САНКТ-ПЕТЕРБУРГСКИЙ
ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

ПРИОРИТЕТНЫЙ
НАЦИОНАЛЬНЫЙ ПРОЕКТ
"ОБРАЗОВАНИЕ"

Проект «Инновационная образовательная среда в классическом университете»

Пилотный проект № 22 «Разработка и внедрение
инновационной образовательной программы «Прикладные математика и физика»»

Физический факультет

Кафедра
Вычислительной физики

С.Л. Яковлев, Е.А. Яревский

Численные методы для дифференциальных уравнений в частных производных

Учебно-методическое пособие

.

Санкт Петербург
2007 г.

- Рецензент: зав. кафедрой высшей математики и математической физики, д.ф.-м.н., проф. Буслаев В.С.

- Печатается по решению методической комиссии физического факультета СПбГУ.

- Рекомендовано Ученым советом физического факультета СПбГУ.

Численные методы для дифференциальных уравнений в частных производных. – СПб., 2007

В учебно-методическом пособии рассмотрены вариационные методы решения дифференциальных уравнений в частных производных. Подробно описан метод конечных элементов: триангуляция, различные типы элементов, оценка погрешностей, стратегии улучшения решений. Приведена краткая информация о методе граничных элементов и спектральных методах.
Пособие предназначено для студентов 5-7-го курсов, аспирантов, соискателей и других обучающихся.

S.L. Yakovlev, E.A. Yarevsky
Numerical methods for the partial differential equations. St Petersburg, 2007

The variational methods for the solution of partial differential equations are considered. The finite element method is presented in details, including the triangulation, various types of finite elements, error estimations and the solution refinements. The boundary element method and the spectral methods are shortly discussed.

1 Introduction

Many physical processes in nature are described by equations that involve physical quantities together with their spatial and temporal partial derivatives. Among such processes are the weather, flow of liquids, deformation of solid bodies, heat transfer, chemical reactions, electromagnetics, quantum evolution, and many others. Equations involving partial derivatives are called partial differential equations (PDEs). For most PDEs we are not able to find their exact solutions, and in most cases the only way to solve PDEs is to approximate their solutions numerically. Numerical methods for PDEs constitute an indivisible part of modern engineering and science.

In these lectures, we describe the modern approaches to the numerical solution of PDEs. We start with the formulation of the abstract minimization and abstract variational problems. We show how PDEs can be reduced to the abstract variational problem. Then we discuss the Galerkin approach, a powerful tool for the reduction of the infinite dimensional problem to the finite dimensional one, and its extensions.

The variational problem and Galerkin approach form the basis for the finite element method (FEM). The FEM is one of the most general and efficient tool for the numerical solution of PDEs. The FEM is based on the spatial subdivision of the physical domain into finite elements, where the solution is approximated via a finite set of polynomial shape functions. In this way the original problem is transformed into a discrete problem for a finite number of unknown coefficients.

In the lectures, we present the general structure of the FEM and analyze in detail how the one-dimensional FEM works. Then we describe the multidimensional FEM: the Lagrange and hierarchical elements, triangulation methods, coordinate transformation. The special attention is paid to the error analysis: we analyze interpolation and integration errors while other sources of errors are also discussed. A priori and a posteriori error estimation are studied, and different adaptive refining strategies are considered.

In last lectures we discuss two other numerical methods employed for the PDE solution. The boundary element method (BEM) is closely related to the FEM and permits an essential reduction of the computational effort when applicable. The spectral method, another kind of the Galerkin-type methods, is based on the idea which is opposite to the FEM approach: the basis functions are chosen to be global and, as a rule, rather complicated. Despite this difference, the spectral methods have also been proved to be very effective for the numerical solution of PDEs.

The area which we deal with in these lectures, is the area of active research both in theory and applications. We tried to highlight here the main ideas of the field, both classical and emerging. Many details and further developments are missed but can be found in the books from the reference list, and articles.

Last but not least: the theory and applications in this area come hand in hand. In order to better understand theoretical results, one should work on real physical and engineering problems. On the other hand, the efficient and accurate calculations are only possible when the underlying mathematical results are understood.

Lecture 1

Abstract minimization problem. Variational inequalities. Abstract variational problem. The Lax-Milgram lemma.

2 The abstract minimization problem

Let us look for a value of u (e.g., *displacement of a mechanical system*) that satisfies the equation

$$u \in U \quad \text{and} \quad J(u) = \inf_{v \in U} J(v).$$

Here V is a space, and U (*allowed displacement*) is a subset of the space V, $J(u)$ is the energy of the system. The definition of the energy depends on the system, but we consider here the simplest (while very important) case of quadratic energy form:

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

In this equation, $a(.,.)$ is the symmetric bilinear form, and f is the linear form. Both these forms are defined on the space V and are continuous there.

Let us now be more precise with the definitions. Namely, let

- V be a linear vector space with the norm $||.||$,
- $a(.,.)$ be a bilinear continuous form $a(.,.) : V \times V \to \mathbf{R}$,
- f be a linear continuous form $f : V \to \mathbf{R}$,
- U be a non-empty subset of the space V.

Then we formulate the **Abstract minimization problem** (AMP) as the problem to find such u that

$$u \in U \quad \text{and} \quad J(u) = \inf_{v \in U} J(v), \tag{1}$$

where the functional $J : V \to \mathbf{R}$ is defined as

$$J(v) = \frac{1}{2} a(v, v) - f(v). \tag{2}$$

Based on these definitions, we can prove the following uniqueness theorem:

Theorem 1. Let us additionally assume that for the abstract variational problem the following properties are satisfied:
(i) the space V is complete (i.e. it is the Banach space),
(ii) the subset U is the closed and convex subset of the space V,
(iii) bilinear form $a(.,.)$ is symmetric and V-elliptic, i.e.

$$\exists \alpha > 0, \quad \forall v \in V \quad a(v,v) \geq \alpha ||v||^2. \tag{3}$$

Then problem (1) has the unique solution.

Proof.

The form $a(.,.)$ gives us a scalar product on the space V. The norm generated by this scalar product is equivalent to the standard norm $||.||$ due to the V-ellipticity and continuity of the bilinear form. The space V equipped with this scalar product is a Hilbert space. According to the Riesz theorem about the linear functional representation [1], there exists an element σf from V such that

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

As the bilinear form is symmetric, we find that

$$J(v) = \frac{1}{2}a(v,v) - a(\sigma f, v) = \frac{1}{2}a(v - \sigma f, v - \sigma f) - \frac{1}{2}a(\sigma f, \sigma f).$$

Therefore, the solution of the AMP is equivalent to the minimization of the distance between the subset U and an element outside U in the $\sqrt{a(.,.)}$ norm. Hence the solution is the projection σf on the subset U in the scalar product $a(.,.)$. According to the projection theorem [1], such element exists and is unique as the subset U is closed and convex. This completes the proof. ■

Equivalent variational problems

Theorem 2. The vector u is the solution of the AMP (1) if and only if

$$u \in U \ \textit{and} \ \forall v \in U \quad a(u, v-u) \geq f(v-u) \tag{4}$$

in the general case;

$$u \in U \ \textit{and} \ \forall v \in U \quad a(u,v) \geq f(v); \quad a(u,u) = f(u) \tag{5}$$

if the subset U is the convex cone with the vertex at 0;

$$u \in U \textit{ and } \forall v \in U \quad a(u,v) = f(v) \tag{6}$$

if the subset U is the convex subspace.

Equation (6) is usually called *the variational equation*, while the equations (4),(5) are called *the variational inequalities*.

Proof.

Let us prove the equations one by one starting from Eq.(4). The solution u is the projection σf on the subset U. As the subset U is convex, the projection is completely characterized with the property

$$u \in U \text{ and } \forall v \in U a(\sigma f - u, v - u) \leq 0.$$

(in other words, the angle between the vectors is obtuse). Let us rewrite the last expression as

$$a(u, v - u) \geq a(\sigma f, v - u) = f(v - u),$$

so we get Eq.(4).

Now, let the subset U be the convex cone with the vertex at 0. Then the vector $u + v$ belongs to the subset U as soon as v belongs U. Substituting v with $u + v$ in Eq.(4), we get

$$\forall v \in U \quad a(u,v) \geq f(v),$$

and also $a(u,u) \geq f(u)$. Considering Eq.(4) for $v = 0$, we find that

$$a(u,u) \leq f(u).$$

From these two inequalities it follows that $a(u,u) = f(u)$, that finally proves Eq.(5).

When the subset U is the subspace, we write down Eq.(5) for both vectors v and $-v$, and find

$$\left.\begin{array}{l} a(u,v) \geq f(v) \\ a(u,v) \leq f(v) \end{array}\right\}$$

that leads to the equality $a(u,v) = f(v)$ for an arbitrary vector $v \in U$. ■

Remark 1. The projection operator is linear if and only if the subset U is the subspace. Therefore, the problems corresponding to the variational inequalities are, generally speaking, the nonlinear problems.

Remark 2. We have the linear problem for the subspace due to the fact that functional (2) is quadratic. If this is not true, the problem may not be linear.

3 The abstract variational problem

Let us now introduce the *Abstract Variational Problem* (AVP) without any references to the functional J. Namely:

Find the vector u such that

$$u \in U \text{ and } \forall v \in U \quad a(u, v-u) \geq f(v-u), \tag{7}$$

in general case; or find the vector u such that

$$u \in U \text{ and } \begin{cases} \forall v \in U \quad a(u,v) \geq f(v), \\ \qquad a(u,u) = f(u), \end{cases} \tag{8}$$

if the subset U is the convex cone with the vertex at 0; or find the vector u such that

$$u \in U \text{ and } \forall v \in U \quad a(u,v) = f(v), \tag{9}$$

if the subset U is the subspace.

These problems have the unique solutions provided that the conditions of Theorem 1 are satisfied.

One can ask if these conditions can be weakened. In the general situation, they can not. However, we can additionally assume that the space V is the Hilbert space. As we shall see below, this assumption is very often met in the applications. Then we may not require the symmetry of the bilinear form, but the AVP still has the unique solution. It is very important to stress that the AMP DOES NOT have the unique solution under the same conditions!

Let us now consider the corresponding theory. For the sake of simplicity we assume that $U = V$. The main result for the nonsymmetric forms is given by the *Lax-Milgram lemma*:

Theorem 3. Let V be the Hilbert space, the bilinear form

$$a(.,.) : V \times V \to \mathbf{R}$$

be continuous and V-elliptic, and the linear form $f : V \to \mathbf{R}$ be continuous. Then the AVP: find the vector u such that

$$u \in V \text{ and } \forall v \in V \quad a(u,v) = f(v), \tag{10}$$

has the unique solution.

Proof.

As the bilinear form is continuous, there exists such constant M that

$$\forall u, \quad v \in V \quad |a(u,v)| \leq M||u||||v||.$$

Let us denote as V' the space conjugated to V, and let $||.||^*$ denote the norm in the space V'. For any vector $u \in V$, the linear form $v \in V \to a(u,v)$ is continuous, therefore there exists the unique element $Au \in V'$ such that

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

The continuity of the bilinear form can now be written as

$$||Au||^* = \sup_{v \in V} \frac{|Au(v)|}{||v||} \leq M||u||.$$

This means that the linear operator $A : V \to V'$ is continuous and its norm is estimated as

$$||A||_{\mathcal{L}(V;V')} \leq M.$$

Let $\tau : V \to V'$ be the Riesz mapping. By definition,

$$\forall f \in V' \quad \forall v \in V \quad f(v) = ((\tau f, v)),$$

where $((.,.))$ is the scalar product in the V space. The solution of variational problem (10) is equivalent to the solution of the equation $\tau Au = \tau f$. This equation has the unique solution if for some values of the parameter $\rho > 0$, the mapping

$$v \in V \rightarrow v - \rho(\tau Av - \tau f) \in V \tag{11}$$

is squeezing. We can write down the estimation

$$||v - \rho\tau Av||^2 = ||v||^2 - 2\rho((\tau Av, v)) + \rho^2||\tau Av||^2 \leq$$

$$\leq (1 - 2\rho\alpha + \rho^2 M^2)||v||^2,$$

where we used the following inequalities:

$$((\tau Av, v)) = Av(v) = a(v, v) \geq \alpha||v||^2,$$

$$||\tau Av|| = ||Av||^* \leq ||A||||v|| \leq M||v||.$$

These inequalities immediately follow from the V-ellipticity of the bilinear form $a(.,.)$ and the continuity of A. Therefore, mapping (11) is squeezing for $\rho \in (0, 2\alpha/M^2)$. That completes the proof. ■

We would like to stress that the AVP is the main problem we study through these lectures. It gives the representation of PDEs which is very convenient for many numerical methods, and for the FEM in particular.

4 The Green's formulas

In order to apply the abstract theory developed in the previous sections to specific PDEs, we need to define appropriate spaces and to find tools to work with them. In this section we sketch this information. More detailed discussion can be found, for example, in the books [1, 2].

Let Ω be the bounded closed domain in $\mathbb{R}^n$ with the smooth boundary Γ. We will use the following functional spaces over the domain Ω:

- the space of the continuous functions $C(\Omega)$ with the norm

$$||v||_{C(\Omega)} = \max_{x \in \Omega} |v(x)|$$

- the space of the square integrable functions $L^2(\Omega)$ with the norm

$$||v||_{0,\Omega} = \left(\int_\Omega |v(x)|^2 \, dx \right)^{1/2}$$

- the Hardy space $H^1(\Omega)$ of the continuous differentiable functions with the norm

$$||v||_{1,\Omega} = \left(\sum_{|\alpha| \le 1} \int_\Omega |\partial^\alpha v(x)|^2 \, dx \right)^{1/2}$$

- the subspace $H_0^1(\Omega)$ of the Hardy space $H^1(\Omega)$. This subspace consists of functions which are equal to zero at the boundary:

$$H_0^1(\Omega) = \{ v \in H^1(\Omega) : \quad v(x) = 0 \text{ for } x \in \Gamma \}.$$

The important tool to work with the PDEs is the Green's formulas. They give the natural multidimensional generalization for the integration by parts for the one-dimensional integral. The fundamental Green's formula reads

$$\int_\Omega u \partial_i v dx = - \int_\Omega \partial_i u v dx + \int_\Gamma u v \nu_i d\gamma \tag{12}$$

where $i \in [1, n]$, and the functions $u, v \in H^1(\Omega)$. In the latter equation, ν is the unit normal vector, and $d\gamma$ is the measure defined on the boundary.

Substituting the function u with $\partial_i u$ in Eq.(12) and summing up over $i = 1 \ldots n$, we find

$$\int_\Omega \sum_{i=1}^n \partial_i u \partial_i v dx = - \int_\Omega \Delta u v dx + \int_\Gamma \partial_\nu u v d\gamma \tag{13}$$

for any $u \in H^2(\Omega)$, $v \in H^1(\Omega)$. We denote here

$$\partial_\nu = \sum_{i=1}^n \nu_i \partial_i.$$

In the same way, substituting v with $\partial_i v$, we get

$$\int_\Omega u \Delta v dx = - \int_\Omega \sum_{i=1}^n \partial_i u \partial_i v dx + \int_\Gamma u \partial_\nu v d\gamma \tag{14}$$

for any functions $u \in H^1(\Omega)$, $v \in H^2(\Omega)$. Substraction Eq.(14) from Eq.(13), we find:

$$\int_\Omega (u\Delta v - \Delta uv)dx = \int_\Gamma (u\partial_\nu v - \partial_\nu uv)d\gamma. \tag{15}$$

Lecture 2

Examples of the second order boundary problem. The Galerkin and Ritz methods. The orthogonality of errors and Cea's lemma. Beyond the Galerkin method.

5 Examples of the second order boundary problem. Boundary conditions.

In order to show how the general abstract methods developed in Lecture 1 work, we analyze here two examples of the boundary problem for the second order PDE.

Example 1. The Dirichlet problem.

For this problem, we choose the spaces U and V (see Lecture 1) to be identical and equal to $H_0^1(\Omega)$:

$$V = U = H_0^1(\Omega).$$

The bilinear functional is chosen to be

$$A(u,v) = \int_\Omega \left(\sum_{i=1}^n \partial_i u(x)\, \partial_i v(x) + a(x)u(x)v(x) \right) dx,$$

and the linear functional is defined as

$$F(v) = \int_\Omega f(x)v(x)\, dx.$$

We also assume that the functions $a(x)$ and $f(x)$ satisfy the following conditions:

$$a \in C(\Omega), \quad a(x) \text{ bounded and } a(x) \geq 0 \text{ in } \Omega,$$
$$f \in L^2(\Omega).$$

The analysis of the problem.

Twice using the Cauchey-Buniakovsky inequality $|(u,v)| \leq ||u||\,||v||$ for arbitrary $u, v \in H^1(\Omega)$, we find the following estimation:

$$|A(u,v)| \leq \sum_{i=1}^n ||\partial_i u||_{0,\Omega}||\partial_i v||_{0,\Omega} + ||a||_{C(\Omega)}||u||_{0,\Omega}||v|||_{0,\Omega} \leq$$

$$\leq \max\{1, |a|_{C(\Omega)}\}\, ||u||_{1,\Omega}\, ||v||_{1,\Omega}.$$

Therefore, we have proved that the bilinear functional $A(u, v)$ is continuous in the Hardy space $H^1(\Omega)$. It is worth noting that the appropriate choice of the space is vital for the applicability of the abstract theory. For example, the same functional $A(u, v)$ is NOT continuous in the space of the square integrable functions $L^2(\Omega)$, even for the smooth functions.

As the function $a(x)$ is positive, we can write for any $v \in H^1(\Omega)$

$$A(v, v) \geq \int_\Omega \sum_{i=1}^{n} (\partial_i v)^2\, dx.$$

Using the *Poincare-Friedrichs' inequality* for the bounded domain Ω [1]

$$||v||_{0,\Omega} \leq C(\Omega) \left(\int_\Omega \sum_{i=1}^{n} (\partial_i v)^2\, dx \right)^{1/2} \qquad \forall v \in H_0^1(\Omega),$$

we find

$$A(v, v) \geq \frac{1}{2} \int_\Omega \sum_{i=1}^{n} (\partial_i v)^2\, dx + \frac{1}{2C^2(\Omega)} ||v||_{0,\Omega}^2 \geq \min\left\{\frac{1}{2}, \frac{1}{2C^2(\Omega)}\right\} ||v||_{1,\Omega}^2.$$

Therefore, the functional $A(v, v)$ is also $H_0^1(\Omega)$-elliptic.

Let us now estimate the linear functional $F(v)$. For any function $v \in H^1(\Omega)$ we can write:

$$|F(v)| \leq ||f||_{0,\Omega} ||v||_{0,\Omega} \leq ||f||_{0,\Omega} ||v||_{1,\Omega},$$

so the linear functional is also continuous.

At this point, we have proved all the assumptions of the Theorems 1.1 and 1.2. Therefore, there exists the unique function $u \in H_0^1(\Omega)$ which gives the minimum to the functional

$$J(u) = \frac{1}{2} \int_\Omega \left(\sum_{i=1}^{n} (\partial_i u(x))^2 + a(x) u^2(x) \right) dx - \int_\Omega f(x) u(x)\, dx \qquad (16)$$

on the space $H_0^1(\Omega)$. Due to Theorem 1.2, the same function $u(x)$ also satisfies the variational equation

$$\forall v \in H_0^1(\Omega) \quad \int_\Omega \left(\sum_{i=1}^n \partial_i u(x) \partial_i v(x) + a(x)u(x)v(x) \right) dx = \int_\Omega f(x)v(x)\,dx. \tag{17}$$

Connection between the abstract problems and the PDEs. The function $u(x)$ solves minimization problem (16) and variational problem (17). These problems are stated in terms of functions defined on the finite-dimensional domain. So it is natural to look for the PDEs corresponding to these problems. In order to do this, let us introduce the space of smooth finite functions $\mathcal{D}(\Omega)$. We can apply the Green's formula to the bilinear functional and get

$$\int_\Omega \sum_{i=1}^n \partial_i u(x) \partial_i v(x)\,dx = -\int_\Omega \Delta u(x)v(x)dx + \int_\Gamma \partial_\nu uvd\gamma \tag{18}$$

for any functions such that $u \in H^2(\Omega)$, $v \in H^1(\Omega)$. Then for an arbitrary smooth function $\phi \in \mathcal{D}(\Omega)$ we can write:

$$A(u,\phi) = \int_\Omega (-\Delta u(x) + a(x)u(x))\phi\,dx \equiv < -\Delta u + au, \phi >,$$

and

$$f(\phi) =< f, \phi > .$$

Hence we see from variational equation (17) that the function u is the solution in the space $\mathcal{D}'(\Omega)$ of the partial differential equation

$$\begin{aligned} -\Delta u + au &= f \text{ in } \Omega \\ u &= 0 \text{ on } \Gamma. \end{aligned} \tag{19}$$

Therefore, the initially stated problem is equivalent to the homogenious Dirichlet problem. Boundary conditions (19) are called *the essential boundary conditions.*

We would like also to note that the (more general) inhomogenious Dirichlet problem

$$\begin{aligned} -\Delta u + au &= f \text{ in } \Omega \\ u &= g(x) \text{ on } \Gamma \end{aligned}$$

can be reduced to problem (19) with the substitution $u \mapsto u - \tilde{u}$, where the function $\tilde{u}$ is chosen in such a way that

$$\tilde{u}|_\Gamma = g(x).$$

Example 2. The Neumann problem.
For this example, we again chose the identical spaces U and V

$$V = U = H^1(\Omega).$$

The bilinear functional coincides with the one in Example 1:

$$A(u,v) = \int_\Omega \left(\sum_{i=1}^n \partial_i u(x) \partial_i v(x) + a(x)u(x)v(x) \right) dx,$$

but the linear functional contains the integral over the boundary:

$$F(v) = \int_\Omega f(x)v(x)\,dx + \int_\Gamma gv\,d\gamma.$$

The functions a, f and g satisfy the following conditions:

$$a \in C(\Omega), \text{ is bounded and } a \geq a_0 > 0 \text{ on } \Omega,$$

$$f \in L^2(\Omega), \quad g \in L^2(\Gamma).$$

The analysis of the problem.
Repeating the same arguments as in Example 1, we can see that the functional $A(u,v)$ is bounded. On the other hand, for any $v \in H^1(\Omega)$ we can estimate

$$A(v,v) = \int_\Omega \sum_{i=1}^n (\partial_i v(x))^2\,dx + \int_\Omega a(x)v^2(x)\,dx \geq \min\{1, a_0\}||v||^2_{1,\Omega}.$$

Therefore, the functional $A(v,v)$ is also $H^1(\Omega)$-elliptic. In the contrast to the Example 1, the ellipticity here is guaranteed by the positivity of the function a.

The linear form $v \in H^1(\Omega) \to \int_\Gamma gv\,d\gamma$ is bounded due to the existence of the trace of the function g on the boundary:

$$\left| \int_\Gamma gv\,d\gamma \right| \leq ||g||_{L^2(\Gamma)} ||v||_{L^2(\Gamma)} \leq C(\Omega) ||g||_{L^2(\Gamma)} ||v||_{1,\Omega}.$$

So we can see that all the conditions of Theorem 1.1 are satisfied. Therefore, there exists the unique function $u \in H^1(\Omega)$ which delivers the minimum of the functional

$$J(v) = \frac{1}{2} \int_\Omega \left(\sum_{i=1}^n (\partial_i v(x))^2 + a(x) v^2(x) \right) dx - \int_\Omega f(x) v(x)\,dx - \int_\Gamma gv\,d\gamma$$

on the Hardy space $H^1(\Omega)$.

According to Theorem 1.2, the function u gives also the solution of the variational problem: for any $\forall v \in H^1(\Omega)$

$$\int_\Omega \left(\sum_{i=1}^n \partial_i u \partial_i v + auv \right) dx = \int_\Omega fv\,dx + \int_\Gamma gv\,d\gamma.$$

Connection between the abstract problems and the PDEs. For the smooth enough solutions (e.g. $u \in H^2(\Omega)$) we can use the Green's formula and get:

$$a(u,v) = \int_\Omega (-\Delta u + au) v\,dx + \int_\Gamma \partial_\nu u v\,d\gamma =$$
$$= \int_\Omega fv\,dx - \int_\Gamma gv\,d\gamma.$$

For the functions v which are equal to zero at the boundary, we get:

$$-\Delta u + au = f \text{ in } \Omega. \qquad (20)$$

Now, as any function u must satisfy Eq.(20), we find that

$$\forall v \in H^1(\Omega) \quad \int_\Gamma \partial_\nu u v\,d\gamma = \int_\Gamma gv\,d\gamma,$$

and therefore

$$\partial_\nu u = g \text{ on } \Gamma. \qquad (21)$$

Eqs.(20,21) define *the inhomogenious Neumann problem.* It is called *the homogenious problem* when $g = 0$. The boundary conditions in the homogenious Neumann problem are called *the natural boundary conditions.* The term

$$\int_\Gamma gv\, d\gamma$$

is a way in order to introduce additional boundary conditions.

6 The Galerkin and Ritz methods

Let us consider the standard variational problem, i.e. the problem to find the function $u \in V$ such that

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

We assume below that all the conditions of the Lax-Milgram lemma are satisfied. Typically, the space V is an infinite dimensional space, $\dim V = \infty$, so we can hardly find the exact solution of the variational problem. As we can only operate with the finite dimensional spaces in the computer simulations, it is very natural to look for the solution in the finite dimensional subspaces V_N of the space V, $\dim V_N = N$. It is clear, that this solution will not be the exact solution but only the approximate one.

Example 1. As the example of such finite dimensional spaces, we mention here the space $P(k)$ of all polynomials degree less or equal k over $\mathbb{R}^n$. Depending on n, the dimension of this space is equal to

$$\dim P(k) = k + 1 \text{ in } R^1,$$

$$\dim P(k) = (k + 1)(k + 2)/2 \text{ in } R^2,$$

$$\dim P(k) = (k + 1)(k + 2)(k + 3)/6 \text{ in } R^3. \quad \blacksquare$$

When we use any finite dimensional approximation, there appear few issues which we should address in order to get reliable numerical results. Namely,

- Existence and uniqueness of the finite dimensional solution.
- Convergence: is it true that

  $$||u - u_N|| \to 0 \text{ when } N \to \infty \text{ ?}$$

- Convergence speed: is it true that

  $$||u - u_N|| \sim C/N^p \text{ when } N \to \infty,$$

  and what is the value p?

In order to accurately resolve these issues, we first need to appropriately approximate the space V. Let $\{V_n\}_{n=1}^{\infty} \subset V$ be a sequence of subspaces V such that $\overline{\bigcup_{i=1}^{\infty} V_n} = V$, where $V_n \subset V_{n+1} \subset V$, and $\dim V_n = N_n < \infty$. Now we can define *the discrete problem*: to find *a discrete solution* $u_n \in V_n$ such that

$$A(u_n, v) = F(v) \quad \text{for all} \quad v \in V_n. \tag{22}$$

It is important to note that if the variational problem satisfies the assumptions of the Lax-Milgram lemma in the space V, it also does in its finite-dimensional subspace. This immediately means that the discrete problem also has the unique solution. This approximation approach, and specifically problem (22), is called *the Galerkin method.*

If the form $a(u, v)$ is symmetric, we can consider the discrete minimization problem instead of the variational problem:

$$J(u_N) = \inf_{v_N \in V_N} \left(\frac{1}{2} a(v_N, v_N) - f(v_N) \right). \tag{23}$$

This approach and problem (23) is called *the Ritz method.* Comparing the Theorems 1.1 and 1.2, we can see that the Galerkin method leads to the same equation as the Ritz method if the bilinear form is symmetric and the same basis is chosen for u_N and v_N in Eq.(22).

In our lectures, we mainly use the Galerkin method. So let us now consider in more detail how this method works. As the subspace U_N is finite dimensional, there exists a basis $\phi_k|_{k=1}^N$ of the length N there. Any function u_N can be expanded in terms of this basis as $u_N(x) = \sum_{k=1}^N \alpha_k \phi_k$ with some coefficients α_k. The variational equation has to be satisfied for every function $v_N \in V_N$, for example, for the functions belonging to an arbitrary basis $\psi_k|_{k=1}^N$. Writing down these N equations, we get

$$\sum_{k=1}^N A(\phi_k, \psi_i)\alpha_k = f(\psi_i), \quad i = 1 \dots N. \tag{24}$$

It is convenient to write these equations as the matrix equation:

$$\hat{A}\hat{\alpha} = \hat{f}, \quad \text{where} \quad \hat{A}_{ik} = A(\phi_k, \psi_i), \quad \hat{\alpha}_i = \alpha_i, \quad \hat{f}_i = f(\psi_i). \tag{25}$$

The matrix $\hat{A}$ is called *the stiffness matrix*, and the vector $\hat{f}$ is called *the load vector*. Hence, the solution of the variational equation is reduced to the solution of the linear system of the algebraic equations (25). We remind that due to the Lax-Milgram lemma, the solution of this system always exists and is unique.

It is very important to note that the existence and uniqueness of the solution of the discrete problem do not guarantee the convergence of this solution to the exact one when $N \to \infty$. Let us consider the following

Example 2. We are looking for the solution of the ordinary differential equation on the interval $[0, 1]$:

$$-\frac{d^2}{d^2x}u(x) = 2 \tag{26}$$

among the functions in $H_0^1([0, 1])$. As the problem is symmetric, we can use the Ritz method. As the sequence of the finite dimensional spaces, we choose the linear hulls of the functions $\phi_k(x) = \sin(2\pi kx)$ which belong to $H_0^1([0, 1])$. Any function from $v_N \in V_N$ can be represented as a linear combination $v_N(x) = \sum_{k=1}^N \alpha_k \phi_k(x)$. In order to find the expansion coefficients α_k, we should minimize the following functional:

$$J(v_N) = \frac{1}{2}\int_0^1 (v_N'(x))^2 dx - \int_0^1 2\sum_{k=1}^N \alpha_k \phi_k dx. \tag{27}$$

The second integral in this expression is equal to zero, and the first one is nonnegative. Therefore, its minimum is equal to zero if zero can be achieved. In our case, it is possible as the integral is zero if all $\alpha_k = 0$. Hence the solution $v_N(x) \equiv 0$ for arbitrary N. On the other hand, the exact solution of the problem (26) is equal to $v(x) = x(1-x)$. This means that the approximation error does not depend on N and does not approach zero when N goes to infinity. ■

6.1 The least square method

The Galerkin method is a general and powerful approach to the variational problems. In order to illustrate this fact, we show here how the well-known *Least Square method* can be derived from the Galerkin method. As usually, let us consider the variational problem

$$(Au_N, v_N) = (f, v_N). \tag{28}$$

Let us now replace the functions v_N with the functions Av_N. This always can be done when the operator A has the inverse. Then problem (28) can be rewritten as

$$\begin{aligned}(Au_N, Av_N) &= (f, Av_N),\\ (A^*Au_N, v_N) &= (A^*f, v_N). \qquad (29)\end{aligned}$$

The operator A^*A is symmetric, so we can apply the Ritz method to the problem (29). Its solution is found as the minimum of the functional

$$\frac{1}{2}(A^*Av_N, v_N) - (A^*f, v_N) = \frac{1}{2}(Av_N, Av_N) - (f, Av_N) =$$

$$= \frac{1}{2}(Av_N - f, Av_N - f) - \frac{1}{2}(f, f) = \frac{1}{2}||Av_N - f||^2 - ||f||^2.$$

So we can see that the function which minimizes this functional coincides with that minimizing the LS functional, i.e. $||Av_N - f||^2$. Therefore, with the special choice of the projection functions, the Galerkin method gives the same results as the LS method does.

7 The orthogonality of errors and Cea's lemma

For the elliptic problems considered in the previous lectures, the error $e_n = u - u_n$ of the solution of the discrete problem exhibits the orthogonality property.

Lemma. Let $u \in V$ be the exact solution and u_n be the solution of the discrete problem. Then

$$A(u - u_n, v) = A(e_n, v) = 0 \quad \textit{for all} \quad v \in V_n. \tag{30}$$

Proof of this lemma is very simple. As u_n is the solution of the discrete problem, it satisfies

$$A(u_n, v) = f(v) \quad \text{for all} \quad v \in V_n.$$

For the exact problem, the variational equation has to be satisfied for all $v \in V$, and particularly for v from the subspace V_n:

$$A(u, v) = f(v) \quad \text{for all} \quad v \in V_n \subset V.$$

Subtracting these two equations and using the linearity of the form, we get the proof. ■

It is interesting to discuss the geometrical meaning of this orthogonality. If the form $A(.,.)$ is symmetric, we can introduce an energetic inner product

$$(u, v)_e = A(u, v),$$

and $(e_n, v)_e = 0$ for all $v \in V_n$. So the error e_n is orthogonal to the Galerkin subspace V_n in the energetic inner product. Hence u_n is the orthogonal projection of $u \in V$ onto V_n and thus is the nearest element

$$||u - u_n||_e = \inf_{v \in V_n} ||u - v||_e.$$

Theorem (Cea's lemma). Let V be a Hilbert space, $a(.,.) : V \times V \to \mathbb{R}$ bilinear bounded V-elliptic form. $u \in V$ is the solution of the exact problem, *and subspace $V_n \subset V$, $u_n \in V_n$. Let C_b and C_{el} be the continuity and V-ellipticity constants of a, respectively. Then*

$$||u - u_n||_V \leq \frac{C_b}{C_{el}} \inf_{v \in V_n} ||u - v||_V.$$

The Cea's lemma is important as it shows the independence of the errors on the specific choice of the basis in the subspace V_n. While its proof is not complicated, we won't give it here.

The lemma can also be used to prove the convergence of the discrete solution to the exact one.

Theorem. Let the conditions of the Cea's lemma be satisfied. Additionally, let the sequence of the subspaces V_n be such that $V_1 \subset V_2 \ldots \subset V_n \ldots \subset V$ and

$$\overline{\bigcup_{i=1}^{\infty} V_n} = V. \qquad (31)$$

Then

$$\lim_{n \to \infty} ||u - u_n||_V = 0.$$

Proof. Let $u \in V$ be the exact solution. According to Eq.(31), it is possible to find a sequence $\{v_n\}$, $v_n \in V_n$ such that $\lim_{n \to \infty} ||u - v_n||_V = 0$. Let $u_n \in V_n$ be the discrete solutions. Then we can estimate the corresponding error with the Cea's lemma as

$$||u - u_n||_V \leq \frac{C_b}{C_{el}} \inf_{v \in V_n} ||u - v||_V \leq \frac{C_b}{C_{el}} ||u - v_n||_V$$

for all n. As the r.h.s. goes to zero when n goes to infinity, the same does the l.h.s. that completes the proof. ■

8 Beyond the Galerkin method

While the Galerkin method is the general and powerful numerical approach, sometimes it becomes necessary to extend it beyond its applicability limits. The main idea is to choose different spaces U and V, and then extend the space V of the test functions. In principle, the space V can be as big as the space conjugated to U. As the result, we have the general linear functionals instead of the integrals over the domain. This approach constitutes the essence of the weighted residual method.

8.1 The weighted residual method

Here we discuss the method in details. Let us consider the time-depended PDE. It can be written as

$$\begin{aligned} L(u) &= 0, \quad \text{PDE symbol}, \\ I(u) &= 0, \quad \text{initial conditions}, \\ B(u) &= 0, \quad \text{boundary conditions}. \end{aligned} \tag{32}$$

We are looking for the approximate solution u_a. Then Eqs.(32) are not satisfied exactly, we generally have the *residuals* R:

$$L(u_a) = R, \quad I(u_a) = R_I, \quad B(u_a) = R_B.$$

We can however choose the approximate solution in such a way that some of these residuals are zero. The different choices have their own names:

$$\begin{array}{ccl} (i) & R = 0, & \text{boundary methods}, \\ (ii) & R_B = 0, & \text{internal methods}, \\ (iii) & \text{all residuals nonzero}, & \text{mixed methods}. \end{array}$$

Let us concentrate on the internal methods. We are looking for the solution in the form

$$u_a(\vec{x}, t) = u_0(\vec{x}, t) + \sum_{j=1}^{N} a_j(t) \varphi_j(\vec{x}). \tag{33}$$

We suppose that $u_0(\vec{x}, t)$ and $\varphi_j(\vec{x})$ are chosen such that $R_I = R_B = 0$. The functions $\varphi_j(\vec{x})$ are known, we need to calculate the expansion coefficients $a_j(t)$. We find the ODE for them if we require

$$W_k(R) = 0, \quad k = 1 \ldots N,$$

where W_k are linear functionals. Different sets of these functionals lead to different computational methods. When these functionals can be represented as the integrals with some functions,

$$(R, w_k(\vec{x})) = 0, \quad k = 1 \ldots N, \tag{34}$$

the functions $w_k(\vec{x})$ are called the weighted functions. The convergence of the approximate solution to the exact one is the convergence in average.

8.2 The discrete method of weighted residuals

The scalar product in equation (34) was represented by an integral, so it was continuous. In the practical calculations, very often we cannot calculate integrals exactly and substitute them with the quadrature sums. Hence it is natural to abandon the integrals from the very beginning, and use the finite sum instead of the integrals in the definition of the scalar product:

$$(f, g) = \sum_{i=1}^{N} f_i g_i.$$

This approach is called the discrete method of weighted residuals.

8.3 Particular weighted residuals type methods

Choosing different functions $w_k(\vec{x})$, we arrive to different weighted residuals type methods. Some of them were first invented without any connection to residuals, and have their own names. Let us consider few methods of this type.

1. **The subdomain method**

In this approach, we divide the equation domain into a number of subdomains D_j. Then we define the functions $w_k(\vec{x})$ as

$$w_k(\vec{x}) = \begin{cases} 1, & \vec{x} \text{ inside } D_k, \\ 0, & \vec{x} \text{ outside } D_k. \end{cases}$$

This method is, in fact, the finite volume method.

2. **Collocation method**

In this method, the functions $w_k(\vec{x})$ are defined as

$$w_k(\vec{x}) = \delta(\vec{x} - \vec{x}_k)$$

for a set of points $\vec{x}_k$. This means that at these points $R(\vec{x}_k) = 0$. Most of the finite difference methods can be described in this way. In the other way, we can choose zeros of the Chebyshev polynomials as the set of points. According to the properties of the Chebyshev polynomials, we will then minimize the maximal error.

3. **The least square method**

If we choose

$$w_k = \frac{\partial R}{\partial a_k},$$

where a_k are the coefficients from equation (33), equation (34) will give a minimum of the (R, R) functional. This is the well known result for the stationary equation while its application for non-stationary equation is not straightforward.

4. **Method of moments**

This method can be obtained when we choose

$$w_k(x) = x^k.$$

5. **The Galerkin method**

The Galerkin method can be obtained if we choose the weighted functions coinciding with the basis functions, $w_k(\vec{x}) = \varphi_k(\vec{x})$. On the other hand, one should be careful: in literature there appears a tendency to call the weighted residual method with any smooth functions as the Galerkin method.

Lecture 3

The main features of the finite element method (FEM). One-dimensional FEM. Lagrange elements, Hermite elements. Hierarchical elements.

Starting from this lecture, we shall discuss the computational Galerkin approaches. We will mainly focus on the finite element method and on the boundary element method which is closely related to FEM. In the last lecture, we will discuss the spectral methods. In some sense, these two types of methods represent opposite cases among all varieties of methods. In the FEM, the basis is chosen to be local, i.e. the basis functions are non-zero in a very small part of the whole domain. For the spectral methods, the basis functions are global (i.e. non-zero everywhere in the domain) and very frequently orthogonal to each other.

9 The main features of the FEM

In the Galerkin method for the PDE, we look for the solution of the variational problem: to find the function $u \in V$ such that for all $v \in V$

$$\int_\Omega \left(\sum_{i=1}^n \partial_i u(x) \partial_i v(x) + a(x)u(x)v(x) \right) dx = \int_\Omega f(x)v(x)\,dx.$$

We will construct the discrete problem by choosing the appropriate subspaces V_N of the space V. Different choices give us different methods. One of them is the FEM which is usually characterized by the following features:

FEM 1. The triangulation $\mathcal{F}_h$ of the domain Ω. This means that Ω is divided into a finite number of subdomains (elements) K_i such that

1. $\Omega = \bigcup\limits_{K \in \mathcal{F}_h} K$.
2. The boundaries ∂K_i are piecewise smooth.
3. $(K_i \backslash \partial K_i) \bigcap (K_j \backslash \partial K_j) = \emptyset$.

This triangulation is often called the mesh.

Let us choose the finite dimensional functional space V_h which is called the space of the *finite elements*. Then we define the projection spaces

$$P_{K_i} = \{v_{h|K} : v_h \in V_h\}.$$

The next feature describe these spaces, namely

FEM 2. The space P_{K_i} consists of polynomials (or, more generally, of functions close to polynomials in some sense).

This choice assures the convergence properties and the simple calculation of the algebraic system coefficients. In fact, the discrete problem is solved with the expansion onto the basis $\{w_k(x)\}_{k=1}^N$ in V_h. The solution is written as

$$u_h(x) = \sum_{k=1}^{N} u_k w_k(x)$$

and is found from the linear system

$$\sum_{k=1}^{N} u_k \int\limits_{\Omega} \left(\sum_{i=1}^{n} \partial_i w_k(x) \partial_i w_l(x) + a(x) w_k(x) w_l(x) \right) dx$$

$$= \int_{\Omega} f(x) w_l(x)\, dx, \quad 1 \leq l \leq N.$$

Therefore, the choice of simple $w_k(x)$ essentially facilitates the integral calculations in the latter equation.

FEM 3. The existence in the space V_h of (at least) one basis consisting of functions with “minimal” support.

Such basis always exists but we require here that this basis can be simply and explicitly described.

Among different families of the FEM, there exists one important particular case of the FEM which is called the *conformal FEM*. The conformal FEM has the following properties:

- the space V_h is a subspace of V,

- the linear and bilinear forms in the variational equation coincide with the exact (non-approximated) forms.

The conformal FEM is the simplest method to study and analyze. However, a number of application may lead to the non-conformal FEMs. Among different reasons for the loss of the conformal property, we point out few possibilities:

- the boundary of Ω is curved. Then the triangulation cannot be done exactly, and the approximated domain Ω_h is used.
- the function from V_h have jumps.
- the linear and bilinear forms are calculated approximately (e.g. numerical quadratures are used).

The analysis for these FEMs becomes more complicated, so we first study the conformal FEM and then analyze how the loss of conformity affects our results.

10 The one-dimensional FEM

In order to show how the FEM works for the real problems, we start with the simplest, one-dimensional case. We shall analyze the Dirichlet problem

$$-(p(x)u'(x))' + q(x)u(x) = f(x)u(x), \quad 0 \leq x \leq 1,$$

$$u(0) = u(1) = 0,$$

where the functions $p(x)$, $q(x)$, $f(x)$ are smooth, and $p(x) > 0$, $q(x) > 0$ on the interval $[0, 1]$. The corresponding variational problem is formulated as the problem to find $u(x)$ such that

$$\int\limits_0^1 \left(u'(x)p(x)v'(x) + q(x)u(x)v(x)\right) dx = \int\limits_0^1 f(x)v(x)dx \qquad (35)$$

for all $v(x) \in H_0^1([0,1])$. The corresponding minimization problem reads: to find $u(x)$ which delivers the minimum to the functional

$$\int\limits_0^1 \big(p(x)u'^2(x) + q(x)u^2(x) - 2f(x)u(x)\big)\, dx. \tag{36}$$

The results of Lecture 1 confirm that the solutions of problems (35) and (36) exist, are unique and coincide with each other.

Let us now construct the FEM equations for this problem. The first step is the triangulation of the domain. For any one-dimensional domain, it is very easy: we divide the interval into number of elements

$$x_0 < x_1 < \ldots < x_N.$$

The second step is the choice of the basis functions. They can be constructed with the Lagrange or Newton interpolation. The linear polynomials are the same for both interpolation types:

$$\phi_j(x) = \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}}, & \text{for } x_{j-1} \le x < x_j, \\ \frac{x_{j+1} - x}{x_{j+1} - x_j}, & \text{for } x_j \le x < x_{j+1}, \\ 0, & \text{for all other } x. \end{cases}$$

In order to construct higher-order polynomials, we use here the Lagrange interpolation. For the quadratic case, we add into each interval $K_j = [x_{j-1}, x_j]$ of the length $h_j = x_j - x_{j-1}$ its middle point $x_{j-1/2}$ and construct three basis functions

$$\phi_j(x) = \begin{cases} 1 + 3(\frac{x - x_j}{h_j}) + 2(\frac{x - x_j}{h_j})^2, & \text{for } x_{j-1} \le x < x_j, \\ 1 - 3(\frac{x - x_j}{h_{j+1}}) + 2(\frac{x - x_j}{h_{j+1}})^2, & \text{for } x_j \le x < x_{j+1}, \\ 0, & \text{for all other } x, \end{cases}$$

$$\phi_{j-1/2}(x) = \begin{cases} 1 - 4\left(\frac{x - x_{j-1/2}}{h_j}\right)^2, & \text{for } x_{j-1} \le x < x_j, \\ 0, & \text{for all other } x. \end{cases}$$

We can see that this basis is convenient as the overlapping of the basis functions is rather small.

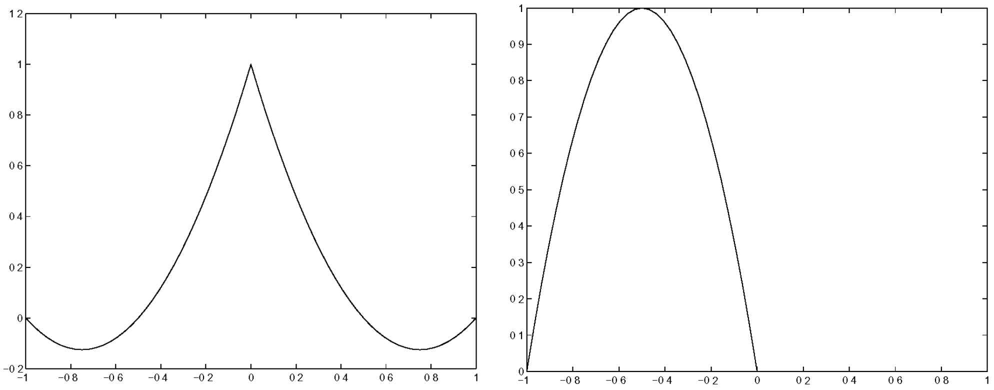


Figure 1: The Lagrange basis functions. On the left, for the boundary $x = 0$, on the right, for the middle point $x = -0.5$. The nodes are chosen to be $x_{j-1} = -1$, $x_{j-1/2} = -0.5$, $x_j = 0$, $x_{j+1} = 1$.

We can also see that the constructed basis satisfies the Lagrange property:

$$\phi_j(x_k) = \begin{cases} 1, & \text{for } j = k \\ 0, & \text{for } j \neq k \end{cases}, \quad j, k = 0, 1/2, 1, \ldots N.$$

The elemental functions

The elemental functions are the projection of the basis functions ϕ_j onto each element K_j. For the quadratic basis, there are three elemental functions:

$$N_{j-1,j}(x) = 1 - 3\left(\frac{x - x_{j-1}}{h_j}\right) + 2\left(\frac{x - x_{j-1}}{h_j}\right)^2,$$

$$N_{j-1/2,j}(x) = 1 - 4\left(\frac{x - x_{j-1/2}}{h_j}\right)^2,$$

$$N_{j,j}(x) = 1 + 3\left(\frac{x - x_j}{h_j}\right) + 2\left(\frac{x - x_j}{h_j}\right)^2.$$

In the general case, the elemental functions $N_{k,j}$ are non-zero only on their own interval, or, in other words, only if the node k belongs to the element K_j. This ensures the small overlap between different elemental functions.

In order to define the elemental functions of the order p, let us introduce the *canonical element* $-1 \leq \xi \leq 1$ which is connected to each *physical element*

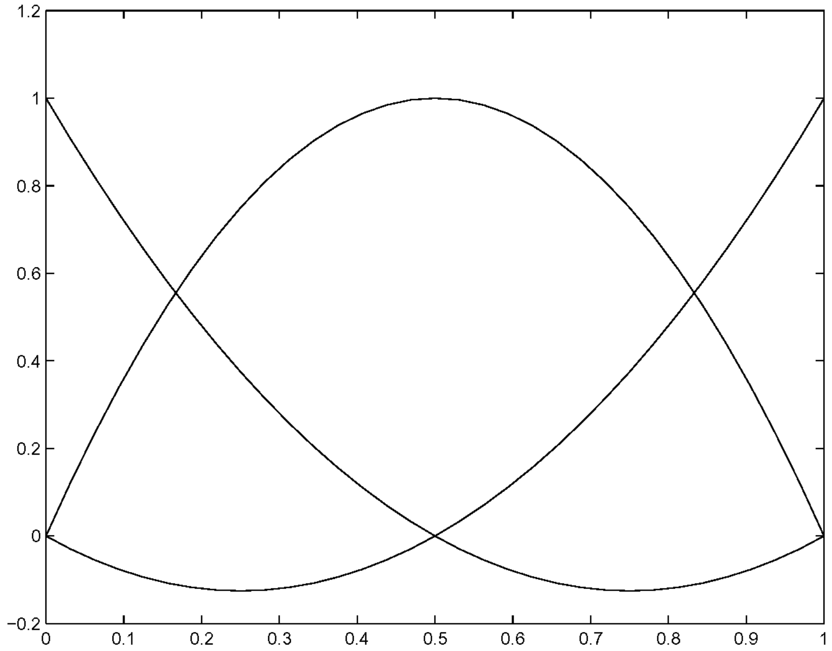


Figure 2: The Lagrange elemental functions on the interval [0,1].

through

$$x(\xi) = \frac{1-\xi}{2} x_{j-1} + \frac{1+\xi}{2} x_j \quad \in K_j.$$

Let us introduce $p+1$ nodes on the canonical element

$$-1 = \xi_0 < \xi_1 < \ldots < \xi_{p-1} < \xi_p = 1.$$

The nodes on the physical element K_j can be easily derived as

$$x_{j-1+i/p} = \frac{1-\xi_i}{2} x_{j-1} + \frac{1+\xi_i}{2} x_j \quad i = 0, 1, \ldots p.$$

Now we define the elemental functions $N_{k,c}$ on the canonical element as

$$N_{k,c}(\xi) = \prod_{i=0, i \neq k}^{p} \frac{\xi - \xi_i}{\xi_k - \xi_i}.$$

It is clear that they satisfy the following properties

$$N_{k,c}(\xi_i) = \delta_{ki}, \quad \deg N_{k,c}(\xi) = p.$$

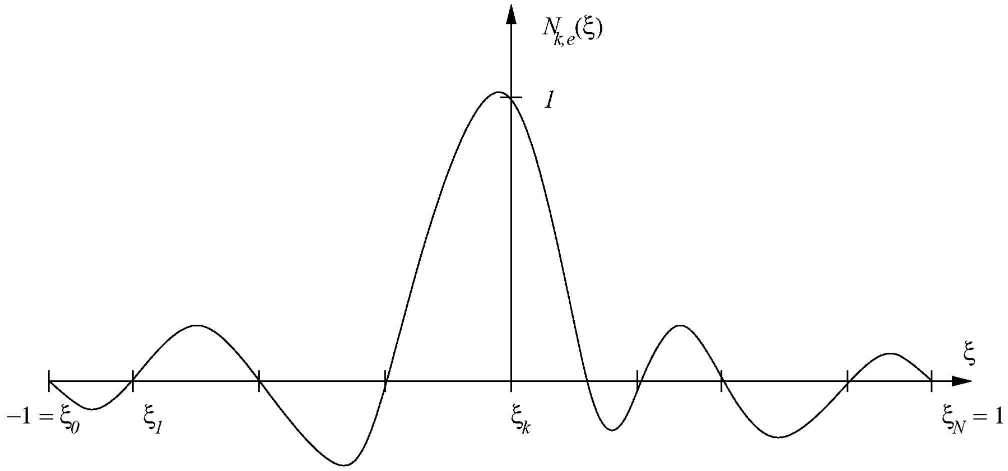


Figure 3: The Lagrange elemental function on the canonical element.

Here are few examples of the elemental functions on the canonical element

1. The linear functions, $p = 1$:

$$N_{0,c}(\xi) = \frac{1-\xi}{2}, \quad N_{1,c}(\xi) = \frac{1+\xi}{2}.$$

1. The quadratic functions, $p = 2$, $\xi_1 = 0$:

$$N_{0,c}(\xi) = \frac{\xi(\xi-1)}{2}, \quad N_{1,c}(\xi) = 1 - \xi^2, \quad N_{2,c}(\xi) = \frac{\xi(\xi+1)}{2}.$$

The hierarchical basis.

For the Lagrange basis, we need to construct all the functions for each order p from the very beginning. The idea of the hierarchical basis is different: we construct the basis of the order $p+1$ by adding a new function to the basis of the order p. This way, we keep our previous calculations meaningful and only need to add few matrix elements.

So, the quadratic hierarchical basis on K_j is constructed as

$$U^2(x) = U^1(x) + c_{j-1/2} N^2_{j-1/2,j}(x), \tag{37}$$

where $U^1(x)$ is the linear basis

$$U^1(x) = c_{j-1}N^1_{j-1,j}(x) + c_j N^1_{j,j}(x),$$

$$N^1_{j-1,j}(x) = \begin{cases} \frac{x_j - x}{h_j}, & \text{for } x \in K_j \\ 0, & \text{for all other } x \end{cases},$$

$$N^1_{j,j}(x) = \begin{cases} \frac{x - x_{j-1}}{h_j}, & \text{for } x \in K_j \\ 0, & \text{for all other } x \end{cases}.$$

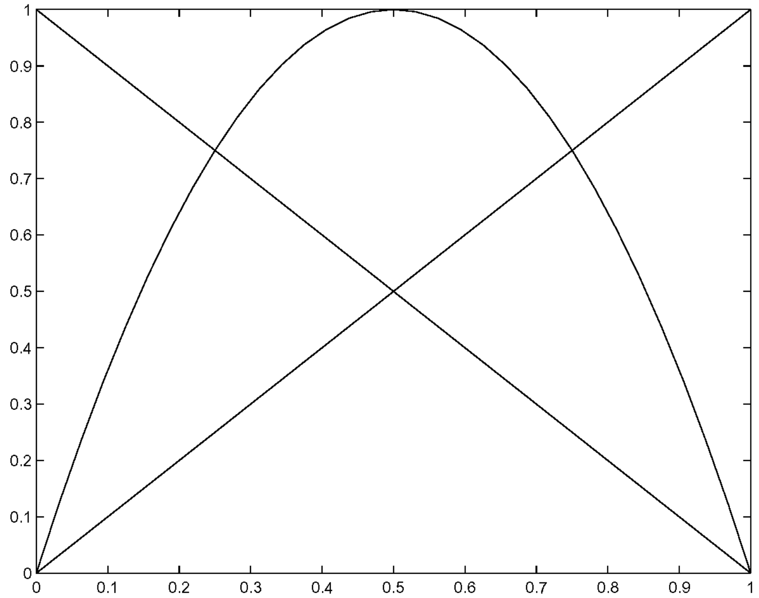


Figure 4: The elemental functions for the hierarchical basis on the interval [0,1].

We would like to stress that the quadratic function $N^2_{j-1/2,j}(x)$ here is not uniquely defined. The only properties it has to satisfy are:
1. It is the quadratic polynomial
2. It is continuous everywhere
3. It is zero outside the element K_j.

For example, we can choose it as

$$N^2_{j-1/2,j}(x) = \begin{cases} 1 - 4\left(\frac{x - x_{j-1/2}}{h_j}\right)^2, & \text{for } x \in K_j \\ 0, & \text{for all other } x \end{cases}. \tag{38}$$

It follows from representation (37) that

$$c_{j-1} = U^2(x_{j-1}), \quad c_j = U^2(x_j),$$
$$U^2(x_{j-1/2}) = \frac{c_{j-1} + c_j}{2} + c_{j-1/2}.$$

Using Eq.(38), we can find

$$c_{j-1/2} = -\frac{h_j^2}{8}(U^2)''(x_{j-1/2}).$$

Considering the difference between the Lagrange and hierarchical basises, we arrive at the following definition:

Def.1 The Lagrange basis element is the element for which all the degrees of freedom (i.e. expansion coefficients) are defined only through the *values* of the basis functions at the nodes.

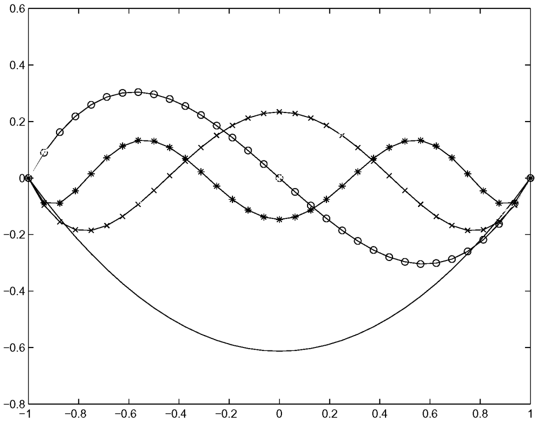


Figure 5: The hierarchical elemental functions of the order 2, 3, 4, and 6 on the canonical element.

Def.2 The Hermite basis element is the element for which at least one degree of freedom is defined through the *derivative value* of the basis functions at some nodes.

There also exist other types of elements which do not connect their degrees of freedom with any values at nodes. As a rule, those are elements of the higher order, and that is a way to exclude the need for high derivative calculations.

The hierarchical elements of the higher order p are constructed similarly to the quadratic ones

$$U(\xi) = c_{-1}N^1_{-1}(\xi) + c_1 N^1_1(\xi) + \sum_{i=2}^{p} c_i N^i(\xi).$$

The elemental functions $N^1_{-1}(\xi)$ and $N^1_1(\xi)$ are defined as

$$N^1_{-1}(\xi) = \frac{1-\xi}{2}, \quad N^1_1(\xi) = \frac{1+\xi}{2}.$$

The requirements for $N^i(\xi)$ are the correct polynomial degree and zero values at the interval ends.

Lecture 4

An example of the FEM for the one-dimensional boundary problem.

The approximation errors.

11 An example of the FEM for the one-dimensional boundary problem

Let us consider the hierarchical basis of the order p on the canonical element. The elemental function on the interval $[-1, 1]$ can be written as

$$U(\xi) = c_{-1}N^1_{-1}(\xi) + c_1 N^1_1(\xi) + \sum_{i=2}^{p} c_i N^i(\xi). \tag{39}$$

The values of the basis functions on the boundaries of the interval are equal to

$$N^1_{-1}(-1) = 1, \quad N^1_{-1}(1) = 0,$$

$$N^1_1(-1) = 0, \quad N^1_1(1) = 1,$$

$$N^i(-1) = 0, \quad N^i(1) = 0, \quad i = 2...p.$$

The basis functions can be recalculated from the canonical element to an arbitrary physical element $[x_{j-1}, x_j]$ with the linear transformation

$$x(\xi) = \frac{1-\xi}{2}x_{j-1} + \frac{1+\xi}{2}x_j. \tag{40}$$

Let us consider the quadratic hierarchical basis

$$N^1_{-1}(\xi) = \frac{1-\xi}{2}, \quad N^1_1(\xi) = \frac{1+\xi}{2}, \quad N^2(\xi) = \frac{3}{2\sqrt{6}}(\xi^2 - 1). \tag{41}$$

The quadratic function obviously meets the boundary conditions.

Example. We will analyze the following simple boundary problem:

$$-pu''(x) + qu(x) = f(x), \quad p > 0, q > 0, \tag{42}$$

$$0 \leq x \leq 1, \quad u(0) = u(1) = 0.$$

The functions $p(x)$ and $q(x)$ are chosen to be constants for sake of simplicity. We also prefer concentrate now on the ideas of the method rather than on the technical details. As we already did, let us first convert boundary problem (42) into the variational problem. Namely, we will look for the function $u(x)$ such that for all functions $v(x)$ the variational equation

$$\int_0^1 (u'(x)pv'(x) + u(x)qv(x))\,dx = \int_0^1 f(x)v(x)\,dx \tag{43}$$

is satisfied.

Now we rewrite the integral over the interval $[0, 1]$ as the sum over all intervals $[x_{j-1}, x_j]$, and substitute instead of $u(x)$ and $v(x)$ the elemental functions (39) recalculated from the canonical element with transformation (40). As the result, we get

$$\sum_{j=1}^{N} \left[A_j^S(U, V) + A_j^M(U, V) - (f, V)_j\right] = 0 \quad \forall V. \tag{44}$$

In this expression we used the following notations: the internal energy A_j^S,

$$A_j^S(U, V) = \int_{x_{j-1}}^{x_j} pU'(x)V'(x)\,dx = \frac{2p}{h_j}\int_{-1}^{1} \frac{dU(\xi)}{d\xi}\frac{dV(\xi)}{d\xi}\,d\xi;$$

the inertial (or "external") energy A_j^M,

$$A_j^M(U, V) = \int_{x_{j-1}}^{x_j} qU(x)V(x)\,dx = \frac{qh_j}{2}\int_{-1}^{1} U(\xi)V(\xi)\,d\xi;$$

and the load vector,

$$(f, V)_j = \int_{x_{j-1}}^{x_j} f(x)V(x)\,dx = \frac{h_j}{2}\int_{-1}^{1} f(x(\xi))V(\xi)\,d\xi.$$

The functions $U(\xi)$ and $V(\xi)$ can be written as the scalar products in two different ways:

$$U(\xi) = \left[c_{j-1}, c_j, c_{j-1/2}\right] \begin{bmatrix} N^1_{-1} \\ N^1_1 \\ N^2 \end{bmatrix} = \left[N^1_{-1}, N^1_1, N^2\right] \begin{bmatrix} c_{j-1} \\ c_j \\ c_{j-1/2} \end{bmatrix},$$

$$V(\xi) = \left[d_{j-1}, d_j, d_{j-1/2}\right] \begin{bmatrix} N^1_{-1} \\ N^1_1 \\ N^2 \end{bmatrix} = \left[N^1_{-1}, N^1_1, N^2\right] \begin{bmatrix} d_{j-1} \\ d_j \\ d_{j-1/2} \end{bmatrix}.$$

With these representations, we can write the internal energy in a more convenient way

$$A^S_j(U, V) = \left[c_{j-1}, c_j, c_{j-1/2}\right] K_j \begin{bmatrix} d_{j-1} \\ d_j \\ d_{j-1/2} \end{bmatrix},$$

where the matrix K_j is defined by the basis functions and can be easily calculated:

$$K_j = \frac{2p}{h_j} \int\limits_{-1}^{1} \frac{d}{d\xi} \begin{bmatrix} N^1_{-1} \\ N^1_1 \\ N^2 \end{bmatrix} \frac{d}{d\xi} \left[N^1_{-1}, N^1_1, N^2\right] d\xi =$$

$$= \frac{2p}{h_j} \int\limits_{-1}^{1} \begin{bmatrix} -1/2 \\ 1/2 \\ \xi\sqrt{3/2} \end{bmatrix} \left[-1/2, 1/2, \xi\sqrt{3/2}\right] d\xi =$$

$$= \frac{2p}{h_j} \int\limits_{-1}^{1} \begin{bmatrix} 1/4 & -1/4 & -\xi\sqrt{3/8} \\ -1/4 & 1/4 & \xi\sqrt{3/8} \\ -\xi\sqrt{3/8} & \xi\sqrt{3/8} & 3\xi^2/2 \end{bmatrix} d\xi =$$

$$= \frac{p}{h_j} \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

In the same way, we can derive the expression for the energy A^M_j:

$$A^M_j(U, V) = \left[c_{j-1}, c_j, c_{j-1/2}\right] M_j \begin{bmatrix} d_{j-1} \\ d_j \\ d_{j-1/2} \end{bmatrix},$$

where the matrix M_j is calculated as

$$M_j = \frac{qh_j}{2} \int\limits_{-1}^{1} \begin{bmatrix} N^1_{-1} \\ N^1_1 \\ N^2 \end{bmatrix} \left[N^1_{-1}, N^1_1, N^2\right] d\xi =$$

$$= \frac{qh_j}{6} \begin{bmatrix} 2 & 1 & -\sqrt{3/2} \\ 1 & 2 & -\sqrt{3/2} \\ -\sqrt{3/2} & -\sqrt{3/2} & 6/5 \end{bmatrix}.$$

The load vector can also be written as the scalar product:

$$(f, V)_j = \mathbf{l}_j \left[d_{j-1}, d_j, d_{j-1/2}\right]^\top,$$

where

$$\mathbf{l}_j = \frac{h_j}{2} \int\limits_{-1}^{1} f(x(\xi)) \left[N^1_{-1}, N^1_1, N^2\right] d\xi.$$

In order to calculate the integral in the latter expression, we use the linear approximation for the function $f(x)$:

$$f(x) \approx N^1_{-1}(\xi) f_{j-1} + N^1_1(\xi) f_j = [f_{j-1}, f_j] \begin{bmatrix} N^1_{-1} \\ N^1_1 \end{bmatrix}, \text{ where } f_j = f(x_j).$$

The required approximation order depends on the order p of the finite elements employed, and must be higher for the larger values of p. For the vector $\mathbf{l}_j$, we get

$$\mathbf{l}_j = \frac{h_j}{2} \int\limits_{-1}^{1} [f_{j-1}, f_j] \begin{bmatrix} N^1_{-1} \\ N^1_1 \end{bmatrix} \left[N^1_{-1}, N^1_1, N^2\right] d\xi =$$

$$= \frac{h_j}{6} \begin{bmatrix} 2f_{j-1} + f_j \\ f_{j-1} + 2f_j \\ -\sqrt{3/2}(f_{j-1} + f_j) \end{bmatrix}^\top.$$

Now we have all the representations for the energy matrices K_j, M_j, and the load vector $\mathbf{l}_j$ on the jth element. In order to get *global* matrices and vectors, we should take into account the boundary conditions in Eq.(42). In the case considered, we just require that the functions which are non-zero on the interval boundaries, have zero expansion coefficients:

$$c_0 = c_N = d_0 = d_N = 0.$$

As the rest of the functions are equal to zero at $x = 0$ and $x = 1$ by construction, the boundary conditions (42) are satisfied.

In order to make the formulas more transparent, we will presume below $h_j \equiv h$. It is important to notice that the structure of the global matrix essentially depends on the function numbering! If we change the order of the elemental functions, the matrix is also changed. We will use here the following numbering

$$\mathbf{c} = [c_{1/2}, c_1, c_{3/2}, ...c_{N-1}, c_{N-1/2}]. \tag{45}$$

The coefficients c_0 and c_N are not included because of the boundary conditions. The matrices and the vector calculated above are rewritten for numbering (45) and for the equal intervals as

$$K_j = \frac{p}{h}\begin{bmatrix} 1 & 0 & -1 \\ 0 & 2 & 0 \\ -1 & 0 & 1 \end{bmatrix},$$

$$M_j = \frac{qh}{6}\begin{bmatrix} 2 & -\sqrt{3/2} & 1 \\ -\sqrt{3/2} & 6/5 & -\sqrt{3/2} \\ 1 & -\sqrt{3/2} & 2 \end{bmatrix},$$

$$\mathbf{l}_j = \frac{h}{6}\begin{bmatrix} 2f_{j-1} + f_j \\ -\sqrt{3/2}(f_{j-1} + f_j) \\ f_{j-1} + 2f_j \end{bmatrix}^{\top}.$$

The global matrices can be constructed by combining the elemental matrices for the coefficient triplets $[c_j, c_{j+1/2}, c_{j+1}]$. They overlap by elements with the integer indices:

$$\mathbf{K} = \frac{p}{h}\begin{bmatrix} 2 & 0 & 0 & & & \\ 0 & 1+1 & 0 & -1 & & \\ 0 & 0 & 2 & 0 & & \\ & -1 & 0 & 1+1 & 0 & -1 \\ & & 0 & 0 & 2 & 0 \\ & & & -1 & 0 & 1+1 \end{bmatrix},$$

$$\mathbf{M} = \frac{qh}{6}\begin{bmatrix} 6/5 & -\sqrt{3/2} & 0 & & & \\ -\sqrt{\frac{3}{2}} & 2+2 & -\sqrt{3/2} & 1 & & \\ 0 & -\sqrt{3/2} & 6/5 & -\sqrt{\frac{3}{2}} & & \\ & 1 & -\sqrt{3/2} & 4 & -\sqrt{\frac{3}{2}} & 1 \\ & & & -\sqrt{\frac{3}{2}} & 6/5 & -\sqrt{\frac{3}{2}} \end{bmatrix},$$

$$\mathbf{l} = \frac{h}{6}\begin{bmatrix} -\sqrt{3/2}(f_0+f_1) \\ f_0+4f_1+f_2 \\ -\sqrt{3/2}(f_1+f_2) \\ f_1+4f_2+f_3 \end{bmatrix}^\top .$$

One of the advantages of the hierarchical basis is that the basis contains approximations with lower orders. So we can easily get the linear approximation ($p = 1$) from the matrices derived. We just need to exclude the rows and columns corresponding to the half-integer indexes, and immediately get

$$\mathbf{K} = \frac{p}{h}\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \end{bmatrix},$$

$$\mathbf{M} = \frac{qh}{6}\begin{bmatrix} 4 & 1 & & \\ 1 & 4 & 1 & \\ & 1 & 4 & 1 \end{bmatrix},$$

$$\mathbf{l} = \frac{h}{6}\begin{bmatrix} f_0+4f_1+f_2 \\ f_1+4f_2+f_3 \end{bmatrix}^\top .$$

Finally, we can write variational problem (43) as

$$\mathbf{c}(\mathbf{K}+\mathbf{M})\mathbf{d}^\top = \mathbf{l}\,\mathbf{d}^\top .$$

As this equation have to be satisfied for all $\mathbf{d}$, we should solve the linear matrix problem

$$(\mathbf{K}+\mathbf{M})\mathbf{c}^\top = \mathbf{l}^\top . \tag{46}$$

The matrix $\mathbf{K}+\mathbf{M}$ has few properties which highly facilitate the solution of the linear system. It is

- symmetric,

- positively defined,
- banded with the band width $2p+1$.

The numerical example

Let us consider the problem

$$-u''(x)+u(x)=x, \quad 0 \le x \le 1, \;\; u(0)=u(1)=0.$$

It has the exact analytical solution

$$u(x)=x-\frac{\sinh x}{\sinh 1},$$

which can be used to control the accuracy. The accuracy of the numerical solution

$$e(x)=u(x)-U(x)$$

can be characterized by its various norms. We consider here the following norms

$$|e|_\infty = \max_{0\le j\le N} |e(x_j)|, \quad |e'|_\infty = \max_{0\le j\le N} |e'(x_j)|,$$

$$||e||_0 = \left(\int_0^1 e^2(x)\,dx\right)^{1/2},$$

$$||e||_A = \sqrt{A(e,e)} = \left(\int_0^1 \left[p(e')^2+qe^2\right]dx\right)^{1/2},$$

N	$\|e\|_\infty$	$\|e\|_\infty/h^2$	$\|e'\|_\infty$	$\|e'\|_\infty/h$
4	0.269(-3)	0.430(-2)	0.111(0)	0.444
8	0.688(-4)	0.441(-2)	0.589(-1)	0.471
16	0.172(-4)	0.441(-2)	0.303(-1)	0.485
32	0.432(-5)	0.442(-2)	0.154(-1)	0.492
64	0.108(-5)	0.442(-2)	0.775(-2)	0.496
128	0.270(-6)	0.442(-2)	0.389(-2)	0.498

Figure 6: The linear approximation, the maximal errors.

N	$\|\|e\|\|_0$	$\|\|e\|\|_0/h^2$	$\|\|e\|\|_A$	$\|\|e\|\|_A/h$
4	0.265(-2)	0.425(-1)	0.390(-1)	0.156
8	0.656(-3)	0.426(-1)	0.195(-1)	0.157
16	0.167(-3)	0.427(-1)	0.979(-2)	0.157
32	0.417(-4)	0.427(-1)	0.490(-2)	0.157
64	0.104(-4)	0.427(-1)	0.245(-2)	0.157
128	0.260(-5)	0.427(-1)	0.122(-2)	0.157

Figure 7: The linear approximation, the average errors.

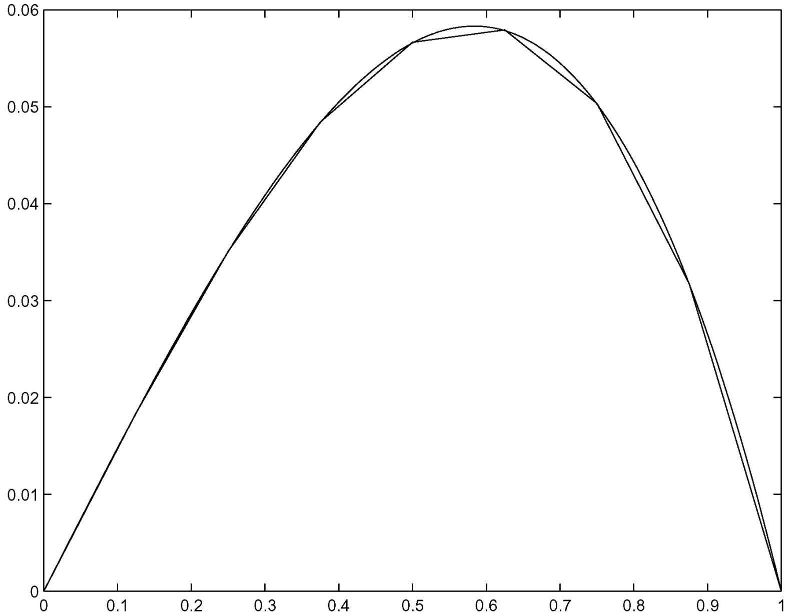


Figure 8: The exact solution and the linear approximation for 8 elements.

N	Linear			Quadratic		
	DOF	$\|\|e\|\|_0$	$\|\|e\|\|_0/h^2$	DOF	$\|\|e\|\|_0$	$\|\|e\|\|_0/h^3$
4	3	0.265(-2)	0.425(-1)	7	0.126(-3)	0.807(-2)
8	7	0.656(-3)	0.426(-1)	15	0.158(-4)	0.809(-2)
16	15	0.167(-3)	0.427(-1)	31	0.198(-5)	0.809(-2)
32	31	0.417(-4)	0.427(-1)			

Figure 9: The linear and quadratic approximations, the average errors.

12 The approximation errors

As we could see in the numerical example above, the error of the numerical solution behaves in the regular way. For the linear approximation, the error for the solution is proportional to h^2, and the derivative error is proportional to h. For the quadratic approximation, the convergence is one order of h better. In this section we will analyze the approximation errors in different norms and find general estimation for the convergence speed.

We start with the estimations on the canonical element

$$U(\xi) = \sum_{k=0}^{p} c_k N_k(\xi).$$

As the solution is unique, we can expand it with any basis. We will use the Lagrange basis.

I. The point estimate.

Theorem 1. Let $u(\xi) \in C^{p+1}[-1,1]$, then for any $\xi \in [-1,1]$ there exists $\zeta(\xi) \in (-1,1)$ such that the error

$$e(\xi) = \frac{u^{(p+1)}(\zeta)}{(p+1)!} \prod_{i=0}^{p} (\xi - \xi_i).$$

The proof of this theorem can be found for example in book [6]. Additionally we will use the fact that $\zeta(\xi)$ is a smooth function of ξ.

12.1 The linear approximation

We start with the simplest case of the linear approximation. Here $\xi_0 = -1$, $\xi_1 = 1$, and the error can be written according to Theorem 4-1 as

$$e(\xi) = \frac{u''(\zeta(\xi))}{2}(\xi+1)(\xi-1).$$

We can estimate it from above

$$|e(\xi)| \leq \frac{1}{2} \max_{-1\leq\xi\leq1} |u''\xi)| \max_{-1\leq\xi\leq1} |\xi^2 - 1| \leq \frac{1}{2} \max_{-1\leq\xi\leq1} |u''\xi)|.$$

Introducing the notation

$$||f||_{\infty,j} := \max_{x\in[x_{j-1},x_j]} |f(x)|$$

and using the relationship

$$\frac{d^2u(\xi)}{d\xi^2} = \frac{h_j^2}{4}\frac{d^2u(\xi)}{dx^2}, \tag{47}$$

we get

$$||e||_{\infty,j} \leq \frac{h_j^2}{8}||u''||_{\infty,j},$$

$$||e||_{\infty} \leq \frac{h^2}{8}||u''||_{\infty}. \tag{48}$$

Here

$$||e||_{\infty} = \max_{1\leq j\leq N} ||e||_{\infty,j} = \max_{[x_0,x_N]} |e(x)|,$$

$$h = \max_{1\leq j\leq N} h_j.$$

So we get the error estimate through the maximal length of the element and the largest derivative value.

The average estimate.

Here we will find the average error estimates, i.e. the estimates in the L^2 space. First we calculate the integral over each physical element:

$$\int_{x_{j-1}}^{x_j} e^2(x)\,dx = \frac{h_j}{2}\int_{-1}^{1}\left[\frac{u''(\zeta(\xi))}{2}(\xi^2-1)\right]^2 d\xi,$$

$$\int_{x_{j-1}}^{x_j} e^2(x)\,dx \leq \frac{h_j}{8}\int_{-1}^{1}\left[u''(\zeta(\xi))\right]^2 d\xi,$$

that can be written as

$$||e||_{0,j}^2 \leq \frac{h_j}{8}\int_{-1}^{1}\left[u''(\zeta(\xi))\right]^2 d\xi.$$

The function $\zeta(\xi) \in (-1,1)$ is monotonic on this interval. Hence, making the variable substitution, we get

$$||e||_{0,j}^2 \leq \frac{h_j}{8}\int_{\zeta(-1)}^{\zeta(1)}\left[u''(\zeta(\xi))\right]^2 \frac{d\zeta(\xi)}{|\zeta'(\xi)|} \leq$$

$$\leq \frac{h_j}{8} \int_{-1}^{1} \left[u''(\zeta(\xi))\right]^2 \frac{d\zeta(\xi)}{|\zeta'(\xi)|} \leq C^2 \frac{h_j}{8} \int_{-1}^{1} \left[u''(\zeta)\right]^2 d\zeta.$$

Changing coordinates to the physical ones with Eq.(47), we have

$$||e||_{0,j}^2 \leq= C^2 \frac{h_j^4}{64} \int_{x_{j-1}}^{x_j} \left[u''(x)\right]^2 dx = C^2 \frac{h_j^4}{64} ||u''||_{0,j}^2.$$

Summing up over all elements and calculating the square root, we get

$$||e|| \leq \tilde{C} h^2 ||u''||. \tag{49}$$

The error estimate in H^1 space.

The norm in the Hardy space includes also the derivative, so we need additionally estimate the derivative error

$$e'(\xi) = u'(\xi) - U'(\xi).$$

By differentiating the expression for $e(\xi)$ we find

$$e'(\xi) = u''(\xi)\xi + \frac{u'''(\zeta)}{2} \frac{d\zeta}{d\xi}(\xi^2 - 1).$$

Therefore, the integral of the derivative square over the physical element is calculated as

$$||e'||_{0,j}^2 = \int_{x_{j-1}}^{x_j} \left[\frac{de(x)}{dx}\right]^2 dx = \frac{2}{h_j} \int_{-1}^{1} \left[u''(\xi)\xi + \frac{u'''(\zeta)}{2} \frac{d\zeta}{d\xi}(\xi^2 - 1)\right]^2 d\xi.$$

Using the boundness of $d\zeta/d\xi$, in the way similar to the L^2 case we estimate

$$||e'||_{0,j}^2 \leq C h_j^2 ||u''||_{0,j}^2$$

and finally

$$||e'||_0^2 \leq \tilde{C} h^2 ||u''||_0^2.$$

Summarizing the estimates for the function and the derivative, we find the estimate in the H^1 space

$$||e||_1^2 = ||e||_0^2 + ||e'||_0^2 \leq C_1 h^4 ||u''||_0^2 + C_2 h^2 ||u''||_0^2,$$

$$||e||_1 \leq C h ||u''||_0.$$

12.2 The approximation of the degree p

The question about the approximation with a higher polynomial degree is more technically involved than the linear case. The main result gives the following theorem

Theorem 4-2. Let us choose the interval triangulation $[a = x_0 < x_1 < x_2 ... < x_N = b]$. *Let* $U(x)$ *be the polynomial of the degree less than or equal to* p *on all elements* $[x_{j-1}, x_j]$ *and* $U(x) \in H^1[a, b]$. *We also assume that* $U(x)$ *interpolates* $u(x) \in H^{p+1}[a, b]$ *in such a way that this interpolation is exact for any polynomial of the degree not exceeding* p. *Then there exists such constant* C_p *that*

$$||u - U||_0 \leq C_p h^{p+1} ||u^{(p+1)}||_0$$

and

$$||u - U||_1 \leq C_p h^p ||u^{(p+1)}||_0,$$

where $h = \max_j(h_j)$.

Proof is not given here, only a sketch. It is based on Theorem 4-1

$$|e(\xi)| \leq \max_{\zeta \in [a,b]} \left[\frac{|u^{(p+1)}(\zeta)|}{(p+1)!} \left| \prod_{i=0}^{p} (\xi - \xi_i) \right| \right] \leq \frac{||u^{(p+1)}||_\infty}{(p+1)!} \left| \prod_{i=0}^{p} (\xi - \xi_i) \right|.$$

The estimate of

$$\prod_{i=0}^{p} (\xi - \xi_i)$$

for the elements $[\xi_{k-1}, \xi_k]$ with the same length is given in book [6]

$$|\prod_{i=0}^{p} (\xi - \xi_i)| \leq (p - k + 1)! k! h^{p+1}.$$

Then the proof is performed in the way similar to the linear case. ■

Lecture 5

The main stages of the multidimensional FEM application. The Lagrange elements it the triangle. The Lagrange elements it the rectangle.

13 The main stages of the multidimensional FEM application

As well as for the one-dimensional case, we can formulate few stages of the FEM construction in the multidimensional case. We would like to note that their relative complexity and importance are not necessary the same as for one-dimensional case. These stages are:

- **The variational formulation of the problem in the domain Ω.**
  The problem should be formulated as the variational equation
  $$A(v,u) = (v,f) + < v, \beta > \qquad \text{for all } v \in \mathcal{H}. \tag{50}$$
  In this equation we have explicitly included the boundary conditions with the function β. The accurate work with the boundary condition is considerably more complicated here comparing to the one dimensional case, mainly due to the domain triangulation.

- **The domain triangulation.**
  The triangulation in the one-dimensional case is trivial as any interval can be exactly divided into an arbitrary number of elements. In the multidimensional case, the triangulation is a complicated geometrical problem. Furthermore, this problem often cannot be solved *exactly.* For example, the domain with a curved boundary $\partial\Omega$ cannot be exactly triangulated into triangles or rectangles. Therefore, we have here a new type of errors, so-called triangulation errors.

- **The construction of the elemental stiffness matrices and load vectors.**
  In the same manner as for the one-dimensional case, we reduce the

integral over the whole domain Ω in Eq.(50) to the sum of integrals over all physical elements

$$\sum_{i=1}^{N} [A_i(V,U) - (V,f)_i - < V, \beta >_i] = 0 \quad \forall V. \tag{51}$$

For example, for the equation

$$-\nabla\left(p(x,y)\nabla v(x,y)\right) + q(x,y)v(x,y) = f(x,y),$$

the functionals in Eq.(51) can be written as

$$A_i(V,U) = \int\int_{\Omega_i} (V_x p(x,y)U_x + V_y p(x,y)U_y + Vq(x,y)U)dx\,dy,$$

$$(V,f)_i = \int\int_{\Omega_i} Vf\,dx\,dy,$$

$$< V, \beta >_i = \int\limits_{\partial\Omega_i \cap \partial\tilde{\Omega}} V\beta\,ds.$$

It is important that the integration in the last integral is performed over the approximation of the domain boundary $\partial\tilde{\Omega}$, not over the boundary $\partial\Omega$ itself. As we have already noted, they do not necessarily coincide.

- **The construction of the global stiffness matrix and load vector.**
  As well as for the one-dimensional case, we should number all the basis functions and combine the global stiffness matrix and the load vector from the elemental matrices. For the one-dimensional case, we had a simple and effective numbering according to the increase of the coordinate. In the multidimensional case this numbering is missed but we have a few different strategies to number the basis functions.

- **The solution of the linear algebraic system.**
  The solution of the linear algebraic system is the last stage of the FEM

construction. It hardly differs from the one-dimensional case. As the variational equation

$$\mathbf{d}^\top\left[(\mathbf{K}+\mathbf{M})\mathbf{c}-\mathbf{l}\right]=0$$

has to be satisfied for all $\mathbf{d}$, we need to solve the linear algebraic system

$$(\mathbf{K}+\mathbf{M})\mathbf{c}=\mathbf{l},$$

where $\mathbf{K}$, $\mathbf{M}$ are the matrix of inertial and internal energy, and $\mathbf{l}$ is the load vector.

As well as in the one-dimensional case, we start the analysis of the FEM with the construction of the global basis functions. In the multidimensional case, the basis functions can be attributed to various classes of objects. For example, in the three-dimensional case $\mathbb{R}^3$ the basis functions can be connected with vertexes, edges, and faces. Consequently, the elemental functions ϕ_j attributed to jth object, are non-zero only on the elements which contain the object j.

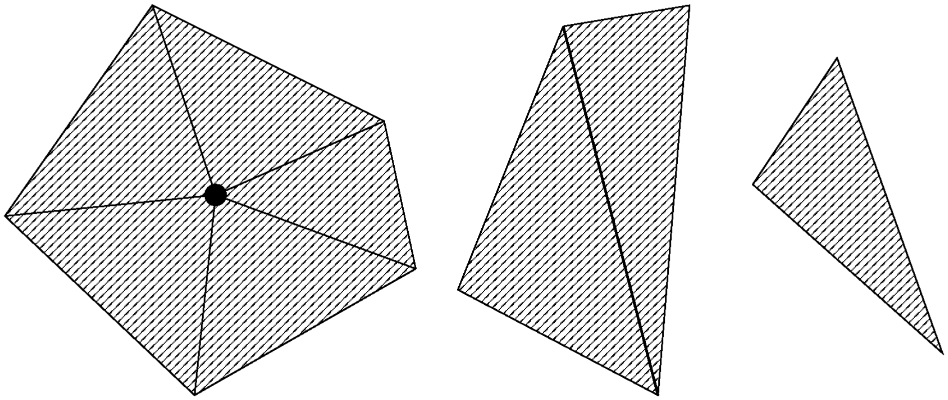


Figure 10: The domains where the functions attributed to the vertex, the edge and the element are non-zero.

14 The Lagrange elements on the triangle

We start our construction for the two-dimensional case. It is relatively simple but still contains all the typical problems of the multidimensional FEM.

Let us first construct the linear Lagrange elements. In order to do this we consider the triangle with the vertexes 1, 2, 3 having coordinates (x_j, y_j), $j = 1, 2, 3$. The functions under construction have to be linear and have to satisfy

$$N_j(x_k, y_k) = \delta_{jk}, \quad j, k = 1, 2, 3. \tag{52}$$

As an arbitrary linear function in $\mathbb{R}^2$ can be written as

$$N_j(x, y) = a + bx + cy, \quad x, y \in \Omega_e,$$

we only need to find coefficients a, b and c. We can write condition (52) as

$$\begin{bmatrix} 1 & x_j & y_j \\ 1 & x_k & y_k \\ 1 & x_l & y_l \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad k \neq l \neq j.$$

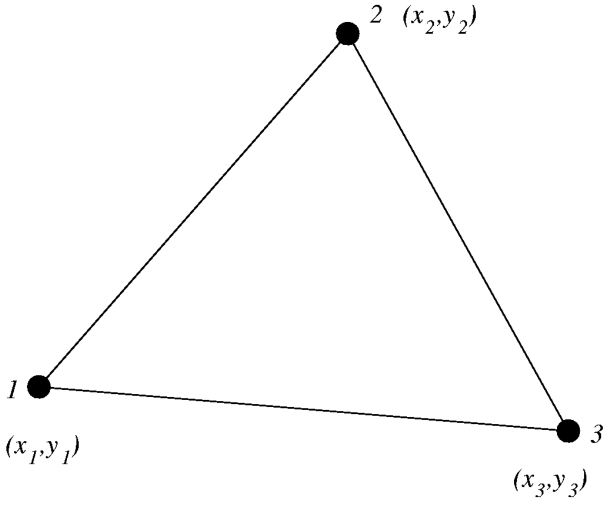


Figure 11: The triangle element with the vertices 1, 2, 3.

The solution can be found in terms of the Cramer's formulas:

$$N_j(x,y) = \frac{D_{kl}(x,y)}{C_{jkl}}, \quad k \neq l \neq j,$$

where

$$D_{kl} = \det \begin{bmatrix} 1 & x & y \\ 1 & x_k & y_k \\ 1 & x_l & y_l \end{bmatrix}, \quad C_{jkl} = \det \begin{bmatrix} 1 & x_j & y_j \\ 1 & x_k & y_k \\ 1 & x_l & y_l \end{bmatrix}.$$

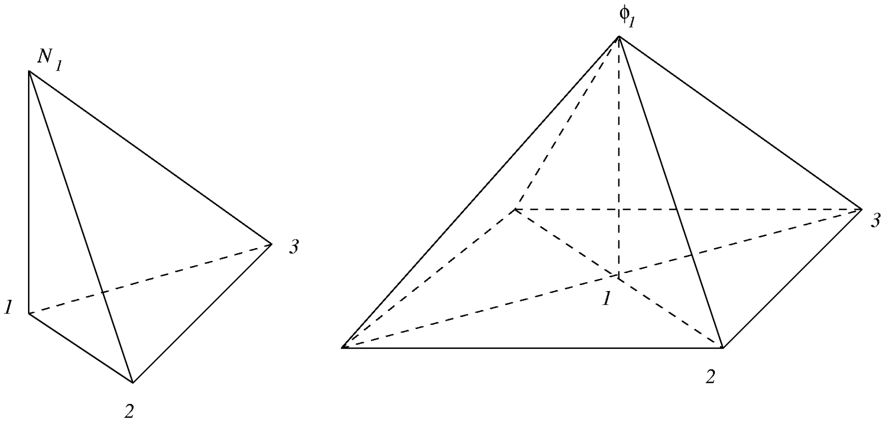


Figure 12: The elemental function N_1 and the basis function ϕ_1.

The restriction of the function $U(x,y)$ on the element e $(x,y \in \Omega_e)$ is

$$U(x,y) = c_1 N_1(x,y) + c_2 N_2(x,y) + c_3 N_3(x,y).$$

Condition (52) defines the coefficients c_j

$$c_j = U(x_j, y_j), \quad j = 1,2,3.$$

14.1 The Lagrange elements of the order p

Let us define n_p functions $N_j(x,y)$:

$$j = 1, 2 ... n_p = \frac{(p+1)(p+2)}{2},$$

$$N_j(x,y) = \sum_{i=1}^{n_p} a_i q_i(x,y) = \mathbf{a}^\top \mathbf{q}(x,y),$$

where

$$\mathbf{q}^\top(x,y) = [1, x, y, x^2, xy, y^2, ..., y^p].$$

For example, for the quadratic element $p = 2$, $n_2 = 6$ and

$$\mathbf{q}^\top(x,y) = [1, x, y, x^2, xy, y^2].$$

It is important to have all n_p terms as only for this set the maximal polynomial degree is conserved under the shift and rotation transformations.

For the sake of convenience, we introduce nodes in order to number the elemental function inside of each element, see Fig. 13.

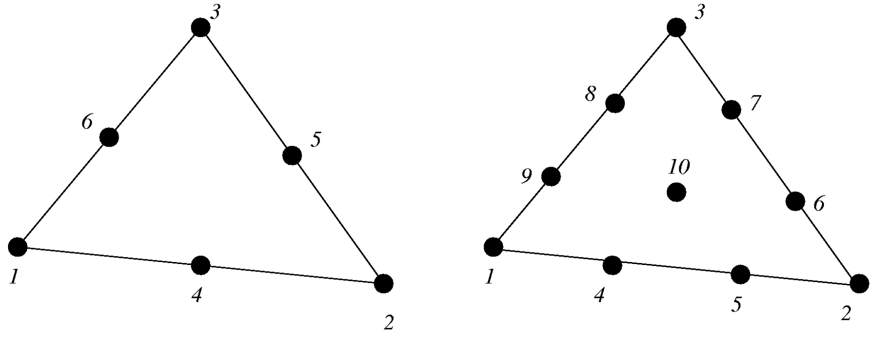


Figure 13: The nodes for the quadratic and cubic approximations.

$$N_j = a_1 + a_2 x + a_3 y + a_4 x^2 + a_5 xy + a_6 y^2.$$

The Lagrange interpolation conditions have to be satisfied at all nodes:

$$N_j(x_k, y_k) = \delta_{jk}, \quad j, k = 1, 2, ..., 6.$$

$$\phi_j = \bigcup_{e=1} N_{j,e}(x,y).$$

The function on the edge is defined by the points which belong to this edge only that assures the continuity of the basis functions.

We can calculate the coefficients a_i exactly the same way as for the linear element. While the idea of these calculations is quite simple, technically they become very involved. So we will use few different ideas to make this calculation more transparent.

The first idea is the use of the canonical element and the *coordinate transformations*. Additionally to the physical (x, y)-plane we introduce the computational (ξ, η)-plane, and construct a transformation which relates these two planes. Then we define the elemental functions on the canonical element in the computational plane and recalculate them to an arbitrary physical element with the coordinate transformation.

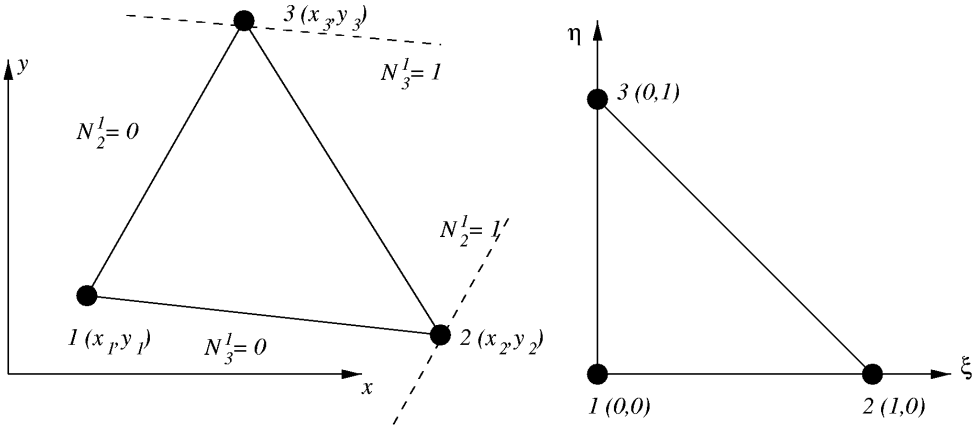


Figure 14: The mapping of an arbitrary triangle onto the canonical element.

The coordinate transformation can be calculated as follows. The equation for the line connecting vertexes (1) and (3) reads $N_2(x, y) = 0$. The line which is parallel to $N_2(x, y) = 0$ and crossing the vertex (2) is defined as $N_2(x, y) = 1$. So the mapping of the physical line $N_2(x, y) = 0$ into the computational line $\xi = 0$ is given by

$$\xi = N_2(x, y).$$

In the same way, the mapping of the physical line connecting the vertexes

(1) and (2) is given by

$$\eta = N_3(x, y).$$

Taking into account the definition of the functions $N_j(x, y)$ we have

$$\xi = \frac{\det \begin{bmatrix} 1 & x & y \\ 1 & x_1 & y_1 \\ 1 & x_3 & y_3 \end{bmatrix}}{\det \begin{bmatrix} 1 & x_2 & y_2 \\ 1 & x_1 & y_1 \\ 1 & x_3 & y_3 \end{bmatrix}}, \quad \eta = \frac{\det \begin{bmatrix} 1 & x & y \\ 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{bmatrix}}{\det \begin{bmatrix} 1 & x_3 & y_3 \\ 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{bmatrix}}. \tag{53}$$

As noted above, transformations (53) are used to recalculate the elemental functions from the computational coordinates into the physical coordinates.

2. Baricenteric coordinates.

Here the idea is to introduce a new convenient coordinate system. As such system, we chose the triple of the $\{N_1, N_2, N_3\}$ values. These coordinates are called baricenteric (or triangle) coordinates. And yes, they are redundant! We will use the following notations for the baricenteric coordinates:

$$\zeta_1 = N_1, \quad \zeta_2 = N_2, \quad \zeta_3 = N_3.$$

It is easy to see that the transformation from the baricenteric coordinates to the physical ones is given by

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{bmatrix}.$$

The last row describes the redundance of the coordinates. As the transformation is linear, it suffices to check the two first rows at the vertexes:

$$\text{node (1): } (\zeta_1, \zeta_2, \zeta_3) = (1, 0, 0),$$

$$\text{node (2): } (\zeta_1, \zeta_2, \zeta_3) = (0, 1, 0),$$

$$\text{node (3): } (\zeta_1, \zeta_2, \zeta_3) = (0, 0, 1).$$

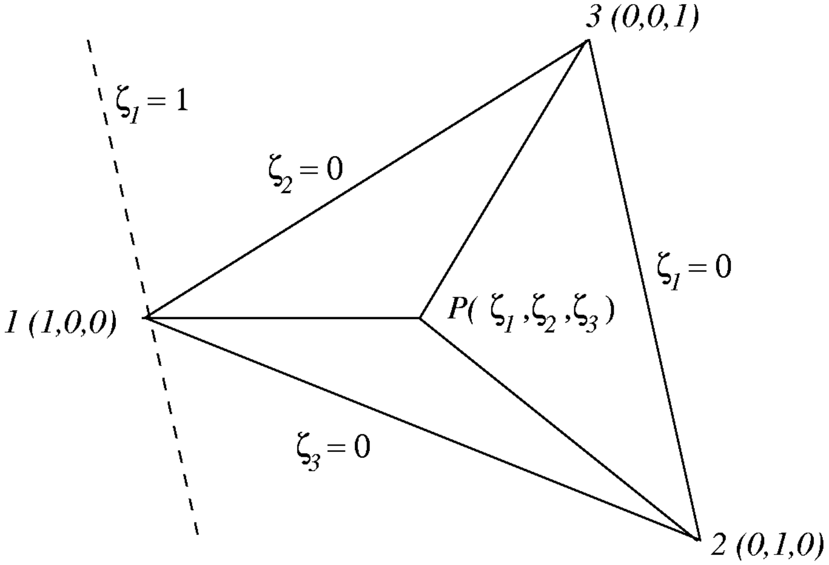


Figure 15: The baricenteric coordinate system.

The baricenteric coordinates can also be defined in the geometrical way. Namely, they are given by the ratios of the triangle squares A_{ijk}

$$\zeta_1 = \frac{A_{P23}}{A_{123}}, \quad \zeta_2 = \frac{A_{P31}}{A_{123}}, \quad \zeta_3 = \frac{A_{P12}}{A_{123}}.$$

It is possible to show that these two definitions are equivalent.

Example for the cubic elemental functions $p = 3$.

The Lagrange elemental functions are constructed by choosing the product of the baricenteric coordinates in such a way that they are equal to zero in all points of the set except of one. Then we normalize the result to make it equal to one at the non-zero point. For example,

$$N_1^3(x,y) = \frac{9}{2}\zeta_1(\zeta_1 - 1/3)(\zeta_1 - 2/3) = \frac{9}{2}N_1^1(N_1^1 - 1/3)(N_1^1 - 2/3),$$

$$N_4^3(x,y) = \frac{27}{2}\zeta_1\zeta_2(\zeta_1 - 1/3) = \frac{27}{2}N_1^1N_2^1(N_1^1 - 1/3),$$

$$N_{10}^3(x,y) = 27\zeta_1\zeta_2\zeta_3 = 27N_1^1N_2^1N_3^1.$$

Some of these functions are plotted in Fig. 16.

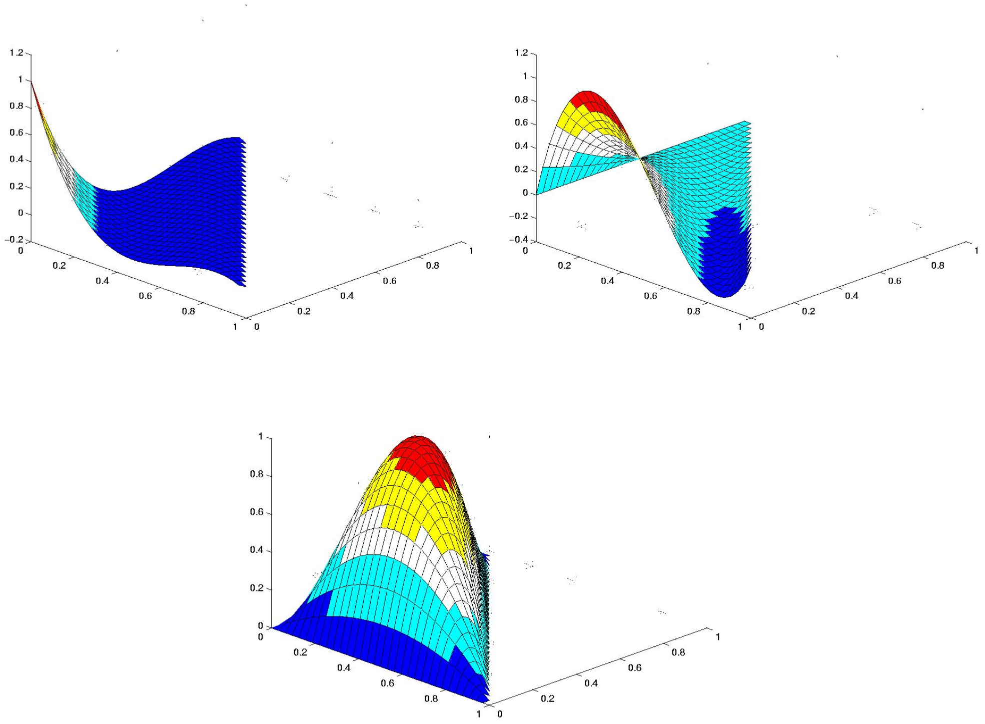


Figure 16: The cubic Lagrange elemental functions on the canonical triangle.

15 The Lagrange elements on the rectangle

The triangle has the minimal number of the edges (correspondingly, the surfaces in $\mathbb{R}^3$), so it is optimal to define the continuous basis functions. In other cases, and especially for appropriate domains, the rectangle can appear to be more convenient. So, here we discuss the Lagrange elements on the rectangle. First, we define the canonical square as

$$\{(\xi,\eta) : -1 \le \xi, \eta \le 1\}.$$

The multidimensional polynomials are constructed as the direct (tensor) product of the one-dimensional polynomials.

The bilinear Lagrange functions

$$U(\xi,\eta) = c_{1,1}N_{1,1}(\xi,\eta) + c_{2,1}N_{2,1}(\xi,\eta) + c_{2,2}N_{2,2}(\xi,\eta) + c_{1,2}N_{1,2}(\xi,\eta).$$

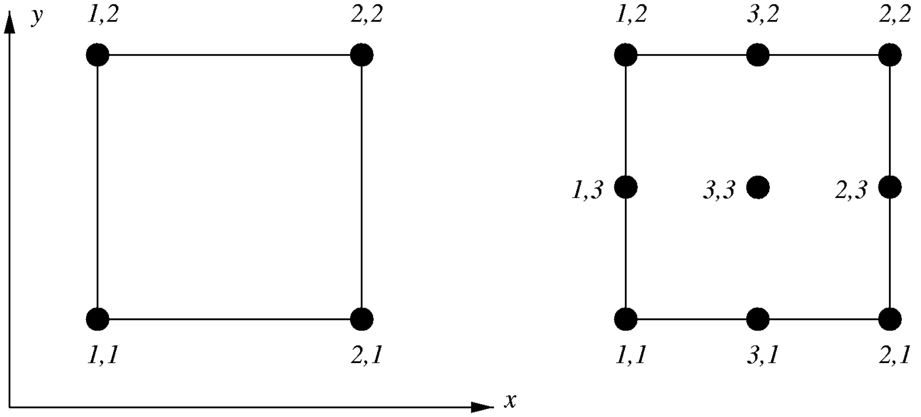


Figure 17: The canonical square for the bilinear and biquadratic polynomials.

As usual, for the Lagrange functions the following properties are satisfied

$$N_{i,j}(\xi_k, \eta_l) = \delta_{ik}\delta_{jl}, \quad U(\xi_k, \eta_l) = c_{k,l}.$$

As we stated above, the functions $N_{i,j}(\xi, \eta)$ for the rectangle are the products of the one-dimensional polynomials:

$$N_{i,j}(\xi, \eta) = N_i^1(\xi)N_j^1(\eta),$$

where

$$N_1^1(\xi) = \frac{1-\xi}{2}, \quad N_2^1(\xi) = \frac{1+\xi}{2}, \quad -1 \le \xi \le 1.$$

The bilinear functions $N_{i,j}(\xi, \eta)$ can be represented as

$$N_{i,j}(\xi, \eta) = a_1 + a_2\xi + a_3\eta + a_4\xi\eta.$$

We notice here that these functions contain also the quadratic term. We shall discuss that term later.

The functions $N_{i,j}(\xi, \eta)$ are the elemental functions. The basis functions are constructed by merging of four functions from the neighboring physical rectangles, which correspond to the same vertex. As the elemental functions are linear on all edges, their values on the edges are defined by the values at two corresponding vertexes only. Therefore, the basis function are continuous.

The biquadratic Lagrange elemental functions Any function $U(\xi, \eta)$ on the rectangle is approximated with the biquadratic basis as

$$U(\xi, \eta) = \sum_{i=1}^{3} \sum_{j=1}^{3} c_{i,j} N_{i,j}(\xi, \eta),$$

where

$$N_{i,j}(\xi, \eta) = N_i^2(\xi) N_j^2(\eta), \quad i, j = 1, 2, 3.$$

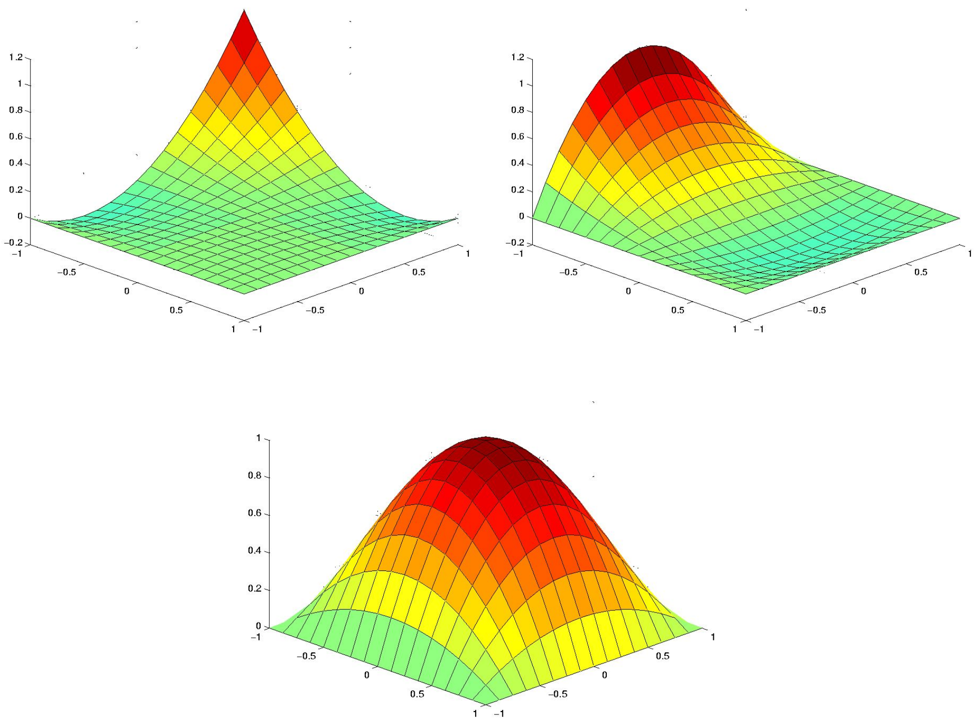


Figure 18: The biquadratic elemental functions for the vertex, edge and surface.

The one-dimensional elemental functions are given by

$$N_1^2(\xi) = -\xi(1 - \xi)/2, \quad N_2^2(\xi) = \xi(1 + \xi)/2,$$

$$N_3^2(\xi) = (1 - \xi^2), \quad -1 \leq \xi \leq 1.$$

So we can see that the biquadratic function $N_{i,j}(\xi, \eta)$ is generally written as

$$N_{i,j}(\xi, \eta) = a_1 + a_2\xi + a_3\eta + a_4\xi^2 + a_5\xi\eta + a_6\eta^2 +$$

$$+a_7\xi^2\eta + a_8\xi\eta^2 + a_9\xi^2\eta^2.$$

It contains some terms of the order higher than two, they ale collected in the second line. However, an arbitrary polynomial of the second degree only can be represented by $N_{i,j}(\xi, \eta)$. According to Theorem 4-2, we have here the approximation of the second order. The higher degree terms do not improve the approximation. In fact, they are redundant, and might even result in the degradation of the numerical accuracy and stability.

We have constructed the basis function on the computational plane, and need to transform them into the physical plane. The mapping of the canonical square onto the physical rectangle (x_{ij}, y_{ij}) is given with the functions N_i^1

$$\begin{bmatrix} x \\ y \end{bmatrix} = \sum_{i=1}^{2} \sum_{j=1}^{2} \begin{bmatrix} x_{ij} \\ y_{ij} \end{bmatrix} N_i^1(\xi) N_j^1(\eta).$$

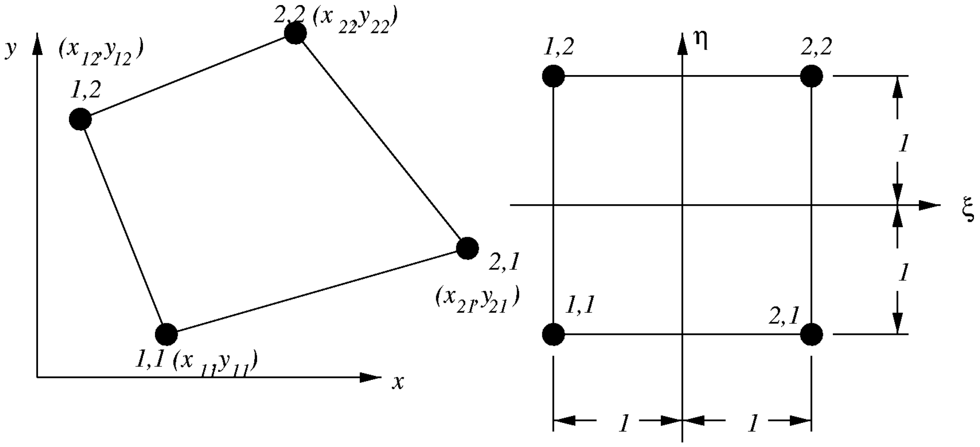


Figure 19: The mapping of the canonical square onto the physical rectangle.

For example, the edge $\eta = -1$ maps onto the edge $(x_{11}, y_{11}) - (x_{21}, y_{21})$:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_{11} \\ y_{11} \end{bmatrix} \frac{1-\xi}{2} + \begin{bmatrix} x_{21} \\ y_{21} \end{bmatrix} \frac{1+\xi}{2}. \tag{54}$$

It is important that the vertexes (x_{12}, y_{12}) and (x_{22}, y_{22}) do not enter Eq.(54) and do not affect this transformation. Therefore, the continuity of the basis functions is preserved with this mapping.

Lecture 6

Multidimensional FEM. The hierarchical finite elements. Three-dimensional elements. The interpolation errors.

16 The hierarchical elements

In order to study the methods for the construction of the multidimensional hierarchical finite elements, we restrict ourselves to the simpler case of the rectangles. We describe the hierarchical basis for the square. These functions can be then recalculated for an arbitrary rectangle with the transformation described in the previous lecture.

The order p basis consists of the different types of functions. Those functions are linked with the vertices, edges and the center of the square. Let us describe them.

Elemental functions for the vertices (4 bilinear functions):

$$N^1_{ij}(\xi,\eta) = N^1_i(\xi)N^1_j(\eta), \quad i,j = 1,2.$$

Elemental functions for the edges (linked with the center of edges, $4(p-1)$ *functions existing for* $p \geq 2$*)*:

$$N^k_{31}(\xi,\eta) = N^1_1(\eta)N^k(\xi), \quad N^k_{13}(\xi,\eta) = N^1_1(\xi)N^k(\eta),$$

$$N^k_{32}(\xi,\eta) = N^1_2(\eta)N^k(\xi), \quad N^k_{23}(\xi,\eta) = N^1_2(\xi)N^k(\eta).$$

In the last equations the index k spans the values $k = 2, 3...p$. The function $N^k(y)$ are defined as

$$N^k(y) = \sqrt{\frac{2k-1}{2}} \int_{-1}^{y} P_{k-1}(t)\, dt.$$

Here $P_{k-1}(t)$ is the Legendre polynomial of the degree $k-1$. It is easy to check that $N^k(-1) = N^k(1) = 0$, so each elemental function for the edges is identically zero on three edges out of four. Few of these functions are plotted in Fig. 20.

Elemental functions linked with the center of the square (3,3). These $(p-2)(p-3)/2$ *functions existing for* $p \geq 4$*)* can be described in terms of the function N_{33}^{400}:

$$N_{33}^{400} = (1-\xi^2)(1-\eta^2).$$

For $p = 4$, only this function enters the expansion. For the polynomial degrees $p = 5, 6$, the hierarchical functions are defined as

$$N_{33}^{510} = N_{33}^{400} P_1(\xi), \quad N_{33}^{501} = N_{33}^{400} P_1(\eta),$$

$$N_{33}^{620} = N_{33}^{400} P_2(\xi), \quad N_{33}^{611} = N_{33}^{400} P_1(\xi) P_1(\eta), \quad N_{33}^{602} = N_{33}^{400} P_2(\eta).$$

The upper index $k\lambda\mu$ of the function consists of the total polynomial degree and the degrees of the additional polynomials in ξ and η coordinates so that $k = \lambda + \mu + 4$. Few of these functions are plotted in Fig. 21. For the higher polynomial degrees $p > 6$, the functions are introduced in the similar way.

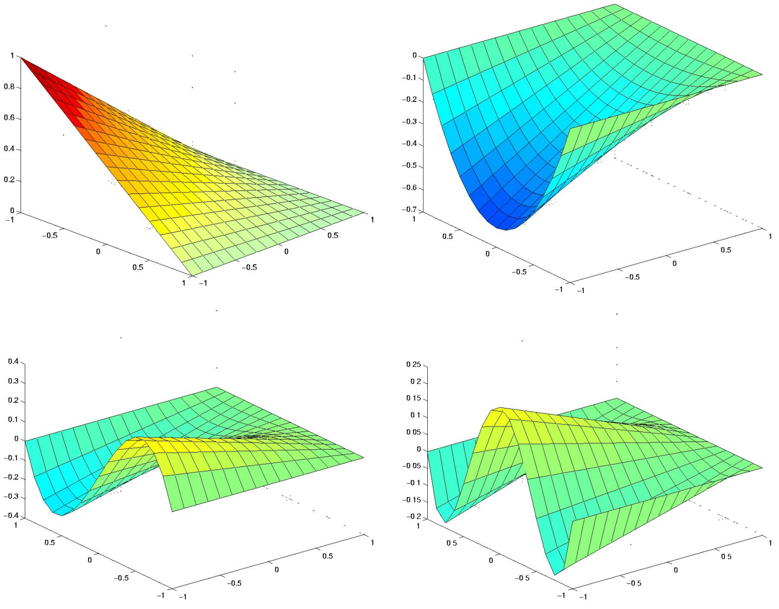


Figure 20: The hierarchical functions for the vertex and for the edges ($k = 1, 2, 3, 4$).

The solution in terms of the hierarchical basis is expanded as follows:

$$U(\xi,\eta) = \sum_{i=1}^{2}\sum_{j=1}^{2} c_{ij}^{1} N_{ij}^{1} + \sum_{k=2}^{p}\left[\sum_{j=1}^{2} c_{3j}^{k} N_{3j}^{k} + \sum_{i=1}^{2} c_{i3}^{k} N_{i3}^{k}\right] + \sum_{k=4}^{p}\sum_{\lambda+\mu=k-4} c_{33}^{k\lambda\mu} N_{33}^{k\lambda\mu}.$$

The total number of the functions in this representation can be calculated as

$$4 + 4(p-1)_{+} + \frac{(p-2)_{+}(p-3)_{+}}{2}, \quad \text{where} \quad q_{+} = \max(q,0).$$

It is interesting to compare the total number of the function for three basises of the order p: the direct product basis and the hierarchical basis in the canonical square, and the minimally admissible set which coincides with the basis in the canonical triangle. These number are presented in the table:

degree p	1	2	3	4
the triangle basis	3	6	10	15
the direct product basis	4	9	16	25
the hierarchical basis	4	8	12	17

One can see that both basises for the square are not optimal with respect to the number of functions. For the low polynomial degrees $p = 1, 2$, they are very similar, and the number of the functions essentially exceeds that for the triangle. For higher p, however, quality of the hierarchical basis improves and the number of the functions asymptotically converges to the optimal value. Quality of the direct product basis stays the same, the number of the functions essentially exceeds the optimal value.

17 The three-dimensional elements

17.1 The tetrahedron

For the three-dimensional case, the number of polynomials of the degree p is equal to

$$n_p = \frac{(p+1)(p+2)(p+3)}{6}.$$

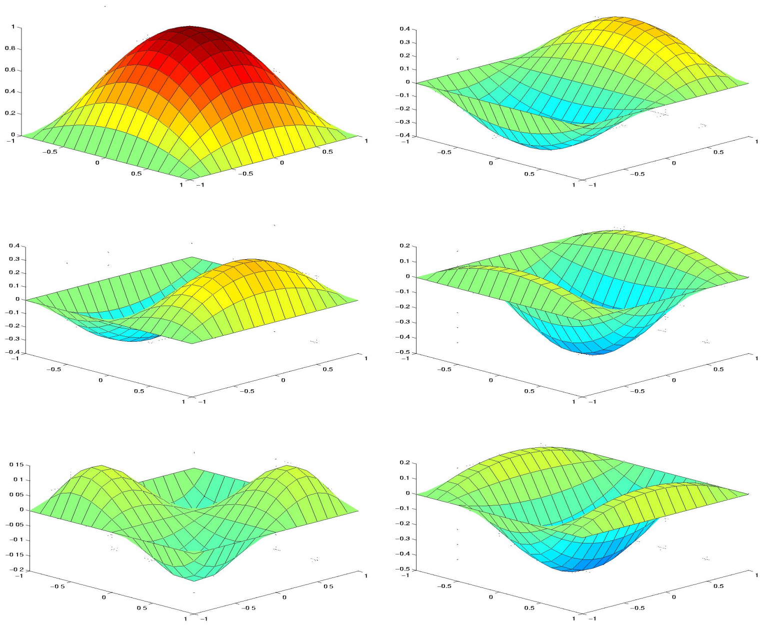


Figure 21: The hierarchical elemental functions linked with the center of the square $N_{33}^{k\lambda\mu}$, $\lambda + \mu = k - 4$ $(k = 4, 5, 6)$.

The construction of the elemental functions is done in the same way as for the two-dimensional case so we give here the formulas for the linear Lagrange elements only.

For the tetrahedron, the Lagrange conditions

$$N_j(x_k, y_k, z_k) = \delta_{jk}, \quad j, k = 1, 2, 3, 4,$$

lead to the following representation for the linear functions:

$$N_j(x, y, z) = \frac{D_{klm}(x, y, z)}{C_{jklm}}.$$

Here $(jklm)$ is a permutation of (1,2,3,4) numbers, and the determinants D_{klm} and C_{jklm} are written as

$$D_{klm} = \det \begin{bmatrix} 1 & x & y & z \\ 1 & x_k & y_k & z_k \\ 1 & x_l & y_l & z_l \\ 1 & x_m & y_m & z_m \end{bmatrix}, \quad C_{jklm} = \det \begin{bmatrix} 1 & x_j & y_j & z_j \\ 1 & x_k & y_k & z_k \\ 1 & x_l & y_l & z_l \\ 1 & x_m & y_m & z_m \end{bmatrix}.$$

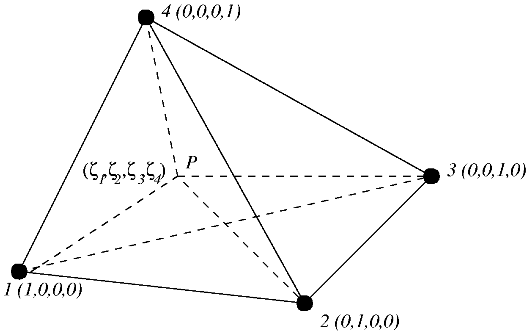


Figure 22: The nodes for the linear elemental function on the tetrahedron and barycentric coordinates.

The projection of the solution U on the element can be expanded in terms of these functions as

$$U(x, y, z) = \sum_{j=1}^{4} c_j N_j(x, y, z).$$

The barycentric coordinates (also known as the volume coordinates) are defined as

$$\zeta_j = N_j(x, y, z), \quad j = 1, 2, 3, 4.$$

These coordinates also give the volumes of the corresponding tetrahedrons with the vertex at the point P:

$$\zeta_1 = \frac{V_{P234}}{V_{1234}}, \quad \zeta_2 = \frac{V_{P134}}{V_{1234}}, \quad \zeta_3 = \frac{V_{P124}}{V_{1234}}, \quad \zeta_4 = \frac{V_{P123}}{V_{1234}}.$$

As in the two-dimensional case, the barycentric coordinates are redundant.

The inverse coordinate transformation can be defined with the formula

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \end{bmatrix}.$$

The last line in this equation shows us the redundancy of the barycentric coordinates.

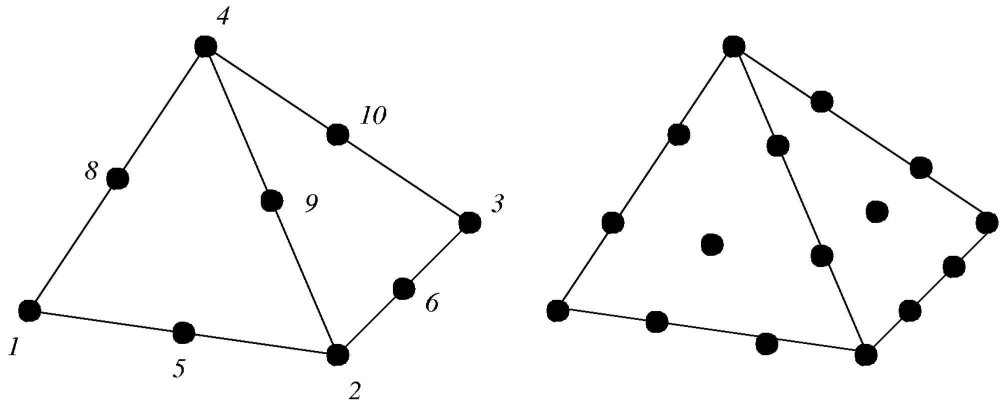


Figure 23: The nodes for the quadratic and cubic elemental functions on the tetrahedron.

17.2 The cube

The canonical cube is defined as

$$\{(\xi, \eta, \zeta) : \, -1 \leq \xi, \eta, \zeta \leq 1\}.$$

The elemental function for the node (ijk) is written as the direct product of the one-dimensional linear functions:

$$N_{ijk} = N_i(\xi) N_j(\eta) N_k(\zeta).$$

18 The interpolation error

The estimation of the interpolation errors will be done in two steps:

- the interpolation error estimation with polynomials on the canonical element
- the recalculation of the interpolation errors from the canonical to the physical element.

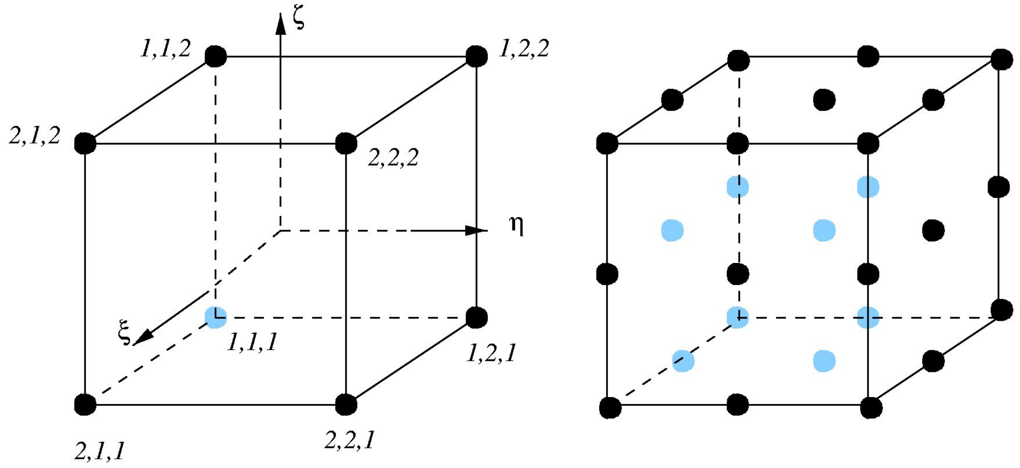


Figure 24: The nodes for the tri-linear tri-quadratic elemental functions on the cube.

For the sake of simplicity, let us consider the two-dimensional case. We expand a function on the canonical element in terms of the elemental functions

$$U(\xi,\eta)=\sum_{j=1}^{n} c_j N_j(\xi,\eta). \tag{55}$$

Then the interpolation error on the canonical element is estimated with the following

Theorem 1: Let p be the maximal integer number, such that Eq.(55) is exact for any polynimial of the degree p. Then for the canonical element Ω_0 there exists $C > 0$ such that

$$|u-U|_{s,\Omega_0} \le C|u|_{p+1,\Omega_0},$$

$$\forall u \in H^{p+1}(\Omega_0), \quad s = 0, 1 ... p+1.$$

where $|u|_{s,\Omega_0}$ stands for the Sobolev quasinorm.

In order to recalculate the interpolation error onto the physical element Ω_e, we need the transformation of the canonical triangle

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{bmatrix} =$$

$$= \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1-\xi-\eta \\ \xi \\ \eta \end{bmatrix}.$$

The Jacobian determinant of this transformation

$$J_e = \begin{bmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{bmatrix}$$

is easy to calculate. It is equal to

$$\det J_e = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1).$$

For applications, it is convenient to express this Jacobian in terms of geometric characteristics of the triangle. The corresponding result can be formulated as

Lemma: Let h_e be the longest side of the element Ω_e, and let α_e be its smallest angle. Then

$$\frac{h_e^2}{2}\sin(\alpha_e) < \det J_e < h_e^2 \sin(\alpha_e).$$

Also, one can prove the following

Theorem 2: Let $\theta(x,y) \in H^s(\Omega_e)$, $\tilde{\theta}(x,y) \in H^s(\Omega_0)$, where Ω_0 is the canonical element. Then there exist constants c_s, C_s such that

$$c_s \sin^{s-1/2}\alpha_e h_e^{s-1}|\theta|_{s,\Omega_e} \le |\tilde{\theta}|_{s,\Omega_0} \le C_s \sin^{-1/2}\alpha_e h_e^{s-1}|\theta|_{s,\Omega_e}.$$

From Theorems 1 and 2 we can derive the main results for the interpolation error estimates. They are formulated separately for the domains divided into triangles and rectangles.

The interpolation error for the triangles. *Let Ω be divided into the triangle elements Ω_e. Let h be the longest side in the mesh, and let α be the smallest angle. For the interpolation of the order p there exists a constant $C > 0$ (which is independent on $u \in H^{p+1}$ and the mesh) such that*

$$|u - U|_s \leq \frac{Ch^{p+1-s}}{\sin^s \alpha}|u|_{p+1}, \quad \forall u \in H^{p+1}(\Omega),\ s = 0, 1.$$

The interpolation error for the rectangles. *Let Ω be divided into the rectangle elements Ω_e. Let h be the longest side in the mesh, and let β be the smallest side ratio for the rectangles. For the interpolation of the order p there exists a constant $C > 0$ (which is independent on $u \in H^{p+1}$ and the mesh) such that*

$$|u - U|_s \leq \frac{Ch^{p+1-s}}{\beta^s}|u|_{p+1}, \quad \forall u \in H^{p+1}(\Omega),\ s = 0, 1.$$

Let us introduce now an important family of the meshes.

Definition The family of the finite elements is called *regular* if all angles in the mesh are separated from 0 and π (β is separated from 0) when the size of the elements goes to zero.

For the regular families, the above theorems can be simplified and combine into one result. Namely,

The interpolation error theorem for the regular family. *Let Ω be divided into elements Ω_e from the regular family. Let h be the longest side in the mesh. For the interpolation of the order p there exists a constant $C > 0$ (which is independent on $u \in H^{p+1}(\Omega)$ and the mesh) such that*

$$|u - U|_s \leq Ch^{p+1-s}|u|_{p+1}, \quad s = 0, 1. \tag{56}$$

It is important to note that the above results are valid for the solutions which are smooth enough. If this is not true, the convergence gets worse.

Namely, if the solution $u \in H^{q+1}(\Omega)$, $q < p$, then error estimate (56) has to be replaced with the estimate

$$|u - U|_s \leq Ch^{q+1-s}|u|_{q+1}, \quad s = 0, 1.$$

So one can see that for the non-smooth solution, the convergence speed is bounded by the smoothness of the solution, and the use of the higher degree polynomials becomes meaningless.

Lecture 7

Multidimensional FEM. The triangulation. The coordinate transformations.

19 The triangulation

In contrast to the one-dimensional case, the problem of dividing an arbitrary multi-dimensional domain into some number of elements (i.e. **the triangulation problem**) is rather complicated and requires a special consideration. This problem becomes even more complicated as we restrict ourselves to the regular families of the finite elements that is necessary to achieve an acceptable accuracy for the solution (see the previous lecture). The triangulation problem belongs to the field of the *computational geometry.* Depending on the properties of the domain and the accuracy required, the time spent for the triangulation can amount for an appreciably part of the whole FEM calculation time.

In this section, we briefly review ideas of the main methods used for the triangulation. More information can be found in books [5, 7].

19.1 Division into subdomains, Cock's method

The triangulation here is done in three steps:

- Description of the domain geometry
- Division of the domain into subdomains
- Division of the subdomains into finite elements

The main stage is the second stage. Here we put a number of points on the domain boundary and then join these points by straight lines. Then these points can be recalculated into curved domain with the coordinate transformation, see Fig. 25. The crossing points can be calculated as

$$u_{ij} = \frac{(u_{i,j_{max}} - u_{i,j_{min}})(v_{i_{min},j} + u_{i,j_{min}})}{1 - (u_{i,j_{max}} - u_{i,j_{min}})(v_{i_{max},j} + v_{i_{min},j})},$$

$$v_{ij} = \frac{(v_{i_{max},j} - v_{i_{min},j})(u_{i,j_{min}} + v_{i_{min},j})}{1 - (u_{i,j_{max}} - u_{i,j_{min}})(v_{i_{max},j} + v_{i_{min},j})}.$$

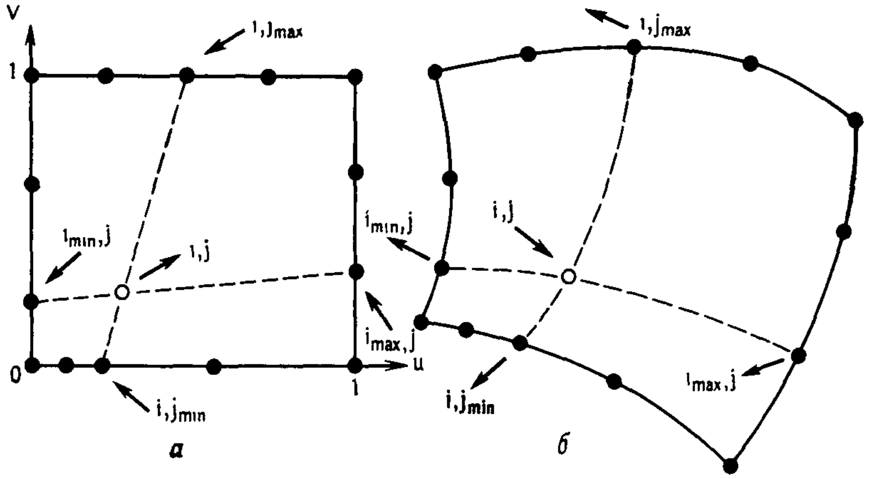


Figure 25: Division of the tetragon (Cock's method).

The corresponding subdomains can be chosen much smaller than the full initial domain, so they are easier to split into finite elements.

19.2 The triangulation by the grid covering

In this approach, we first prepare a regular grid consisting of triangles (or squares, tetrahedrons, cubes etc). Then we cover our domain of interest by this grid, and get a good triangulation of the main part of the domain except of its boundary. The nodes and sides of our mesh will not lie on the boundary. In order to correct this drawback, we can do the following:

1. move the nearest nodes to the boundary.

2. calculate the intersection points of the grid and the boundary, and add them to the nodes.

In both cases, we get more complicated polygons, which we should then divide into simpler ones.

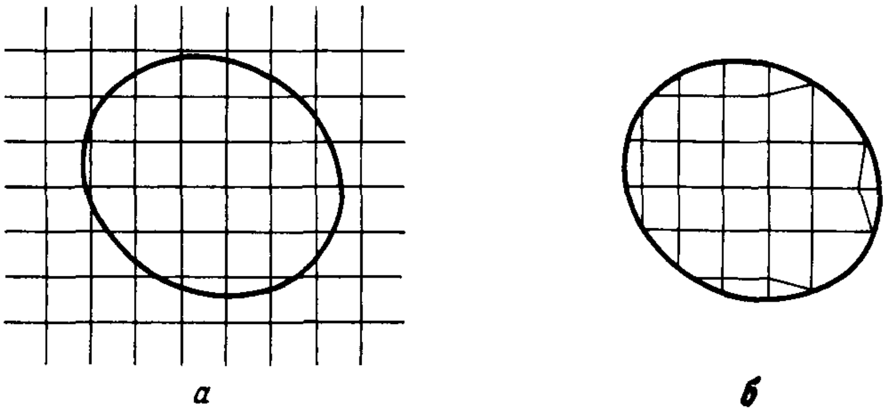


Figure 26: The triangulation by the grid covering.

Even with this adjustment, the average size of the elements is the same everywhere, that can lead to a rough description of the boundary. In order to get a more accurate approximation around the boundary, the technique of so-called quadratic (octal in the three-dimensional space) tree is used. The idea is to divide the rectangles crossing the boundary into 4 (8) smaller rectangles which give better approximation of the boundary, see Fig. 27. The procedure can be repeated when necessary.

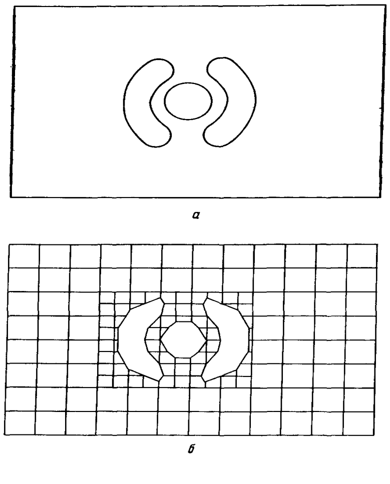


Figure 27: Using of the quadratic tree.

19.3 Frontal propagation

This approach is mainly used for the two-dimensional problems. We start from elements sitting on one side of the boundary, and add new elements layer by layer, propagating the front of the triangulated part of the domain, see Fig. 28. Typically, the elements used for this construction are chosen to be close to the equilateral triangle. In order to use this approach for non-simply connected domains, we need to modify it. This modification can be easily done with introducing of artificial cuts, Fig. 29. Depending on the connectivity of the domain, we may need one or a few cuts.

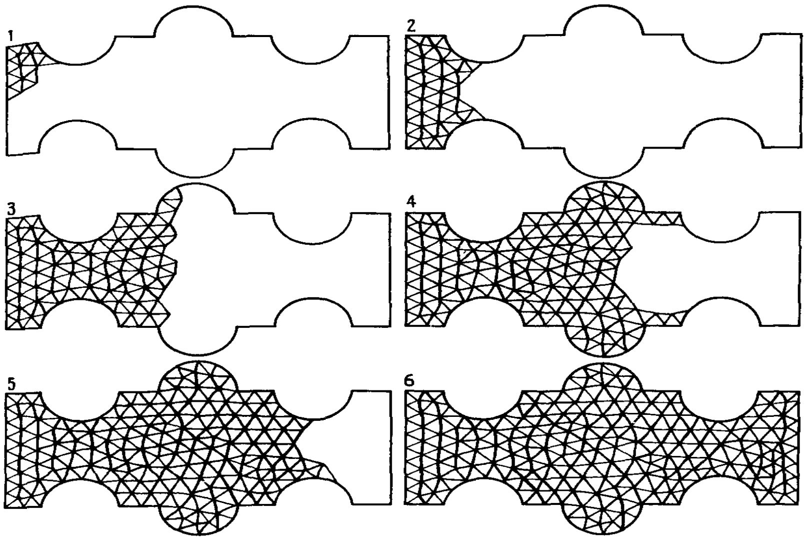


Figure 28: An example of the triangulation with the frontal propagation algorithm.

19.4 Triangulation with the layer covering

This technique is suitable for the three-dimensional domains. The idea of the method is the propagation of the triangulated surface along a line in the three-dimensional space. That's why this method is also called the 2.5 dimensional method.

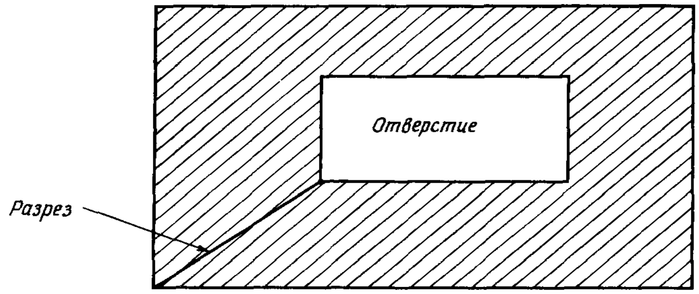


Figure 29: Non-simply connected domain with the cut.

The triangulated surface smoothly changes during this propagation. In order to get the three-dimensional mesh, we add to the new layer the nodes of the moved previous layer, see Fig. 30. As we have two surfaces divided into triangles, we can naturally divide the space between them into the tetrahedrons.

19.5 The global triangulation into triangles and tetrahedrons (the Delone method)

This method is the computational geometry method for the triangulation of the domain and/or for the mesh refining. The method is described as the iterative procedure. Namely:

- We start from the initial triangulation D_0 of the entire domain. It consists of 2 triangles in R^2 and 8 tetrahedrons in R^3.

- The next triangulation D_i is constructed by adding the new node i and combining it with the previous triangulation D_{i-1} in such a way that:
  1. the elements those circumscribed circle (sphere) contains the node i are excluded
  2. the edges (surfaces) are jointed by lines (surfaces) with the node i
  3. the elements which do not have the intersection with the domain anymore are excluded.

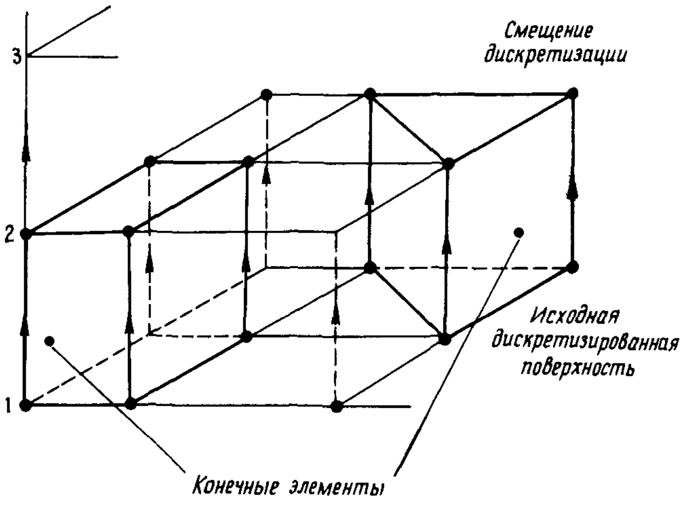


Figure 30: The propagation of the triangulated surface.

It can be proved that the Delone method for the triangulation has the following advantages:

- the triangles and tetrahedrons constructed have "on average" the acceptable shape (i.e. the angles are not very close to zero).
- new nodes can be added during the triangulation process.

For the triangulation of a domain with the Delone method, the following steps are usually used:

- the first set of the nodes is appropriately chosen on the domain boundary.
- the triangulation based on these nodes is constructed.
- the relatively big elements are eliminated from the mesh by adding a node into their barycenters and applying the Delone procedure.
- the previous step is repeated until the prescribed size of all elements is reached.

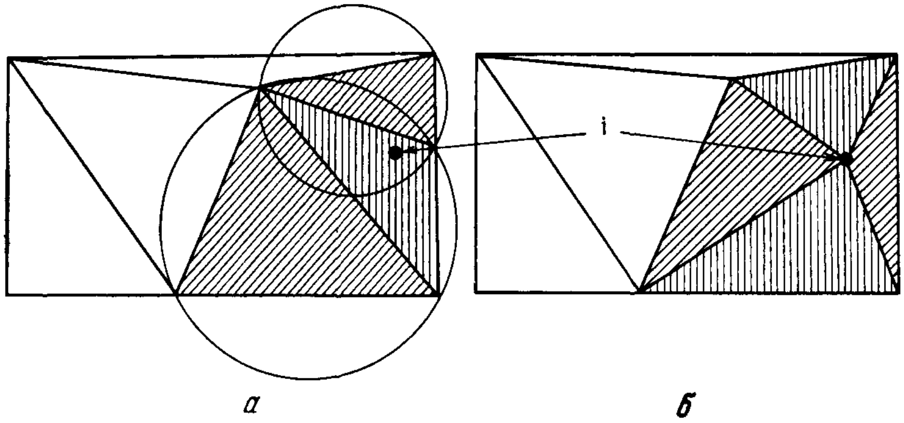


Figure 31: One step of the iterative triangulation with the Delone method.

While the Delone method give good elements on average, some elements in the mesh still may have bad shapes. In order to improve the triangulation quality, we need to optimize the size of the elements and make their shapes better.

The size optimization includes the exclusion of the big elements (by adding nodes as described above) and the exclusion of very small elements. The shape improvement is done through the exclusion of very "plane" elements. One chooses a polygon which circumscribes the undesired element, and moves the internal node into the element barycenter, see Fig. 32. This procedure is clearly ambiguous, so its efficiency strongly depends on the realization details.

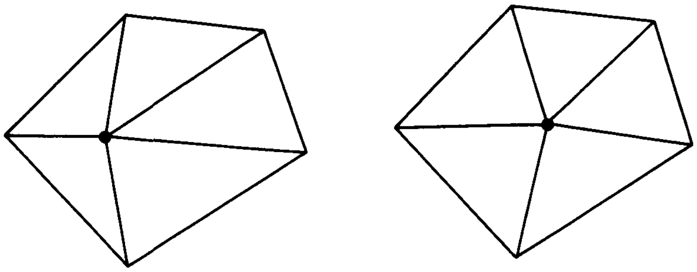


Figure 32: The exclusion of the plane element.

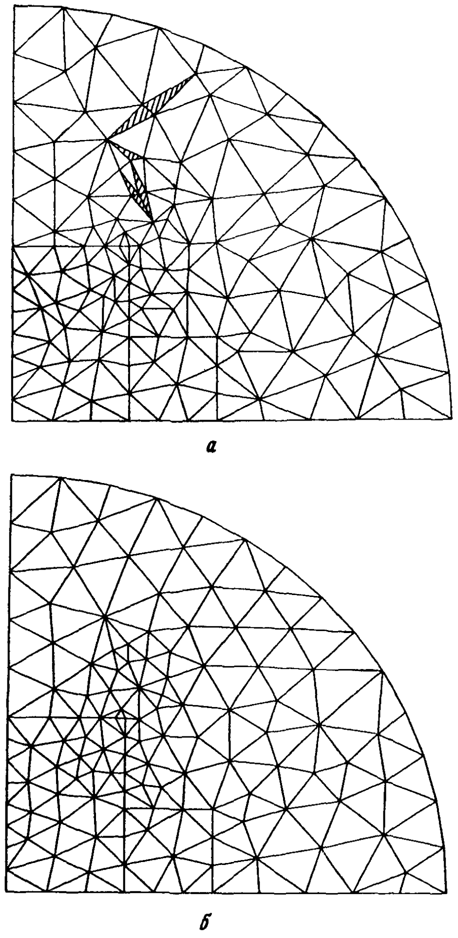


Figure 33: The example of the triangulation improvement.

20 The coordinate transformation

We define the elemental functions on the canonical element as this is a relatively easy procedure. However, in order to solve the variational problem, we need the basis on the physical elements which have various positions, shapes and sizes. Therefore, we need a coordinate transformation from the canonical to the physical coordinates. We can chose different transformations but it is natural to require that these transformations are

- easy to calculate,
- continuity preserving,
- invertible (i.e. the Jacobian is non-singular). For example, for the transformations in the plain we require

$$\det \begin{bmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{bmatrix} \neq 0.$$

As the elemental functions are the polynomials, the natural choice for the coordinate transformations are the piecewise-polynomial functions. Depending on the relations between the degrees of the elemental functions n_{func} and of the transformations n_{trans}, these transformations are called
subparametric if $n_{trans} < n_{func}$,
isoparametric if $n_{trans} = n_{func}$,
and superparametric if $n_{trans} > n_{func}$.

In the previous lectures, we have already used the linear and bilinear transformations. They obviously satisfy the first two requirements, but the non-singularity has to be checked separately. Let us consider here few examples for different types of the coordinate transformations.

Example 1. The bilinear transformation of the square with quadratic elemental functions (the subparametric transformation).

The restriction of the solution on the element is written in terms of the biquadratic functions $N^2_{ij}(\xi, \eta)$

$$U(\xi, \eta) = \sum_{i=1}^{3} \sum_{j=1}^{3} c_{ij} N^2_{ij}(\xi, \eta),$$

where

$$N^2_{ij}(\xi, \eta) = N^2_i(\xi) N^2_j(\eta), \quad i, j = 1, 2, 3,$$

$$N^2_i(\xi) = \begin{cases} -\xi(1-\xi)/2, & i = 1, \\ \xi(1+\xi)/2, & i = 2, \\ 1 - \xi^2, & i = 3. \end{cases}$$

The linear transformation can be written in terms of the bilinear elemental

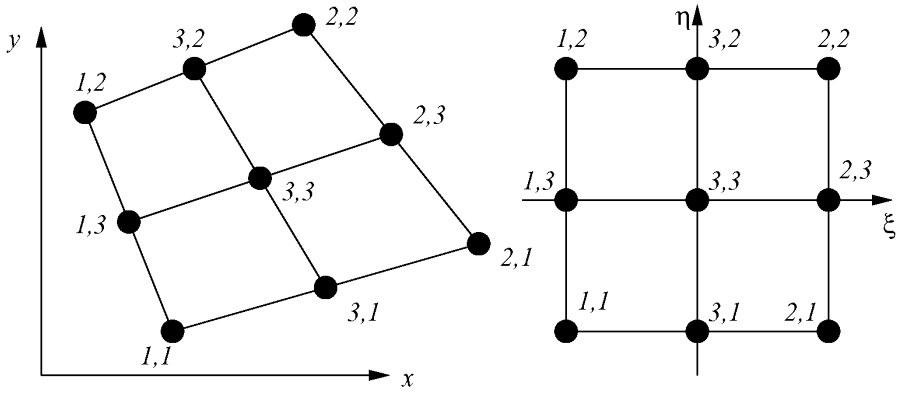


Figure 34: The bilinear transformation of the square with the biquadratic functions.

function as

$$\begin{bmatrix} x(\xi,\eta) \\ y(\xi,\eta) \end{bmatrix} = \sum_{i=1}^{2}\sum_{j=1}^{2} \begin{bmatrix} x_{ij} \\ y_{ij} \end{bmatrix} N_{ij}^{1}(\xi,\eta),$$

where

$$N_{ij}^{1}(\xi,\eta) = N_i^1(\xi)N_j^1(\eta), \quad i,j = 1,2,$$

$$N_i^1(\xi) = \begin{cases} (1-\xi)/2, & i=1, \\ (1+\xi)/2, & i=2. \end{cases}$$

It can be checked that this transformation is non-singular, if all the angles of the physical quadrangle are less than π, i.e. the physical quadrangle is convex.

Example 2. The biquadratic transformation of the square with quadratic elemental functions (the isoparametric transformation).

The basis elemental functions are the same as in example 1. The transformation is written as

$$\begin{bmatrix} x(\xi,\eta) \\ y(\xi,\eta) \end{bmatrix} = \sum_{i=1}^{3}\sum_{j=1}^{3} \begin{bmatrix} x_{ij} \\ y_{ij} \end{bmatrix} N_{ij}^{2}(\xi,\eta).$$

It is important to notice that the physical element is not a quadrangle anymore. Its sides are curved (quadratic) lines.

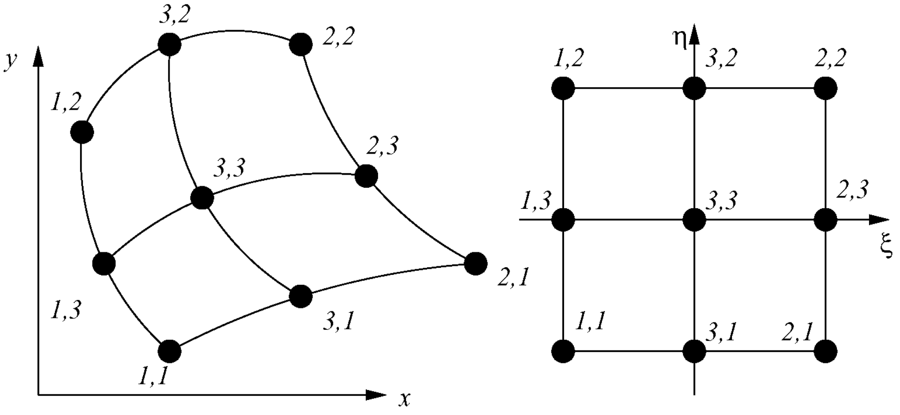


Figure 35: The biquadratic transformation of the square with the biquadratic functions.

Example 3. The quadratic transformation of the triangle with quadratic elemental functions (the isoparametric transformation).

In the general case, the quadratic transformation is written as

$$\begin{bmatrix} x(\xi,\eta) \\ y(\xi,\eta) \end{bmatrix} = \sum_{i=1}^{6} \begin{bmatrix} x_i \\ y_i \end{bmatrix} N_i^2(\xi,\eta).$$

Here the functions $N_i^2(\xi,\eta)$ are defined as

$$N_i^2 = 2\zeta_i(\zeta_i - 1/2), \quad i = 1,2,3,$$

$$N_4^2 = 4\zeta_1\zeta_2, \quad N_5^2 = 4\zeta_2\zeta_3, \quad N_6^2 = 4\zeta_3\zeta_1,$$

and

$$\zeta_1 = 1 - \xi - \eta, \quad \zeta_2 = \xi, \quad \zeta_3 = \eta.$$

Let us consider a specific transformation when the only one side can be curved while other two are straight lines. It means that points 4 and 6 keep their places in the center of the corresponding sides:

$$x_4 = (x_1 + x_2)/2, \quad y_4 = (y_1 + y_2)/2,$$

$$x_6 = (x_1 + x_3)/2, \quad y_6 = (y_1 + y_3)/2.$$

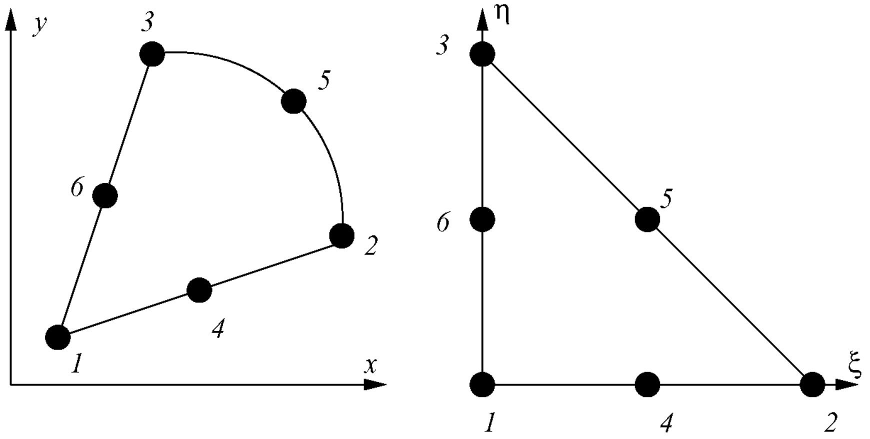


Figure 36: The quadratic transformation of the triangle with two straight lines.

The total transformation can be written as

$$\begin{bmatrix} x(\xi,\eta) \\ y(\xi,\eta) \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} (1-\xi-\eta) + \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \xi(1-2\eta) +$$

$$+ \begin{bmatrix} x_3 \\ y_3 \end{bmatrix} \eta(1-2\xi) + \begin{bmatrix} x_5 \\ y_5 \end{bmatrix} 4\xi\eta. \tag{57}$$

It is clear that the transformations of the sides (1-2) and (1-3) are linear:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} (1-\xi) + \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \xi,$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} (1-\eta) + \begin{bmatrix} x_3 \\ y_3 \end{bmatrix} \eta,$$

respectively. The Jacobians of these transformations are constants and therefore non-singular. However, a singularity may appear at the third (curved) side, and this depends on the position of the (x_5, y_5) point.

Let us make the required transformation in two steps:.

1. The linear transformation to the canonical element with the curved size. The Jacobian determinant for the linear transformation is a constant so it is

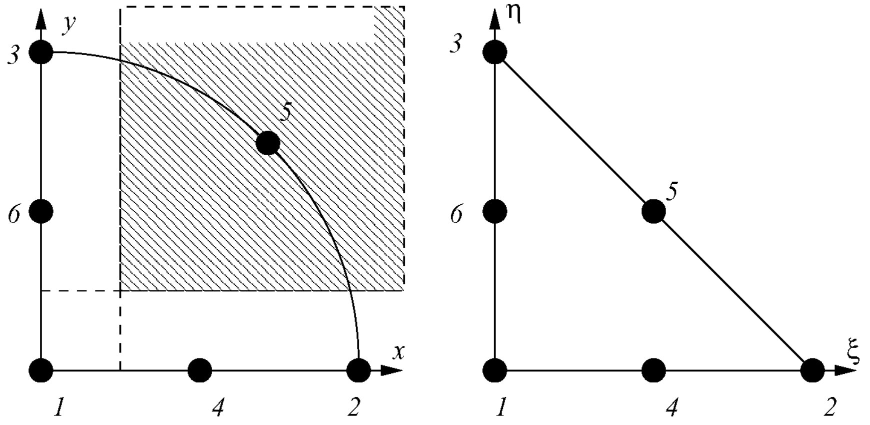


Figure 37: The linearization of the canonical triangle.

always non-singular.
2. Make the curved size straight.

Substituting the values $(x_1, y_1) = (0, 0)$, $(x_2, y_2) = (1, 0)$, $(x_3, y_3) = (0, 1)$ into the general transformation (57), we have

$$\begin{bmatrix} x(\xi,\eta) \\ y(\xi,\eta) \end{bmatrix} = \begin{bmatrix} \xi(1-2\eta) \\ \eta(1-2\xi) \end{bmatrix} + 4\xi\eta \begin{bmatrix} x_5 \\ y_5 \end{bmatrix}.$$

The Jacobian matrix $\mathbf{J}(\xi,\eta)$ is calculated as

$$\begin{bmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{bmatrix} = \begin{bmatrix} 1-2\eta+4x_5\eta & -2\xi+4x_5\xi \\ -2\eta+4y_5\eta & 1-2\xi+4y_5\xi \end{bmatrix},$$

and its determinant is equal to

$$\det \mathbf{J}(\xi,\eta) = 1 + (4x_5 - 2)\eta + (4y_5 - 2)\xi.$$

As the Jacobian determinant is the linear function of ξ and η, we can check the singularity of the transformation on the vertices only:

$$\det \mathbf{J}(0,0) = 1, \quad \det \mathbf{J}(1,0) = 4y_5 - 1, \quad \det \mathbf{J}(0,1) = 4x_5 - 1.$$

The transformation is non-singular (i.e. the determinant is not equal to zero) iff the position of point (5) satisfies the following restrictions:

$$x_5 > 1/4, \quad y_5 > 1/4. \tag{58}$$

So we can see that the curved triangle can be both convex and concave but it cannot be too concave as it should satisfy condition (58).

Lecture 8

Numerical integration in the FEM. The discretization and perturbation errors.

21 The numerical integration in the FEM

As we have already studied, in order to find the matrix elements of the local and global matrices we should calculate the quadratic forms on the elemental functions:

$$I = \int_{\Omega_e} A(U_i(\mathbf{x}), U_j(\mathbf{x}))\, d\mathbf{x}.$$

Let us discuss this calculation for the two-dimensional FEM. Making use of the coordinate transformations to the canonical element, we find for the integral the following representation:

$$I = \int\!\!\int\limits_{\Omega_0} \alpha(\xi,\eta) N_s(\xi) N_t^\top(\eta) \det{(\mathbf{J_e})} d\xi d\eta.$$

Here, the derivative is taken with respect to one of coordinates $s, t \in \{\emptyset, \xi, \eta\}$, and

$$N^\top = [N_1, N_2, ...N_{n_p}].$$

Therefore, the integral is the matrix whose dimension is equal to the number of elemental functions n_p:

$$I = \int\!\!\int\limits_{\Omega_0} \begin{bmatrix} (N_1)_s(N_1)_t & (N_1)_s(N_2)_t & \\ (N_2)_s(N_1)_t & (N_2)_s(N_2)_t & \\ & & \ddots \end{bmatrix} \alpha(\xi,\eta) \det{(\mathbf{J_e})} d\xi d\eta. \qquad (59)$$

In some special cases, integral (59) can be computed analytically. In the general case, however, we should calculate it numerically.

In the framework of the FEM, we can stress the following properties of the numerical approach:

- The numerical integration gives *exact* results for the simple cases (i.e. simple coordinate transformations and a simple function $\alpha(\xi,\eta)$). In many applications, these functions as well as elemental functions are low degree polynomials, so it is easy to find a numerical integration formula which produces the exact result.

- In order to reach the best convergence speed, the exact calculation of integrals is *not necessary*, see below.

Before the discussion of the details of the numerical integration in the FEM, we recall the basic facts about the quadrature formulas. The quadrature formula is a finite sum which gives an approximate value of the integral:

$$I = \int\int_{\Omega_0} f(\xi, \eta) d\xi d\eta \approx \sum_{i=1}^{n} W_i f(\xi_i, \eta_i).$$

It is said that the quadrature formula is the formula of the order q, if for an arbitrary $q+1$ times differentiable function $f(\xi, \eta) \in H^{q+1}(\Omega_0)$, the following estimation is satisfied:

$$|I - \sum_{i=1}^{n} W_i f(\xi_i, \eta_i)| \leq C||f^{(q+1)}||.$$

In the one-dimensional case,

$$I = \int_{-1}^{1} f(\xi) d\xi \approx \sum_{i=1}^{n} W_i f(\xi_i),$$

the most useful quadrature formulas are the Newton-Cotes formulas of the n order, and the Gauss-Legendre formulas of the $2n - 1$ order. The error estimations and the examples of the higher order formulas can be found in the standard books on the numerical methods.

21.1 The quadrature formulas for the square

We start the discussion about the multidimensional case with a simpler example of the canonical square. For this case, the simplest quadrature formula

n	$\pm\xi_i$	W_i
1	0	2
2	$1/\sqrt{3}$	1
3	0	8/9
	$\sqrt{3/5}$	5/9

Table 1: The nodes and weights for the Gauss-Legendre formulas of the low orders.

is given as the tensor product of the one-dimensional quadrature formulas:

$$I = \int_{-1}^{1}\int_{-1}^{1} f(\xi,\eta)d\xi d\eta \approx \int_{-1}^{1}\sum_{i=1}^{n} W_i f(\xi_i,\eta)d\eta =$$

$$= \sum_{i=1}^{n} W_i \int_{-1}^{1} f(\xi_i,\eta)d\eta \approx \sum_{i=1}^{n}\sum_{j=1}^{m} W_i\tilde{W}_j f(\xi_i,\eta_j).$$

It is worth mentioning that the quadrature formulas constructed in this way

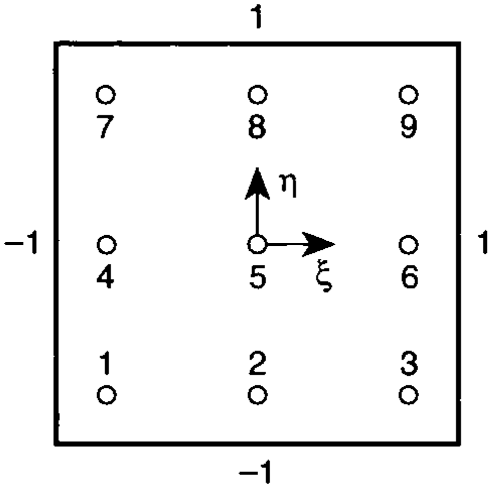


Figure 38: The integration nodes for $n = 3$ in the square. The formula is exact for the polynomials of the fifth degree.

are *not optimal.* They contain more nodes than minimally possible amount.

The difference becomes especially pronounced for the higher order elements in the three-dimensional space.

		R^2		R^3	
n	Order	N_{tens}	N_{opt}	N_{tens}	N_{opt}
1	1	1	1	1	1
2	3	4	4	8	5
3	5	9	7	27	14
4	7	16	12	64	30

Table 2: The number of nodes n for the one-dimensional formulas, the formula orders and the node number for the tensor-type N_{tens} and optimal N_{opt} formulas.

21.2 The quadrature formulas for the triangles

For the low order quadrature formulas, it is possible to derive them with the undetermined coefficient approach. We assume that the formula can be written in the following way

$$\iint\limits_{\Omega_0} f(\xi,\eta)d\xi d\eta = W_1 f(\xi_1,\eta_1) + E.$$

As the formula contains three coefficient, it can be exact for all linear functions. So we substitute the following functions

$$f(\xi,\eta) = [1,\xi,\eta]^\top,$$

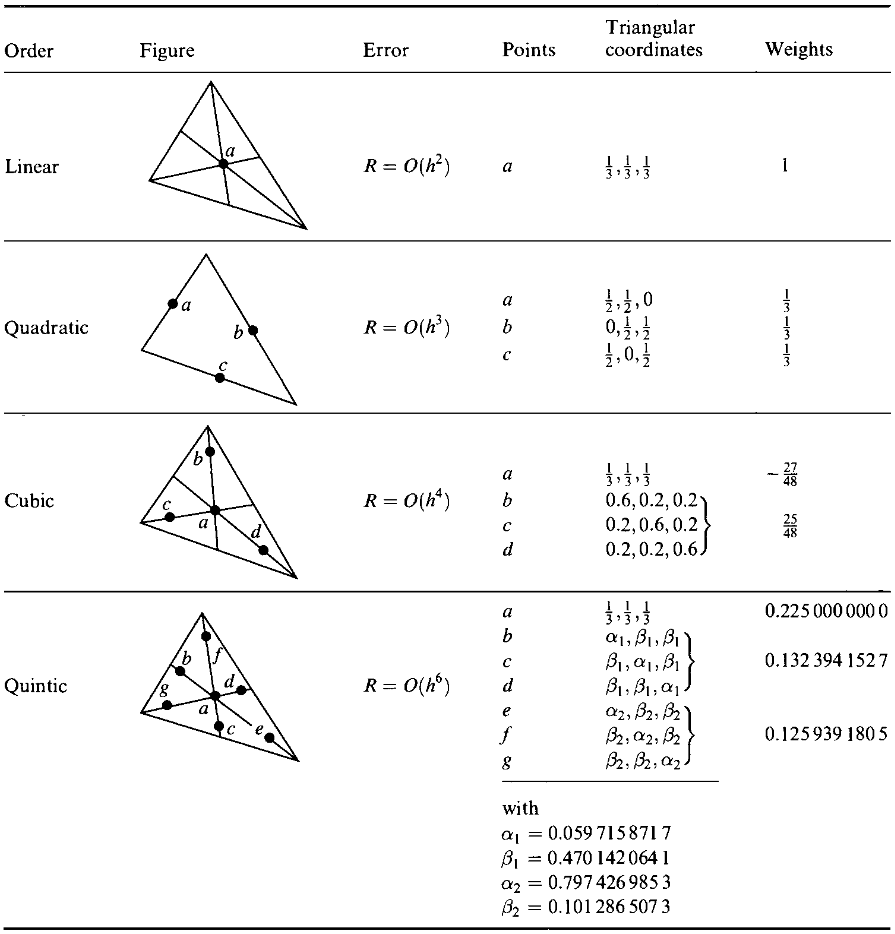
and calculate the integrals

$$\int\limits_0^1 \int\limits_0^{1-\xi} \begin{bmatrix} 1 \\ \xi \\ \eta \end{bmatrix} d\eta d\xi = \begin{bmatrix} 1/2 \\ 1/6 \\ 1/6 \end{bmatrix} = \begin{bmatrix} W_1 \\ W_1\xi_1 \\ W_1\eta_1 \end{bmatrix}.$$

We can see now that the solution exists as we are able to find all undetermined coefficients: $W_1 = 1/2$, $\xi_1 = \eta_1 = 1/3$. The corresponding quadrature formula is written as

$$\iint\limits_{\Omega_0} f(\xi,\eta)d\xi d\eta = \frac{1}{2} f(1/3,1/3) + O(f'').$$

The quadrature formulas of the higher order are presented on figure (39) for the triangles and on figure (40) for the tetrahedrons.



Order	Figure	Error	Points	Triangular coordinates	Weights
Linear		$R = O(h^2)$	a	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	1
Quadratic		$R = O(h^3)$	a	$\frac{1}{2}, \frac{1}{2}, 0$	$\frac{1}{3}$
			b	$0, \frac{1}{2}, \frac{1}{2}$	$\frac{1}{3}$
			c	$\frac{1}{2}, 0, \frac{1}{2}$	$\frac{1}{3}$
Cubic		$R = O(h^4)$	a	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$-\frac{27}{48}$
			b	$0.6, 0.2, 0.2$	$\frac{25}{48}$
			c	$0.2, 0.6, 0.2$	
			d	$0.2, 0.2, 0.6$	
Quintic		$R = O(h^6)$	a	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	0.225 000 000 0
			b	$\alpha_1, \beta_1, \beta_1$	0.132 394 152 7
			c	$\beta_1, \alpha_1, \beta_1$	
			d	$\beta_1, \beta_1, \alpha_1$	
			e	$\alpha_2, \beta_2, \beta_2$	0.125 939 180 5
			f	$\beta_2, \alpha_2, \beta_2$	
			g	$\beta_2, \beta_2, \alpha_2$	
			with		
			$\alpha_1 = 0.059\,715\,871\,7$		
			$\beta_1 = 0.470\,142\,064\,1$		
			$\alpha_2 = 0.797\,426\,985\,3$		
			$\beta_2 = 0.101\,286\,507\,3$		

Figure 39: The quadrature formulas for the triangles.

22 The discretization and perturbation errors

We have already found that for the interpolation errors, the following theorem is satisfied:

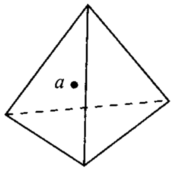
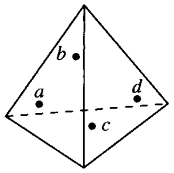
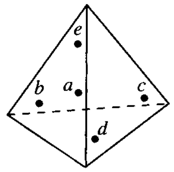
No.	Order	Figure	Error	Points	Tetrahedral coordinates	Weights
1	Linear		$R = O(h^2)$	a	$\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$	1
2	Quadratic		$R = O(h^3)$	a b c d	$\alpha, \beta, \beta, \beta$ $\beta, \alpha, \beta, \beta$ $\beta, \beta, \alpha, \beta$ $\beta, \beta, \beta, \alpha$ $\alpha = 0.585\,410\,20$ $\beta = 0.138\,196\,60$	$\frac{1}{4}$
3	Cubic		$R = O(h^4)$	a b c d e	$\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$ $\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$ $\frac{1}{6}, \frac{1}{2}, \frac{1}{6}, \frac{1}{6}$ $\frac{1}{6}, \frac{1}{6}, \frac{1}{2}, \frac{1}{6}$ $\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{2}$	$-\frac{4}{5}$ $\frac{9}{20}$

Figure 40: The quadrature formulas for the tetrahedrons.

Let Ω be divided into elements Ω_e from the regular family. Let h be the longest side in the mesh. For the interpolation of the order p there exists a constant $C > 0$ (which is independent on $u \in H^{p+1}(\Omega)$ and the mesh) such that

$$|u - U|_s \leq Ch^{p+1-s}|u|_{p+1}, \quad s = 0, 1. \tag{60}$$

The convergence (60) is getting worse when the solution u is not smooth enough, see Lecture 6.

However, error estimation (60) does not take into account all possible errors in the FEM. Namely, it does not account for

- numerical integration errors,
- boundary condition approximation errors,
- boundary triangulation errors.

Let us discuss now the results which describe these kinds of errors.

22.1 The numerical integration errors

We start from the initial variational problem

$$A(u,v) = (f,v) \quad \forall v \in H,$$

and apply the Galerkin method to this problem assuming the exact calculation of the integrals:

$$A(U,V) = (f,V) \quad \forall V \in H^N$$

Now we will account for the approximate integration. This means that both the bilinear and linear forms are changed:

$$A_*(U^*,V) = (f,V)_* \quad \forall V \in H^N$$

in the way we discussed above:

$$(f,V)_* = \sum_{e=1}^{N_\Delta} (f,V)_{e,*} = \sum_{e=1}^{N_\Delta} \sum_{k=1}^{n} W_k V(x_k,y_k) f(x_k,y_k).$$

For this problem, the following results can be proved.

Theorem 1. Let $A(u,v)$ and $A_*(U,V)$ be the bilinear forms, the form A be the continuous form, and A_* be positively defined, i.e. there exist two constants α and β such that

$$|A(u,v)| \leq \alpha ||u||_1 ||v||_1, \quad \forall u,v \in H,$$

$$A_*(U,U) \geq \beta ||U||_1^2, \quad \forall U \in H^N.$$

Then

$$||u - U^*||_1 \leq C \left\{ ||u - V||_1 + \right.$$
$$\left. + \sup_{W \in H^N} \frac{|A(W,V) - A_*(W,V)|}{||W||_1} + \sup_{W \in H^N} \frac{|(f,W) - (f,W)_*|}{||W||_1} \right\}, \quad \forall V \in H^N.$$

One can see that this theorem gives us a possibility to separate two types of errors: the interpolation errors with the exact integration described by

Eq.(60), and the errors erasing from the numerical integration. The numerical integration errors themselves are covered by the next theorem.

Theorem 2. Let $\mathbf{J}(\xi,\eta)$ *be the Jacobian of the transformation from the* (ξ,η) *into* (x,y) *plane. Let* Δ_h *be the regular FEM family. Let* $\det(\mathbf{J}(\xi,\eta))W_x(\xi,\eta)$ *and* $\det(\mathbf{J}(\xi,\eta))W_y(\xi,\eta)$ *be the piecewise polynomials of the degree less or equal to* r_1*, and* $\det(\mathbf{J}(\xi,\eta))W(\xi,\eta)$ *be the piecewise polynomials of the degree less or equal to* r_0*. Then:*
1. If the quadrature formula in (ξ,η) *is exact for polynomials of the* r_1+r *degree, then*

$$\frac{|A(W,V)-A_*(W,V)|}{||W||_1} \leq Ch^{r+1}||V||_{r+2}, \quad \forall V,W \in H^N.$$

2. If the quadrature formula in (ξ,η) *is exact for polynomials of the* r_1+r-1 *degree, then*

$$\frac{|(f,W)-(f,W)_*|}{||W||_1} \leq Ch^{r+1}||f||_{r+1}, \quad \forall W \in H^N.$$

Here we have the estimation of the numerical integration inaccuracy for both bilinear and linear forms. Let us consider two examples of the application of theorem 2.

Example 1. Let the coordinate transformation be the linear one. Then the Jacobian determinant $\det(\mathbf{J}(\xi,\eta))$ is a constant. Let also the space H^N consist of the piecewise polynomials of the degree p (i.e. we consider the FEM of the order p with the linear coordinate transformation). For this case, we can see that $r_1 = p-1$, $r_0 = p$. According to the interpolation theorem, we find

$$||u-V||_1 = O(h^p)$$

Let now assume that we use the quadrature formula of the order ρ. Then $\rho = r_1 + r$ and $\rho = r_0 + r - 1$, and we find

$$r = \rho - p + 1.$$

Applying Theorem 2, we find:

$$\frac{|A(W,V)-A_*(W,V)|}{||W||_1} \leq Ch^{\rho-p+2}||V||_{\rho-p+3}.$$

$$\frac{|(f,W)-(f,W)_*|}{||W||_1} \leq Ch^{\rho-p+2}||f||_{\rho-p+2}.$$

Depending on the values of p and ρ, we can have the different cases:
1. If $\rho = 2(p-1)$ then $r = p-1$. All errors have the same order with respect to h:

$$||u - U^*||_1 = O(h^p).$$

This situation is optimal, as we acquire the best convergence speed possible, but we do not spend excessive effort for the numerical integration.
2. If $\rho > 2(p-1)$ then $r > p-1$. The interpolation error is the main error:

$$||u - U^*||_1 = O(h^p).$$

The integration error has a higher order with respect to h and does not contribute to the last estimation. Again, we have the best convergence speed possible, but the additional numerical integration effort may be meaningless.
3. If $\rho < 2(p-1)$ then $r < p-1$. The integration error is the main error:

$$||u - U^*||_1 = O(h^{\rho-p+2}).$$

In this case, we do reach the convergence speed which is possible for the chosen FEM. Even worse, if $\rho \leq p-2$ (i.e. we use the low order quadrature formula for the higher order polynomials), the FEM approximation does not converge to the exact result when $h \to 0$. ■

Example 2. Let us consider the isoparametric elements. For them, the degree of the coordinate transformations coincides with the order of the elemental functions. Then one can check that the optimal integration order ρ is found as

$$\rho = 4(p-1).$$

■

22.2 The boundary condition approximation errors

Here we just present the basic result for this kind of errors. Namely, if the integration is exact, and there are no boundary triangulation errors, then for the degree p polynomials one can find that

$$||u - U||_1 \leq \left\{ h^p ||u||_{p+1} + h^{p+1/2} ||u||_{p+1} \right\}, \tag{61}$$

for the solution $u \in H^{p+1}(\Omega)$. If the boundary of the domain Ω is not smooth (e.g. it contains edges), then the solution may not belong to the space $H^{p+1}(\Omega)$, and estimation (61) is not valid.

22.3 The boundary triangulation errors

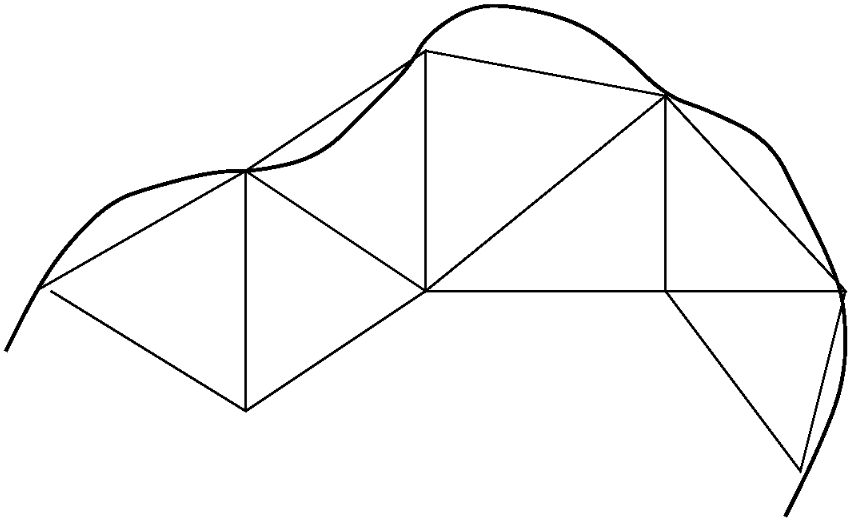


Figure 41: The approximation of the domain boundary with polynomials.

The specific problem for this type of errors is due to the fact that the space of approximating functions do not necessarily belong to the full initial functional space. This requires a special consideration of the problem.

A few results are derived for this type of errors. For example, for the linear approximation

$$||u - U||_1 = O(h),$$

while for the quadratic one

$$||u - U||_1 = O(h^{3/2}).$$

The detailed analysis shows that errors of this type are localized in a vicinity $\partial\Omega$, while outside that vicinity they are not influenced by the boundary approximation.

Another known result is the following one: for the approximation with the degree p polynomials,

$$||u - U||_1 = O(h^p),$$

if the distance between the exact $\partial\Omega$ and approximated $\partial\tilde{\Omega}$ boundaries is proportional to h^{p+1}. Therefore, in order to reach the optimal convergence, we should approximate the boundary by the degree p polynomials. This means, that the optimal convergence near the domain boundary can be achieved with the isoparametric finite elements.

Lecture 9

A priori and a posteriori error estimations. Superconvergence. Adaptive solution refinement. h-, p-, and hp-refinement.

23 A priori and a posteriori error estimations

For the error estimations in all branches of the numerical analysis including the FEM there widely use two types of methods: *a priori* and *a posteriori* methods. According to their names, a priori methods can give some information about the accuracy of the solution prior to the solution itself is actually computed. A posteriori methods require a solution (or, typically, few solutions) to be calculated, then they can measure its accuracy.

In the frame of the FEM, a priori error estimations are well-known. In the most regular cases, they can be described as

$$|u - U|_s \leq Ch^{p+1-s}|u|_{p+1}, \tag{62}$$

$$s = 0, 1, \quad u \in H^{p+1}(\Omega),$$

where h is the maximal finite element volume. In Eq.(62), the approximation U for the exact solution u is constructed in the domain Ω, and the error estimations are given for the solution and its derivatives. The advantage of these estimations is that we know them before any solution is constructed. On the other hand, we do not know as a rule the value of the constant C in the r.h.s. and the norm of the exact solution, and, therefore, cannot estimate the error quantitatively.

As we have already mentioned, a posteriori estimations employ already known approximate solutions in order to estimate errors quantitatively. One can find different a posteriori estimators, so it is worthwhile to formulate the requirements to get useful estimators. As a rule, the applicable estimator must

- give accurate error estimation for arbitrary meshes and polynomial degrees,

- be computationally effective
- be able to work with different norms of the solution.

The main types of a posteriori estimations are based on

- the extrapolation of the solutions,
- the calculation of the solution residuals,
- the solution recovery.

Let us analyze these estimators.

23.1 Estimators based on the solution extrapolation

Space extrapolation

Let us construct two approximate solutions $U_h^p(x)$ and $U_{h/2}^p(x)$, corresponding to the finite element volumes h and $h/2$. Then we assume that the error estimation in Eq.(62) is in fact exact:

$$u(x) - U_h^p(x) = C_{p+1}h^{p+1} + O(h^{p+2}) \tag{63}$$

$$u(x) - U_{h/2}^p(x) = C_{p+1}\left(\frac{h}{2}\right)^{p+1} + O(h^{p+2})$$

Substracting last equations from each other, we get

$$U_{h/2}^p(x) - U_h^p(x) = C_{p+1}h^{p+1}(1 - \frac{1}{2^p+1}) + O(h^{p+2}).$$

Ignoring the higher order terms with respect to h, we find the error estimation

$$u(x) - U_h^p(x) \approx \frac{U_{h/2}^p(x) - U_h^p(x)}{1 - 1/2^{p+1}}.$$

So we get the approximate expression for the solution $U_h^p(x)$ inaccuracy in terms of the solution itself and another solution calculated on the twice finer mesh. This approach to the error estimation construction is called Richardson's extrapolation, or h-extrapolation.

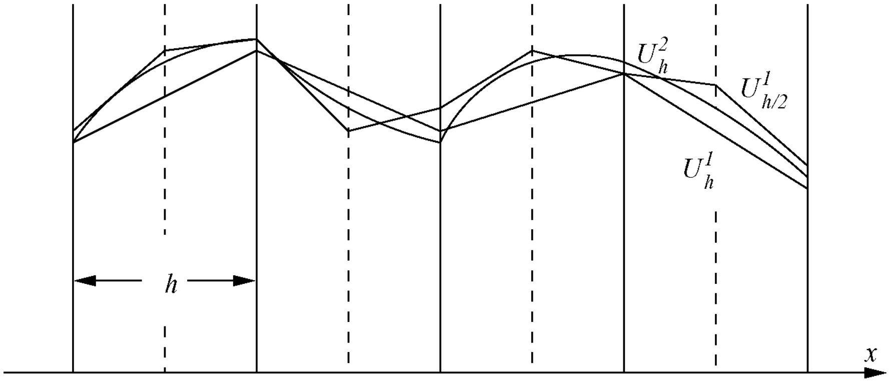


Figure 42: The solutions for the linear elements U_h^1 and $U_{h/2}^1$, and for the quadratic elements U_h^2.

Extrapolation on the polynomial degree

Let us supplement Eq.(62) with the equation for the solution error $U_h^{p+1}(x)$ corresponding to the same element volume but the higher polynomial degree:

$$u(x) - U_h^{p+1}(x) = C_{p+2}h^{p+2} + O(h^{p+3}).$$

Then adding to and substracting from Eq.(62) the solution U_h^{p+1}, we get

$$u(x) - U_h^p(x) = \left[u(x) - U_h^{p+1}(x)\right] + \left[U_h^{p+1}(x) - U_h^p(x)\right] =$$

$$= \left[U_h^{p+1}(x) - U_h^p(x)\right] + O(h^{p+2}).$$

Neglecting the higher terms with respect to h, we can calculate the inaccuracy as

$$u(x) - U_h^p(x) \approx \left[U_h^{p+1}(x) - U_h^p(x)\right].$$

This approach to the error estimation is called p-extrapolation.

23.2 Estimations based on the solution residuals

In this approach, we construct the approximate solution U, and then use it for the calculation of the local error. The general idea is the following: we use the boundary conditions extracted from the approximate solution for each finite element, and solve our problem for each finite element *separately* with higher accuracy. As the big problem is split into many small independent problems, the computational cost of these auxiliary problems is low. The differences between the solution U and calculated local solutions are called *residuals*. The residuals give us estimations for the inaccuracies on each element. One can get a more accurate solution on the element in two ways: splitting the element into number of subelements, or increasing the polynomial degree on the same element. Both ways can be efficiently employed, but we discuss here the second approach in more details.

Let us look for the solution U of the variational problem in the domain in the space H^p of the degree p polynomials:

$$A(U,V) = (f,V), \quad \forall V.$$

For the exact solution u, on each element it is satisfied

$$A_e(u,v) = (f,v)_e, \quad \forall v \in H(\Omega_e),$$

and the residual e is defined as

$$u = U + e.$$

Combining these relations, we can derive the equation for the residual:

$$A_e(e,v) = (f,v)_e - A_e(U,v).$$

In order to calculate the residual, we substitute e and v with their finite-dimensional approximations E, V in an appropriate space:

$$A_e(E,V) = (f,V)_e - A_e(U,V) \quad \forall V \in H^L(\Omega_e) \tag{64}$$

When the space H^L contains the polynomials of the same degrees as U, then $H^L = H^p$ and therefore $E = 0$. Usually one chooses $H^L = H^{p+1}$, and then we can approximately assume $E \approx e$. In order to calculate the residuals over the entire domain, we should solve equation (64) for all elements. As these equations are mutually independent (in contrast to the initial equation!), the computational cost of these solutions is low.

Before we discuss the approaches based on the solution recovery, we analyze an important property of the FEM, namely the idea of superconvergency.

24 Superconvergency in the FEM

In order to illustrate this idea, we consider a simple model, namely the 2nd order one-dimensional differential equation:

$$-u''(x) + \beta u(x) + q = 0.$$

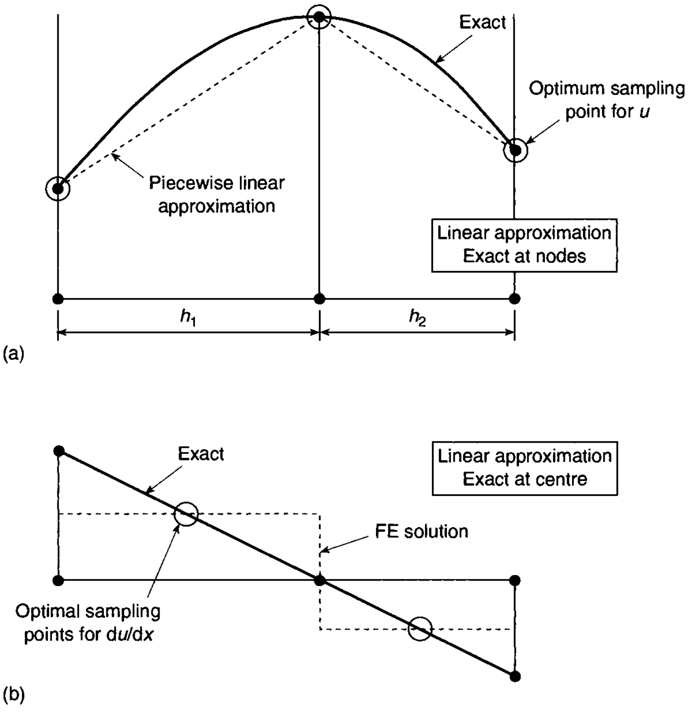

Exact
Optimum sampling point for u
Piecewise linear approximation
Linear approximation Exact at nodes
h_1
h_2
(a)
Exact
Linear approximation Exact at centre
FE solution
Optimal sampling points for du/dx
(b)

Figure 43: The optimal points for the functions and derivatives, linear elements.

Let us construct its solution with the linear and quadratic finite elements, and plot the exact and approximate solutions on Figs. 43,44.

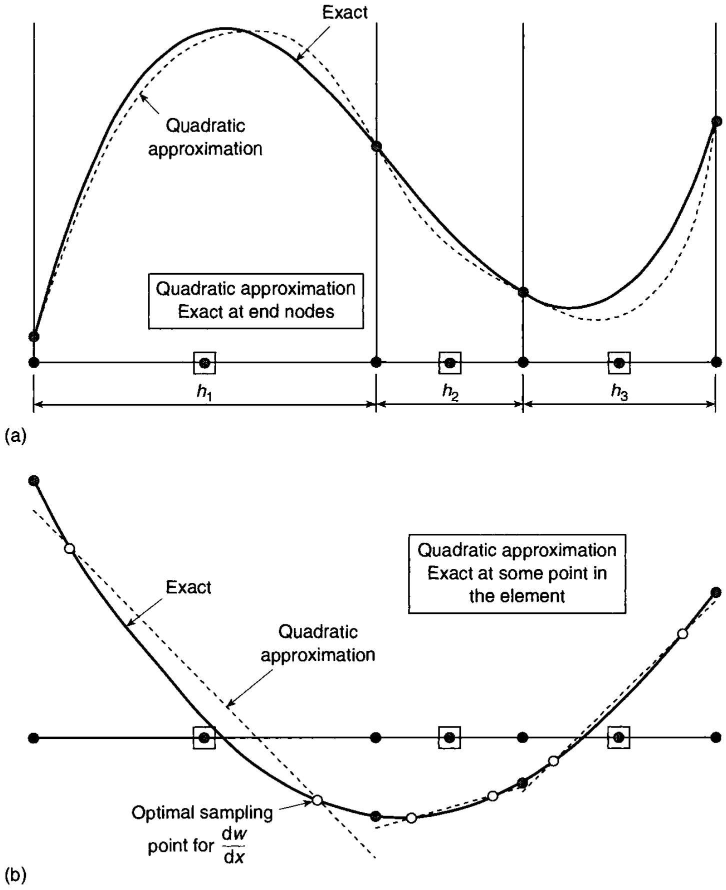


Figure 44: The optimal points for the functions and derivatives, quadratic elements.

One can notice that there exist special points where the accuracy is much better than at an arbitrary point of the interval. Those points exist both for the function and the derivative, and these points are different for the function and the derivative. The solution errors are considerably lower at the nodes of the elements, while the solution derivative errors vanish at some internal points. Similar observations led to the introduction of the *superconvergency*, i.e. the convergency which is faster than that guaranteed by the general interpolation theorem. However, there is no contradiction here: the general theorem assures the convergence rate for *all* points of the interval while the superconvergence is present for some *specific* points only.

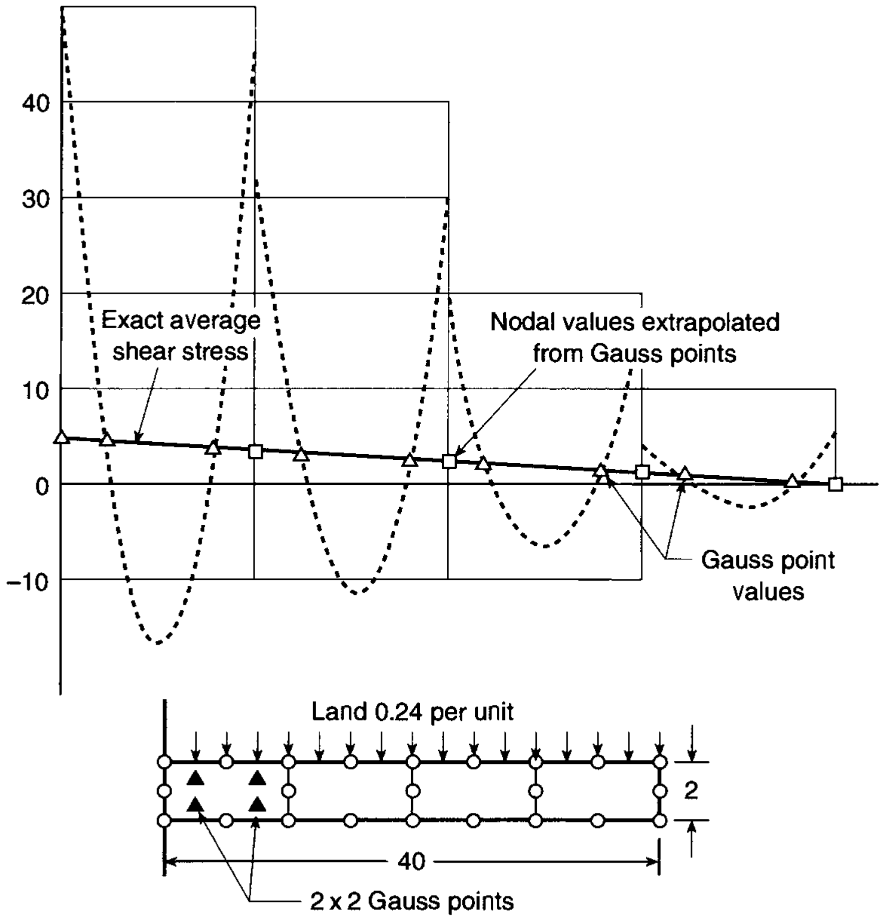


Figure 45: The solution of the two-dimensional problem with four biquadratic elements.

By a careful analysis of the polynomial approximation, one can prove the following statements:

- at the superconvergent points, we gain one additional convergency order both for the function and its derivative,
- for the one-dimensional case, the optimal convergence points for the functions are the element boundaries, and those for the derivatives are

the nodes of the Gauss-Legendre quadrature formulas of the appropriate order.

- for the multidimensional case, the situation differs for the rectangles and triangles. For the rectangles, the superconvergent points exist and are found as the direct product of the one-dimensional points. For the triangles, there exist *optimal* points but they are not superconvergent!

Now we discuss how the superconvergent properties can be used for the construction of the error estimations based on the solution recovery. Let us consider the two-dimensional problem, and restore the solution U^* with the least square method:

$$U^* = \mathbf{pa} = [1, x, y, ...y^p]\mathbf{a}, \quad \mathbf{a} = [a_1, a_2, ...a_m]^\top.$$

We are looking for the coefficient set $\mathbf{a}$, which can be found with the minimization of the sum of the difference squares at the superconvergent points (x_k, y_k):

$$\Pi = \sum_{k=1}^{n} \left(U(x_k, y_k) - \mathbf{p}_k\mathbf{a}\right)^2, \quad \mathbf{p}_k = \mathbf{p}(x_k, y_k).$$

The solution $\mathbf{a}$ is found with use of the standard least square formulas:

$$\mathbf{a} = A^{-1}\mathbf{b},$$

where

$$A = \mathbf{p}_k^\top \mathbf{p}_k, \quad \mathbf{b} = \sum_{k=1}^{n} \mathbf{p}_k^\top U(x_k, y_k).$$

Knowing the coefficients $\mathbf{a}$, we can restore the solution U^* at an arbitrary point. It is worth mentioning that the number of points n cannot be less than the number of polynomials in $\mathbf{p}$, therefore, in our calculations we will always use the values of the solution in the neighboring elements. This means, that this approach is nonlocal, in contrast to the methods based on the residuals.

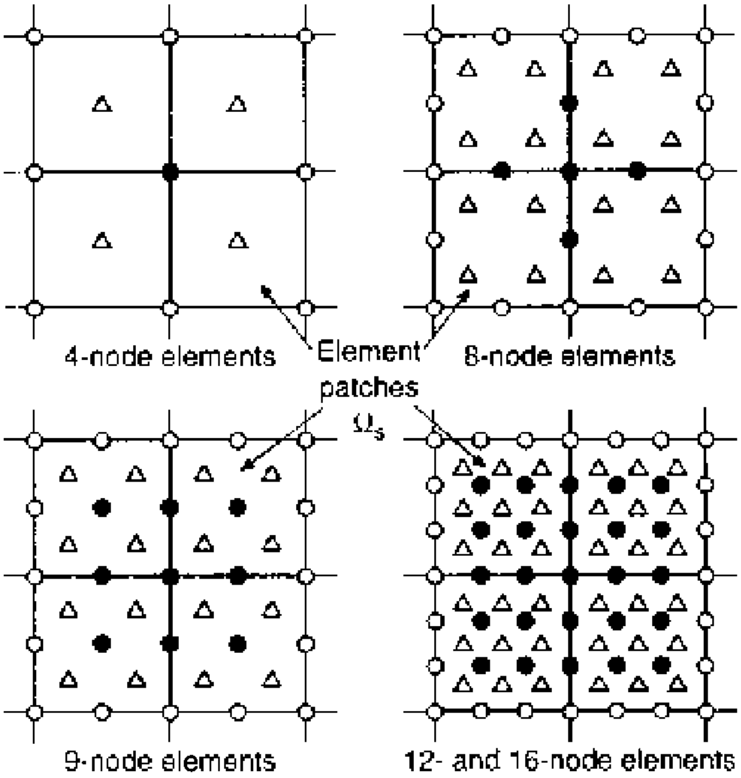


Figure 46: The solution recovery based on superconvergent points.

Neglecting the difference between the exact solution and the better approximation, we can estimate the inaccuracy as

$$|u - U| \approx |U^* - U|.$$

It is important that the superconvergency can also be used for other purposes, for example for the solution refinement. The derivative refinement is especially attractive as then the derivative accuracy will be the same as the solution accuracy.

25 The adaptive solution refinement. h-, p-, and hp-refinement

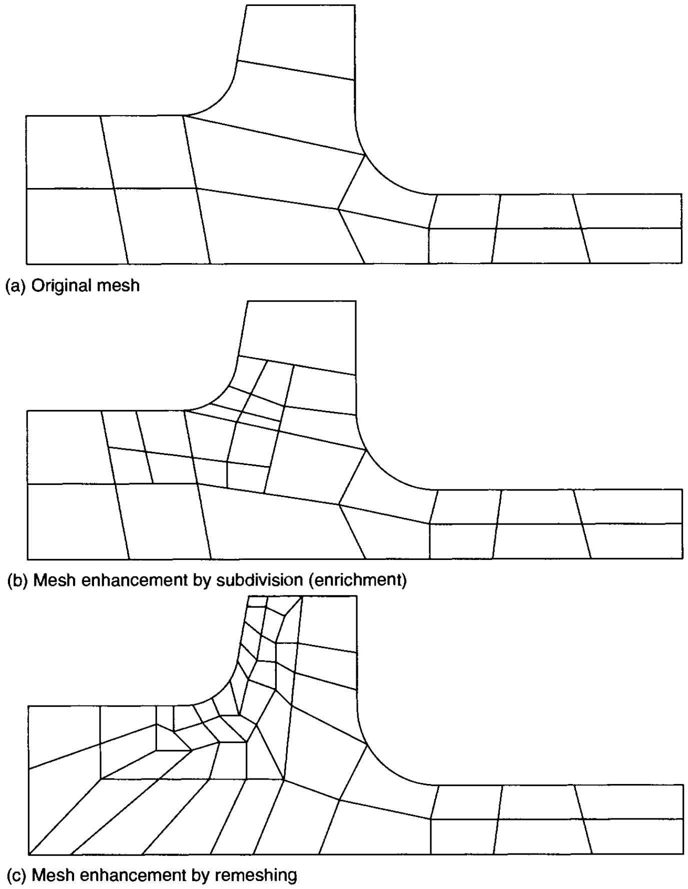
The construction of the FEM solution of the variational problem on one chosen mesh is, as a rule, not enough to get the solution of the initial physical problem. We should check the accuracy of the solution, and we are not guaranteed that the accuracy is satisfactory. Therefore, the way of the typical solution of the physical problem with the FEM (and equally with any other numerical approach) consists of the following steps:

- the choice of the mesh and elemental functions (i.e. h and p are fixed),
- the solution construction on the chosen mesh,
- the solution analysis and the error calculation,
- changing of the mesh and/or elemental functions.

These sequence can be repeated few times when necessary.

The appropriate choice of the mesh and elemental functions can be done in different ways. There exist two main types of the solution refinement:
1. h-refinement:
The type of the finite elements is fixed. The size of the elements can change: it can be decreased in some places while increased in other places.
2. p-refinement:
The size of the elements is fixed, the polynomial degree can be increased (or decreased) on specific elements. As a rule, the hierarchical elements are used.

The types of h-refinement:
1. The element subdivision (enrichment).
Elements with the big local error are subdivided into subelements while elements with small error can be enriched with neighboring elements. Here there might be two kind of problems preventing us from using this approach:
– dangling nodes,
– the complicated structure of the unified elements.
2. The total remeshing.



(a) Original mesh

(b) Mesh enhancement by subdivision (enrichment)

(c) Mesh enhancement by remeshing

Figure 47: Different types of h-refinement.

We calculate the local errors for the old mesh, predict the new element size based on the error analysis, and calculate the completely new mesh. The following problems can appear:

– the new mesh generation can be computationally costly.

– it is hard to transfer the information about the solution from the old mesh to the new one.

3. r-refinement.

The number and the size of elements is fixed but their positions in the domain can change. This approach can be used for the description of moving fronts.

The types of p-refinement:

1. The uniform increase of the polynomial degrees in all elements.
2. The local increase of the polynomial degrees where necessary with the use of the hierarchical elements.

It is clear that the refinement procedure depends on the norm chosen for the error calculations. The different norms may result in different meshes and elemental functions. Let us give one example of the refinement criterium. We will require that the relative error $\eta = ||e||/||u||$ would not exceed a prescribed value:

$$\eta \leq \overline{\eta}.$$

The optimal mesh is such a mesh that the errors on all elements are the same. Then the acceptable error is calculated as

$$\overline{\eta}||u|| \approx \overline{\eta} \left(||U||^2 + ||e||^2\right)^{1/2}.$$

We require that

$$||e||_k < \overline{\eta} \left(\frac{||U||^2 + ||e||^2}{m}\right)^{1/2} = \overline{e}_m,$$

where m is the number of the elements. The value ξ_k,

$$\xi_k = ||e||_k / \overline{e}_m,$$

defines the acceptability of the error on the element. We should refine the solution on a specific element, if

$$\xi_k > 1.$$

If we are going to completely remesh the triangulation, we should require that $\xi_k \leq 1$ for all new elements. We can assume that the errors behave accordingly to the interpolation theorem, i.e.

$$||e||_k \sim h_k^p.$$

Then the new element size can be chosen as

$$h_{new} = \min_k \left(\xi_k^{-1/p} h_k\right).$$

25.1 *hp*-refinement

The *hp*-refinement combines both *h*- and *p*-refinement together. There may exist different strategies of the *hp*-refinement. Let us present one example [8].
1. First step: *h*-refinement using low degree polynomials ($p = 1$ or 2) until an acceptable accuracy is reached (e.g. 5% or less, depending on the problem). The errors are distributed regularly over elements.
2. Second step: the uniform *p*-refinement for all elements.
The second step can be performed easier if we know p value which is necessary in order to reach the accuracy required. We can do this with the p extrapolation. We assume that the error $||e||$ behaves as

$$||e|| \leq CN^{-\beta}, \tag{65}$$

where N is the number of the degrees of freedom in the approximated solution. We should calculate the error squares for three polynomial degrees

$$||e||^2 = ||u||^2 - ||U_q||^2 = C^2 N_q^{-2\beta}, \quad q = p-2, p-1, p.$$

Here we have three equation with three unknowns $||u||$, C and β. Excluding C and β, we get a nonlinear equation:

$$\frac{||u||^2 - ||U_p||^2}{||u||^2 - ||U_{p-1}||^2} = \left(\frac{||u||^2 - ||U_{p-1}||^2}{||u||^2 - ||U_{p-2}||^2} \right)^{\frac{\log{(N_{p-1}/N_p)}}{\log{(N_{p-2}/N_{p-1})}}}.$$

Getting the solution of this equation $||u||$, we can then find C and β, and estimate the error for an arbitrary q from equation (65).

Lecture 10

The boundary element method (BEM). Indirect formulation of the BEM. The direct BEM.

26 The boundary element method (BEM)

The finite element method can be successfully applied to the various problems in arbitrary domains, with various boundary conditions and differential operators. However, the algebraic problems resulting from the FEM are rather big, and their solutions cannot always be easily calculated. Due to this fact, one always looks for other methods for the solution of similar problems. For a specific class of problems, there has been found another method, the boundary element method (BEM), which results in considerably smaller algebraic problems.

The main idea of the BEM is quite natural. The initial partial differential equation in a domain is rewritten as the integral equation on the domain boundary. We solve this integral equation, so we find the solution on the boundary. Then, if necessary, we restore the solution inside the entire domain. As the dimension of the integral equation is less by one compare to the dimension of the initial partial differential equation, the size of the algebraic problem and the computational cost are considerably smaller.

There exist different ways to write down the BEM equations. We discuss here two BEM approaches, namely the so-called direct and indirect formulations on the simple example of the Laplace equation in the two-dimensional domain.

27 Indirect formulation of the BEM

We study here the Laplace equation defined in the domain Ω. The boundary conditions on the domain boundary can be chosen as the Dirichlet condition on the part Γ_1 and the Neumann condition on the part Γ_2, where we require $\Gamma = \Gamma_1 + \Gamma_2$. Let us denote the solution of the problem by $u(x)$ and its derivative by $q(x) = \partial u(x)/\partial n$. Then our problem can be written as follows:

$$\begin{cases} \Delta u(x) = 0, & x \in \Omega \\ u(x) = \bar{u}(x), & x \in \Gamma_1 \\ q(x) = \bar{q}(x), & x \in \Gamma_2. \end{cases} \tag{66}$$

Here the vector $\vec{n}$ is the external normal unit vector to the boundary. As the function $u(x)$ satisfies the Laplace equation, it is the harmonic function. It is known that there exists the distribution of a potential on the boundary corresponding to each harmonic function. On the other hand, such distribution defines a harmonic function. Having these facts in mind, let us first discuss the Dirichlet and the Neumann problems separately.

27.1 The Neumann problem

It is know that the solution of the Neumann problem for the Laplace equation can be written as the single layer potential with a density $\sigma(x)$:

$$u(x) = \int_\Gamma \sigma(\xi) G(\xi, x) d\Gamma(\xi), \quad x \in \Omega. \tag{67}$$

Here, $G(\xi, x)$ is the fundamental solution of the Laplace equation:

$$G(\xi, x) = \begin{cases} 1/|\xi - x| & \text{in } R^3, \\ \log(1/|\xi - x|) & \text{in } R^2. \end{cases} \tag{68}$$

Calculating the derivative of representation (67) over the external normal, we arrive to

$$q(x) = -\alpha\pi\sigma(x) + \int_\Gamma \sigma(\xi) \frac{\partial G(\xi, x)}{\partial n(x)} d\Gamma(\xi), \quad x \in \Gamma, \tag{69}$$

where $\alpha = 1$ for R^2 and $\alpha = 2$ for R^3. As the l.h.s. of this equation is known, we have the Fredholm equation of the second kind for the unknown function $\sigma(x)$. When we solve this equation, we can restore the solution $u(x)$ over the entire domain Ω by representation (67).

It is worth reminding that the solution of equation (69) exists only if the Gauss condition

$$\int_\Gamma q(x) d\Gamma(x) = 0$$

is satisfied.

27.2 The Dirichlet problem

The solution of the Dirichlet problem in the domain can be written as the double layer potential with unknown density $\mu(x)$:

$$u(x) = \int_\Gamma \mu(\xi) \frac{\partial G(\xi, x)}{\partial n(x)} d\Gamma(\xi), \quad x \in \Omega. \tag{70}$$

Restricting this equation on the boundary and taking into account the jump of the double layer potential, we can write equation (70) as

$$u(x) = -\alpha\pi\mu(x) + \int_\Gamma \mu(\xi) \frac{\partial G(\xi, x)}{\partial n(x)} d\Gamma(\xi), \quad x \in \Gamma,$$

Again, this equation is the Fredholm equation of the second kind. The function $\mu(x)$ can be found from this equation, and then the solution over entire domain can be restored with representation (70).

One can note that one can also look for the solution of the Dirichlet problem in terms of the single layer potential:

$$u(x) = \int_\Gamma \sigma(\xi) G(\xi, x) d\Gamma(\xi), \quad x \in \Omega.$$

Due to the continuity of the single layer potential on the boundary, we get the relationship

$$u(x) = \int_\Gamma \sigma(\xi) G(\xi, x) d\Gamma(\xi), \quad x \in \Gamma.$$

This equation is the Fredholm equation of the first kind with respect to the function $\sigma(x)$. As a rule, these equations are badly conditioned, and solving them is a uneasy task. In this specific case, however, the kernel of the integral operator is singular that leads to the well-conditioned equation.

28 The direct BEM

In the indirect BEM, we need to introduce new unknowns: the single and double layer potentials. In some areas of research (e.g. in the electrostatics) these potentials have a clear physical sense. In many problems, however, they are just artificial auxiliary function without any physical sense. For these problems, we would prefer the formulation of the BEM without any additional functions. This BEM version exists and called the direct BEM. The equations of the direct BEM are written in terms of the solution and its derivative only. In order to get such equations, one can use the third Green's formula or the residual approach.

We are looking for the solution of equation (66). Applying the third Green's formula for an arbitrary function $u^*(\xi, x)$, we can directly get from the Laplace equation

$$\int_\Omega \Delta u^*(\xi, x)u(x)d\Omega = -\int_\Gamma q(x)u^*(\xi, x)d\Gamma(x) + \int_\Gamma u(x)q^*(\xi, x)d\Gamma(x), \quad (71)$$

where

$$q^*(\xi, x) = \frac{\partial u^*(\xi, x)}{\partial n(x)}.$$

Let us now choose the fundamental solution G of the Laplace equation,

$$\Delta G(\xi, x) = -2\alpha\pi\delta(\xi - x),$$

as the function $u^*(\xi, x)$. Substituting it into equation (71), we get

$$2\alpha\pi u(\xi) + \int_\Gamma u(x)\frac{\partial G(\xi, x)}{\partial n(x)}d\Gamma(x) = \int_\Gamma q(x)G(\xi, x)d\Gamma(x). \quad (72)$$

Therefore, we have now in our equation the sum of two layer potentials but their density are expressed in terms of the solution of the equation. Depending on the boundary conditions, this equation can be the Fredholm equation of the first or the second kind. It can also take into account the mixed boundary conditions.

The close investigation of the direct BEM equation shows that the requirements for the smoothness of the boundary can be weakened here. For example, the presence of angles and edges is here allowed.

28.1 How to employ the BEM numerically

We describe here how the direct BEM can be used for the construction of the numerical solution of the problem. This approach consists of the following steps:

- the subdivision of the boundary into elements,
- writing down the connection between the function and its derivative at the nodes,
- numerical calculation of the integrals,
- taking into account the boundary conditions and reduction of the equation to the matrix problem,
- restoring the function values inside the domain if necessary.

We will discuss the same equation (66). Let us subdivide the boundary part Γ_1 into N_1 elements, and the part Γ_2 into N_2 elements. If we choose for the approximation the simplest case of the piecewise constant polynomials ($p = 0$), then the functions q and u are constants on each element, and one of them is necessarily known. If we write equation (72) for the ith element, we get

$$c_i u_i + \int_\Gamma u(x) q^*(x) d\Gamma(x) = \int_\Gamma q(x) u^*(\xi, x) d\Gamma(x).$$

Using the subdivision of the boundary into the elements, we can rewrite the last equation as

$$c_i u_i + \sum_{j=1}^{N} \int_{\Gamma_j} u(x) q^*(x) d\Gamma(x) = \sum_{j=1}^{N} \int_{\Gamma_j} q(x) u^*(\xi, x) d\Gamma(x).$$

As the functions are constants on each element, we simplify the last equation as

$$c_i u_i + \sum_{j=1}^{N} \left(\int_{\Gamma_j} q^*(x) d\Gamma(x) \right) u_j = \sum_{j=1}^{N} \left(\int_{\Gamma_j} u^*(\xi, x) d\Gamma(x) \right) q_j. \qquad (73)$$

The integrals in the last equation should be calculated in some way. As the functions in these integrals are known, they can sometimes be calculated analytically but in the general case the numerical integration is required. There are no problems for the calculation of integrals with $i \neq j$. Here, the Gauss-Legendre quadrature formula with the few nodes can be employed. If $i = j$, then there is an (integrable!) singularity in the integral so the special care should be taken. Depending on the problem, accuracy and boundary, one can use the high order Gauss-Legendre formulas, specially derived formulas for the singular integrals, or put effort into the analytical integration.

Let us denote the integral values as

$$\hat{H}_{ij} = \int_{\Gamma_j} q^*(x) d\Gamma(x), \quad G_{ij} = \int_{\Gamma_j} u^*(\xi, x) d\Gamma(x).$$

Then we also add the non-integral term into the matrix H:

$$H_{ij} = \begin{cases} \hat{H}_{ij}, & i \neq j, \\ \hat{H}_{ij} + c_i, & i = j. \end{cases}$$

With these notation, equation (73) can be written as

$$\sum_{j=1}^{N} H_{ij} u_j = \sum_{j=1}^{N} G_{ij} q_j, \quad i = 1 \ldots N.$$

In the matrix form, it reads

$$HU = GQ.$$

As the unknowns in this equation, we have N_1 values of the function u and N_2 values of its derivative q. Together, we have $N_1 + N_2 = N$ unknowns, that coincide with the number of linear equations. If we combine all the unknown values of the function and derivative in one vector Y, and gather all known values into the vector F, we get the standard matrix problem

$$AY = F.$$

Here the matrix A is constructed from two matrices H and G. The solution, vector Y, give us all not jet known values of the function and derivative on the boundary. At any internal point of the domain, we can restore the solution as

$$u_i = \left(\sum_{j=1}^{N} G_{ij} q_j - \sum_{j=1}^{N} \hat{H}_{ij} u_j \right) / c_i.$$

Lecture 11

The spectral methods. Fast calculations in the spectral methods.

29 The spectral methods

Computationally, the Galerkin method can be realized in different ways. In the previous lectures, we accurately discussed the FEM approach. In that approach, the basis functions chosen to be local, they are non zero on very few elements only. The error estimation for the FEM in general case be written as

$$\forall v \in H^{k+1}(K) \quad |e|_{m,K} \leq C\frac{h_K^{k+1}}{\rho_K^m}|v|_{k+1,K}, \quad 0 \leq m \leq k+1, \tag{74}$$

where h_K is the diameter of the element K, $h_K = \mathrm{diam}(K)$, and

$$\rho_K = \sup\{\mathrm{diam}(S) : \quad S \text{ is insphere of } K\}.$$

When the finite element family is regular, $h_K \sim C\rho_K$, and expression (74) becomes the standard error estimate $|e|_{m,K} \leq Ch_K^{k+1-m}|v|_{k+1,K}$.

One can see from equation (74) that the solution error decreases as a power of the element diameter. If we have a fixed number of the elements in the mesh, we can perform the solution refinement by increasing the polynomial degree k. In this case, the error estimation with respect to k will be *exponential* so the convergence is much faster. The idea described leads to the *spectral method* approach.

As a rule in the spectral methods, we do not divide the entire domain into elements, so basis functions are *global.* In this sense, the mesh consists of one element only. We can make calculations easier if the basis functions are orthogonal. Of course, we should accurately derive the error estimations as estimate (74) cannot be formally employed (the domain diameter is not necessary less than 1). We illustrate the results of this type with one example.

Example 1. Let us consider the equation of non-viscous convection:

$$\frac{\partial u}{\partial t} + \sum_i \left(V_i\frac{\partial u}{\partial x_i} + \frac{\partial u}{\partial x_i}(V_i u)\right) = 0. \tag{75}$$

Here the functions $V_i(x)$ are known, and u is periodic on ∂R. Let us denote by u_a the approximated solution with N basis functions. Then one can prove that

$$||u(t) - u_a(t)||_{H^0,R} \leq CN^{-k}||u(0)||_{H^{k+1}}. \tag{76}$$

Here $u(0)$ is the initial condition, the constant C is independent of N and $u(0)$. Hence the convergence can be very fast if the initial and boundary conditions are smooth enough. ■

The choice of the basis functions strongly affects the accuracy of the spectral methods. Based on the known applications, we present frequently employed basis functions in table 3.

Basis functions	Properties
Eigenfunctions	the solution of a similar problem
Fourier expansion	periodic boundary conditions, infinitely differentiable
Legendre polynomials	non-periodicity
Tchebyshev polynomials	non-periodicity, the minimax principle

Table 3: The basis functions for the spectral methods.

Let us discuss these basis sets in more detail. If we go from the bottom to the top of the table, we have more restrictive requirements for the sets. On the other hand, if those requirements are satisfied, the convergence speed is better for the topper set.

1. **Eigenfunctions** of a similar problem which can be effectively solved. The boundary conditions of the problem under investigation have to be satisfied. The exact solution has to be infinitely differentiable.

2. **The Fourier expansion.** For the application of the Fourier expansion, the boundary conditions must be periodic, then the convergence speed is exponential. If they are not, the Fourier expansion make them periodic with a jump on one of the boundaries (the Gibbs effect, see Fig. 48.).

So due to this jump the convergence speed degrades to $O(N^{-1})$ for the sine Fourier expansion. For the cosine expansion, the convergence speed is a little bit better, namely $O(N^{-2})$ everywhere except of a boundary vicinity, where it is $O(N^{-1})$ again.

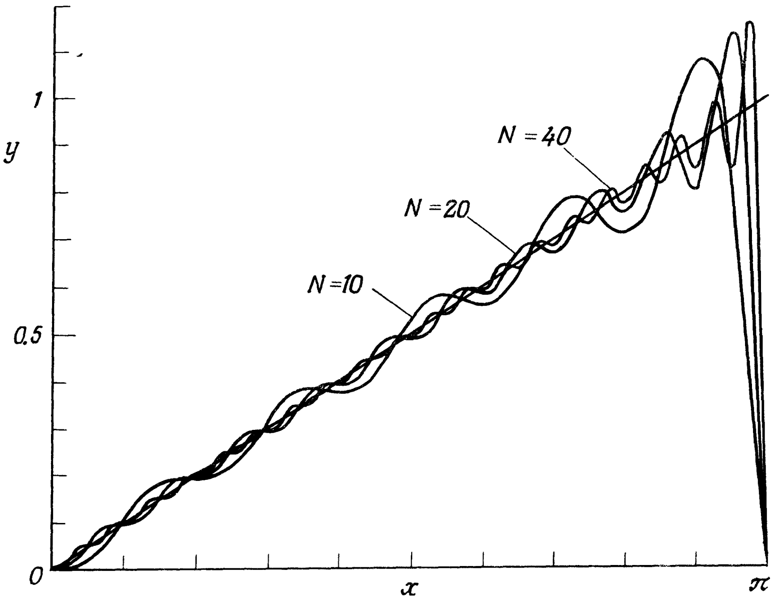


Figure 48: The approximation of the function $y = x/\pi$ with the sine Fourier expansion with N terms.

If the jumps of the solution appear in the interior, they result in the same speed degradation. Even worse, for some examples of hyperbolic type equations the expansion may not converge at all.

3. **The Legendre polynomials.** For the expansion in terms of the Legendre polynomials, we are not restricted with the periodic boundary conditions. The convergence speed for this expansion is also exponential, $O(N^{-k})$. If the solution is not smooth and jumps are present, the convergence degradation does also appear. The global convergence speed reduces to $O(N^{-1})$ while in a vicinity of the boundaries the situation gets even worse as one can only guarantee the $O(N^{-1/2})$ speed. This slow convergence is clearly a disadvantage of this expansion.

3. **The Chebyshev polynomials.** For the smooth solution, the standard for the spectral method convergence rate $O(N^{-k})$ is achieved. The non-periodic boundary conditions can be taken into account without any sacrifice. If the solution has jumps, the convergence speed is lowered to

$O(N^{-1})$ everywhere including boundaries. The additional advantage of this expansion is that the maximal error is close to the achievable minimum (the minimax principle). The disadvantage of this expansion is that in the standard formulation (when the basis function space coincide with the projection space, $U = V$) we get the non-diagonal mass matrix: the Chebyshev polynomials are non-orthogonal. In order to overcome this problem, we can use the generalized formulation with two sets in different spaces:

$$\{T_j(x)\} \quad \text{and} \quad \{T_j(x)/(1-x^2)^{1/2}\}.$$

Example 2. Let us consider an example of the spectral expansion used for Burgers' equation. This equation is often used for describing of the shock waves. We consider the equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} - \frac{1}{\mathrm{Re}}\frac{\partial^2 u}{\partial x^2} = 0 \tag{77}$$

with the boundary condition

$$u(-1,t) = 1, \qquad u(1,t) = 0,$$

and the initial condition

$$u(x,0) = \begin{cases} 1, & x \le 0, \\ 0, & x > 0. \end{cases}$$

This equation contains the nonlinear term $u\frac{\partial u}{\partial x}$ but we will not use any special technique to deal with this term. As the boundary conditions are non periodic, it is natural to use the Legendre or Chebyshev polynomials. We will employ the Legendre expansion here:

$$u(x,t) = \sum_{j=0}^{N} a_j(t)P_j(x), \tag{78}$$

where $P_j(x)$ is the Legendre polynomial of the order j, orthogonal on $[-1,1]$. Using the standard Galerkin approach, i.e. substituting the expansion into the equation and projecting onto the basis set function, we get the following system of the ordinary differential equation:

$$M\frac{d}{dt}\vec{A} + (B + C)\vec{A} = 0. \tag{79}$$

The matrix M is given by

$$M_{jk} = \delta_{jk}\frac{2}{2k+1},$$

and is diagonal due to the orthogonality of the Legendre polynomials. The matrix C is given as

$$C_{jk} = \frac{1}{\mathrm{Re}}\left(\frac{dP_j}{dx}, \frac{dP_k}{dx}\right)$$

and is filled completely. The matrix B

$$B_{jk} = \sum_{i=1}^{N} a_i \left(P_j\frac{dP_i}{dx}, P_k\right)$$

describes the nonlinear term, so it depends on the solution $\vec{A}$. The initial conditions can also be found with the Galerkin procedure, and are written as the matrix equation

$$M\vec{A} = D, \quad \text{where} \quad d_k = \int_{-1}^{0} P_k(x)dx.$$

If we solve the system of equation (79) accurately for the time variable, we can investigate the influence of the spectral expansion on the accuracy of the solution. The differences between the exact and approximated solution are plotted on Fig. 49 for few expansion lengths N. One can see that this difference rapidly vanishes with N. ■

30 Fast calculations in the spectral methods

For the realistic problems one should use sufficiently large number N of the expansion terms in order to get the accurate solution. This makes the

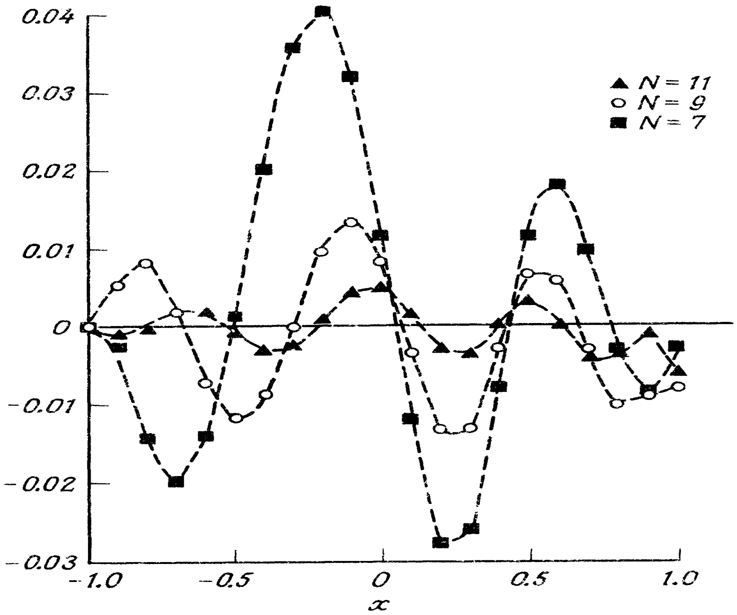


Figure 49: The errors for Burgers' equation for Re = 10.

method non-effective. In order to keep high efficiency, one should use special calculation technique.

Let us consider the time-dependent equation

$$\frac{\partial u}{\partial t} = L(u). \tag{80}$$

If we use the spectral method for the approximation of the coordinate part $L(u)$ of the equation, we should sum N coefficients for the linear terms, N^2 coefficients for quadratic terms etc. These values should be compared to 3-5 coefficients typical to the finite difference and finite element methods. Hence we need special approaches to perform the fast summation in the frame of the spectral method.

Recurrent relations

When we use the spectral expansion for the solution of the equation

$$u(x,t) = \sum_{j=1}^{N} a_j(t)\varphi_j(x),$$

we often need its derivative. It can be expressed as

$$\frac{\partial u}{\partial x}(x,t) = \sum_{j=1}^{N} a_j(t) \frac{\partial \varphi_j(x)}{\partial x}.$$

However, the computation can be done more effective if the derivative is expressed in terms of the same functions $\varphi_j(x)$:

$$\frac{\partial u}{\partial x}(x,t) = \sum_{j=1}^{N} b_j(t)\varphi_j(x),$$

For the specific operator and for the specific basis functions, one can find a relation between coefficients $a_j(t)$ and $b_j(t)$ and then use this relation in the calculations.

Example 3. Let us consider few examples. Let the operator be $L(u) = \partial u/\partial x$, and let the basis set be the Chebyshev polynomials $\{T_j\}$. Then one finds

$$b_j = 2 \sum_{\substack{p=j+1 \\ p+j \text{ is odd}}}^{N} p a_p, \quad j = 1 \dots N-1, \qquad b_0 = \sum_{\substack{p=1 \\ p \text{ is odd}}}^{N} p a_p. \tag{81}$$

For the second derivative $L(u) = \partial^2 u/\partial x^2$, the corresponding relations are given by

$$b_j = \sum_{\substack{p=j+2 \\ p+j \text{ is even}}}^{N} p(p^2 - j^2) a_p, \quad j = 1 \dots N-2, \qquad b_0 = \frac{1}{2} \sum_{\substack{p=2 \\ p \text{ is even}}}^{N} p^3 a_p.$$

■

The latter formulas are rather useful for the orthogonal functions as they exclude the need for the basis set derivatives. Another advantage is a possibility to use recurrent relations. For example, it is known for the Chebyshev polynomials that

$$2T_j = \frac{1}{j+1} T'_{j+1} - \frac{1}{j-1} T'_{j-1}.$$

Using this relationship, we can rewrite equation (81) as

$$b_j = b_{j+2} + 2(j+1)a_{j+1}, \quad 1 \le j \le N-1, \qquad b_0 = 0.5b_2 + a_1, \tag{82}$$

and $b_N = b_{N+1} = 0$. Now we can see that the calculations of all coefficients b_k for $\partial u/\partial x$ requires $O(N^2)$ steps with equation (81), and only $O(N)$ steps with equation (82). It is clear that similar representation can be derived for other operators and different basis sets.

Nonlinear terms

The calculation of nonlinear terms with the spectral representation takes a large number of steps, e.g. N^3 steps for the quadratic nonlinear term. For big values of N, this considerably slows down computations. In order to resolve this problem, we can change the *representation*: we will work with the functions represented as a set of spectral coefficient (i.e. coefficients a_k) wherever possible. Of course, this is possible if we know a very fast way to transform functions from the initial representation into the spectral one and back. It means that we should be able to calculate the sum

$$u(x_l) = \sum_{j=1}^{N} a_j \varphi_j(x_l), \quad l = 1 \dots N,$$

and the integral

$$a_k = \int_R u(x)\varphi_k(x)dx, \quad k = 1 \dots N,$$

more efficiently than $O(N^2)$ prerequested by the latter formulas. Fortunately, such a way is known: these calculations can be efficiently done with the Fast Fourier Transform (FFT). It only takes $O(N \log N)$ steps to calculate these transforms. Furthermore, the fast transforms (FT) can also be derived for other sets of functions, e.g. for the Legendre polynomials.

Example 4. Let us briefly describe here the scheme for the solution of Burgers' equation with use of the FTs. We analyze one time step, so let the coefficients a_j^n be known for step n. Then the following sequence of operations is performed:

- $$u_a^n(x_l) = \sum_j a_j^n \varphi_j(x_l), \quad l = 1 \dots 2N, \quad FT,\ O(2N \log 2N) \text{ ops},$$

- $b_j^{(1)n}$ are calculated from a_j^n with the recurrence relations, $O(2N)$ ops,

- $$\frac{\partial u_a^n(x_l)}{\partial x} = \sum_j b_j^{(1)n} \varphi_j(x_l), \quad l = 1 \dots 2N, \quad FT,\ O(2N \log 2N) \text{ ops},$$

- $$w^n(x_l) = u_a^n(x_l) \frac{\partial u_a^n}{\partial x}(x_l), \quad l = 1 \dots 2N, \quad O(2N) \text{ ops},$$

- $$d_k^n = \int_R w^n \varphi_k dx, \quad FT,\ O(2N \log 2N) \text{ ops},$$

- $$s_k^n = \sum_j \left(\varphi_k, \frac{\partial^2 \varphi_j}{\partial x^2} \right) a_j^n,$$

  with the recurrence relations, $O(N)$ ops,

- $$\frac{da_k^{n+1}}{dt} = d_k^n - \frac{s_k^n}{\mathrm{Re}}, \quad O(N) \text{ ops},$$

- $$\frac{a_k^{n+1}}{dt} = a_k^n + f\left(\frac{da_k^{n+1}}{dt} \right), \quad O(N) \text{ ops}.$$

So we can see that at any step of the procedure we maximally perform $O(2N \log 2N)$ operations. Hence this more complicated structure is far more efficient in the computational sense than the straightforward realization described in Example 2. This fast approach can be naturally generalized for many parabolic equations. ■

References

[1] Solin P., *Partial Differential Equations and the Finite Element Method*, Wiley, 2005.

[2] Ciarlet P., *The finite element method for elliptic problems*, North-Holland publishing company, 1978.

[3] Mitchell A.R., Wait R., *The finite element method in partial differential equations*, Wiley Interscience publication, 1977.

[4] Fletcher C.A.J., *Computational Galerkin methods*, Springer-Verlag, 1984.

[5] Flaherty J.E., *Finite element analysis*, Renssellaer lecture notes, 2000.

[6] Buslov V.A., Yakovlev S.L., *Numerical methods I. The analysis of functions*, SPbGU, 2001 (in Russian).

[7] Sabonnadiere J.-C., Coulomb J.-L., *The finite element method and CAD*, Mir, 1989 (in Russian).

[8] Zienkiewicz O.C., Taylor R.L., *The finite element method. Vol. 1. The basis*, 2000.

Contents