdot q_{i+3} \cdot \phi_{i+2} \cdot \phi_{i+3} \Big] \, dV \end{aligned}$$

Apply the formulas (6.31)

$$\int_{\text{tet}} v^2_{(i)(i+3)}\, dV = q_i^2 \cdot \frac{V_{\text{tet}}}{10} + q_{i+1}^2 \cdot \frac{V_{\text{tet}}}{10} + q_{i+2}^2 \cdot \frac{V_{\text{tet}}}{10} + q_{i+3}^2 \cdot \frac{V_{\text{tet}}}{10} + 2 \cdot q_i \cdot q_{i+1} \cdot \frac{V_{\text{tet}}}{20} + 2 \cdot q_i \cdot q_{i+2} \cdot \frac{V_{\text{tet}}}{20} + 2 \cdot q_i \cdot q_{i+3} \cdot \frac{V_{\text{tet}}}{20} + 2 \cdot q_{i+1} \cdot q_{i+2} \cdot \frac{V_{\text{tet}}}{20} + 2 \cdot q_{i+1} \cdot q_{i+3} \cdot \frac{V_{\text{tet}}}{20} + 2 \cdot q_{i+2} \cdot q_{i+3} \cdot \frac{V_{\text{tet}}}{20}$$

Simplify the expression

$$\int_{\text{tet}} v^2_{(i)(i+3)}\, dV = \frac{V_{\text{tet}}}{20} \cdot \Big[2 \cdot q_i^2 + 2 \cdot q_{i+1}^2 + 2 \cdot q_{i+2}^2 + 2 \cdot q_{i+3}^2 + 2 \cdot q_i \cdot q_{i+1} + 2 \cdot q_i \cdot q_{i+2} + 2 \cdot q_i \cdot q_{i+3} + 2 \cdot q_{i+1} \cdot q_{i+2} + 2 \cdot q_{i+1} \cdot q_{i+3} + 2 \cdot q_{i+2} \cdot q_{i+3}\Big]$$

We introduce the notation for the elements of the local damping matrix of the tetrahedron

$$\begin{aligned} c_{(i)(i)} &= \frac{V_{\text{tet}}}{10} \\ c_{(i+1)(i+1)} &= \frac{V_{\text{tet}}}{10} \\ c_{(i+2)(i+2)} &= \frac{V_{\text{tet}}}{10} \\ c_{(i+3)(i+3)} &= \frac{V_{\text{tet}}}{10} \\ c_{(m)(n)} = c_{(n)(m)} &= \frac{V_{\text{tet}}}{20}, \quad m \neq n, \quad m, n \in \{i, i+1, i+2, i+3\} \end{aligned} \tag{6.32}$$

Thus, the local damping matrix for the tetrahedral element has the form

$$\mathbf{C}_{\text{tet}} = \begin{bmatrix} c_{(i)(i)} & c_{(i)(i+1)} & c_{(i)(i+2)} & c_{(i)(i+3)} \\ c_{(i+1)(i)} & c_{(i+1)(i+1)} & c_{(i+1)(i+2)} & c_{(i+1)(i+3)} \\ c_{(i+2)(i)} & c_{(i+2)(i+1)} & c_{(i+2)(i+2)} & c_{(i+2)(i+3)} \\ c_{(i+3)(i)} & c_{(i+3)(i+1)} & c_{(i+3)(i+2)} & c_{(i+3)(i+3)} \end{bmatrix} = \frac{V_{\text{tet}}}{20} \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}. \tag{6.33}$$

The local damping matrix is symmetric. The global damping matrix $\mathbf{C}$ is obtained by summing the contributions from all tetrahedral elements of the mesh using the assembly method: the elements of the local matrices are added to the corresponding elements of the global matrix according to the global node numbering. The dimension of the global damping matrix is $N \times N$, where $N$ is the total number of mesh nodes.

# Load vector 1D

Now let us compute the load vector for the one-dimensional case. The load vector is related to the integral of the product of the source function and the trial function. Consider the integral for a single segment with vertices $x_i, x_{i+1}$

$$\int_{x_i}^{x_{i+1}} f(x) \cdot \upsilon \, dx. \tag{6.34}$$

The trial function on the segment has the form $\upsilon(x) = q_i \cdot \phi_i(x) + q_{i+1} \cdot \phi_{i+1}(x)$. To evaluate the integral, it is necessary to interpolate the source function $f(x)$ on the segment $[x_i, x_{i+1}]$ by a linear function. Let us represent $f(x)$ in the form $\widetilde{f}(x) = e_1 + e_2 \cdot x$. The coefficients $e_1$ and $e_2$ are determined from the interpolation conditions at the nodes

$$\begin{cases} e_1 + e_2 \cdot x_i = f_i \\ e_1 + e_2 \cdot x_{i+1} = f_{i+1} \end{cases} \tag{6.35}$$

Solving this system analogously to (5.8), we obtain

$$\begin{cases} e_1 = \dfrac{f_i \cdot x_{i+1} - f_{i+1} \cdot x_i}{x_{i+1} - x_i} \\ e_2 = \dfrac{f_{i+1} - f_i}{x_{i+1} - x_i} \end{cases} \tag{6.36}$$

Let us substitute the interpolated source function and the trial function into (6.34). Taking into account (5.7) and (5.8) from the section on hat functions

$$\int_{x_i}^{x_{i+1}} f(x) \cdot \upsilon \, dx = \int_{x_i}^{x_{i+1}} (e_1 + e_2 \cdot x) \cdot [q_i \cdot (a_i + b_i \cdot x) + q_{i+1} \cdot (a_{i+1} + b_{i+1} \cdot x)] \, dx.$$

Expand the brackets and split the integral into parts

$$\int_{x_i}^{x_{i+1}} f(x) \cdot \upsilon \, dx = \int_{x_i}^{x_{i+1}} \Big[ e_1 \cdot (q_i \cdot a_i + q_{i+1} \cdot a_{i+1}) + e_1 \cdot (q_i \cdot b_i + q_{i+1} \cdot b_{i+1}) \cdot x + e_2 \cdot (q_i \cdot a_i + q_{i+1} \cdot a_{i+1}) \cdot x + e_2 \cdot (q_i \cdot b_i + q_{i+1} \cdot b_{i+1}) \cdot x^2 \Big] \, dx$$

Let us compute each of the integrals

$$\int_{x_i}^{x_{i+1}} f(x) \cdot \upsilon \, dx = e_1 \cdot (q_i \cdot a_i + q_{i+1} \cdot a_{i+1}) \cdot x \Big|_{x_i}^{x_{i+1}} + [e_1 \cdot (q_i \cdot b_i + q_{i+1} \cdot b_{i+1}) + e_2 \cdot (q_i \cdot a_i + q_{i+1} \cdot a_{i+1})] \cdot \frac{x^2}{2} \Big|_{x_i}^{x_{i+1}} + e_2 \cdot (q_i \cdot b_i + q_{i+1} \cdot b_{i+1}) \cdot \frac{x^3}{3} \Big|_{x_i}^{x_{i+1}}$$

Substituting the expressions for $e_1, e_2$ from (6.36) and using the relations (5.8), after simplification we obtain

$$\int_{x_i}^{x_{i+1}} f(x) \cdot \upsilon \, dx = \frac{(f_i + f_{i+1}) \cdot (x_{i+1} - x_i)}{2} \cdot (q_i \cdot a_i + q_{i+1} \cdot a_{i+1}) + \frac{x_{i+1} - x_i}{6} \cdot [f_i \cdot x_i + f_{i+1} \cdot x_{i+1} + (f_i + f_{i+1}) \cdot (x_i + x_{i+1})] \cdot (q_i \cdot b_i + q_{i+1} \cdot b_{i+1})$$

Further transformations taking into account (5.8) lead to the final result

$$\int_{x_i}^{x_{i+1}} f(x) \cdot \upsilon \, dx = \frac{x_{i+1} - x_i}{6} \cdot [q_i \cdot (2 \cdot f_i + f_{i+1}) + q_{i+1} \cdot (f_i + 2 \cdot f_{i+1})] .$$

Expand and regroup the terms

$$\int_{x_i}^{x_{i+1}} f(x) \cdot \upsilon \, dx = l_{(i)(i+1)} \cdot \left[ \frac{2 \cdot f_i + f_{i+1}}{6} \cdot q_i + \frac{f_i + 2 \cdot f_{i+1}}{6} \cdot q_{i+1} \right],$$

where $l_{(i)(i+1)} = x_{i+1} - x_i$ is the length of the segment.
We introduce the notation for the elements of the local load vector of the segment

$$\begin{aligned} r_i &= \frac{l_{(i)(i+1)}}{6} \cdot (2 \cdot f_i + f_{i+1}) \\ r_{i+1} &= \frac{l_{(i)(i+1)}}{6} \cdot (f_i + 2 \cdot f_{i+1}) \end{aligned} \tag{6.37}$$

Thus, the local load vector for a one-dimensional element has the form

$$\mathbf{R} = \begin{bmatrix} r_i \\ r_{i+1} \end{bmatrix} = \frac{l_{(i)(i+1)}}{6} \begin{bmatrix} 2 \cdot f_i + f_{i+1} \\ f_i + 2 \cdot f_{i+1} \end{bmatrix}. \tag{6.38}$$

The global load vector $\mathbf{R}$ is obtained by summing the contributions from all segments of the mesh using assembly: the elements of the local vectors are added to the corresponding elements of the global vector according to the global numbering of the nodes. The dimension of the global load vector equals $N$, where $N$ is the total number of nodes in the mesh. It is important to note that for interior nodes the contributions from two adjacent elements are summed, as a result of which the element of the global load vector for an interior node $j$ takes the form

$$R_j = \frac{l_{(j-1)(j)}}{6} \cdot (f_{j-1} + 2 \cdot f_j) + \frac{l_{(j)(j+1)}}{6} \cdot (2 \cdot f_j + f_{j+1}), \tag{6.39}$$

where $l_{(j-1)(j)} = x_j - x_{j-1}$ and $l_{(j)(j+1)} = x_{j+1} - x_j$ are the lengths of the adjacent segments. For boundary nodes the element contributions are assembled in the same way, but the corresponding components of the system may be modified when boundary conditions are imposed.

For example, for a mesh with nodes $0, 1, \ldots, i, i+1, \ldots, N$ the global load vector, assembled from the local contributions by rule (6.39) for interior nodes, takes the form (before boundary conditions are applied)

$\mathbf{R}$

$$= \begin{bmatrix} \frac{l_{(0)(1)}}{6} \cdot (2 \cdot f_0 + f_1) \\ \frac{l_{(0)(1)}}{6} \cdot (f_0 + 2 \cdot f_1) + \frac{l_{(1)(2)}}{6} \cdot (2 \cdot f_1 + f_2) \\ \vdots \\ \frac{l_{(i-1)(i)}}{6} \cdot (f_{i-1} + 2 \cdot f_i) + \frac{l_{(i)(i+1)}}{6} \cdot (2 \cdot f_i + f_{i+1}) \\ \frac{l_{(i)(i+1)}}{6} \cdot (f_i + 2 \cdot f_{i+1}) + \frac{l_{(i+1)(i+2)}}{6} \cdot (2 \cdot f_{i+1} + f_{i+2}) \\ \vdots \\ \frac{l_{(N-1)(N)}}{6} \cdot (f_{N-1} + 2 \cdot f_N) \end{bmatrix}.$$

# Load vector 2D

Now we compute the load vector for the two-dimensional case. The load vector is related to the integral of the product of the source function and the trial function. Consider the integral for a single triangle with vertices $(x_i, y_i), (x_{i+1}, y_{i+1}), (x_{i+2}, y_{i+2})$

$$\int_{\triangle} f(x,y) \cdot \upsilon \, dS. \tag{6.40}$$

The trial function on the triangle has the form $\upsilon_{(i)(i+2)}(x,y) = q_i \cdot \phi_i(x,y) + q_{i+1} \cdot \phi_{i+1}(x,y) + q_{i+2} \cdot \phi_{i+2}(x,y)$. To evaluate the integral, the source function $f(x,y)$ must be interpolated on the triangle $\triangle$ by a linear function. We represent $f(x,y)$ in the form $\widetilde{f}(x,y) = e_1 + e_2 \cdot x + e_3 \cdot y$. The coefficients $e_1, e_2, e_3$ are determined from the interpolation conditions at the nodes

$$\begin{cases} e_1 + e_2 \cdot x_i + e_3 \cdot y_i = f_i \\ e_1 + e_2 \cdot x_{i+1} + e_3 \cdot y_{i+1} = f_{i+1} \\ e_1 + e_2 \cdot x_{i+2} + e_3 \cdot y_{i+2} = f_{i+2} \end{cases} \tag{6.41}$$

Since the hat functions $\phi_n$ equal one at their own node and zero at the others, the linear interpolant of the source function, expressed through them, takes the form

$$\widetilde{f}(x,y) = f_i \cdot \phi_i(x,y) + f_{i+1} \cdot \phi_{i+1}(x,y) + f_{i+2} \cdot \phi_{i+2}(x,y), \tag{6.42}$$

where $f_i, f_{i+1}, f_{i+2}$ are the values of the source function at the vertices of the triangle.

Substitute the interpolated source function and the trial function into (6.40). We take into account (6.25) from the section on hat functions

$$\int_{\triangle} f(x,y) \cdot \upsilon \, dS = \int_{\triangle} \left( \sum_n f_n \cdot \phi_n \right) \cdot \left( \sum_m q_m \cdot \phi_m \right) dS$$
$$= \sum_m \sum_n q_m \cdot f_n \cdot \int_{\triangle} \phi_m \cdot \phi_n \, dS,$$

where the indices $m$ and $n$ run over the vertices of the triangle $\{i, i+1, i+2\}$.

We use the relations (6.26) for the integrals of products of hat functions, derived when computing the damping matrix, where $S_\triangle$ is the area of the triangle, which is computed by formula (6.9). After collecting like terms we obtain

$$\int_\triangle f(x,y)\cdot v\,dS = \frac{S_\triangle}{12}\cdot\Big[q_i\cdot(2\cdot f_i + f_{i+1} + f_{i+2}) + q_{i+1}\cdot(f_i + 2\cdot f_{i+1} + f_{i+2}) + q_{i+2}\cdot(f_i + f_{i+1} + 2\cdot f_{i+2})\Big]$$

We introduce the notation for the elements of the local load vector of the triangle

$$\begin{aligned} r_i &= \frac{S_\triangle}{12}\cdot(2\cdot f_i + f_{i+1} + f_{i+2}) \\ r_{i+1} &= \frac{S_\triangle}{12}\cdot(f_i + 2\cdot f_{i+1} + f_{i+2}) \\ r_{i+2} &= \frac{S_\triangle}{12}\cdot(f_i + f_{i+1} + 2\cdot f_{i+2}) \end{aligned} \tag{6.43}$$

Thus, the local load vector for a two-dimensional triangular element has the form

$$\mathbf{R}_\triangle = \begin{bmatrix} r_i \\ r_{i+1} \\ r_{i+2} \end{bmatrix} = \frac{S_\triangle}{12}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}\begin{bmatrix} f_i \\ f_{i+1} \\ f_{i+2} \end{bmatrix}. \tag{6.44}$$

Note that the resulting load vector coincides with the product of the local damping matrix (6.28) and the vector of nodal source values, that is $\mathbf{R}_\triangle = \mathbf{C}_\triangle\cdot\mathbf{f}$.

The global load vector $\mathbf{R}$ is obtained by summing the contributions from all triangular elements of the mesh using the assembly method: the elements of the local vectors are added to the corresponding elements of the global vector according to the global node numbering. The dimension of the global load vector equals $N$, where $N$ is the total number of nodes in the mesh. It is important to note that for interior nodes the contributions from all adjacent triangular elements containing the given node are summed. For boundary nodes the element contributions are assembled in the same way, but the corresponding components of the system may be modified when boundary conditions are imposed.

## Load vector 3D

Now we compute the load vector for the three-dimensional case. The load vector is related to the integral of the product of the source function and the trial function. Consider the integral for a single tetrahedron with vertices

$$(x_i, y_i, z_i),\quad (x_{i+1}, y_{i+1}, z_{i+1}),\quad (x_{i+2}, y_{i+2}, z_{i+2}),\quad (x_{i+3}, y_{i+3}, z_{i+3})$$

$$\int_{\text{tet}} f(x,y,z)\cdot v\,dV. \tag{6.45}$$

The trial function on the tetrahedron has the form $\upsilon_{(i)(i+3)}(x,y,z) = q_i \cdot \phi_i(x,y,z) + q_{i+1} \cdot \phi_{i+1}(x,y,z) + q_{i+2} \cdot \phi_{i+2}(x,y,z) + q_{i+3} \cdot \phi_{i+3}(x,y,z)$. To evaluate the integral, the source function $f(x,y,z)$ must be interpolated on the tetrahedron by a linear function. Let us represent $f(x,y,z)$ in the form $\widetilde{f}(x,y,z) = e_1 + e_2 \cdot x + e_3 \cdot y + e_4 \cdot z$. The coefficients $e_1, e_2, e_3, e_4$ are determined from the interpolation conditions at the nodes

$$\begin{cases} e_1 + e_2 \cdot x_i + e_3 \cdot y_i + e_4 \cdot z_i = f_i \\ e_1 + e_2 \cdot x_{i+1} + e_3 \cdot y_{i+1} + e_4 \cdot z_{i+1} = f_{i+1} \\ e_1 + e_2 \cdot x_{i+2} + e_3 \cdot y_{i+2} + e_4 \cdot z_{i+2} = f_{i+2} \\ e_1 + e_2 \cdot x_{i+3} + e_3 \cdot y_{i+3} + e_4 \cdot z_{i+3} = f_{i+3} \end{cases} \tag{6.46}$$

Since the hat functions $\phi_n$ equal one at their own node and zero at the others, the linear interpolant of the source function, expressed through them, takes the form

$$\widetilde{f}(x,y,z) = f_i \cdot \phi_i + f_{i+1} \cdot \phi_{i+1} + f_{i+2} \cdot \phi_{i+2} + f_{i+3} \cdot \phi_{i+3}, \tag{6.47}$$

where $f_i, f_{i+1}, f_{i+2}, f_{i+3}$ are the values of the source function at the vertices of the tetrahedron.

Substitute the interpolated source function and the trial function into (6.45). We take into account (6.30) from the section on hat functions

$$\int_{\text{tet}} f(x,y,z) \cdot \upsilon \, dV = \int_{\text{tet}} \left( \sum_n f_n \cdot \phi_n \right) \cdot \left( \sum_m q_m \cdot \phi_m \right) dV = \sum_m \sum_n q_m \cdot f_n \cdot \int_{\text{tet}} \phi_m \cdot \phi_n \, dV,$$

where the indices $m$ and $n$ run over the vertices of the tetrahedron $\{i, i+1, i+2, i+3\}$.

We use the relations (6.31) for the integrals of products of hat functions, derived when computing the damping matrix, where $V_{\text{tet}}$ is the volume of the tetrahedron, which is computed by formula (6.15). After collecting like terms we obtain

$$\begin{aligned} \int_{\text{tet}} f(x,y,z) \cdot \upsilon \, dV &= \frac{V_{\text{tet}}}{20} \cdot \Big[ q_i \cdot (2 \cdot f_i + f_{i+1} + f_{i+2} + f_{i+3}) \\ &+ q_{i+1} \cdot (f_i + 2 \cdot f_{i+1} + f_{i+2} + f_{i+3}) \\ &+ q_{i+2} \cdot (f_i + f_{i+1} + 2 \cdot f_{i+2} + f_{i+3}) \\ &+ q_{i+3} \cdot (f_i + f_{i+1} + f_{i+2} + 2 \cdot f_{i+3}) \Big] \end{aligned}$$

We introduce the notation for the elements of the local load vector of the tetrahedron

$$
\begin{aligned}
r_i &= \frac{V_{\text{tet}}}{20} \cdot (2 \cdot f_i + f_{i+1} + f_{i+2} + f_{i+3}) \\
r_{i+1} &= \frac{V_{\text{tet}}}{20} \cdot (f_i + 2 \cdot f_{i+1} + f_{i+2} + f_{i+3}) \\
r_{i+2} &= \frac{V_{\text{tet}}}{20} \cdot (f_i + f_{i+1} + 2 \cdot f_{i+2} + f_{i+3}) \\
r_{i+3} &= \frac{V_{\text{tet}}}{20} \cdot (f_i + f_{i+1} + f_{i+2} + 2 \cdot f_{i+3})
\end{aligned}
\tag{6.48}
$$

Thus, the local load vector for a three-dimensional tetrahedral element has the form

$$
\mathbf{R}_{\text{tet}} = \begin{bmatrix} r_i \\ r_{i+1} \\ r_{i+2} \\ r_{i+3} \end{bmatrix} = \frac{V_{\text{tet}}}{20} \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} f_i \\ f_{i+1} \\ f_{i+2} \\ f_{i+3} \end{bmatrix}. \tag{6.49}
$$

Note that the resulting load vector coincides with the product of the local damping matrix (6.33) and the vector of nodal source values, that is $\mathbf{R}_{\text{tet}} = \mathbf{C}_{\text{tet}} \cdot \mathbf{f}$.

The global load vector $\mathbf{R}$ is obtained by summing the contributions from all tetrahedral elements of the mesh using the assembly method: the elements of the local vectors are added to the corresponding elements of the global vector according to the global node numbering. The dimension of the global load vector equals $N$, where $N$ is the total number of nodes in the mesh. It is important to note that for interior nodes the contributions from all adjacent tetrahedral elements containing the given node are summed. For boundary nodes the element contributions are assembled in the same way, but the corresponding components of the system may be modified when boundary conditions are imposed.

# Chapter 7

# Examples

## Thermal field 1D, Dirichlet

### Example: a transient Dirichlet problem

Consider a concrete example of a transient Dirichlet problem. Take a rod on the segment $[\alpha, \beta] = [0, 1]$ with thermal diffusivity $a^2 = 0.05$. The rod is initially cold, and the end temperatures are not imposed at once but grow over time: the ends heat up linearly from zero to 20 and 80, respectively, over the first 5 seconds and are then held. Inside the rod we place a point heat source $f(x) = P \cdot \delta(x - x_0)$ of power $P = 10$ at the point $x_0 = 0.35$. The problem statement reads

$$\begin{cases} \dfrac{\partial T(x,t)}{\partial t} = a^2 \cdot \dfrac{\partial^2 T(x,t)}{\partial x^2} + P \cdot \delta(x - x_0), \\ T(x,0) = 0, \\ T(\alpha,t) = \varphi_\alpha(t), \\ T(\beta,t) = \varphi_\beta(t), \end{cases} \tag{7.1}$$

where the end temperatures depend on time

$$\varphi_\alpha(t) = \begin{cases} 4t, & 0 \le t \le 5, \\ 20, & t > 5, \end{cases} \qquad \varphi_\beta(t) = \begin{cases} 16t, & 0 \le t \le 5, \\ 80, & t > 5. \end{cases}$$

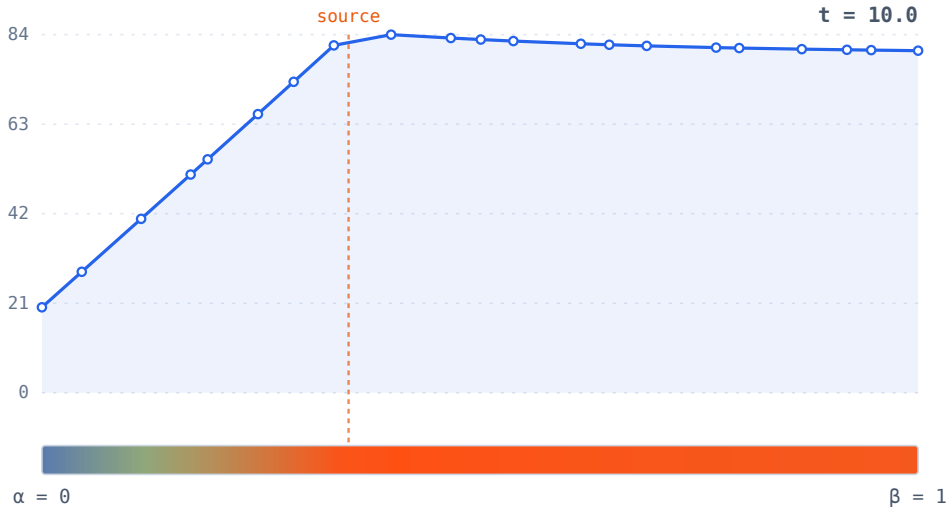


**Fig. 7.1.** Transient Dirichlet problem: heating of the rod over time. The blue curve is the temperature profile $T(x,t)$, the orange marker is the point source.
(interactive version on the website)

# Thermal field 2D, Dirichlet

## Example: a plate with a heat source

Consider a two-dimensional example with an internal heat source. Take a square plate $\Omega = [0,1] \times [0,1]$ with thermal diffusivity $a^2 = 0.004$. The edges of the plate are held at zero temperature (Dirichlet condition), and four point heat sources are placed near the corners, $f(x,y) = \sum_{k=1}^{4} P_k \cdot \delta(x - x_k,\, y - y_k)$. Initially the plate is heated non-uniformly. Over time, the initial heat escapes through the cold edges while the sources sustain hot regions, and the solution settles to a steady profile. The problem statement reads

$$\begin{cases} \dfrac{\partial T(x,y,t)}{\partial t} = a^2 \cdot \left( \dfrac{\partial^2 T(x,y,t)}{\partial x^2} + \dfrac{\partial^2 T(x,y,t)}{\partial y^2} \right) + \sum_{k=1}^{4} P_k \cdot \delta(x - x_k,\, y - y_k), \\ T(x,y,0) = T_0(x,y), \\ T|_{\partial\Omega} = 0. \end{cases} \tag{7.2}$$

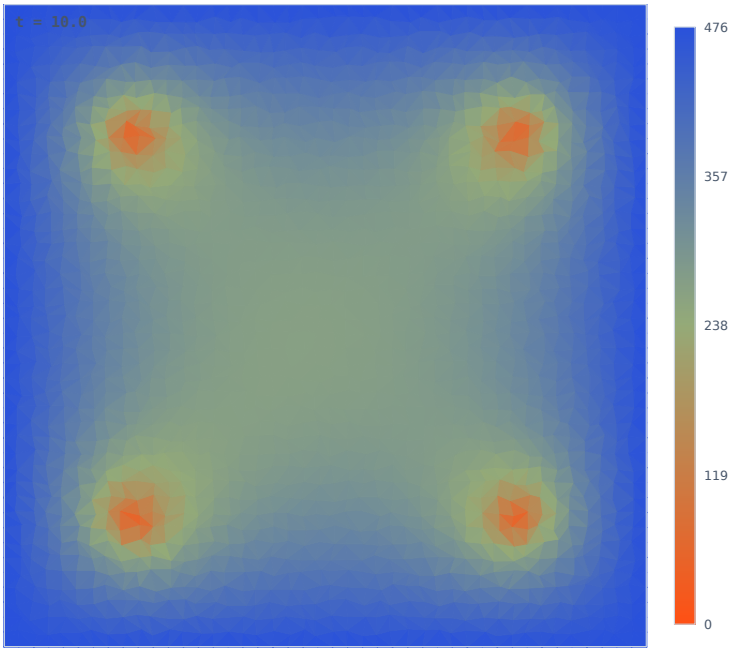


**Fig. 7.2.** The 2D Dirichlet problem with an internal heat source. The colour shows the temperature $T(x, y, t)$ at each point of the plate; the scale is shown on the right.

(interactive version on the website)

# Magnetic field 2D

Consider magnetostatics — the stationary field of a rectangular permanent magnet, uniformly magnetized ($\mathbf{M} = \text{const}$). There are no free currents, so the field is potential: $\nabla \times \mathbf{H} = 0$, hence $\mathbf{H} = -\nabla\varphi$. Substituting this into $\nabla \cdot \mathbf{B} = 0$ with $\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M})$, we obtain a single elliptic equation for the scalar potential (with the Laplace operator) and a Dirichlet condition on the outer boundary — the field decays far away

$$\begin{cases} \dfrac{\partial^2\varphi}{\partial x^2} + \dfrac{\partial^2\varphi}{\partial y^2} = \sigma(x, y), \quad (x, y) \in \Omega, \\ \varphi\big|_{\partial\Omega} = 0. \end{cases} \tag{7.3}$$

On the left is the Laplace operator ($\nabla^2 = \Delta$). The right-hand side is the source $\sigma = \nabla \cdot \mathbf{M}$: inside a uniform magnet it vanishes, while on the pole faces it gives a bound surface charge (essentially the normal "flux" of the magnetization $\mathbf{M} \cdot \mathbf{n}$), which depends on which face we are on

$$\sigma = \mathbf{M} \cdot \mathbf{n} = \begin{cases} +M, & \text{right face (N)}, \\ -M, & \text{left face (S)}, \\ 0, & \text{top and bottom faces.} \end{cases}$$

Equation (7.3) is a Poisson equation, i.e. an elliptic boundary value problem; for elliptic equations and reducing transient problems to them see Appendix N. We solve it numerically by the finite element method on a triangular mesh: the weak form is $\int_\Omega \nabla\varphi \cdot \nabla v \, d\Omega = \int_{\text{magnet}} \mathbf{M} \cdot \nabla v \, d\Omega$.

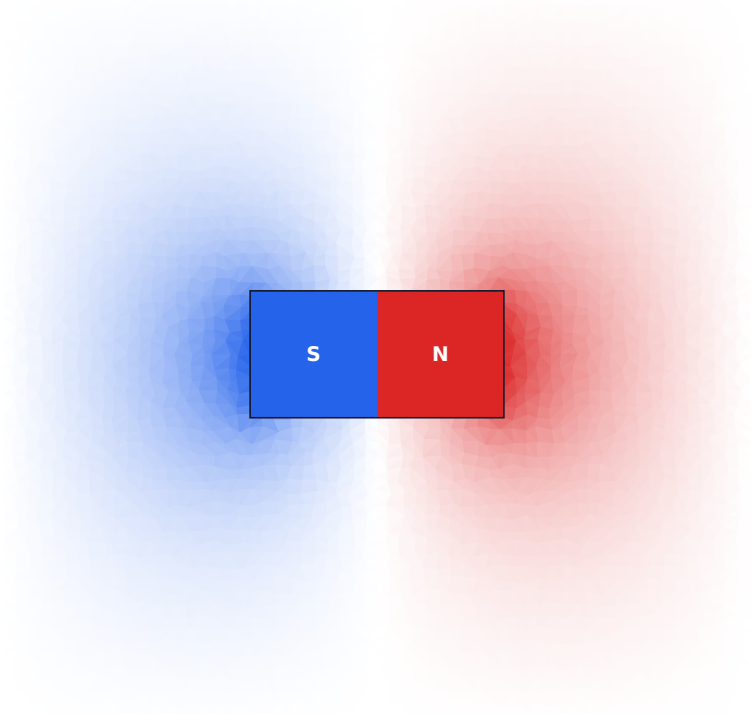


**Fig. 7.3.** Numerical (FEM) field of a rectangular magnet (N — red pole, S — blue). The colour shows the potential map $\varphi$; the $\nabla$ button overlays the field vectors.
(interactive version on the website)

## Electric field of a three-phase cable 2D

Consider the cross-section of a three-core cable. Inside a grounded metallic sheath (held at potential $\varphi = 0$) sit three conductive cores — phases A, B and C, symmetric at 120°. Between the cores and the sheath lies a charge-free dielectric, so the electrostatic potential obeys the Laplace equation, while each core is an equipotential at its instantaneous phase voltage (a Dirichlet condition):

$$\begin{cases} \dfrac{\partial^2 \varphi}{\partial x^2} + \dfrac{\partial^2 \varphi}{\partial y^2} = 0, & (x, y) \in \Omega, \\ \varphi\big|_{\Gamma_k} = V_k(t), \quad k = A, B, C, \\ \varphi\big|_{\partial\Omega} = 0. \end{cases} \tag{7.4}$$

Here $\Omega$ is the dielectric (the region between the cores and the sheath), $\Gamma_k$ are the core boundaries, $\partial\Omega$ is the inner surface of the sheath. The core voltages form a symmetric three-phase system. For a grid with line voltage 380 V the phase RMS is 220 V, i.e. the amplitude is $U_m = 220\sqrt{2} \approx 311$ V:

$$\begin{aligned} V_A &= U_m \sin\theta, \\ V_B &= U_m \sin\!\left(\theta - \tfrac{2\pi}{3}\right), \qquad \theta = \omega t. \\ V_C &= U_m \sin\!\left(\theta + \tfrac{2\pi}{3}\right), \end{aligned} \tag{7.5}$$

Since $V_A + V_B + V_C = 0$, the field pattern flows continuously while staying symmetric. We solve problem (7.4) numerically by the finite element method on a triangular mesh: the weak form is $\int_\Omega \nabla\varphi \cdot \nabla v \, d\Omega = 0$ for the prescribed $\varphi$ on cores and sheath. Each

instant $\theta$ has its own right-hand side (its own potential ratios), so the series of boundary value problems below is solved one after another, mimicking the run of the sinusoids.

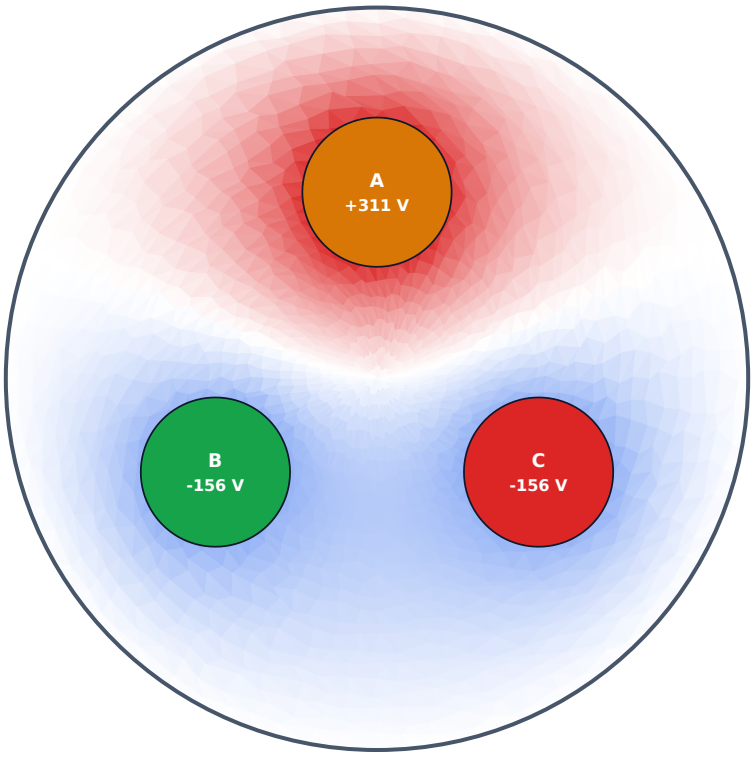


**Fig. 7.4.** Numerical (FEM) electrostatic field of a three-phase cable. The colour shows the potential map $\varphi$ (blue — minus, red — plus), the cores show the instantaneous voltages. The $\nabla$ button overlays the field vectors $\mathbf{E} = -\nabla\varphi$, the ▶ button runs the animation of the sinusoids.

(interactive version on the website)

# Thermal field 3D, Dirichlet

## Example: heating a ball with sources

Consider a three-dimensional transient example. Take a ball $\Omega = \{\,\mathbf{r} : |\mathbf{r}| \le R\,\}$ with thermal diffusivity $a^2 = 0.3$. The surface of the ball is held at zero temperature (Dirichlet condition), and six point heat sources are placed symmetrically inside, $f(\mathbf{r}) = \sum_{k=1}^{6} P_k\,\delta(\mathbf{r} - \mathbf{r}_k)$ (at the vertices of an octahedron). Initially the ball is cold; over time the sources heat it up while heat escapes through the cold surface, and the solution settles to a steady profile. The problem statement reads

$$\begin{cases} \dfrac{\partial T}{\partial t} = a^2\left(\dfrac{\partial^2 T}{\partial x^2} + \dfrac{\partial^2 T}{\partial y^2} + \dfrac{\partial^2 T}{\partial z^2}\right) + \sum_{k=1}^{6} P_k\,\delta(\mathbf{r} - \mathbf{r}_k), \quad \mathbf{r} \in \Omega, \\ T(\mathbf{r}, 0) = 0, \\ T|_{\partial\Omega} = 0. \end{cases} \tag{7.6}$$

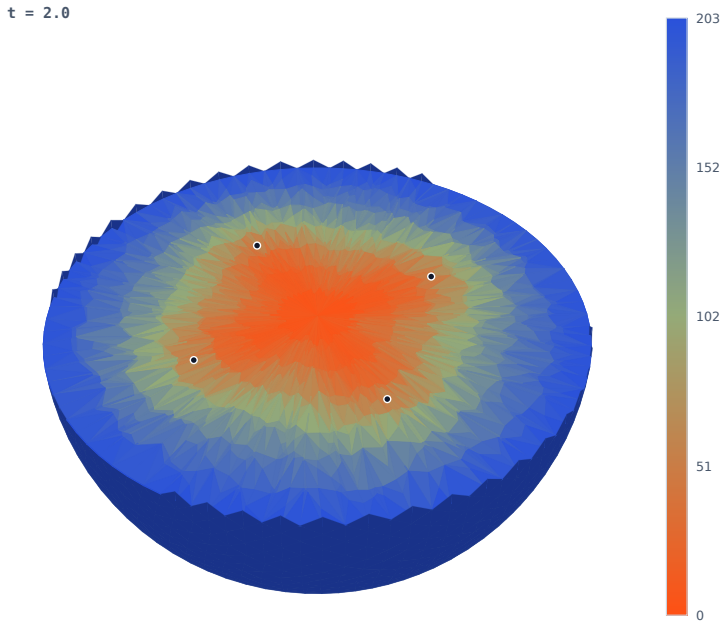


**Fig. 7.5.** A 3D transient Dirichlet problem: heating of a ball by six internal sources. The colour shows the temperature $T(\mathbf{r}, t)$.

(interactive version on the website)

# Appendix A

# Fourier Series

Here we recall, in the most general form, how a function is expanded into a series in an orthogonal system of functions — we use this expansion everywhere we solve boundary value problems by separation of variables. Suppose in the domain $G$ there is given a system of functions $\varphi_1(M), \varphi_2(M), \ldots$, pairwise orthogonal with the weight function $\rho(M) > 0$, that is

$$\iiint_G \rho(M) \cdot \varphi_n(M) \cdot \varphi_m(M)\, dG = 0, \quad n \neq m. \tag{A.1}$$

The function $f(M)$ is called expandable in this system if it can be represented by a convergent series

$$f(M) = \sum_{n=1}^{\infty} c_n \cdot \varphi_n(M). \tag{A.2}$$

The coefficients $c_n$ (they are called Fourier coefficients) are found thanks to orthogonality. Let us multiply both sides of (A.2) by $\rho(M) \cdot \varphi_m(M)$ and integrate over the domain $G$. Due to orthogonality (A.1) only one term survives on the right-hand side — the one with $n = m$:

$$\iiint_G \rho(M) \cdot f(M) \cdot \varphi_n(M)\, dG = c_n \cdot \iiint_G \rho(M) \cdot \varphi_n^2(M)\, dG, \tag{A.3}$$

whence the expansion coefficients are determined uniquely:

$$c_n = \frac{\iiint_G \rho(M) \cdot f(M) \cdot \varphi_n(M)\, dG}{\iiint_G \rho(M) \cdot \varphi_n^2(M)\, dG}. \tag{A.4}$$

The denominator in formula (A.4) is the squared norm of the function $\varphi_n$:

$$\|\varphi_n\|^2 = \iiint_G \rho(M) \cdot \varphi_n^2(M)\, dG. \tag{A.5}$$

Formula (A.4) is universal: whatever the orthogonal system, the expansion coefficients in it are computed in the same way. The weight $\rho(M)$ is determined by the coordinate system in which the variables separate: it is the factor with which the coordinate enters the volume element (the Jacobian) and with respect to which orthogonality holds for the system $\{\varphi_n\}$. For the main systems it equals:

| Coordinate system | $\rho$ |
|---|---|
| Cartesian (1D) | 1 |
| polar ($r$) | $r$ |
| spherical ($r$) | $r^2$ |

# Appendix B

# Special functions

When solving boundary value problems by separation of variables, the partial differential equation splits into ordinary differential equations in each coordinate. Depending on the coordinate system and the geometry of the domain, these equations lead to different families of orthogonal functions.

**The Bessel functions of the first kind** $J_m(r)$. They arise when separating variables in polar and cylindrical coordinates: the radial part of the Helmholtz equation reduces to the Bessel equation

$$r^2 \cdot \frac{d^2y(r)}{dr^2} + r \cdot \frac{dy(r)}{dr} + (r^2 - m^2) \cdot y(r) = 0.$$

The graphs of the first orders are shown in figure (B.1).

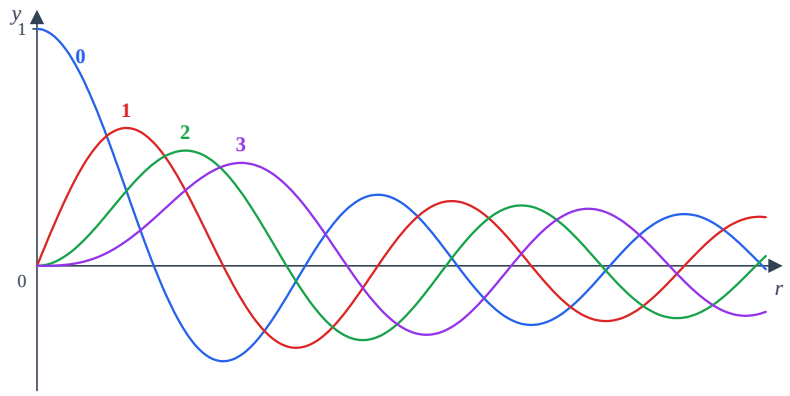


**Fig. B.1.** The Bessel functions of the first kind $J_m(r)$ of orders $m = 0, 1, 2, 3$.

**The Neumann functions** $Y_m(r)$ (Bessel functions of the second kind). They arise as the second linearly independent solution of the same Bessel equation. Unlike $J_m$, they grow unboundedly in absolute value near the axis: $Y_m(r) \to -\infty$ for $r \to 0$. Therefore, in problems where the domain includes the axis $r = 0$, they are discarded by the requirement that the solution be bounded, but in annular domains (not containing the axis) they enter the solution on a par with $J_m$. The graphs of the first orders are shown in figure (B.2).

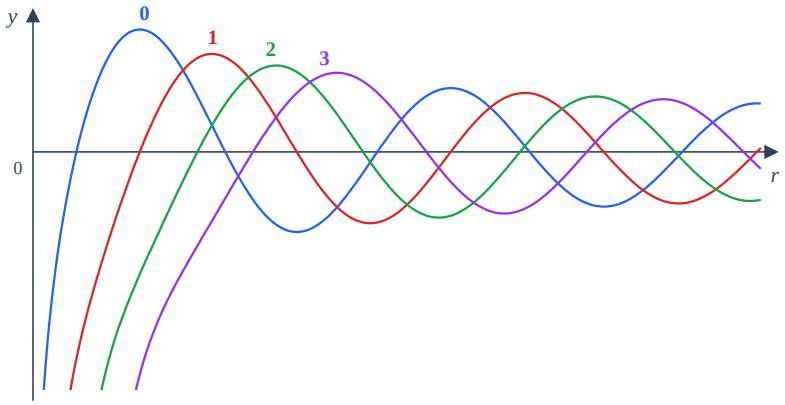


**Fig. B.2.** The Neumann functions $Y_m(r)$ of orders $m = 0, 1, 2, 3$.

**The spherical Bessel functions** $j_k(r)$. They appear in the radial part of the Helmholtz equation in spherical coordinates

$$r^2 \cdot \frac{d^2y(r)}{dr^2} + 2 \cdot r \cdot \frac{dy(r)}{dr} + \big(r^2 - k \cdot (k+1)\big) \cdot y(r) = 0.$$

The substitution $R(r) = \dfrac{1}{\sqrt{r}} \cdot \widehat{R}(r)$ reduces it to the Bessel equation of half-integer order, so the solutions are expressed in terms of ordinary Bessel functions

$$j_k(r) = \sqrt{\frac{\pi}{2 \cdot r}} \cdot J_{k+1/2}(r).$$

As in the cylindrical case, $j_k$ oscillate with decreasing amplitude; their graphs are shown in figure (B.3).

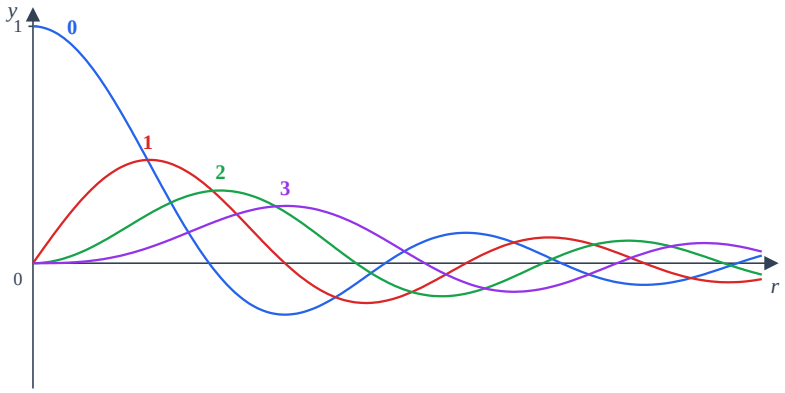


**Fig. B.3.** The spherical Bessel functions $j_k(r)$ of orders $k = 0, 1, 2, 3$.

**The Legendre polynomials** $P_k(z)$. They arise from the angular (polar) part of the Helmholtz equation in spherical coordinates. After the substitution $z = \cos(\theta)$ and for $m = 0$ the equation takes the form of the Legendre equation

$$(1 - z^2) \cdot \frac{d^2y(z)}{dz^2} - 2 \cdot z \cdot \frac{dy(z)}{dz} + k \cdot (k+1) \cdot y(z) = 0,$$

whose solutions bounded on the interval $z \in [-1, 1]$ are the Legendre polynomials of degree $k$. They form an orthogonal basis on $[-1, 1]$; the first few are shown in figure (B.4).

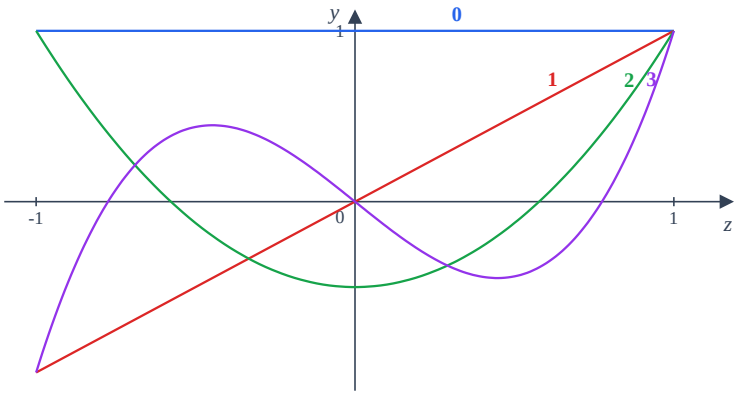


**Fig. B.4.** The Legendre polynomials $P_k(z)$ of degrees $k = 0, 1, 2, 3$.

**The associated Legendre functions** $P_k^{(m)}(z)$. If in the polar part $m \neq 0$, the Legendre equation generalizes to the associated equation

$$(1 - z^2) \cdot \frac{d^2y(z)}{dz^2} - 2 \cdot z \cdot \frac{dy(z)}{dz} + \left[k \cdot (k + 1) - \frac{m^2}{1 - z^2}\right] \cdot y(z) = 0,$$

whose solutions are the associated Legendre functions $P_k^{(m)}(z)$, expressed in terms of the Legendre polynomials

$$P_k^{(m)}(z) = (1 - z^2)^{m/2} \cdot \frac{d^m}{dz^m} P_k(z).$$

For $m = 0$ they turn into the Legendre polynomials $P_k(z)$, and together with the azimuthal factor $\cos(m \cdot \phi)$ or $\sin(m \cdot \phi)$ form the angular part of the eigenfunction.

# Appendix C

# Bessel function norm (Dirichlet)

To compute the norm of the solution of the equation for the geometry, we need to compute the norm of the Bessel function, which is defined as follows

$$\int_0^{\mu} r \cdot J_m^2(r)\, dr, \tag{C.1}$$

where $\mu$ is one of the solutions of the equation $J_m(\mu) = 0$.
The following integrals are known

$$\int r^{m+1} \cdot J_m(r)\, dr = r^{m+1} \cdot J_{m+1}(r) + C, \tag{C.2}$$

$$\int r^{-m} \cdot J_{m+1}(r)\, dr = -r^{-m} \cdot J_m(r) + C, \tag{C.3}$$

as well as the formulas for the derivatives of the Bessel function

$$r \cdot \frac{dJ_m(r)}{dr} = m \cdot J_m(r) - r \cdot J_{m+1}(r), \quad \text{as } m = 0 \;\Rightarrow\; \frac{dJ_0(r)}{dr} = -J_1(r), \tag{C.4}$$

$$r \cdot \frac{dJ_m(r)}{dr} = -m \cdot J_m(r) + r \cdot J_{m-1}(r), \quad \text{as } m = 0 \;\Rightarrow\; \frac{dJ_0(r)}{dr} = J_{-1}(r). \tag{C.5}$$

Let us try to compute the norm (C.1) by integration by parts:

$$\int_0^{\mu} r \cdot J_m^2(r)\, dr = \left| \begin{array}{l} u = r^{-m} \cdot J_m(r), \quad v = r^{m+1} \cdot J_{m+1}(r) \\ du = -m \cdot r^{-m-1} \cdot J_m(r)\, dr + r^{-m-1} \cdot [m \cdot J_m(r) - r \cdot J_{m+1}(r)]\, dr \\ dv = r^{m+1} \cdot J_m(r)\, dr \end{array} \right|$$

$$\int_0^{\mu} r \cdot J_m^2(r)\, dr = r \cdot J_m(r) \cdot J_{m+1}(r)\Big|_0^{\mu} + \int_0^{\mu} r \cdot J_{m+1}^2(r)\, dr = \int_0^{\mu} r \cdot J_{m+1}^2(r)\, dr.$$

We integrate the resulting integral by parts as well:

$$\int_0^\mu r \cdot J_{m+1}^2(r)\,dr = \left| \begin{array}{ll} u = J_{m+1}^2(r), & du = 2 \cdot J_{m+1}(r) \cdot \frac{1}{r} \cdot [-(m+1) \cdot J_{m+1}(r) + r \cdot J_m(r)]\ dr \\ dv = r\,dr, & v = r^2/2 \end{array} \right|$$

$$\int_0^\mu r \cdot J_{m+1}^2(r)\,dr = \frac{\mu^2}{2} \cdot J_{m+1}^2(\mu) + (m+1) \cdot \int_0^\mu r \cdot J_{m+1}^2(r)\,dr - \int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\,dr.$$

We integrate the last integral by parts:

$$\int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\,dr = \left| \begin{array}{l} u = r^{m+2} \cdot J_m(r), \quad v = -r^{-m} \cdot J_m(r) \\ du = (m+2) \cdot r^{m+1} \cdot J_m(r) + r^{m+1} \cdot [m \cdot J_m(r) - r \cdot J_{m+1}(r)]\ dr \\ dv = r^{-m} \cdot J_{m+1}(r)\,dr \end{array} \right|$$

$$u \cdot v\Big|_0^\mu = -r^2 \cdot J_m^2(r)\Big|_0^\mu = 0,$$

$$du \cdot v = -2 \cdot (m+1) \cdot r \cdot J_m^2(r) + r^2 \cdot J_m(r) \cdot J_{m+1}(r),$$

$$\int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\,dr = 2 \cdot (m+1) \cdot \int_0^\mu r \cdot J_m^2(r)\,dr - \int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\,dr,$$

$$\int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\,dr = (m+1) \cdot \int_0^\mu r \cdot J_m^2(r)\,dr.$$

Let us substitute the last equality into the previously obtained expression for the integral $\int_0^\mu r \cdot J_{m+1}^2(r)\,dr$. Taking into account the equality $\int_0^\mu r \cdot J_m^2(r)\,dr = \int_0^\mu r \cdot J_{m+1}^2(r)\,dr$, the terms with $(m+1)$ cancel, and we finally obtain:

$$\int_0^\mu r \cdot J_m^2(r)\,dr = \frac{\mu^2}{2} \cdot J_{m+1}^2(\mu). \tag{C.6}$$

# Appendix D

# Bessel function norm (Neumann)

To compute the norm of the solution of the equation for the geometry, we need to compute the norm of the Bessel function, which is defined as follows

$$\int_0^\mu r \cdot J_m^2(r)\, dr, \tag{D.1}$$

where $\mu$ is one of the solutions of the equation $\left.\frac{dJ_m(r)}{dr}\right|_{r=\mu} = 0$.

From formula (C.4) for $r = \mu$, where $J'_m(\mu) = 0$, we obtain the relation $m \cdot J_m(\mu) = \mu \cdot J_{m+1}(\mu)$, which is used below.

Let us try to compute the norm (D.1) by integration by parts:

$$\int_0^\mu r \cdot J_m^2(r)\, dr = \left| \begin{array}{l} u = r^{-m} \cdot J_m(r), \quad v = r^{m+1} \cdot J_{m+1}(r) \\ du = -m \cdot r^{-m-1} \cdot J_m(r)\, dr + r^{-m-1} \cdot [m \cdot J_m(r) - r \cdot J_{m+1}(r)]\ dr \\ dv = r^{m+1} \cdot J_m(r)\, dr \end{array} \right|$$

The boundary term, unlike the Dirichlet case, does not vanish:

$$r \cdot J_m(r) \cdot J_{m+1}(r)\Big|_0^\mu = \mu \cdot J_m(\mu) \cdot J_{m+1}(\mu) = m \cdot J_m^2(\mu).$$

$$\int_0^\mu r \cdot J_m^2(r)\, dr = m \cdot J_m^2(\mu) + \int_0^\mu r \cdot J_{m+1}^2(r)\, dr.$$

We integrate the resulting integral by parts as well:

$$\int_0^\mu r \cdot J_{m+1}^2(r)\, dr = \left| \begin{array}{ll} u = J_{m+1}^2(r), & du = 2 \cdot J_{m+1}(r) \cdot \frac{1}{r} \cdot [-(m+1) \cdot J_{m+1}(r) + r \cdot J_m(r)]\ dr \\ dv = r\, dr, & v = r^2/2 \end{array} \right|$$

$$\begin{aligned} \int_0^\mu r \cdot J_{m+1}^2(r)\, dr = \frac{m^2}{2} \cdot J_m^2(\mu) & \\ + (m+1) \cdot \int_0^\mu r \cdot J_{m+1}^2(r)\, dr & \\ - \int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\, dr. & \end{aligned}$$

We integrate the last integral by parts:

$$\int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\,dr = \left| \begin{array}{l} u = r^{m+2} \cdot J_m(r), \quad v = -r^{-m} \cdot J_m(r) \\ du = (m+2) \cdot r^{m+1} \cdot J_m(r) + r^{m+1} \cdot [m \cdot J_m(r) - r \cdot J_{m+1}(r)]\,dr \\ dv = r^{-m} \cdot J_{m+1}(r)\,dr \end{array} \right|$$

$$u \cdot v\Big|_0^\mu = -r^2 \cdot J_m^2(r)\Big|_0^\mu = -\mu^2 \cdot J_m^2(\mu),$$

$$du \cdot v = -2 \cdot (m+1) \cdot r \cdot J_m^2(r) + r^2 \cdot J_m(r) \cdot J_{m+1}(r),$$

$$\int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\,dr = -\mu^2 \cdot J_m^2(\mu) + 2 \cdot (m+1) \cdot \int_0^\mu r \cdot J_m^2(r)\,dr - \int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\,dr,$$

$$\int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\,dr = -\frac{\mu^2}{2} \cdot J_m^2(\mu) + (m+1) \cdot \int_0^\mu r \cdot J_m^2(r)\,dr.$$

Substituting the obtained equalities successively into one another, we express the norm:

$$\int_0^\mu r \cdot J_{m+1}^2(r)\,dr = \frac{1}{m} \cdot \int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\,dr - \frac{m}{2} \cdot J_m^2(\mu),$$

$$\int_0^\mu r \cdot J_m^2(r)\,dr = m \cdot J_m^2(\mu) + \frac{1}{m} \cdot \int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\,dr - \frac{m}{2} \cdot J_m^2(\mu),$$

$$\int_0^\mu r \cdot J_m^2(r)\,dr = \frac{m}{2} \cdot J_m^2(\mu) + \frac{1}{m} \cdot \int_0^\mu r^2 \cdot J_m(r) \cdot J_{m+1}(r)\,dr,$$

$$\int_0^{\mu} r \cdot J_m^2(r)\,dr = \frac{m}{2} \cdot J_m^2(\mu) - \frac{\mu^2}{2 \cdot m} \cdot J_m^2(\mu) + \frac{m+1}{m} \cdot \int_0^{\mu} r \cdot J_m^2(r)\,dr,$$

$$\frac{1}{m} \cdot \int_0^{\mu} r \cdot J_m^2(r)\,dr = -\frac{m}{2} \cdot J_m^2(\mu) + \frac{\mu^2}{2 \cdot m} \cdot J_m^2(\mu).$$

Finally we obtain:

$$\int_0^{\mu} r \cdot J_m^2(r)\,dr = \left(\mu^2 - m^2\right) \cdot \frac{J_m^2(\mu)}{2}. \tag{D.2}$$

# Appendix E

# The Legendre equation

When solving the equation for the geometry in spherical coordinates, the following equation for the polar angle arises

$$\frac{d}{dz}\left((1-z^2)\cdot\frac{d\Theta(z)}{dz}\right)+\left[\gamma_1^2-\frac{m^2}{1-z^2}\right]\cdot\Theta(z)=0,\quad -1<z<1,\quad \text{(E.1)}$$

where $\Theta(z)$ is the unknown function, $m$ is a nonnegative integer, $\gamma_1^2$ is the separation constant. Let us expand the derivative in the first term

$$(1-z^2)\cdot\frac{d^2\Theta(z)}{dz^2}-2\cdot z\cdot\frac{d\Theta(z)}{dz}+\left[\gamma_1^2-\frac{m^2}{1-z^2}\right]\cdot\Theta(z)=0.$$

Note that for $m=0$ the equation takes the form of the Legendre equation

$$(1-z^2)\cdot\frac{d^2\Theta(z)}{dz^2}-2\cdot z\cdot\frac{d\Theta(z)}{dz}+\gamma_1^2\cdot\Theta(z)=0,\quad -1<z<1.$$

This equation has solutions bounded on the interval $-1<z<1$ only for the eigenvalues $\gamma_{1k}^2=k\cdot(k+1)$, $k\in(0..\infty)$; these solutions are given by the Rodrigues formula

$$\Theta_k(z)=P_k(z)=\frac{1}{2^k\cdot k!}\cdot\frac{d^k}{dz^k}\left(z^2-1\right)^k,\quad \text{(E.2)}$$

where $P_k(z)$ are the Legendre polynomials. We will look for solutions of equation (E.1) for the same eigenvalues $\gamma_{1k}^2=k\cdot(k+1)$ — then it can be written as

$$(1-z^2)\cdot\frac{d^2\Theta_{km}(z)}{dz^2}-2\cdot z\cdot\frac{d\Theta_{km}(z)}{dz}+\left[k\cdot(k+1)-\frac{m^2}{1-z^2}\right]\cdot\Theta_{km}(z)=0,\quad k\in(0..\infty).\quad \text{(E.3)}$$

Let us make the substitution $\Theta_{km}(z) = (1-z^2)^{\frac{m}{2}} \cdot \widehat{\Theta}_{km}(z)$ into equation (E.3). The first derivative takes the form

$$\frac{d}{dz}\left[(1-z^2)^{\frac{m}{2}} \cdot \widehat{\Theta}_{km}(z)\right] = (1-z^2)^{\frac{m}{2}} \cdot \frac{d\widehat{\Theta}_{km}(z)}{dz} - m \cdot z \cdot (1-z^2)^{\frac{m}{2}-1} \cdot \widehat{\Theta}_{km}(z).$$

To compute the second derivative, we first compute

$$\frac{d}{dz}\left[\frac{z}{1-z^2}\right] = \frac{1}{1-z^2} + \frac{2 \cdot z^2}{(1-z^2)^2} = \frac{1+z^2}{(1-z^2)^2}.$$

Then the second derivative can be written as

$$\frac{d^2}{dz^2}\left[(1-z^2)^{\frac{m}{2}} \cdot \widehat{\Theta}_{km}(z)\right] = (1-z^2)^{\frac{m}{2}} \cdot \frac{d^2\widehat{\Theta}_{km}(z)}{dz^2} - m \cdot z \cdot (1-z^2)^{\frac{m}{2}-1} \cdot \frac{d\widehat{\Theta}_{km}(z)}{dz} - m \cdot \frac{z}{1-z^2} \cdot \left[(1-z^2)^{\frac{m}{2}} \cdot \frac{d\widehat{\Theta}_{km}(z)}{dz} - m \cdot z \cdot (1-z^2)^{\frac{m}{2}-1} \cdot \widehat{\Theta}_{km}(z)\right] - m \cdot \frac{1+z^2}{(1-z^2)^2} \cdot (1-z^2)^{\frac{m}{2}} \cdot \widehat{\Theta}_{km}(z).$$

After collecting like terms it takes the form

$$\frac{d^2}{dz^2}\left[(1-z^2)^{\frac{m}{2}} \cdot \widehat{\Theta}_{km}(z)\right] = (1-z^2)^{\frac{m}{2}} \cdot \frac{d^2\widehat{\Theta}_{km}(z)}{dz^2} - \frac{2 \cdot m \cdot z}{1-z^2} \cdot (1-z^2)^{\frac{m}{2}} \cdot \frac{d\widehat{\Theta}_{km}(z)}{dz} + m \cdot \frac{m \cdot z^2 - 1 - z^2}{(1-z^2)^2} \cdot (1-z^2)^{\frac{m}{2}} \cdot \widehat{\Theta}_{km}(z).$$

Let us substitute everything into equation (E.3), cancelling along the way $(1-z^2)^{\frac{m}{2}}$:

$$(1-z^2) \cdot \left[\frac{d^2\widehat{\Theta}_{km}(z)}{dz^2} - \frac{2 \cdot m \cdot z}{1-z^2} \cdot \frac{d\widehat{\Theta}_{km}(z)}{dz} + m \cdot \frac{m \cdot z^2 - 1 - z^2}{(1-z^2)^2} \cdot \widehat{\Theta}_{km}(z)\right] - 2 \cdot z \cdot \left[\frac{d\widehat{\Theta}_{km}(z)}{dz} - \frac{m \cdot z}{1-z^2} \cdot \widehat{\Theta}_{km}(z)\right] + \left[k \cdot (k+1) - \frac{m^2}{1-z^2}\right] \cdot \widehat{\Theta}_{km}(z) = 0.$$

Let us group the coefficients of the derivatives:

$$\zeta_1 \cdot \frac{d^2\widehat{\Theta}_{km}(z)}{dz^2} + \zeta_2 \cdot \frac{d\widehat{\Theta}_{km}(z)}{dz} + \zeta_3 \cdot \widehat{\Theta}_{km}(z) = 0,$$

$$\zeta_1 = 1 - z^2,$$

$$\zeta_2 = -2 \cdot m \cdot z - 2 \cdot z = -2 \cdot (m+1) \cdot z,$$

$$\zeta_3 = m \cdot \frac{m \cdot z^2 - 1 - z^2}{1-z^2} + \frac{2 \cdot m \cdot z^2}{1-z^2} + k \cdot (k+1) - \frac{m^2}{1-z^2} =$$
$$\frac{m^2 \cdot z^2 - m - m \cdot z^2 + 2 \cdot m \cdot z^2 - m^2}{1-z^2} + k \cdot (k+1) = -\frac{m^2 \cdot (1-z^2) + m \cdot (1-z^2)}{1-z^2} +$$
$$k \cdot (k+1) = -m^2 - m + k \cdot (k+1) = (k-m) \cdot (k+m+1),$$

thus, after the substitution, equation (E.3) takes the form

$$(1-z^2)\cdot\frac{d^2\widehat{\Theta}_{km}(z)}{dz^2}-2\cdot(m+1)\cdot z\cdot\frac{d\widehat{\Theta}_{km}(z)}{dz}+(k-m)\cdot(k+m+1)\cdot\widehat{\Theta}_{km}(z)=0. \quad \text{(E.4)}$$

Let us take equation (E.3), set $m=0$ and take into account that in this case its solutions are the Legendre polynomials $P_k(z)$ — we obtain the following equation

$$(1-z^2)\cdot\frac{d^2P_k(z)}{dz^2}-2\cdot z\cdot\frac{dP_k(z)}{dz}+k\cdot(k+1)\cdot P_k(z)=0, \quad \text{(E.5)}$$

which we differentiate $m$ times using the Leibniz formula, defined as follows

$$\frac{d^m}{dz^m}\left(u(z)\cdot v(z)\right)=\sum_{i=0}^{m}\binom{m}{i}\cdot\frac{d^{m-i}}{dz^{m-i}}u(z)\cdot\frac{d^i}{dz^i}v(z),$$

where $\binom{m}{i}=\frac{m!}{i!\cdot(m-i)!}$ is the binomial coefficient.
Let us start with the first term of equation (E.5), taking $v(z)=1-z^2$

$$\frac{d^0}{dz^0}v=1-z^2;\quad \frac{d^1}{dz^1}v=-2\cdot z;\quad \frac{d^2}{dz^2}v=-2;\quad \frac{d^i}{dz^i}v=0,\quad i\in(3..\infty),$$

now let us take $v(z)=-2\cdot z$, we obtain

$$\frac{d^0}{dz^0}v=-2\cdot z;\quad \frac{d^1}{dz^1}v=-2;\quad \frac{d^i}{dz^i}v=0,\quad i\in(2..\infty),$$

for the case $v(z)=k\cdot(k+1)$ everything is trivial

$$\frac{d^0}{dz^0}v=k\cdot(k+1);\quad \frac{d^i}{dz^i}v=0,\quad i\in(1..\infty).$$

The binomial coefficients are respectively equal to

$$\binom{m}{0}=1;\quad \binom{m}{1}=m;\quad \binom{m}{2}=\frac{m\cdot(m-1)}{2}.$$

The result of differentiating the first term of equation (E.5) equals

$$(1-z^2)\cdot\frac{d^{m+2}}{dz^{m+2}}P_k(z)-2\cdot z\cdot m\cdot\frac{d^{m+1}}{dz^{m+1}}P_k(z)-m\cdot(m-1)\cdot\frac{d^m}{dz^m}P_k(z),$$

for the second term

$$-2\cdot z\cdot\frac{d^{m+1}}{dz^{m+1}}P_k(z)-2\cdot m\cdot\frac{d^m}{dz^m}P_k(z),$$

for the third term

$$k \cdot (k+1) \cdot \frac{d^m}{dz^m} P_k(z),$$

let us add all three terms

$$(1-z^2) \cdot \frac{d^{m+2}}{dz^{m+2}} P_k(z) - 2 \cdot (m+1) \cdot z \cdot \frac{d^{m+1}}{dz^{m+1}} P_k(z) + (k-m) \cdot (k+m+1) \cdot \frac{d^m}{dz^m} P_k(z).$$

The left-hand side of equation (E.5) is identically zero, hence so is its $m$-fold derivative:

$$(1-z^2) \cdot \frac{d^{m+2}}{dz^{m+2}} P_k(z) - 2 \cdot (m+1) \cdot z \cdot \frac{d^{m+1}}{dz^{m+1}} P_k(z) + (k-m) \cdot (k+m+1) \cdot \frac{d^m}{dz^m} P_k(z) = 0.$$

The resulting equation coincides with equation (E.4), whence $\widehat{\Theta}_{km}(z) = \frac{d^m}{dz^m} P_k(z)$, which means the solution of equation (E.1) has the form

$$\Theta_{km}(z) = P_k^{(m)}(z) = (1-z^2)^{\frac{m}{2}} \cdot \frac{d^m}{dz^m} P_k(z), \tag{E.6}$$

where $\gamma_{1k}^2 = k \cdot (k+1)$ are the eigenvalues, $P_k^{(m)}(z)$ are the associated Legendre polynomials. For $m > k$ the derivative $\frac{d^m}{dz^m} P_k(z)$ vanishes, so nontrivial solutions exist only for $k \geq m$.

# Appendix F

# Norm of the Legendre polynomials

To compute the norm of the solution of the equation for the geometry, we need to compute the norm of the associated Legendre polynomial, which is defined as follows

$$\int_{-1}^{1} \left[P_k^{(m)}(z)\right]^2 \, dz, \tag{F.1}$$

where $P_k^{(m)}(z) = (1-z^2)^{\frac{m}{2}} \cdot \frac{d^m}{dz^m} P_k(z)$ is the associated Legendre polynomial. Let us substitute this definition into the integral (F.1):

$$\int_{-1}^{1} \left[P_k^{(m)}(z)\right]^2 \, dz = \int_{-1}^{1} (1-z^2)^m \left[\frac{d^m}{dz^m} P_k(z)\right]^2 \, dz.$$

Let us substitute the Rodrigues formula (E.2) into the integral and take out everything that does not depend on $z$:

$$\int_{-1}^{1} (1-z^2)^m \left[\frac{d^m}{dz^m} P_k(z)\right]^2 \, dz$$
$$= \frac{1}{2^{2k} \cdot (k!)^2} \cdot \int_{-1}^{1} (1-z^2)^m \cdot \left[\frac{d^{k+m}}{dz^{k+m}} \left(z^2-1\right)^k\right]^2 \, dz. \tag{F.2}$$

Clearly, $m \leq k$: otherwise the order of the derivative $k+m$ exceeds the degree of the polynomial $2 \cdot k$, and the integrand is identically zero. First let us compute the auxiliary integral corresponding to the case $m = 0$

$$\frac{1}{2^{2k} \cdot (k!)^2} \cdot \int_{-1}^{1} \left[\frac{d^k}{dz^k} \left(z^2-1\right)^k\right]^2 \, dz. \tag{F.3}$$

Let us integrate it by parts:

$$\int_{-1}^{1} \left[\frac{d^k}{dz^k} \left(z^2-1\right)^k\right]^2 \, dz$$
$$= \left| \begin{array}{ll} u = \frac{d^k}{dz^k} \left(z^2-1\right)^k, & du = \frac{d^{k+1}}{dz^{k+1}} \left(z^2-1\right)^k \, dz \\ dv = \frac{d^k}{dz^k} \left(z^2-1\right)^k \, dz, & v = \frac{d^{k-1}}{dz^{k-1}} \left(z^2-1\right)^k \end{array} \right|$$

The boundary terms are zero: the derivatives $\frac{d^i}{dz^i}\left(z^2-1\right)^k$ for $i<k$ vanish at the endpoints of the interval. After applying the integration by parts formula $k$ times (each application changes the sign) we obtain

$$\frac{1}{2^{2k}\cdot(k!)^2}\cdot\int_{-1}^{1}\left[\frac{d^k}{dz^k}\left(z^2-1\right)^k\right]^2 dz = \frac{(-1)^k}{2^{2k}\cdot(k!)^2}\cdot\int_{-1}^{1}\left(z^2-1\right)^k\cdot\frac{d^{2k}}{dz^{2k}}\left(z^2-1\right)^k dz,$$

now from Newton's binomial formula it becomes obvious that $\frac{d^{2k}}{dz^{2k}}\left(z^2-1\right)^k=(2k)!$: the polynomial has maximal degree $2\cdot k$, a derivative of that order kills all terms except the leading one, and its coefficient equals $(2k)!$.

It remains to compute the integral

$$\int_{-1}^{1}\left(z^2-1\right)^k dz = \left|\begin{array}{ll} u=\left(z^2-1\right)^k, & du=2\cdot k\cdot z\cdot\left(z^2-1\right)^{k-1}dz \\ dv=dz, & v=z \end{array}\right|$$

obviously, the boundary term equals zero, therefore

$$\int_{-1}^{1}\left(z^2-1\right)^k dz=-2\cdot k\cdot\int_{-1}^{1}z^2\cdot\left(z^2-1\right)^{k-1}dz,$$

let us replace $z^2$ by $z^2-1+1$ and obtain

$$\int_{-1}^{1}\left(z^2-1\right)^k dz=-\frac{2\cdot k}{2\cdot k+1}\cdot\int_{-1}^{1}\left(z^2-1\right)^{k-1}dz,$$

applying this recurrence formula another $k-1$ times, we obtain

$$\int_{-1}^{1}\left(z^2-1\right)^k dz=(-1)^k\cdot\frac{2\cdot k}{2\cdot k+1}\cdot\frac{2\cdot(k-1)}{2\cdot(k-1)+1}\cdot\frac{2\cdot(k-2)}{2\cdot(k-2)+1}\cdot\cdots\cdot\frac{2\cdot 1}{2\cdot 1+1}\cdot 2,$$

$$\int_{-1}^{1}\left(z^2-1\right)^k dz=(-1)^k\cdot 2\cdot\frac{2^k\cdot k!\cdot 2^k\cdot k!}{(2\cdot k+1)!}.$$

As a result, the factors $(-1)^k$ cancel, and we obtain

$$\frac{1}{2^{2k}\cdot(k!)^2}\cdot\int_{-1}^{1}\left[\frac{d^k}{dz^k}\left(z^2-1\right)^k\right]^2 dz=\frac{2}{2\cdot k+1}. \tag{F.4}$$

Let us write the well-known recurrence formula for the associated Legendre polynomials, which will be needed for the next integration by parts

$$2\cdot m\cdot z\cdot P_k^{(m)}(z) = \sqrt{1-z^2}\cdot\left[P_k^{(m+1)}(z)+(k+m)\cdot(k-m+1)\cdot P_k^{(m-1)}(z)\right], \tag{F.5}$$

for a slight simplification let us denote $\epsilon = (k+m)\cdot(k-m+1)$.
Let us return to the integral (F.2) and perform integration by parts:

$$\int_{-1}^{1}\left(1-z^2\right)^m\cdot\left[\frac{d^m}{dz^m}P_k(z)\right]^2\,dz=\left|\begin{array}{l} u=\left(1-z^2\right)^m\cdot\dfrac{d^m}{dz^m}P_k(z)\\ du=\left(1-z^2\right)^{(m/2)-1}\cdot\left[-2\cdot m\cdot z\cdot P_k^{(m)}(z)+\sqrt{1-z^2}\cdot P_k^{(m+1)}(z)\right]\,dz=\\ \left(1-z^2\right)\dfrac{m-1}{2}\cdot\left[-P_k^{(m+1)}(z)-\epsilon\cdot P_k^{(m-1)}(z)+P_k^{(m+1)}(z)\right]\,dz=\\ -\left(1-z^2\right)\dfrac{m-1}{2}\cdot\epsilon\cdot P_k^{(m-1)}(z)\,dz\\ dv=\dfrac{d^m}{dz^m}P_k(z)\,dz,\\ v=\dfrac{d^{m-1}}{dz^{m-1}}P_k(z)=\left(1-z^2\right)^{\frac{1-m}{2}}\cdot P_k^{(m-1)}(z) \end{array}\right|$$

the boundary term contains the factor $\sqrt{1-z^2}$ and vanishes at the endpoints of the interval, so we obtain

$$\int_{-1}^{1}P_k^{(m)}(z)\cdot P_k^{(m)}(z)\,dz = (k+m)\cdot(k-m+1)\cdot\int_{-1}^{1}P_k^{(m-1)}(z)\cdot P_k^{(m-1)}(z)\,dz,$$

successively applying integration by parts another $m-1$ times and taking into account (F.4), we obtain

$$\int_{-1}^{1}P_k^{(m)}(z)\cdot P_k^{(m)}(z)\,dz=\frac{2}{2\cdot k+1}\cdot(k+m)\cdot(k-m+1)\cdot(k+m-1)\cdot(k-m+2)\cdots,$$

the factors with a plus decrease from $k+m$ to $k+1$, the factors with a minus increase from $k-m+1$ to $k$ (it is important here that $m\le k$ holds), so

$$(k+m)\cdot(k-m+1)\cdot(k+m-1)\cdot(k-m+2)\cdots = \frac{(k+m)!}{k!}\cdot\frac{k!}{(k-m)!}=\frac{(k+m)!}{(k-m)!}.$$

The final norm of the associated Legendre polynomial has the form

$$\int_{-1}^{1}\left[P_k^{(m)}(z)\right]^2\,dz=\frac{2}{2\cdot k+1}\cdot\frac{(k+m)!}{(k-m)!},\quad m\le k. \tag{F.6}$$

# Appendix G

# The Bessel equation in spherical coordinates

When solving the equation for the geometry in spherical coordinates, the following equation for the radius arises

$$r^2 \cdot \frac{d^2R(r)}{dr^2} + 2 \cdot r \cdot \frac{dR(r)}{dr} + \left(r^2 \cdot \gamma^2 - k \cdot (k+1)\right) \cdot R(r) = 0, \quad \text{(G.1)}$$

where $\gamma$ is the eigenvalue, $k$ is a nonnegative integer.
Let us make the substitution $R(r) = \frac{1}{\sqrt{r}} \cdot \widehat{R}(r)$ and carry out the transformations

$$\frac{dR(r)}{dr} = \frac{d}{dr}\left(\frac{1}{\sqrt{r}} \cdot \widehat{R}(r)\right) = -\frac{1}{2 \cdot r \cdot \sqrt{r}} \cdot \widehat{R}(r) + \frac{1}{\sqrt{r}} \cdot \frac{d\widehat{R}(r)}{dr},$$

$$\frac{d^2R(r)}{dr^2} = -\frac{1}{2} \cdot \frac{d}{dr}\left(\frac{1}{r \cdot \sqrt{r}} \cdot \widehat{R}(r)\right) + \frac{d}{dr}\left(\frac{1}{\sqrt{r}} \cdot \frac{d\widehat{R}(r)}{dr}\right),$$

$$-\frac{1}{2} \cdot \frac{d}{dr}\left(\frac{1}{r \cdot \sqrt{r}} \cdot \widehat{R}(r)\right) = \frac{3}{4} \cdot \frac{1}{r^2 \cdot \sqrt{r}} \cdot \widehat{R}(r) - \frac{1}{2} \cdot \frac{1}{r \cdot \sqrt{r}} \cdot \frac{d\widehat{R}(r)}{dr},$$

$$\frac{d}{dr}\left(\frac{1}{\sqrt{r}} \cdot \frac{d\widehat{R}(r)}{dr}\right) = -\frac{1}{2 \cdot r \cdot \sqrt{r}} \cdot \frac{d\widehat{R}(r)}{dr} + \frac{1}{\sqrt{r}} \cdot \frac{d^2\widehat{R}(r)}{dr^2},$$

$$\frac{d^2R(r)}{dr^2} = \frac{1}{\sqrt{r}} \cdot \frac{d^2\widehat{R}(r)}{dr^2} - \frac{1}{r \cdot \sqrt{r}} \cdot \frac{d\widehat{R}(r)}{dr} + \frac{3}{4} \cdot \frac{1}{r^2 \cdot \sqrt{r}} \cdot \widehat{R}(r),$$

$$r^2 \cdot \frac{d^2R(r)}{dr^2} = r \cdot \sqrt{r} \cdot \frac{d^2\widehat{R}(r)}{dr^2} - \sqrt{r} \cdot \frac{d\widehat{R}(r)}{dr} + \frac{3}{4} \cdot \frac{1}{\sqrt{r}} \cdot \widehat{R}(r),$$

$$r^2 \cdot \frac{d^2R(r)}{dr^2} + 2 \cdot r \cdot \frac{dR(r)}{dr} = r \cdot \sqrt{r} \cdot \frac{d^2\widehat{R}(r)}{dr^2} + \sqrt{r} \cdot \frac{d\widehat{R}(r)}{dr} - \frac{1}{4} \cdot \frac{1}{\sqrt{r}} \cdot \widehat{R}(r),$$

we divide by $r \cdot \sqrt{r}$ and write equation (G.1) in terms of the coefficients

$$\zeta_1 \cdot \frac{d^2\widehat{R}(r)}{dr^2} + \zeta_2 \cdot \frac{d\widehat{R}(r)}{dr} + \zeta_3 \cdot \widehat{R}(r) = 0,$$

and now let us determine the coefficients, first transforming the free term

$$\gamma^2 - \frac{k \cdot (k+1)}{r^2} - \frac{1}{4 \cdot r^2} = \gamma^2 - \frac{(k+1/2)^2}{r^2},$$

$$\zeta_1 = 1, \quad \zeta_2 = \frac{1}{r}, \quad \zeta_3 = \gamma^2 - \frac{(k+1/2)^2}{r^2},$$

thus, equation (G.1) takes the form

$$\frac{d^2\widehat{R}(r)}{dr^2} + \frac{1}{r} \cdot \frac{d\widehat{R}(r)}{dr} + \left(\gamma^2 - \frac{(k+1/2)^2}{r^2}\right) \cdot \widehat{R}(r) = 0.$$

We have obtained the Bessel equation with half-integer index $k + 1/2$. Its solutions bounded at zero are the Bessel functions of the first kind $\widehat{R}(r) = J_{k+1/2}(\gamma \cdot r)$, $k \in (0..\infty)$; the solutions of the second kind $Y_{k+1/2}(\gamma \cdot r)$ are infinite at zero, as seen in figure (B.2), and are discarded by the boundedness condition. Thus, the solution of the original equation (G.1) has the form

$$R(r) = \frac{1}{\sqrt{r}} \cdot J_{k+1/2}(\gamma \cdot r), \tag{G.2}$$

where $J_{k+1/2}(\gamma \cdot r)$ is the Bessel function of the first kind with half-integer index. Up to a constant factor it is the spherical Bessel function: $j_k(\gamma \cdot r) = \sqrt{\frac{\pi}{2 \cdot \gamma \cdot r}} \cdot J_{k+1/2}(\gamma \cdot r)$.

# Appendix H

# Spherical Bessel function norm (Neumann)

To compute the norm of the solution of the equation for the geometry, we need to compute the norm of the spherical Bessel function. It differs from the ordinary Bessel function by the factor $\frac{1}{\sqrt{r}}$, and the weight for spherical coordinates is $\rho(r) = r^2$, so we arrive at the same integral as for the ordinary Bessel function

$$\int_0^\mu \left(\frac{1}{\sqrt{r}} \cdot J_{k+1/2}(r)\right)^2 \cdot r^2 \, dr = \int_0^\mu r \cdot J_{k+1/2}^2(r) \, dr, \tag{H.1}$$

where $\mu$ is one of the solutions of the equation $\left.\frac{dJ_{k+1/2}(r)}{dr}\right|_{r=\mu} = \frac{1}{2 \cdot \mu} \cdot J_{k+1/2}(\mu)$, which follows from the Neumann condition for the spherical Bessel function: $\left.\frac{d}{dr}\left(\frac{1}{\sqrt{r}} \cdot J_{k+1/2}(r)\right)\right|_{r=\mu} = 0$.

Let us try to compute the norm (H.1) by integration by parts, taking into account formulas (C.2) and (C.4):

$$\left|\begin{aligned} u &= r^{-k-1/2} \cdot J_{k+1/2}(r), \quad dv = r^{k+3/2} \cdot J_{k+1/2}(r) \, dr \\ du &= \left(-k - \frac{1}{2}\right) \cdot r^{-k-3/2} \cdot J_{k+1/2}(r) \, dr \\ &\quad + r^{-k-3/2} \cdot \left[\left(k + \frac{1}{2}\right) \cdot J_{k+1/2}(r) - r \cdot J_{k+3/2}(r)\right] \, dr \\ du &= -r^{-k-1/2} \cdot J_{k+3/2}(r) \, dr \\ v &= r^{k+3/2} \cdot J_{k+3/2}(r) \end{aligned}\right|$$

$$\int_0^\mu r \cdot J_{k+1/2}^2(r) \, dr = r \cdot J_{k+1/2}(r) \cdot J_{k+3/2}(r)\Big|_0^\mu + \int_0^\mu r \cdot J_{k+3/2}^2(r) \, dr.$$

Let us transform the eigenvalue equation using formula (C.4) for $r = \mu$

$$\mu \cdot \left.\frac{dJ_{k+1/2}(r)}{dr}\right|_{r=\mu} = \left(k+\frac{1}{2}\right)\cdot J_{k+1/2}(\mu) - \mu \cdot J_{k+3/2}(\mu)$$
$$= \frac{1}{2}\cdot J_{k+1/2}(\mu),$$

$$k \cdot J_{k+1/2}(\mu) = \mu \cdot J_{k+3/2}(\mu),$$

and substitute the obtained relation into the boundary term

$$\int_0^\mu r \cdot J_{k+1/2}^2(r)\,dr = k \cdot J_{k+1/2}^2(\mu) + \int_0^\mu r \cdot J_{k+3/2}^2(r)\,dr.$$

We integrate the resulting integral by parts as well:

$$\left|\begin{array}{ll} u = J_{k+3/2}^2(r), & du = 2\cdot J_{k+3/2}(r)\cdot\frac{1}{r}\cdot\left(-(k+3/2)\cdot J_{k+3/2}(r) + r\cdot J_{k+1/2}(r)\right)\,dr \\ dv = r\,dr, & v = r^2/2 \end{array}\right|$$

$$\int_0^\mu r \cdot J_{k+3/2}^2(r)\,dr = \frac{k^2}{2}\cdot J_{k+1/2}^2(\mu) + \left(k+\frac{3}{2}\right)\cdot\int_0^\mu r\cdot J_{k+3/2}^2(r)\,dr - \int_0^\mu r^2\cdot J_{k+1/2}(r)\cdot J_{k+3/2}(r)\,dr.$$

We integrate the last integral by parts:

$$\left|\begin{array}{l} u = r^{k+5/2}\cdot J_{k+1/2}(r), \quad dv = r^{-k-1/2}\cdot J_{k+3/2}(r)\,dr \\ du = \left(k+\frac{5}{2}\right)\cdot r^{k+3/2}\cdot J_{k+1/2}(r) + r^{k+3/2}\cdot\left[\left(k+\frac{1}{2}\right)\cdot J_{k+1/2}(r) - r\cdot J_{k+3/2}(r)\right]\,dr \\ du = (2\cdot k+3)\cdot r^{k+3/2}\cdot J_{k+1/2}(r) - r^{k+5/2}\cdot J_{k+3/2}(r)\,dr \\ v = -r^{-k-1/2}\cdot J_{k+1/2}(r) \end{array}\right|$$

$$\int_0^\mu r^2\cdot J_{k+1/2}(r)\cdot J_{k+3/2}(r)\,dr = -\mu^2\cdot J_{k+1/2}^2(\mu) + (2\cdot k+3)\cdot\int_0^\mu r\cdot J_{k+1/2}^2(r)\,dr - \int_0^\mu r^2\cdot J_{k+1/2}(r)\cdot J_{k+3/2}(r)\,dr,$$

$$\int_0^\mu r^2 \cdot J_{k+1/2}(r) \cdot J_{k+3/2}(r)\, dr = -\frac{\mu^2}{2} \cdot J_{k+1/2}^2(\mu) + \frac{(2 \cdot k + 3)}{2} \cdot \int_0^\mu r \cdot J_{k+1/2}^2(r)\, dr.$$

Substituting the obtained equalities successively into one another, we express the norm:

$$0 = \frac{k^2}{2} \cdot J_{k+1/2}^2(\mu) + \left(k + \frac{1}{2}\right) \cdot \int_0^\mu r \cdot J_{k+3/2}^2(r)\, dr - \int_0^\mu r^2 \cdot J_{k+1/2}(r) \cdot J_{k+3/2}(r)\, dr,$$

$$\int_0^\mu r \cdot J_{k+3/2}^2(r)\, dr = -\frac{k^2}{2 \cdot k + 1} \cdot J_{k+1/2}^2(\mu) + \frac{2}{2 \cdot k + 1} \cdot \int_0^\mu r^2 \cdot J_{k+1/2}(r) \cdot J_{k+3/2}(r)\, dr,$$

$$\int_0^\mu r \cdot J_{k+1/2}^2(r)\, dr = \frac{k \cdot (k+1)}{2 \cdot k + 1} \cdot J_{k+1/2}^2(\mu) + \frac{2}{2 \cdot k + 1} \cdot \int_0^\mu r^2 \cdot J_{k+1/2}(r) \cdot J_{k+3/2}(r)\, dr,$$

$$\int_0^\mu r \cdot J_{k+1/2}^2(r)\, dr = \frac{k \cdot (k+1) - \mu^2}{2 \cdot k + 1} \cdot J_{k+1/2}^2(\mu) + \frac{2 \cdot k + 3}{2 \cdot k + 1} \cdot \int_0^\mu r \cdot J_{k+1/2}^2(r)\, dr,$$

$$\frac{k \cdot (k+1) - \mu^2}{2 \cdot k + 1} \cdot J_{k+1/2}^2(\mu) = -\frac{2}{2 \cdot k + 1} \cdot \int_0^\mu r \cdot J_{k+1/2}^2(r)\, dr.$$

Finally we obtain:

$$\int_0^\mu r \cdot J_{k+1/2}^2(r)\, dr = \frac{\mu^2 - k \cdot (k+1)}{2} \cdot J_{k+1/2}^2(\mu). \tag{H.2}$$

# Appendix I

# The Bubnov–Galerkin functional

Let us write once more the functional for the Euler equation (5.4)

$$I(v) = (L[v], v) - 2 \cdot (f, v).$$

The linear operator $L[v]$ for the parabolic equation we are interested in has the form

$$L[v] = \frac{\partial v}{\partial t} - a^2 \cdot \Delta v.$$

The scalar product $(L[v], v)$ has the form

$$(L[v], v) = \int_M \left( \frac{\partial v}{\partial t} - a^2 \cdot \Delta v \right) \cdot v \, dM = \int_M \frac{\partial v}{\partial t} \cdot v \, dM - a^2 \cdot \int_M \Delta v \cdot v \, dM.$$

The scalar product $(f, v)$ has the form

$$(f, v) = \int_M f \cdot v \, dM.$$

The Ostrogradsky formula for the Laplace operator is known

$$\int_M \Delta v \, dM = \int_S \nabla v \, dS, \tag{I.1}$$

where $S$ is the boundary of the submanifold $M$, and $dS$ is its vector element directed along the outward normal. Let us integrate the second integral by parts

$$\int_M \Delta v \cdot v \, dM = \int_S v \cdot \nabla v \, dS - \int_M \nabla v \cdot \nabla v \, dM. \tag{I.2}$$

Thus, the original functional can be written as

$$I(v) = \int_M \frac{\partial v}{\partial t} \cdot v \, dM - a^2 \cdot \int_S v \cdot \nabla v \, dS + a^2 \cdot \int_M \nabla v \cdot \nabla v \, dM - 2 \cdot \int_M f \cdot v \, dM.$$

Let us give the increment $\widehat{v} + \epsilon \cdot v$, where $\widehat{v}$ is the exact solution, and $\epsilon$ is some small number. The full increment of the functional can be written as

$$I(\widehat{v} + \epsilon \cdot v) - I(\widehat{v}) = \int_M \frac{\partial(\widehat{v} + \epsilon \cdot v)}{\partial t} \cdot (\widehat{v} + \epsilon \cdot v)\, dM$$
$$- a^2 \cdot \int_S (\widehat{v} + \epsilon \cdot v) \cdot \nabla(\widehat{v} + \epsilon \cdot v)\, dS$$
$$+ a^2 \cdot \int_M \nabla(\widehat{v} + \epsilon \cdot v) \cdot \nabla(\widehat{v} + \epsilon \cdot v)\, dM$$
$$- 2 \cdot \int_M f \cdot (\widehat{v} + \epsilon \cdot v)\, dM - \int_M \frac{\partial \widehat{v}}{\partial t} \cdot \widehat{v}\, dM$$
$$+ a^2 \cdot \int_S \widehat{v} \cdot \nabla \widehat{v}\, dS - a^2 \cdot \int_M \nabla \widehat{v} \cdot \nabla \widehat{v}\, dM + 2 \cdot \int_M f \cdot \widehat{v}\, dM.$$

Neglecting the second-order terms proportional to $\epsilon^2$, we obtain

$$I(\widehat{v} + \epsilon \cdot v) - I(\widehat{v}) = \epsilon \cdot \int_M \left[ \frac{\partial \widehat{v}}{\partial t} \cdot v + \frac{\partial v}{\partial t} \cdot \widehat{v} \right] dM$$
$$- \epsilon \cdot a^2 \cdot \int_S \left[ \widehat{v} \cdot \nabla v + v \cdot \nabla \widehat{v} \right] dS$$
$$+ 2 \cdot \epsilon \cdot a^2 \cdot \int_M \nabla \widehat{v} \cdot \nabla v\, dM - 2 \cdot \epsilon \cdot \int_M f \cdot v\, dM.$$

If we take $\epsilon$ out of the brackets and take into account that $\epsilon$ can be negative, while the increment of the functional is always positive, since $\widehat{v}$ is the exact solution, we obtain as a mandatory condition

$$\int_M \left[ \frac{\partial \widehat{v}}{\partial t} \cdot v + \frac{\partial v}{\partial t} \cdot \widehat{v} \right] dM - a^2 \cdot \int_S \left[ \widehat{v} \cdot \nabla v + v \cdot \nabla \widehat{v} \right] dS$$
$$+ 2 \cdot a^2 \cdot \int_M \nabla \widehat{v} \cdot \nabla v\, dM - 2 \cdot \int_M f \cdot v\, dM = 0.$$

Taking into account that $\widehat{v} \approx v$, and cancelling the common factor, we return to the original equation. This means that the numerical solution of the boundary value problem reduces to the minimization of the functional

$$I(v) = \int_M \frac{\partial v}{\partial t} \cdot v\, dM - a^2 \cdot \int_S v \cdot \nabla v\, dS$$
$$+ a^2 \cdot \int_M \nabla v \cdot \nabla v\, dM - 2 \cdot \int_M f \cdot v\, dM \to \min. \quad \text{(I.3)}$$

The factor 2 in front of the linear term $\int_M f \cdot v dM$ is kept deliberately: it is precisely what ensures that the minimum condition $\delta I = 0$ returns the original equation rather than an equation with an extra factor of $1/2$ (when differentiating, the quadratic terms produce a factor of 2, and the linear term must have the same one).

Let us proceed to the separation into stationary and nonstationary problems. It is fundamentally important to perform this separation and the subsequent time discretization at the level of the *equation*, not of the functional: the time derivative is not a self-adjoint operator, so direct substitution of $v = \psi(M) \cdot \phi(t)$ into the functional would lead to incorrect coefficients. Let us write the weak form of the equation $L[u] = f$, obtained above by integration by parts

$$\int_M \frac{\partial u}{\partial t} \cdot v \, dM + a^2 \cdot \int_M \nabla u \cdot \nabla v \, dM - a^2 \cdot \int_S v \cdot \nabla u \, dS = \int_M f \cdot v \, dM, \quad \text{(I.4)}$$

which must hold for an arbitrary trial function $v$.

*The stationary equation with Dirichlet and/or Neumann boundary conditions*

If $\frac{\partial u}{\partial t} = 0$, the time term disappears, and due to the symmetry of the spatial operator the weak form (I.4) is equivalent to the minimization of the functional

$$a^2 \cdot \int_M \nabla v \cdot \nabla v \, dM - 2 \cdot \int_M f \cdot v \, dM - a^2 \cdot \int_S v \cdot \nabla v \, dS \to \min . \quad \text{(I.5)}$$

*The nonstationary equation with Dirichlet and/or Neumann boundary conditions*

We discretize the time derivative by the implicit Euler scheme with step $\Delta t$, applying it to the weak form of the equation: $\frac{\partial u}{\partial t} \approx \frac{u_n - u_{n-1}}{\Delta t}$, where $u_n$ is the solution on the current time layer, and $u_{n-1}$ on the previous one. Substituting this into (I.4) and multiplying by $\Delta t$, we obtain the equation for the time step

$$\int_M (u_n - u_{n-1}) \cdot v \, dM + \Delta t \cdot a^2 \cdot \int_M \nabla u_n \cdot \nabla v \, dM - \Delta t \cdot a^2 \cdot \int_S v \cdot \nabla u_n \, dS = \Delta t \cdot \int_M f \cdot v \, dM.$$

This equation, in turn, is the minimum condition of the functional

$$\int_M v_n^2 \, dM + \Delta t \cdot a^2 \cdot \int_M \nabla v_n \cdot \nabla v_n \, dM - 2 \cdot \int_M [\Delta t \cdot f + v_{n-1}] \cdot v_n \, dM - \Delta t \cdot a^2 \cdot \int_S v_n \cdot \nabla v_n \, dS \to \min . \quad \text{(I.6)}$$

Thus, the nonstationary equation has been reduced to successively solving stationary problems: at each step, from the known $v_{n-1}$ is found $v_n$.

**Matrix form**

We discretize the trial function over the mesh nodes: $v = \sum_i q_i \cdot \phi_i(M)$, where $\phi_i(M)$ are the basis functions, and $\overrightarrow{q} = (q_0, \ldots, q_n)^T$ is the vector of nodal values. The substitution turns each integral of the functional into a quadratic or linear form in $\overrightarrow{q}$, and their coefficients are assembled into matrices:

$$\int_M \nabla v \cdot \nabla v \, dM = \overrightarrow{q}^T \cdot K \cdot \overrightarrow{q},$$
$$\int_M v \cdot v \, dM = \overrightarrow{q}^T \cdot D \cdot \overrightarrow{q}, \tag{I.7}$$
$$\int_M f \cdot v \, dM = \overrightarrow{F}^T \cdot \overrightarrow{q}$$

where $K$ is the stiffness matrix (the integral of the product of the gradients), $D$ is the damping matrix (the integral of the product of the trial functions), $\overrightarrow{F}$ is the load vector (the integral of the source); the matrices $K$ and $D$ are symmetric. The boundary integral $-a^2 \int_S v \cdot \nabla v \, dS$ does not collapse into a matrix: it is a surface term that, in discrete form, enters only the equations of the boundary nodes (for an interior node the trial function vanishes on the boundary) and is determined by the boundary condition itself.

The discrete system is obtained by the Bubnov–Galerkin method: we substitute $v = \sum_i q_i \phi_i$ into the weak form (I.4) and successively take the trial functions $\phi_j$. For the **stationary** equation this gives the system

$$a^2 K \cdot \overrightarrow{q} = \overrightarrow{F} + a^2 \overrightarrow{s}, \tag{I.8}$$

where $\overrightarrow{s}$ is the vector of nodal boundary fluxes. The flux itself is not specified here — what to do with it is determined by the type of boundary condition. Accounting for boundary conditions is a nontrivial task, and it is discussed in detail in separate appendices: "Accounting for Dirichlet boundary conditions" and "Accounting for Neumann boundary conditions".

For the **nonstationary** equation (implicit Euler scheme, see (I.6)) the system at the time step takes the form

$$\left[D + \Delta t \cdot a^2 K\right] \cdot \overrightarrow{q}_n = \Delta t \cdot \overrightarrow{F} + D \cdot \overrightarrow{q}_{n-1} + \Delta t \cdot a^2 \overrightarrow{s}. \tag{I.9}$$

# Appendix J

# Applying the functional

Let us see how the Bubnov–Galerkin functional from Appendix I works in practice: we go through a one-dimensional example from start to finish — from the test functions to the system of linear equations in matrix form.

We solve the heat equation $L[u] = f$, where the parabolic operator $L = \frac{\partial u}{\partial t} - a^2 \Delta u$, on the segment $[x_0, x_4]$. At the ends we impose Neumann boundary conditions (1.14) — the fluxes $\frac{\partial u}{\partial n}\Big|_{x=x_0} = g_0$ and $\frac{\partial u}{\partial n}\Big|_{x=x_4} = g_4$ are prescribed. We bring the Bubnov–Galerkin functional to matrix form and substitute these conditions.

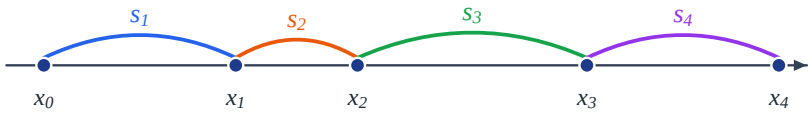


**Fig. J.1.** The example mesh: four simplex segments $s_1, \ldots, s_4$ and five nodes $x_0, \ldots, x_4$; the segment lengths are denoted $l_{(i)(i+1)}$.

Take four simplex segments $s_1, s_2, s_3, s_4$, three interior points $x_1, x_2, x_3$ and two boundary points $x_0, x_4$ (Figure J.1). The test functions on the simplex segments take the form

$$\begin{aligned}
\upsilon_{(0)(1)}(x) &= q_0 \cdot \phi_0(x) + q_1 \cdot \phi_1(x) \\
\upsilon_{(1)(2)}(x) &= q_1 \cdot \phi_1(x) + q_2 \cdot \phi_2(x) \\
\upsilon_{(2)(3)}(x) &= q_2 \cdot \phi_2(x) + q_3 \cdot \phi_3(x) \\
\upsilon_{(3)(4)}(x) &= q_3 \cdot \phi_3(x) + q_4 \cdot \phi_4(x)
\end{aligned}$$

The "hat" functions and their derivatives for the segment simplices are

$$\begin{aligned}
\phi_0(x) &= a_0 + b_0 \cdot x & \phi_0'(x) &= b_0 \\
\phi_1(x) &= a_1 + b_1 \cdot x & \phi_1'(x) &= b_1 \\
\phi_2(x) &= a_2 + b_2 \cdot x & \phi_2'(x) &= b_2 \\
\phi_3(x) &= a_3 + b_3 \cdot x & \phi_3'(x) &= b_3 \\
\phi_4(x) &= a_4 + b_4 \cdot x & \phi_4'(x) &= b_4
\end{aligned}$$

The coefficients take the following values. Let us introduce the notation $l_{(i)(i+1)} = x_{i+1} - x_i$ for the length of the simplex between points $x_i$ and $x_{i+1}$.

The hat-function coefficients for an arbitrary simplex between the points $x_i$ and $x_{i+1}$ are derived exactly as in the "Test functions" section and equal

$$a_i = \frac{x_{i+1}}{l_{(i)(i+1)}} \quad b_i = \frac{-1}{l_{(i)(i+1)}}$$
$$a_{i+1} = \frac{-x_i}{l_{(i)(i+1)}} \quad b_{i+1} = \frac{1}{l_{(i)(i+1)}}$$

For the boundary hats $\phi_0$ and $\phi_4$ one takes the corresponding half. Substituting these coefficients, we write the test functions on each simplex directly.

Let us substitute the coefficients into the test functions and express them in terms of $x$:

For simplex $s_1$:

$$\upsilon_{(0)(1)}(x) = q_0 \cdot \phi_0(x) + q_1 \cdot \phi_1(x) = q_0 \cdot \left(\frac{x_1}{l_{(0)(1)}} - \frac{x}{l_{(0)(1)}}\right) + q_1 \cdot \left(\frac{-x_0}{l_{(0)(1)}} + \frac{x}{l_{(0)(1)}}\right) = \frac{1}{l_{(0)(1)}}\left[q_0 \cdot (x_1 - x) + q_1 \cdot (x - x_0)\right]$$

For simplex $s_2$:

$$\upsilon_{(1)(2)}(x) = q_1 \cdot \phi_1(x) + q_2 \cdot \phi_2(x) = q_1 \cdot \left(\frac{x_2}{l_{(1)(2)}} - \frac{x}{l_{(1)(2)}}\right) + q_2 \cdot \left(\frac{-x_1}{l_{(1)(2)}} + \frac{x}{l_{(1)(2)}}\right) = \frac{1}{l_{(1)(2)}}\left[q_1 \cdot (x_2 - x) + q_2 \cdot (x - x_1)\right]$$

For simplex $s_3$:

$$\upsilon_{(2)(3)}(x) = q_2 \cdot \phi_2(x) + q_3 \cdot \phi_3(x) = q_2 \cdot \left(\frac{x_3}{l_{(2)(3)}} - \frac{x}{l_{(2)(3)}}\right) + q_3 \cdot \left(\frac{-x_2}{l_{(2)(3)}} + \frac{x}{l_{(2)(3)}}\right) = \frac{1}{l_{(2)(3)}}\left[q_2 \cdot (x_3 - x) + q_3 \cdot (x - x_2)\right]$$

For simplex $s_4$:

$$\upsilon_{(3)(4)}(x) = q_3 \cdot \phi_3(x) + q_4 \cdot \phi_4(x) = q_3 \cdot \left(\frac{x_4}{l_{(3)(4)}} - \frac{x}{l_{(3)(4)}}\right) + q_4 \cdot \left(\frac{-x_3}{l_{(3)(4)}} + \frac{x}{l_{(3)(4)}}\right) = \frac{1}{l_{(3)(4)}}\left[q_3 \cdot (x_4 - x) + q_4 \cdot (x - x_3)\right]$$

Let us find the derivatives of the test functions with respect to $x$:

For simplex $s_1$:

$$\upsilon'_{(0)(1)}(x) = \frac{1}{l_{(0)(1)}}(q_1 - q_0)$$

For simplex $s_2$:

$$\upsilon'_{(1)(2)}(x) = \frac{1}{l_{(1)(2)}}(q_2 - q_1)$$

For simplex $s_3$:

$$\upsilon'_{(2)(3)}(x) = \frac{1}{l_{(2)(3)}}(q_3 - q_2)$$

For simplex $s_4$:

$$\upsilon'_{(3)(4)}(x) = \frac{1}{l_{(3)(4)}}(q_4 - q_3)$$

Let us compute the squares of the test functions:
For simplex $s_1$:

$$\upsilon^2_{(0)(1)}(x) = \frac{1}{l^2_{(0)(1)}}\left[q_0 \cdot (x_1 - x) + q_1 \cdot (x - x_0)\right]^2$$
$$= \frac{1}{l^2_{(0)(1)}}\left[q_0^2(x_1 - x)^2 + 2q_0q_1(x_1 - x)(x - x_0) + q_1^2(x - x_0)^2\right]$$

For simplex $s_2$:

$$\upsilon^2_{(1)(2)}(x) = \frac{1}{l^2_{(1)(2)}}\left[q_1 \cdot (x_2 - x) + q_2 \cdot (x - x_1)\right]^2$$
$$= \frac{1}{l^2_{(1)(2)}}\left[q_1^2(x_2 - x)^2 + 2q_1q_2(x_2 - x)(x - x_1) + q_2^2(x - x_1)^2\right]$$

For simplex $s_3$:

$$\upsilon^2_{(2)(3)}(x) = \frac{1}{l^2_{(2)(3)}}\left[q_2 \cdot (x_3 - x) + q_3 \cdot (x - x_2)\right]^2$$
$$= \frac{1}{l^2_{(2)(3)}}\left[q_2^2(x_3 - x)^2 + 2q_2q_3(x_3 - x)(x - x_2) + q_3^2(x - x_2)^2\right]$$

For simplex $s_4$:

$$\upsilon^2_{(3)(4)}(x) = \frac{1}{l^2_{(3)(4)}}\left[q_3 \cdot (x_4 - x) + q_4 \cdot (x - x_3)\right]^2$$
$$= \frac{1}{l^2_{(3)(4)}}\left[q_3^2(x_4 - x)^2 + 2q_3q_4(x_4 - x)(x - x_3) + q_4^2(x - x_3)^2\right]$$

Let us compute the squares of the derivatives of the test functions:

For simplex $s_1$:

$$\left(\upsilon'_{(0)(1)}(x)\right)^2 = \frac{1}{l^2_{(0)(1)}}(q_1 - q_0)^2 = \frac{1}{l^2_{(0)(1)}}\left(q_0^2 - 2q_0q_1 + q_1^2\right)$$

For simplex $s_2$:

$$\left(\upsilon'_{(1)(2)}(x)\right)^2 = \frac{1}{l^2_{(1)(2)}}(q_2 - q_1)^2 = \frac{1}{l^2_{(1)(2)}}\left(q_1^2 - 2q_1q_2 + q_2^2\right)$$

For simplex $s_3$:

$$\left(\upsilon'_{(2)(3)}(x)\right)^2 = \frac{1}{l^2_{(2)(3)}}(q_3 - q_2)^2 = \frac{1}{l^2_{(2)(3)}}\left(q_2^2 - 2q_2q_3 + q_3^2\right)$$

For simplex $s_4$:

$$\left(\upsilon'_{(3)(4)}(x)\right)^2 = \frac{1}{l^2_{(3)(4)}}(q_4 - q_3)^2 = \frac{1}{l^2_{(3)(4)}}\left(q_3^2 - 2q_3q_4 + q_4^2\right)$$

Let us write the linear approximation of the source function $f(x)$ on each simplex according to (6.36). In our example the source acts only on the interior simplices: on the outer ones it is identically zero.

For simplex $s_1$ (between $x_0$ and $x_1$):

Since the heat source density function is identically zero on this interval:

$$\widetilde{f}_{(0)(1)}(x) = 0.$$

For simplex $s_2$ (between $x_1$ and $x_2$):

$$\widetilde{f}_{(1)(2)}(x) = e_1^{(1)(2)} + e_2^{(1)(2)} \cdot x,$$

where:

$$e_1^{(1)(2)} = \frac{f_1 \cdot x_2 - f_2 \cdot x_1}{x_2 - x_1} = \frac{f_1 \cdot x_2 - f_2 \cdot x_1}{l_{(1)(2)}}$$
$$e_2^{(1)(2)} = \frac{f_2 - f_1}{x_2 - x_1} = \frac{f_2 - f_1}{l_{(1)(2)}}$$

For simplex $s_3$ (between $x_2$ and $x_3$):

$$\widetilde{f}_{(2)(3)}(x) = e_1^{(2)(3)} + e_2^{(2)(3)} \cdot x,$$

where:

$$e_1^{(2)(3)} = \frac{f_2 \cdot x_3 - f_3 \cdot x_2}{x_3 - x_2} = \frac{f_2 \cdot x_3 - f_3 \cdot x_2}{l_{(2)(3)}}$$
$$e_2^{(2)(3)} = \frac{f_3 - f_2}{x_3 - x_2} = \frac{f_3 - f_2}{l_{(2)(3)}}$$

For simplex $s_4$ (between $x_3$ and $x_4$):

Since the heat source density function is identically zero on this interval:

$$\widetilde{f}_{(3)(4)}(x) = 0.$$

Now let us write the product of the source function and the corresponding test function for each simplex.

For simplex $s_1$:

Since $\widetilde{f}_{(0)(1)}(x) = 0$, we have:

$$\widetilde{f}_{(0)(1)}(x) \cdot \upsilon_{(0)(1)}(x) = 0.$$

For simplex $s_2$:

$$\widetilde{f}_{(1)(2)}(x) \cdot \upsilon_{(1)(2)}(x) = \left(e_1^{(1)(2)} + e_2^{(1)(2)} \cdot x\right) \cdot \frac{1}{l_{(1)(2)}} \left[q_1 \cdot (x_2 - x) + q_2 \cdot (x - x_1)\right].$$

Substituting the values of $e_1^{(1)(2)}$ and $e_2^{(1)(2)}$:

$$\widetilde{f}_{(1)(2)}(x) \cdot \upsilon_{(1)(2)}(x) = \frac{1}{l_{(1)(2)}^2} \left[(f_1 \cdot x_2 - f_2 \cdot x_1) + (f_2 - f_1) \cdot x\right] \cdot \left[q_1 \cdot (x_2 - x) + q_2 \cdot (x - x_1)\right].$$

Let us transform the first factor:

$$(f_1 \cdot x_2 - f_2 \cdot x_1) + (f_2 - f_1) \cdot x = f_1 \cdot (x_2 - x) + f_2 \cdot (x - x_1).$$

Therefore:

$$\widetilde{f}_{(1)(2)}(x) \cdot \upsilon_{(1)(2)}(x) = \frac{1}{l_{(1)(2)}^2} \left[f_1 \cdot (x_2 - x) + f_2 \cdot (x - x_1)\right] \cdot \left[q_1 \cdot (x_2 - x) + q_2 \cdot (x - x_1)\right].$$

Expanding the brackets:

$$\widetilde{f}_{(1)(2)}(x) \cdot \upsilon_{(1)(2)}(x) = \frac{1}{l_{(1)(2)}^2} \Big[f_1 q_1 (x_2 - x)^2 + f_1 q_2 (x_2 - x)(x - x_1) + f_2 q_1 (x - x_1)(x_2 - x) + f_2 q_2 (x - x_1)^2\Big].$$

For simplex $s_3$:

$$\widetilde{f}_{(2)(3)}(x) \cdot \upsilon_{(2)(3)}(x) = \left(e_1^{(2)(3)} + e_2^{(2)(3)} \cdot x\right) \cdot \frac{1}{l_{(2)(3)}} \left[q_2 \cdot (x_3 - x) + q_3 \cdot (x - x_2)\right].$$

By analogy with simplex $s_2$:

$$\widetilde{f}_{(2)(3)}(x) \cdot \upsilon_{(2)(3)}(x) = \frac{1}{l^2_{(2)(3)}} \left[f_2 \cdot (x_3 - x) + f_3 \cdot (x - x_2)\right] \quad \cdot \left[q_2 \cdot (x_3 - x) + q_3 \cdot (x - x_2)\right].$$

Expanding the brackets:

$$\widetilde{f}_{(2)(3)}(x) \cdot \upsilon_{(2)(3)}(x) = \frac{1}{l^2_{(2)(3)}} \Big[f_2 q_2 (x_3 - x)^2 + f_2 q_3 (x_3 - x)(x - x_2) + f_3 q_2 (x - x_2)(x_3 - x) + f_3 q_3 (x - x_2)^2\Big].$$

For simplex $s_4$:
Since $\widetilde{f}_{(3)(4)}(x) = 0$, we have:

$$\widetilde{f}_{(3)(4)}(x) \cdot \upsilon_{(3)(4)}(x) = 0.$$

Now consider the boundary term. The boundary of the one-dimensional domain consists of two points $x_0$ and $x_4$, so the surface integral reduces to a sum over the boundary nodes:

$$\int_S \upsilon \cdot \nabla \upsilon \, dS = q_0 \cdot \frac{\partial u}{\partial n}\bigg|_{x=x_0} + q_4 \cdot \frac{\partial u}{\partial n}\bigg|_{x=x_4}.$$

Since this example uses Neumann conditions, the normal derivative at the boundary is known: $\frac{\partial u}{\partial n}\Big|_{x=x_0} = g_0$, $\frac{\partial u}{\partial n}\Big|_{x=x_4} = g_4$. This term enters only the equations of the boundary nodes and is moved to the right-hand side as a flux vector; it does not enter the system matrix.

The stiffness integral for each simplex is the integral of the squared derivative of the test function. Since these derivatives are constants, integration reduces to multiplication by the simplex length.

By the local formula from the "Stiffness matrix 1D" section, on each simplex this integral equals $\frac{1}{l_{(i)(i+1)}}$.

The total stiffness integral over the whole domain $M$ is the sum of the integrals over all simplices:

$$\begin{aligned}\int_M (\nabla \upsilon)^2 \, dM &= \sum_{i=1}^{4} \int_{s_i} \left(\upsilon'(x)\right)^2 \, dx \\ &= \frac{1}{l_{(0)(1)}} \left(q_0^2 - 2q_0 q_1 + q_1^2\right) + \frac{1}{l_{(1)(2)}} \left(q_1^2 - 2q_1 q_2 + q_2^2\right) \\ &\quad + \frac{1}{l_{(2)(3)}} \left(q_2^2 - 2q_2 q_3 + q_3^2\right) + \frac{1}{l_{(3)(4)}} \left(q_3^2 - 2q_3 q_4 + q_4^2\right).\end{aligned}$$

The damping integral for each simplex is the integral of the squared test function.

By the local formula from the "Damping matrix 1D" section, the integral of the squared test function on a simplex equals $\frac{l_{(i)(i+1)}}{3}\left(q_i^2 + q_i q_{i+1} + q_{i+1}^2\right)$.

The total damping integral over the whole domain $M$ is the sum of the integrals over all simplices:

$$\begin{aligned}\int_M \upsilon^2\,dM &= \sum_{i=1}^{4}\int_{s_i}\upsilon^2(x)\,dx \\ &= \frac{l_{(0)(1)}}{3}\left(q_0^2+q_0q_1+q_1^2\right)+\frac{l_{(1)(2)}}{3}\left(q_1^2+q_1q_2+q_2^2\right) \\ &\quad+\frac{l_{(2)(3)}}{3}\left(q_2^2+q_2q_3+q_3^2\right)+\frac{l_{(3)(4)}}{3}\left(q_3^2+q_3q_4+q_4^2\right).\end{aligned}$$

The load integral for each simplex is the integral of the product of the source function and the test function.

By the local formula from the "Load vector 1D" section, the integral vanishes on the boundary simplices $s_1$ and $s_4$ (the source is identically zero there), while on $s_2$ and $s_3$ it equals the corresponding local load-vector contribution.

The total load integral over the whole domain $M$ is the sum of the integrals over all simplices:

$$\begin{aligned}\int_M \widetilde{f}(x)\cdot\upsilon(x)\,dM &= \sum_{i=1}^{4}\int_{s_i}\widetilde{f}(x)\cdot\upsilon(x)\,dx \\ &= 0+\frac{l_{(1)(2)}}{6}\left(2f_1q_1+f_1q_2+f_2q_1+2f_2q_2\right) \\ &\quad+\frac{l_{(2)(3)}}{6}\left(2f_2q_2+f_2q_3+f_3q_2+2f_3q_3\right)+0.\end{aligned}$$

Simplifying:

$$\begin{aligned}\int_M \widetilde{f}(x)\cdot\upsilon(x)\,dM &= \frac{l_{(1)(2)}}{6}\left(2f_1q_1+f_1q_2+f_2q_1+2f_2q_2\right) \\ &\quad+\frac{l_{(2)(3)}}{6}\left(2f_2q_2+f_2q_3+f_3q_2+2f_3q_3\right).\end{aligned}$$

Let us write all the integrals obtained in matrix form. Introduce the vector of unknowns $\overrightarrow{q}=\begin{pmatrix}q_0 & q_1 & q_2 & q_3 & q_4\end{pmatrix}^T$.

The stiffness integral can be written in matrix form:

$$\int_M(\nabla\upsilon)^2\,dM=\overrightarrow{q}^T\cdot K\cdot\overrightarrow{q},$$

where $K$ is the stiffness matrix of size $5\times5$:

$$K=\begin{pmatrix}
\frac{1}{l_{(0)(1)}} & -\frac{1}{l_{(0)(1)}} & 0 & 0 & 0 \\
-\frac{1}{l_{(0)(1)}} & \frac{1}{l_{(0)(1)}}+\frac{1}{l_{(1)(2)}} & -\frac{1}{l_{(1)(2)}} & 0 & 0 \\
0 & -\frac{1}{l_{(1)(2)}} & \frac{1}{l_{(1)(2)}}+\frac{1}{l_{(2)(3)}} & -\frac{1}{l_{(2)(3)}} & 0 \\
0 & 0 & -\frac{1}{l_{(2)(3)}} & \frac{1}{l_{(2)(3)}}+\frac{1}{l_{(3)(4)}} & -\frac{1}{l_{(3)(4)}} \\
0 & 0 & 0 & -\frac{1}{l_{(3)(4)}} & \frac{1}{l_{(3)(4)}}
\end{pmatrix}. \qquad \text{(J.1)}$$

Matrix $K$ is symmetric and tridiagonal.

The damping integral can be written in matrix form:

$$\int_M v^2\,dM = \overrightarrow{q}^T \cdot D \cdot \overrightarrow{q},$$

where $D$ is the damping matrix of size $5 \times 5$:

$$D = \begin{pmatrix} \frac{l_{(0)(1)}}{3} & \frac{l_{(0)(1)}}{6} & 0 & 0 & 0 \\ \frac{l_{(0)(1)}}{6} & \frac{l_{(0)(1)}}{3} + \frac{l_{(1)(2)}}{3} & \frac{l_{(1)(2)}}{6} & 0 & 0 \\ 0 & \frac{l_{(1)(2)}}{6} & \frac{l_{(1)(2)}}{3} + \frac{l_{(2)(3)}}{3} & \frac{l_{(2)(3)}}{6} & 0 \\ 0 & 0 & \frac{l_{(2)(3)}}{6} & \frac{l_{(2)(3)}}{3} + \frac{l_{(3)(4)}}{3} & \frac{l_{(3)(4)}}{6} \\ 0 & 0 & 0 & \frac{l_{(3)(4)}}{6} & \frac{l_{(3)(4)}}{3} \end{pmatrix}.$$

Matrix $D$ is also symmetric and tridiagonal.
The load integral can be written in matrix form:

$$\int_M \widetilde{f}(x) \cdot v(x)\,dM = \overrightarrow{F}^T \cdot \overrightarrow{q},$$

where $\overrightarrow{F}$ is the load vector of size $5 \times 1$:

$$\overrightarrow{F} = \begin{pmatrix} 0 \\ \frac{2f_1 \cdot l_{(1)(2)}}{6} + \frac{f_2 \cdot l_{(1)(2)}}{6} \\ \frac{f_1 \cdot l_{(1)(2)}}{6} + \frac{2f_2 \cdot l_{(1)(2)}}{6} + \frac{2f_2 \cdot l_{(2)(3)}}{6} + \frac{f_3 \cdot l_{(2)(3)}}{6} \\ \frac{f_2 \cdot l_{(2)(3)}}{6} + \frac{2f_3 \cdot l_{(2)(3)}}{6} \\ 0 \end{pmatrix}.$$

After simplification:

$$\overrightarrow{F} = \begin{pmatrix} 0 \\ \frac{l_{(1)(2)}}{6}(2f_1 + f_2) \\ \frac{f_1 \cdot l_{(1)(2)}}{6} + \frac{f_2(l_{(1)(2)} + l_{(2)(3)})}{3} + \frac{f_3 \cdot l_{(2)(3)}}{6} \\ \frac{l_{(2)(3)}}{6}(f_2 + 2f_3) \\ 0 \end{pmatrix}.$$

The boundary integral, as shown above, does not fold into a matrix — it gives a flux vector on the right-hand side, nonzero only at the boundary nodes:

$$a^2 \overrightarrow{s} = a^2 \begin{pmatrix} g_0 \\ 0 \\ 0 \\ 0 \\ g_4 \end{pmatrix}.$$

*Stationary equation*
By (I.8), the stationary system has the form

$$a^2 K \cdot \overrightarrow{q} = \overrightarrow{F} + a^2 \overrightarrow{s}.$$

*Non-stationary equation*
By (I.9), the nonstationary system:

$$\left[D + \Delta t \cdot a^2 K\right] \cdot \overrightarrow{q}_n = \Delta t \cdot \overrightarrow{F} + D \cdot \overrightarrow{q}_{n-1} + \Delta t \cdot a^2 \overrightarrow{s}.$$

Notation:

- $K$ — the stiffness matrix $5 \times 5$,
- $D$ — the damping matrix $5 \times 5$,
- $\overrightarrow{F}$ — the load vector $5 \times 1$,
- $\overrightarrow{q}$ — the vector of unknowns $5 \times 1$ for the stationary case,
- $\overrightarrow{q}_n$, $\overrightarrow{q}_{n-1}$ — the vectors of unknowns on the current and previous time layers,
- $a$ — the thermal diffusivity coefficient,
- $\Delta t$ — the time step.

The matrices $K$ and $D$ are symmetric and sparse (tridiagonal). The operator of the stationary system is the full stiffness matrix $a^2 K$; for a pure Neumann problem it is singular ($K \cdot \overrightarrow{1} = 0$), so the solution is defined up to an additive constant and is closed by a normalization condition. The treatment of Neumann conditions is detailed in the appendix on Neumann boundary conditions.

# Appendix K

# Imposing Dirichlet boundary conditions

There is a simple and widely used method for imposing Dirichlet conditions — direct substitution of the known boundary values into the system (the row-and-column elimination method). Suppose a node $i$ carries the Dirichlet condition $q_i = g_i$, where $g_i$ is the known boundary value. Then on the system $A\overrightarrow{q} = \overrightarrow{P}$ we perform the following operations:

1. **Right-hand side modification:** For every row $j \neq i$ subtract the contribution of the known value $q_i$ from the right-hand side: $p_j := p_j - A_{ji} \cdot g_i$.
2. **Zeroing the row and column:** Replace the $i$-th row and $i$-th column of the matrix $A$ with zeros, except for the diagonal element: $A_{ij} := 0, \quad A_{ji} := 0, \quad j \neq i$.
3. **Setting the diagonal element:** Set the diagonal element to one: $A_{ii} := 1$.
4. **Setting the boundary value:** Replace the $i$-th component of the right-hand side with the boundary value: $p_i := g_i$.

After these operations the $i$-th equation of the system takes the form $1 \cdot q_i = g_i$, which guarantees that the boundary condition holds.

Let us return to our one-dimensional example with five nodes $q_0, q_1, q_2, q_3, q_4$. Suppose Dirichlet conditions are prescribed at the boundaries:

$$q_0 = g_0, \quad q_4 = g_4.$$

**Stationary case**

The original system has the form:

$$a^2 K \cdot \overrightarrow{q} = \overrightarrow{F}.$$

The matrix $a^2K$ is full (symmetric and tridiagonal); boundary conditions are imposed by replacing the boundary rows.

Denote $A = a^2K$ and $\overrightarrow{P} = \overrightarrow{F}$. We apply the direct method for the nodes $q_0$ and $q_4$.

*Step 1: Right-hand side modification for the interior nodes*

For the node $q_1$ (index 1):

$$p_1 := p_1 - A_{10} \cdot g_0 = F_1 - a^2 \cdot \left(-\frac{1}{l_{(0)(1)}}\right) \cdot g_0 = F_1 + \frac{a^2 g_0}{l_{(0)(1)}}.$$

For the node $q_3$ (index 3):

$$p_3 := p_3 - A_{34} \cdot g_4 = F_3 - a^2 \cdot \left(-\frac{1}{l_{(3)(4)}}\right) \cdot g_4 = F_3 + \frac{a^2 g_4}{l_{(3)(4)}}.$$

*Steps 2–4: Modifying the matrix and right-hand side for the boundary nodes*
For the node $q_0$ (index 0):

- Zero the first row: $A_{01} := 0$
- Zero the first column: $A_{10} := 0$
- Set the diagonal: $A_{00} := 1$
- Set the right-hand side: $p_0 := g_0$

For the node $q_4$ (index 4):

- Zero the last row: $A_{43} := 0$
- Zero the last column: $A_{34} := 0$
- Set the diagonal: $A_{44} := 1$
- Set the right-hand side: $p_4 := g_4$

In the matrix $a^2K$ the boundary rows and columns (nodes $q_0$ and $q_4$) are zeroed and the diagonal is set to one; the known values are moved to the right-hand side.
The modified right-hand side vector:

$$\widetilde{\overrightarrow{P}} = \begin{pmatrix} g_0 \\ F_1 + \frac{a^2 g_0}{l_{(0)(1)}} \\ F_2 \\ F_3 + \frac{a^2 g_4}{l_{(3)(4)}} \\ g_4 \end{pmatrix}.$$

Note: since the source-density function is identically zero on the simplices $s_1$ and $s_4$, we have $F_0 = 0$ and $F_4 = 0$. However, the components $F_1, F_2, F_3$ may be nonzero.
**Non-stationary case**
For the non-stationary case the system has the form:

$$\left[D + \Delta t \cdot a^2 K\right] \cdot \overrightarrow{q}_n = \Delta t \cdot \overrightarrow{F} + D \cdot \overrightarrow{q}_{n-1}.$$

Denote $A = D + \Delta t \cdot a^2 K$ and $\overrightarrow{P} = \Delta t \cdot \overrightarrow{F} + D \cdot \overrightarrow{q}_{n-1}$. We apply the same direct method.
*Right-hand side modification for the interior nodes*
For the node $q_1$:

$$p_1 := p_1 - A_{10} \cdot g_0 = \Delta t \cdot F_1 + \sum_j D_{1j} q_{n-1,j} - \left(\frac{l_{(0)(1)}}{6} - \frac{\Delta t \cdot a^2}{l_{(0)(1)}}\right) \cdot g_0.$$

For the node $q_3$:

$$p_3 := p_3 - A_{34} \cdot g_4 = \Delta t \cdot F_3 + \sum_j D_{3j} q_{n-1,j} - \left( \frac{l_{(3)(4)}}{6} - \frac{\Delta t \cdot a^2}{l_{(3)(4)}} \right) \cdot g_4.$$

As in the stationary case, the boundary rows and columns of $D + \Delta t\, a^2 K$ are zeroed and the diagonal is set to one, while $q_0 = g_0$ and $q_4 = g_4$ go to the right-hand side.

# Appendix L

# Imposing Neumann boundary conditions

Neumann boundary conditions are handled much more simply than Dirichlet ones: the flux prescribed at the boundary immediately gives the vector $\overrightarrow{s}$ (nonzero only at the boundary nodes), which enters only the right-hand side of the system — the stiffness matrix $a^2K$ need not be changed.

**Stationary case**

For the five-node example (the fluxes $g_0$ and $g_4$ are prescribed; the source on the end simplices is zero, $F_0 = F_4 = 0$) the modified right-hand side vector is:

$$\overrightarrow{P} = \overrightarrow{F} + a^2\overrightarrow{s} = \begin{pmatrix} a^2 g_0 \\ F_1 \\ F_2 \\ F_3 \\ a^2 g_4 \end{pmatrix}.$$

**Non-stationary case**

For the non-stationary problem the operator is $D + \Delta t \cdot a^2 K$, and the flux enters the right-hand side with the factor $\Delta t$:

$$\left[D + \Delta t \cdot a^2 K\right] \cdot \overrightarrow{q}_n = \Delta t \cdot \overrightarrow{F} + D \cdot \overrightarrow{q}_{n-1} + \Delta t \cdot a^2 \overrightarrow{s}.$$

The modified right-hand side vector:

$$\overrightarrow{P} = \Delta t \cdot \overrightarrow{F} + D \cdot \overrightarrow{q}_{n-1} + \Delta t \cdot a^2 \overrightarrow{s} = \Delta t \cdot \overrightarrow{F} + D \cdot \overrightarrow{q}_{n-1} + \begin{pmatrix} \Delta t \cdot a^2 g_0 \\ 0 \\ 0 \\ 0 \\ \Delta t \cdot a^2 g_4 \end{pmatrix}.$$

The operator $D + \Delta t \cdot a^2 K$ is already nonsingular (the damping matrix $D$ is positive definite, $(D + \Delta t \cdot a^2 K)\,\overrightarrow{1} = D\,\overrightarrow{1} \neq 0$), so the non-stationary Neumann problem is solvable even without Dirichlet conditions.

# Appendix M

# Triangulation by divide and conquer

This appendix describes a triangulation algorithm following the divide-and-conquer scheme, with merging of sub-triangulations along common tangents. Historically it was the first one implemented in the project; the main text describes the currently used approach built on the CGAL library. The material below is kept as supplementary: it helps to understand the classical scheme and the difficulties with degenerate configurations that eventually led to the change of approach.

The basis is the divide-and-conquer algorithm from A. V. Skvortsov's monograph "Delaunay Triangulation and Its Applications", with immediate Delaunay restructuring right after each new triangle is built. The algorithm has $O(N \log N)$ average- and worst-case complexity in the number of input points.

## General scheme

The idea follows the classical divide-and-conquer paradigm.

1. The point set is recursively split by a straight line into two roughly equal parts.
2. The recursion continues until each part contains three or four points — such configurations are triangulated explicitly.
3. The resulting sub-triangulations are pairwise stitched together along the common zone between their convex hulls.
4. During stitching, every new edge is checked against the Delaunay condition, and if necessary the pair of adjacent triangles is restructured by a diagonal swap.

Since any partial triangulation produced at intermediate recursion levels is a triangulation of the convex hull of the corresponding subset, the merge step always works with two convex regions. This greatly simplifies the search for tangents and the control of non-intersection.

## Recursive point splitting

The split direction is chosen by the shape of the bounding rectangle: if it is wider than it is tall, the points are sorted and split by the horizontal coordinate, otherwise by the vertical one. This choice shortens the future seam between the sub-triangulations and therefore reduces the number of restructurings.

The sizes of the two parts are chosen so that all leaf subproblems have exactly three or four points. Large sets are split in half, while for a few small sizes a special "3 and $N-3$" split is used, which guarantees that no leaf of the recursion ends up with one or two points. The correctness of this split for any $N \geq 6$ follows from the fact that any $N$ is representable as $3k$, $3(k-1)+4$ or $3(k-2)+4 \cdot 2$.

In formalized form the recursive algorithm looks as follows.

1. If the number of points $N = 3$, build a triangulation of a single triangle.
2. Otherwise, if $N = 4$, build a triangulation of two or three triangles.
3. Otherwise, if $N = 8$, split the set into two parts of four points each, apply the algorithm recursively and merge the results.
4. Otherwise, if $N < 12$, split the set into parts of three and $N-3$ points, apply the algorithm recursively and merge the results.
5. Otherwise ($N \geq 12$) split the set in half, into parts of $\lfloor N/2 \rfloor$ and $\lceil N/2 \rceil$ points, apply the algorithm recursively and merge the results.

The recursion depth is $O(\log N)$, and the total work at each level is linear in the number of points, which gives the overall $O(N \log N)$ complexity in both the average and the worst case (figure Splitting).

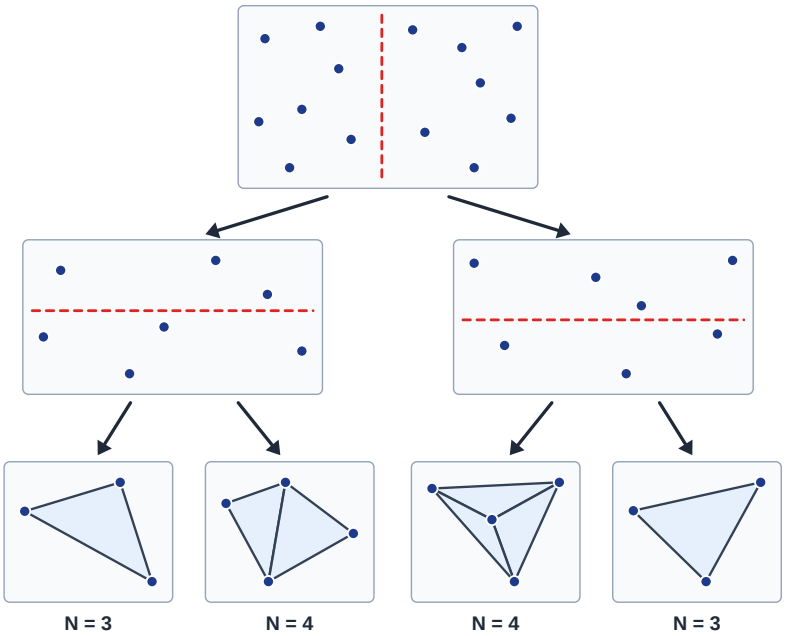


**Fig. M.1.** Splitting the input point set: choosing the direction, the cut, and leaf configurations of three and four points.

## Base configurations

For three points there is a single admissible triangulation — one triangle built directly on those three vertices. For four points two cases are possible, depending on their relative position (figure Base cases).

1. *All four points lie on the convex hull.* The quadrilateral is split by one of its two diagonals. The choice is determined by the Delaunay condition: we check whether one of the vertices lies inside the circumscribed circle of the triangle built on the remaining three. The diagonal for which the condition holds is the one used in the base triangulation.

2. *One point lies inside the triangle formed by the other three.* The interior point is connected to each vertex of the outer triangle, splitting the region into three triangles.

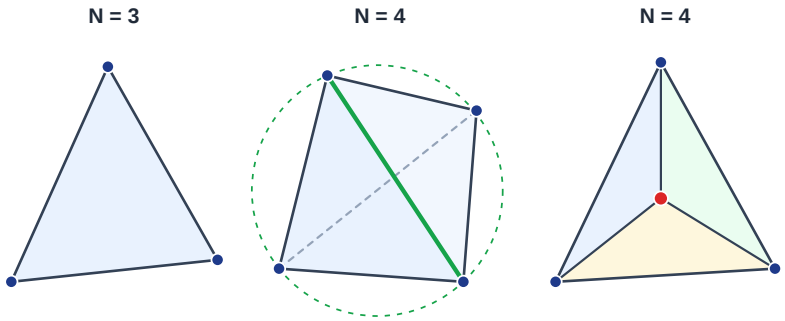


**Fig. M.2.** Base cases of the recursion: the single triangle for three points and the two configurations for four points.

## Merging two sub-triangulations

Merging is the most substantial stage of the algorithm. Suppose we have left and right sub-triangulations, each a Delaunay triangulation of the convex hull of its points. The merge proceeds in three steps.

**Finding common tangents.** First, two common tangents — the lower and the upper — are found on the two convex hulls. An initial approximation is the edge connecting the rightmost point of the left hull with the leftmost point of the right hull; it is then iteratively advanced along the vertices of both hulls until it lies below (or above) all remaining points of the two sets. Between the two tangents remain two chains of boundary vertices — left and right — and the space between them is what has to be filled with triangles.

**Filling the merge zone.** The lower tangent is declared the current *base edge*. On both sides of it the nearest vertices of the two chains are taken, giving two candidates for the next triangle: one with the left vertex and one with the right. A candidate is admissible if it does not invert orientation (the interior angle at the base edge is less than straight) and does not cover already built parts of the mesh. If both candidates are admissible, the one with the larger minimal angle is chosen: such a triangle is "more equilateral" and less likely to be restructured later. The open side of the built triangle becomes the new base edge, and the step repeats until the base edge coincides with the upper tangent (figure Merge).

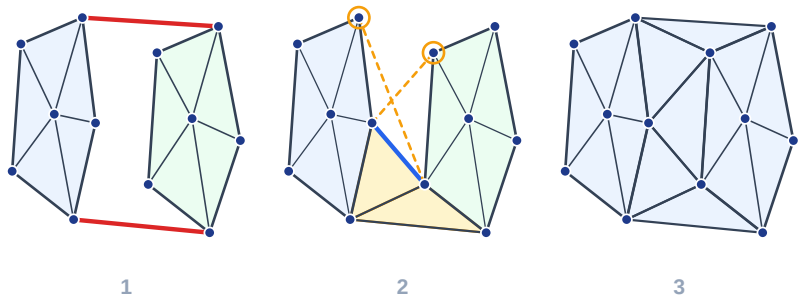


**Fig. M.3.** Merging triangulations: common tangents (1), filling the merge zone from the base edge (2), the final mesh after restructurings (3).

**Immediate local restructuring.** After a new triangle is added, its edges are checked against previously built neighbours: if the quadrilateral formed by the new triangle and its edge neighbour violates the Delaunay condition, a flip is performed. Immediate restructuring removes the need for a separate final pass over the mesh and usually reduces the total number of restructurings.

## Inserting constraint edges

The algorithm above builds the Delaunay triangulation of the convex hull of the points but knows nothing about domain boundaries and hole contours. To guarantee that prescribed edges are present in the mesh, they are inserted into the finished triangulation by a separate edge-recovery operation.

A single edge is inserted as follows: all triangles crossed by the segment are found and removed — this leaves a polygonal cavity in the mesh, which the inserted segment splits into two chains of vertices, an upper and a lower one; each chain is independently re-filled with triangles, and the edge ends up in the mesh (figure Edge insertion). After insertion the edge is marked as a *constraint*: flipping across it is forbidden, so no later restructuring can destroy it, while the Delaunay condition near it may legitimately be violated.

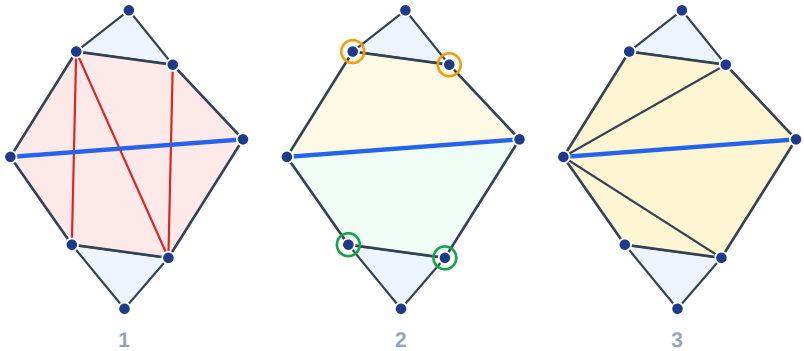


**Fig. M.4.** Inserting a non-crossable edge: removing the crossed triangles, forming two chains, filling with new triangles.

In the current version of the project this very operation is performed by CGAL (the *insert_constraint* method), but the logic inside is the same.

## Why the scheme was abandoned

The divide-and-conquer scheme relies on building convex hulls, finding tangents and controlling orientation — all these operations reduce to sign predicates evaluated in floating-point arithmetic. On "good" inputs (in general position) this works, but on regular, axis-aligned data typical of a square computational domain, *degenerate* configurations arise: many points on one vertical or horizontal line (collinearity) and rectangular quadruples of points lying on one circle (cocircularity). In these cases the predicate value is an exact zero, while floating-point evaluation yields noise of arbitrary sign instead. The tangent search and the merge start to contradict themselves: flat (zero-area) "fan" triangles appear along the boundary and the Delaunay condition is violated. These are precisely the difficulties that led to the switch to the library-based approach with exact predicates described in the main text.

# Appendix N

# Elliptic equations: reduction to a stationary problem

The main text of the book deals with the non-stationary thermal field — a parabolic equation. However, the same finite-element machinery carries over without any changes to a wide class of *elliptic* boundary value problems, in which time is either absent from the formulation altogether, or the transient is of no interest and only the steady (stationary) state of the field is required. This appendix shows how such a problem is reduced to a simpler form and solved by the methods already discussed: assembling the stiffness matrix and iteratively solving a sparse symmetric positive definite system.

## From the parabolic equation to the elliptic one

The original non-stationary equation (1.8) has the form

$$\frac{\partial T(M,t)}{\partial t} = a^2 \cdot \Delta T(M,t) + f(M,t).$$

If only the steady field is of interest, the time derivative vanishes, $\partial T/\partial t = 0$, and the equation loses the time coordinate, turning into the *elliptic* Poisson equation

$$a^2 \cdot \Delta T(M) + f(M) = 0, \qquad \text{or} \qquad -\Delta T(M) = \frac{1}{a^2} f(M). \tag{N.1}$$

In the particular case of no sources ($f \equiv 0$) it degenerates into the homogeneous Laplace equation

$$\Delta T(M) = 0. \tag{N.2}$$

It is essential that the unknown function here depends only on the geometry $M$ and not on time: the problem has turned from an evolutionary one into a purely spatial one.

## Example: the electrostatic field

The canonical elliptic equation is the equation of electrostatics. The electric potential $V(M)$ in a medium without volume charges satisfies the Laplace equation, and in the presence of a charge density $\rho$ — the Poisson equation

$$\Delta V(M) = -\frac{\rho(M)}{\varepsilon}, \tag{N.3}$$

where $\varepsilon$ is the permittivity of the medium [F/m]. Comparison with (N.1) shows a complete correspondence, down to the dimensions: the potential $V$ [V] plays the role of the temperature $T$ [K], the ratio $\rho/\varepsilon$ with the charge density $\rho$ [C/m$^3$] has dimension [V/m$^2$] and plays the role of the source density $f/a^2$ of dimension [K/m$^2$], and the coefficient $a^2$ is replaced by one. Consequently, the very same program that solves the stationary heat conduction problem computes the electrostatic field without modifications.

Consider an illustrative configuration of three conductors — "phases" — each kept at its own voltage $V_1, V_2, V_3$. The conductors impose a Dirichlet condition on their boundaries, $V|_{\partial\Omega_k} = V_k$; the outer boundary of the computational domain is assumed insulated, which corresponds to the homogeneous Neumann condition $\partial V/\partial n = 0$. The presence of at least one conductor makes the finite-element system non-degenerate, and it is solved by the same preconditioned conjugate gradient method.

The computed field picture is shown in figure Field of three conductors: the equipotentials and the field lines form a mutually orthogonal grid; the field lines leave the positive phase and terminate on the negative ones.

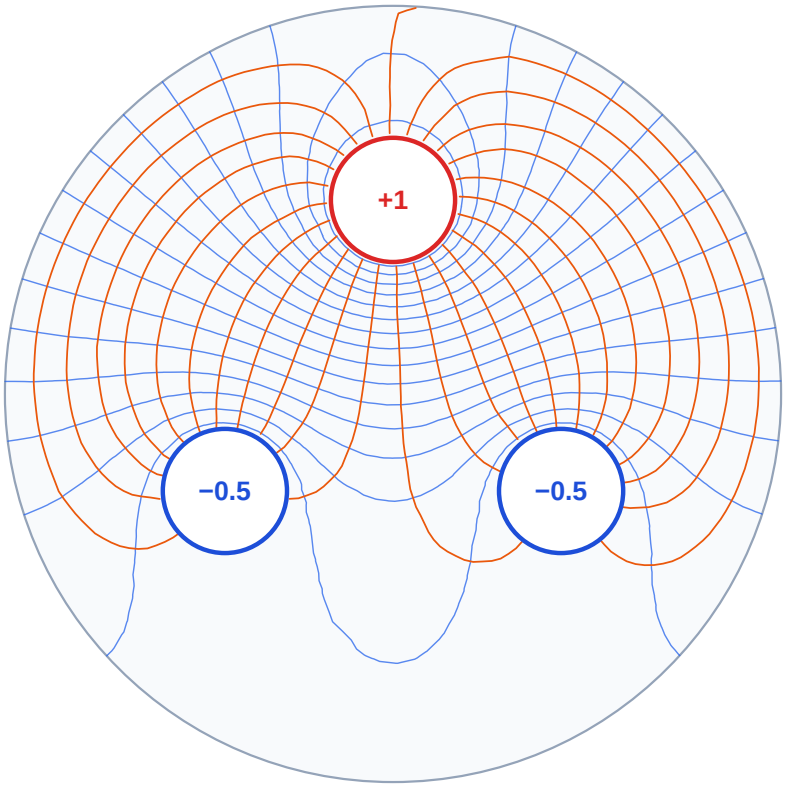


**Fig. N.1.** Field picture of three conductors ($V = +1, -0.5, -0.5$): blue lines — equipotentials, orange — field lines; computed on a mesh of 6486 triangles.

# Bibliography

[1] Angot, A. *Compléments de mathématiques à l'usage des ingénieurs de l'électrotechnique et des télécommunications.* Paris: Masson, 1972.

[2] Kartashov, E. M. *Analytical Methods in the Theory of Heat Conduction in Solids.* Moscow: Vysshaya Shkola, 2001. 550 p. (in Russian)

[3] Strang, G., Fix, G. *An Analysis of the Finite Element Method.* Englewood Cliffs, NJ: Prentice-Hall, 1973. 306 p.

[4] Skvortsov, A. V. *Delaunay Triangulation and Its Applications.* Tomsk: Tomsk University Press, 2002. 128 p. (in Russian)