PDE & FEM TERMINOLOGY.
BASIC PRINCIPLES OF FEM.

Sergey Korotov

Basque Center for Applied Mathematics / IKERBASQUE

http://www.bcamath.org & http://www.ikerbasque.net

Introduction

The analytical solution of problems described by partial differential equations (PDEs) is known only in a few cases on special domains (like balls, cubes, half-spaces, etc). It is often necessary to use some numerical methods to get an approximation of this solution.

One of the most powerful numerical methods for solving PDEs is the Galerkin method.

The standard finite element method (FEM) is, roughly speaking, the Galerkin method with a special choice of basis functions.

The FEM has been developed during last 70 years. The discovery of FEM is usually attributed to Richard Courant [Courant, 1943]. However, in [Ciarlet, Lions] we can find some older references to finite element-like methods. The first monograph on FEM is probably that of Synge [Synge] of 1957.

The notion *element* was introduced in the 1950-th by aerospace engineers performing elasticity computations as they devided a continuum into small pieces called elements. The notion *finite element* was introduced by mathematicians later, in the 1960-th. From that time the theory of FEM has also been rigorously investigated.

R. Courant. Variational methods for problems of equilibrium and vibration. Bull. Amer. Math. Soc. 49 (1943), 1–23.

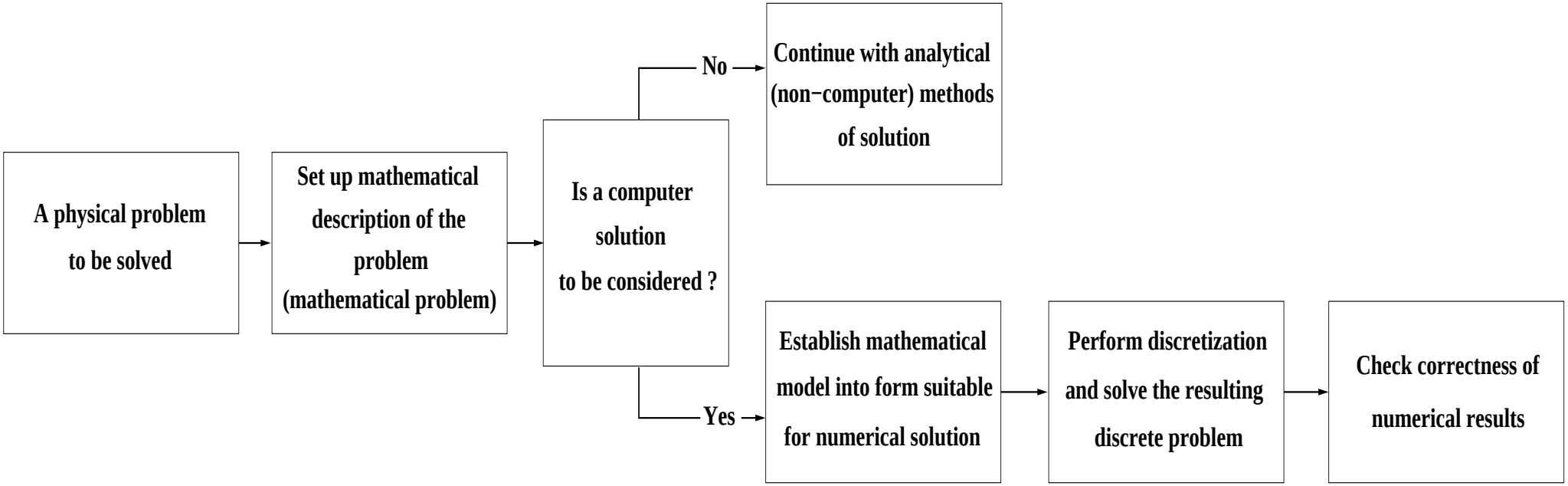
P. G. Ciarlet, J. L. Lions (eds.). Handbook of Numerical Analysis. Vol. II Finite element methods. North-Holland, Amsterdam, 1991.

J. L. Synge. The Hypercircle in Mathematical Physics. Cambridge Univ. Press, Cambridge, 1957.

Nowadays FEM seems to be one of the most efficient numerical methods for solving problems of mathematical physics which are based on variational principles. One may solve by FEM some variational problems which do not correspond to any PDEs (e.g. the obstacle problem).

Moreover, FEM enables perfect description of the examined domain, which was not possible by classical numerical methods (such as the collocation method, finite difference method, etc.).

The flow-chart of numerical simulation:



There are many ways how to set up a mathematical model and then design it into a form suitable for numerical solution. In our case it will be so that we give a variational formulation of the problem in an appropriate function space: We want to find a function minimizing a convex functional over a closed set of admissible functions. The basic idea of the discretization will consist then in transforming the problem formulated in function spaces with infinite dimension into appropriate problems in finite-dimensional spaces.

First we shall remind a variational (weak) formulation of problems described by second order PDEs of elliptic type with some boundary conditions.

This formulation is useful in explaining the mathematical background of FEM and is based on the theory of Sobolev spaces.

Sobolev spaces are extensions of spaces of differentiable functions in the classical sense. Such extensions are natural as the solution of a variational problem usually does not have the classical derivatives.

Model Problem

In a bounded domain Ω with a Lipschitz boundary $\partial\Omega$, consider the following second order PDE for the unknown function $u \in C^2(\overline{\Omega})$:

$$-\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + cu = f \ .$$

Here $c, f \in C(\overline{\Omega})$ and the matrix $\mathcal{A} = (a_{ij}) \in (C^1(\overline{\Omega}))^{d\times d}$ are given. We further assume that

$$a_{ij} = a_{ji} \ , \quad i,j = 1,\ldots,d \ , \qquad c(x) \geq 0 \ , \quad x \in \overline{\Omega} \ ,$$

and that there is a constant $M > 0$ such that

$$\sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \geq M\sum_{i=1}^{d}\xi_i^2 \qquad \forall (\xi_1,\ldots,\xi_d)^T \in R^d, \ x \in \overline{\Omega} \ .$$

The above assumptions ensure that our model problem is *elliptic.*

Three (standard) types of boundary conditions characterizing the *(homogeneous) Dirichlet, Neumann and Newton classical problem:*

$$u = 0 \qquad \text{on } \partial\Omega \ ,$$

$$\frac{\partial}{\partial n_{\mathrm{A}}} u = 0 \qquad \text{on } \partial\Omega \ ,$$

$$\alpha u + \frac{\partial}{\partial n_{\mathrm{A}}} u = 0 \qquad \text{on } \partial\Omega \ .$$

Here $\alpha = \alpha(s) \geq 0$ on $\partial\Omega$ is continuous,

$$\frac{\partial}{\partial n_{\mathrm{A}}} u = \sum_{i,j=1}^{d} a_{ij} \frac{\partial u}{\partial x_j} n_i = n^T \mathcal{A} \operatorname{grad} u$$

denotes the *conormal derivative,* n_i are the components of the unit outward normal to $\partial\Omega$ and $n_{\mathrm{A}} = \mathcal{A}\, n$ is the *conormal.* The vector function $\mathcal{A} \operatorname{grad} u$ on Ω is said to be the *cogradient* and the scalar function $n^T \mathcal{A} \operatorname{grad} u|_{\partial\Omega}$ the *boundary flux.*

The model PDE with one of the above boundary conditions represents the so-called boundary value problem (BVP) (classically formulated). The function $u \in C^2(\overline{\Omega})$ satisfying the PDE and the associated boundary condition is called the *classical solution.*

Such BVP may describe a stationary magnetic, electric, or temperature fields, etc. The coefficients a_{ij} describe physical properties of the medium Ω. If a_{ij} are constants (i.e., independent of x), we call the medium *homogeneous* (otherwise *nonhomogeneous*). If $a_{ii}(x) = a_{11}(x)$ for $i = 2, \ldots, d$, and $a_{ij}(x) = 0$ for $i \neq j$, the medium is said to be *isotropic* (otherwise *anisotropic*). If $\mathcal{A}$ is only diagonal, the medium is *orthotropic.*

In real-life problems the coefficients a_{ij}, c, α are often nonsmooth (they are, e.g., piecewise constant) and we cannot use the equation as it stands, since the classical derivatives in it need not exist.

This is why we introduce a weak formulation of the classical problems, which enables us to consider also nonsmooth coefficients and/or nonsmooth f.

Consider e.g. the Newton classical problem and introduce the *space of test functions* $V := H^1(\Omega)$. Multiplying the PDE by an arbitrary test function $v \in V$ and integrating over Ω we get

$$-\int_\Omega \sum_{i,j} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) v\,dx + \int_\Omega cuv\,dx = \int_\Omega fv\,dx\ .$$

Employing now Green's formula, we find that

$$\int_\Omega \sum_{i,j} a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}\,dx - \int_{\partial\Omega}\sum_{i,j} a_{ij}\frac{\partial u}{\partial x_j}n_i v\,ds + \int_\Omega cuv\,dx = \int_\Omega fv\,dx\ .$$

Using further the boundary condition, we obtain that

$$\int_\Omega \sum_{i,j} a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}\,dx + \int_\Omega cuv\,dx + \int_{\partial\Omega}\alpha uv\,ds = \int_\Omega fv\,dx \quad \forall v \in V\ .$$

We see that any classical solution of the Newton problem (if it exists) satisfies the equation

$$\int_\Omega \sum_{i,j} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i}\, dx + \int_\Omega cuv\, dx + \int_{\partial\Omega} \alpha uv\, ds = \int_\Omega fv\, dx \quad \forall v \in V\ . \quad (*)$$

Let us observe that the integrals in above are well defined even when a_{ij}, $c \in L^\infty(\Omega)$, $\alpha \in L^\infty(\partial\Omega)$, $f \in L^2(\Omega)$, and when prescribed conditions on the coefficients hold almost everywhere (a.e.) in Ω only.

When introducing the space of test functions V we started to use implicitly the *concept of completion.* Instead of looking for $u \in C^2(\overline{\Omega})$ satisfying the PDE and the boundary condition, we shall solve the integral form equation $(*)$. Its solution will be searched in the Hilbert space $H^1(\Omega)$, which is, of course, complete and bigger than $C^2(\overline{\Omega})$. This allows us to employ some useful results from functional analysis (e.g. Lax-Milgram lemma) and apply the finite element method.

Weak Solution

The function $u \in H^1(\Omega)$ is called the *weak (or generalized) solution* of the Newton problem if

$$\int_\Omega \sum_{i,j} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i}\, dx + \int_\Omega cuv\, dx + \int_{\partial\Omega} \alpha uv\, ds = \int_\Omega fv\, dx \quad \forall v \in V\ . \quad (*)$$

For convenience the weak solution is denoted by u as the classical solution. Note that the classical solution from $C^2(\overline{\Omega})$ need not exist. However, if it does exist, it is also the weak solution.

Notice further that the weak formulation $(*)$ contains only the first (generalized) derivatives of u which may be advantageous for approximation.

In order to establish under what conditions the problem posed in a weak form has a unique solution in the Sobolev space $H^1(\Omega)$, we recall some definitions and present two abstract theorems.

Let V be a linear space. A mapping $a(\cdot,\cdot) : V \times V \to R^1$ is called a *bilinear form,* if for any fixed $v \in V$ the mappings $a(v,\cdot) : V \to R^1$ and $a(\cdot, v) : V \to R^1$ are linear.

If, in addition,

$$a(v,w) = a(w,v) \qquad \forall v, w \in V \ ,$$

the bilinear form $a(\cdot,\cdot)$ is said to be *symmetric.*

Theorem 1: Let V be a Banach space equipped with the norm $\|\cdot\|_V$ and let $F : V \to R^1$ be a continuous linear form. Let $a(\cdot,\cdot) : V \times V \to R^1$ be a symmetric and continuous bilinear form, i.e., there exists $C_1 > 0$ such that

$$|a(v,w)| \leq C_1 \|v\|_V \ \|w\|_V \quad \forall v, w \in V.$$

Moreover, we suppose that there exists $C_2 > 0$ such that

$$a(v,v) \geq C_2 \|v\|_V^2 \quad \forall v \in V \qquad (V\text{-ellipticity condition}) \ .$$

Then the problem: Find $u \in V$ such that

$$a(u,v) = F(v) \quad \forall v \in V \qquad (\text{variational equality})$$

has one and only one solution.

P r o o f : Due to the symmetry of $a(\cdot,\cdot)$ and V-ellipticity, the bilinear form $a(\cdot,\cdot)$ is a scalar product on V since $a(v,v) = 0$ implies $v = 0$. The corresponding norm $\sqrt{a(\cdot,\cdot)}$ is, obviously, equivalent to the given norm $\|\cdot\|_V$. Thus V equipped with the scalar product $a(\cdot,\cdot)$ is a Hilbert space. Since $F(\cdot)$ is a continuous linear form, there exists by the Riesz representation theorem a unique element $u = u_F \in V$ such that

$$F(v) = a(u_F, v) \qquad \forall v \in V \ . \qquad \blacksquare$$

Theorem 2: Let all the assumptions of the previous theorem be fulfilled. Then the problem: Find $u \in V$ such that

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

is equivalent to the problem: Find $u \in V$ such that

$$J(u) = \inf_{v \in V} J(v) \ ,$$

where

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

Functional $J(v) = \frac{1}{2}a(v,v) - F(v)$ is called the *energy functional.*

Norm $\sqrt{a(\cdot,\cdot)}$ (equivalent to $\|\cdot\|_V$-norm) is called the *energy norm.*

To apply Theorem 1 to the weak formulation $(*)$ we set

$$V = H^1(\Omega)\ ,$$

$$a(v,w) = \int_\Omega \sum_{i,j} a_{ij} \frac{\partial v}{\partial x_j} \frac{\partial w}{\partial x_i}\, dx + \int_\Omega cvw\, dx + \int_{\partial\Omega} \alpha vw\, ds\ , \qquad v, w \in V\ ,$$

$$F(v) = \int_\Omega fv\, dx, \qquad v \in V$$

and verify whether all the assumptions of the theorem are fulfilled.

Useful Inequalities

Let Ω be a bounded domain with a Lipschitz boundary $\partial\Omega$ and let $v, w \in H^1(\Omega)$. Then

$$v|_{\partial\Omega} \in L^2(\partial\Omega) \ ,$$

$$\|v\|_{L^2(\partial\Omega)} \leq C\|v\|_{H^1(\Omega)} \quad \text{(trace theorem)} \ ,$$

$$\int_\Omega \frac{\partial v}{\partial x_j} w \, dx + \int_\Omega v \frac{\partial w}{\partial x_j} \, dx = \int_{\partial\Omega} v w n_j \, ds \quad \text{(Green's formula)} \ ,$$

$$\|v\|_{H^1(\Omega)} \le C \left(\int_\Omega \sum_{j=1}^d \left(\frac{\partial v}{\partial x_j} \right)^2 dx + \int_{\partial\Omega_0} v^2 \, ds \right)^{\frac{1}{2}} \quad \text{(Friedrichs' inequality)} ,$$

where $\partial\Omega_0 \subset \partial\Omega$ is such that $\operatorname{meas} \partial\Omega_0 > 0$,

$$\|v\|_{H^1(\Omega)} \le C \left(\int_\Omega \sum_{j=1}^d \left(\frac{\partial v}{\partial x_j} \right)^2 dx + \int_B v^2 \, dx \right)^{\frac{1}{2}} \quad \text{(X)} ,$$

where $B \subset \Omega$ is a ball,

$$\|v\|_{H^1(\Omega)} \le C \left(\int_\Omega \sum_{j=1}^d \left(\frac{\partial v}{\partial x_j} \right)^2 dx + \left(\int_\Omega v \, dx \right)^2 \right)^{\frac{1}{2}} \quad \text{(Poincaré's inequality)} .$$

For the proof, see [Dautray, Lions, vol. 2], [Nečas, 1967].

- V is a Hilbert space.

$$a(v,w) = \int_\Omega \sum_{i,j} a_{ij} \frac{\partial v}{\partial x_j} \frac{\partial w}{\partial x_i}\, dx + \int_\Omega cvw\, dx + \int_{\partial\Omega} \alpha vw\, ds, \qquad v, w \in V \ ,$$

$$F(v) = \int_\Omega fv\, dx, \qquad v \in V$$

- The linearity of $F(\cdot)$ and bilinearity of $a(\cdot,\cdot)$ are obvious.

- The continuity of $F(\cdot)$ follows immediately from the Cauchy-Schwarz inequality

$$|F(v)| = \left| \int_\Omega fv\, dx \right| \leq \|f\|_{0,\Omega}\ \|v\|_{0,\Omega} \leq \|f\|_{0,\Omega}\ \|v\|_{1,\Omega}$$

as $\|\cdot\|_{0,\Omega} \leq \|\cdot\|_{1,\Omega}$.

$$a(v,w) = \int_\Omega \sum_{i,j} a_{ij} \frac{\partial v}{\partial x_j} \frac{\partial w}{\partial x_i}\, dx + \int_\Omega cvw\, dx + \int_{\partial\Omega} \alpha vw\, ds, \qquad v, w \in V \ ,$$

- The bilinear form is symmetric as $a_{ij} = a_{ji}$ and continuous as:

$$\begin{aligned} |a(v,w)| \leq & \sum_{i,j=1}^{d} \|a_{ij}\|_{L^\infty(\Omega)} \left\| \frac{\partial v}{\partial x_i} \right\|_{0,\Omega} \left\| \frac{\partial w}{\partial x_j} \right\|_{0,\Omega} \\ & + \|c\|_{L^\infty(\Omega)} \|v\|_{0,\Omega} \|w\|_{0,\Omega} + \|\alpha\|_{L^\infty(\partial\Omega)} \|v\|_{L^2(\partial\Omega)} \|w\|_{L^2(\partial\Omega)} \\ \leq & C \|v\|_{1,\Omega} \|w\|_{1,\Omega} \ , \end{aligned}$$

where we used the trace theorem and the relation

$$\|v\|_{1,\Omega}^2 = \|v\|_{0,\Omega}^2 + \sum_{j=1}^{d} \left\| \frac{\partial v}{\partial x_j} \right\|_{0,\Omega}^2 \ .$$

- What remains is to prove the V-ellipticity condition for

$$a(v,w) = \int_\Omega \sum_{i,j} a_{ij} \frac{\partial v}{\partial x_j} \frac{\partial w}{\partial x_i}\, dx + \int_\Omega cvw\, dx + \int_{\partial\Omega} \alpha vw\, ds,$$

Suppose that $\alpha(x) \geq \alpha_0 > 0$ on some part $\partial\Omega_0 \subset \partial\Omega$ which has a positive measure. Then we can use Friedrichs' inequality, and get

$$a(v,v) \geq M \int_\Omega \sum_{i=1}^{d} \left(\frac{\partial v}{\partial x_i}\right)^2 dx + \alpha_0 \int_{\partial\Omega_0} v^2\, ds \geq C\|v\|^2_{1,\Omega}\ .$$

So, in this case the variational problem $(*)$ has a unique solution by virtue of Theorem 1.

- Another situation.

$$a(v,w) = \int_\Omega \sum_{i,j} a_{ij} \frac{\partial v}{\partial x_j} \frac{\partial w}{\partial x_i}\, dx + \int_\Omega cvw\, dx + \int_{\partial\Omega} \alpha vw\, ds,$$

Let now $c(x) \geq c_0 > 0$ on a ball $B \subset \Omega$. Then we get the V-ellipticity in a view of inequality (X):

$$a(v,v) \geq M \int_\Omega \sum_{i=1}^{d} \left(\frac{\partial v}{\partial x_i} \right)^2 dx + c_0 \int_B v^2\, dx \geq C \|v\|^2_{1,\Omega}\ ,$$

which gives the existence of just one solution $u \in V$.

• We can show that the weak solution of $(*)$ belonging to $C^2(\overline{\Omega})$ is also the classical solution of the Newton problem.

• If $c = 0$ in Ω and simultaneously $\alpha = 0$ on $\partial\Omega$ then we get the so-called Neumann problem, which requires a special treatment (not to be considered now).

• We can similarly handle the Dirichlet problem. For the space of test functions, we choose

$$V = H_0^1(\Omega) = \left\{ v \in H^1(\Omega) \mid v = 0 \text{ on } \partial\Omega \right\} \ .$$

Now from Friedrichs' inequality we immediately find that the associated symmetric bilinear form

$$a(v,w) = \int_\Omega \sum_{i,j} a_{ij} \frac{\partial v}{\partial x_j} \frac{\partial w}{\partial x_i} \, dx + \int_\Omega cvw \, dx \ , \quad v, w \in V \ ,$$

fulfils the V-ellipticity condition.

Consider now the nonhomogeneous mixed boundary conditions:

$$
\begin{aligned}
u &= \bar{u} \qquad \text{on } \Gamma_1 \,, \\
\frac{\partial}{\partial n_{\mathrm{A}}} u &= g \qquad \text{on } \Gamma_2 \,, \\
\alpha u + \frac{\partial}{\partial n_{\mathrm{A}}} u &= g \qquad \text{on } \Gamma_3 \,,
\end{aligned}
$$

where $\bar{u} \in H^1(\Omega)$, $g \in L^2(\Gamma_2 \cup \Gamma_3)$, $\alpha \in L^\infty(\Gamma_3)$, $\alpha \geq 0$, $\Gamma_1, \Gamma_2, \Gamma_3$ are mutually disjoint and are open sets in $\partial\Omega$ (one or two of them may be empty),

$$
\Gamma_0 \cup \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 = \partial\Omega \,,
$$

and Γ_0 (meas $\Gamma_0 = 0$) is a set of those points where one type of boundary condition changes into another.

In order to apply the FEM later, we shall assume from now on that each set Γ_i $(i = 1, 2, 3)$ has a finite number of components.

Let the space of test functions be chosen as follows:

$$V = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_1\} \ .$$

If the problem with the above mixed boundary conditions is not purely Neumann with $c = 0$ then the corresponding variational formulation consists in finding $u = u_0 + \bar{u}$, where $u_0 \in V$ satisfies

$$a(u_0, v) = F(v) \qquad \forall v \in V \ ,$$

and where

$$a(v, w) = \int_\Omega \sum_{i,j} a_{ij} \frac{\partial v}{\partial x_j} \frac{\partial w}{\partial x_i} \, dx + \int_\Omega cvw \, dx$$

$$+ \int_{\Gamma_3} \alpha vw \, ds, \quad v, w \in V \ ,$$

$$F(v) = \int_\Omega fv \, dx + \int_{\Gamma_2 \cup \Gamma_3} gv \, ds - a(\bar{u}, v), \quad v \in V \ .$$

The desired V-ellipticity condition follows again from Friedrichs' inequality or (X).

Note that it may be sometimes difficult to find $\bar{u} \in H^1(\Omega)$ with prescribed values on Γ_1.

Lax-Milgram Lemma

Sometimes we need to solve an elliptic problem whose associated bilinear form $a(\cdot,\cdot)$ is not symmetric. Then, instead of Theorem 1, the following theorem (often called Lax-Milgram lemma in the numerical analysis) is applied.

Theorem 3: Let V be a Hilbert space, let $a(\cdot,\cdot) : V \times V \to R^1$ be a continuous V-elliptic bilinear form and let $F : V \to R^1$ be a continuous linear form. Then the problem: Find $u \in V$ such that

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

has one and only one solution.

If $a(\cdot,\cdot)$ is not symmetric then there is no associated minimization problem as in Theorem 2.

Basic Idea of FEM

Now we shall show how to construct *finite element subspaces* V_h of V.

The FEM in its simplest setting is a Galerkin method characterized by the following basic aspects in the construction of V_h:

(i) a triangulation $\mathcal{T}_h$ is established over the domain $\overline{\Omega}$,

(ii) the functions $v_h \in V_h$ are piecewise polynomials,

(iii) there exists a basis in V_h whose functions have small supports.

• The so-called *discretization parameter* h is "associated" with a characteristic size of triangulations used - it will be defined more precisely later.

Let now V_h be an arbitrary finite-dimensional subspace of V. Then the *Galerkin method* for approximating the solution of the problem: Find $u \in V$ such that

$$a(u, v) = F(v) \qquad \forall v \in V \ ,$$

consists of finding $u_h \in V_h$ such that

$$a(u_h, v_h) = F(v_h) \qquad \forall v_h \in V_h \ . \qquad (\dagger)$$

Applying Lax-Milgram lemma, we observe that this problem has one and only one solution u_h, which we shall call the *discrete solution.*

When $a(\cdot,\cdot)$ is symmetric, the discrete solution is also characterized by the property (see Theorem 2)

$$J(u_h) = \inf_{v_h \in V_h} J(v_h) \ ,$$

where

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

This definition of discrete solution is known as the *Ritz method.*

Note that

$$J(u) \leq J(u_h) \ ,$$

since $V_h \subset V$.

Obviously, we have the following orthogonality relation

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in V_h \ ,$$

which means that the error $u - u_h$ is orthogonal to V_h with respect to the scalar product $a(\cdot, \cdot)$.

This orthogonality relation implies that u_h is the projection of u on V_h with respect to $a(\cdot, \cdot)$ and that

$$a(u - u_h, u - u_h) = \inf_{v_h \in V_h} a(u - v_h, u - v_h) \ ,$$

i.e. the Ritz method yields the best approximation with respect to the energy norm.

Let $a(\cdot,\cdot)$ be not necessarily symmetric. And let $\{v^i\}_{i=1}^m$ be basis in V_h.

We shall look for the discrete solution u_h as a linear combination

$$u_h = \sum_{j=1}^{m} c_j v^j \ .$$

Then by (†) it is true that

$$a\left(\sum_{j=1}^{m} c_j v^j, v^i\right) = F(v^i), \qquad i = 1, \ldots, m \ .$$

This finally leads to the following system of algebraic equations:

$$\sum_{j=1}^{m} a\left(v^j, v^i\right) c_j = F(v^i), \quad i = 1, \ldots, m \ , \qquad (\ddagger)$$

for the unknowns $c_1, \ldots, c_m$.

The matrix $A = \left(a(v^j, v^i)\right)_{i,j=1}^m$ and the vector $\left(F(v^i)\right)_{i=1}^m$ are often called (by reference to elasticity problems) the *stiffness matrix* and the *load vector,* respectively.

- We prove that A is nonsingular.

From the bilinearity of $a(\cdot,\cdot)$ and the V-ellipticity condition we have

$$(A\xi, \xi) = \sum_{i,j} a(v^j, v^i)\xi_i\xi_j = a\left(\sum_j \xi_j v^j, \sum_i \xi_i v^i\right) = a(v, v) \geq$$

$$\geq C_2||v||_V^2 > 0 \quad \forall \xi = (\xi_1, \ldots, \xi_m)^T \in R^m, \quad \xi \neq 0 \ ,$$

where $v = \sum_i \xi_i v^i$, and $v \neq 0$ since $\{v^i\}$ is a basis and $\xi \neq 0$. Hence, the homogeneous equation $A\xi = 0$ cannot have a nontrivial solution $\xi \neq 0$.

The discrete solution u_h is obviously independent of basis functions $v^1, \ldots, v^m \in V_h$, whereas the structure of A depends considerably on $v^1, \ldots, v^m$. Thus, concerning the choice of the basis $\{v^i\}$, it is very important from the numerical point of view that the matrix A possesses as many zero entries as possible.

For instance, if the intersection of the supports of basis functions v^p and v^q has zero measure then $a(v^p, v^q) = 0$. That is why the condition (iii) is required. It may even happen that $a(v^p, v^q) = 0$, $p \neq q$, though the intersection of supports has a positive measure. Thus according to (iii), only a few entries (more precisely $\mathcal{O}(m)$ entries) of the $m \times m$ stiffness matrix A remain different from zero, i.e., A is sparse.

Hence, we need less computer memory and fewer arithmetic operations to solve (‡) than for full matrices which arise, in general, from the classical Galerkin method.

In order to introduce some typical V_h, we first establish a *triangulation* $\mathcal{T}_h$ over domain $\overline{\Omega}$, i.e., we subdivide the set $\overline{\Omega}$ into a finite number of subsets K (called *elements*) in such a way that the following properties hold:

(1) $\overline{\Omega} = \bigcup\limits_{K \in \mathcal{T}_h} K$,

(2) for each $K \in \mathcal{T}_h$, the set K is closed and its interior K^0 is non-empty,

(3) for each distinct $K_1, K_2 \in \mathcal{T}_h$, one has $K_1^0 \cap K_2^0 = \emptyset$,

(4) for each $K \in \mathcal{T}_h$, the boundary ∂K is the Lipschitz one.

Later we shall give further assumptions on $\mathcal{T}_h$. The *discretization parameter* h is the maximum diameter of all $K \in \mathcal{T}_h$. Note that there is a certain ambiguity in the meaning of $\mathcal{T}_h$ as for a given sufficiently small h there may exist many different triangulations $\mathcal{T}_h$. Nevertheless, we will keep this often used notation. Sometimes $\mathcal{T}_h$ is called a partition or a decomposition of $\overline{\Omega}$ into elements.

- From now on we shall write $H^k(K)$ instead of $H^k(K^0)$ for simplicity.

In what follows we shall construct finite element spaces V_h consisting of piecewise polynomial functions over $\mathcal{T}_h$. An important property of such spaces is given in the following theorem.

Theorem 4: Let $\mathcal{T}_h$ be a triangulation (decomposition) of $\overline{\Omega}$ formed by convex elements. Let V_h be a subspace of $L^2(\Omega)$ such that the space

$$P_K = \{v_h|_K \ \ | \ v_h \in V_h\}$$

consists of polynomial functions for any $K \in \mathcal{T}_h$. Then $V_h \subset H^1(\Omega)$ iff $V_h \subset C(\overline{\Omega})$, i.e. a piecewise polynomial function is from $H^1(\Omega)$ iff it is continuous.

P r o o f : So let $V_h \subset H^1(\Omega)$ and let there exist a function $v_h \in V_h$ which is not continuous. Then there exist two adjacent elements $K_1,\ K_2 \in \mathcal{T}_h$ and an open ball $B \subset K_1 \cup K_2$ such that

$$B \cap S \neq \emptyset \quad \text{and} \quad v_h|_{K_1} > v_h|_{K_2} \quad \text{on} \quad B \cap S \ ,$$

where $S = K_1 \cap K_2$. Let $w \in C_0^\infty(\Omega)$ be a "hill" function such that

$$w(x) > 0 \quad \forall x \in B \quad \text{and} \quad w(x) = 0 \quad \forall x \in \Omega \backslash B \ .$$

Denote by $n^j = (n_1^j, \ldots, n_d^j)^T$ the outward unit normal to $\partial K_j,\ j = 1, 2$. We see that S is contained in a hyperplane of R^d as K_1 and K_2 are convex. Therefore n^j is constant on S and there exists $i \in \{1, \ldots, d\}$ such that $n_i^1 \neq 0$.

Now, referring to Green's formula , we arrive at

$$0 = \int_\Omega \frac{\partial v_h}{\partial x_i} w \, dx + \int_\Omega v_h \frac{\partial w}{\partial x_i} \, dx = \sum_{j=1}^{2} \left(\int_{K_j} \frac{\partial v_h}{\partial x_i} w \, dx + \int_{K_j} v_h \frac{\partial w}{\partial x_i} \, dx \right)$$

$$= \sum_{j=1}^{2} \left(- \int_{K_j} v_h \frac{\partial w}{\partial x_i} \, dx + \int_{\partial K_j} v_h|_{K_j} w n_i^j \, ds + \int_{K_j} v_h \frac{\partial w}{\partial x_i} \, dx \right)$$

$$= \int_S \left(v_h|_{K_1} - v_h|_{K_2} \right) w n_i^1 \, ds = n_i^1 \int_{B \cap S} \left(v_h|_{K_1} - v_h|_{K_2} \right) w \, ds \ .$$

But this is a contradiction, since $n_i^1 \neq 0$ and the last integral is positive due to our assumptions upon w and the piecewise polynomial function v_h.

Conversely, let $v_h \in V_h$ be continuous. Then evidently $v_h \in L^2(\Omega)$ and we show that it has also the first generalized derivatives in $L^2(\Omega)$. Since any K has a Lipschitz boundary and $P_K \subset H^1(K)$ for all $K \in \mathcal{T}_h$, we may apply Green's formula for $i \in \{1, \ldots, d\}$ to get

$$\int_K \frac{\partial v_h}{\partial x_i} w \, dx + \int_K v_h \frac{\partial w}{\partial x_i} \, dx = \int_{\partial K} v_h|_K w n_i^K \, ds \qquad \forall w \in C_0^\infty(\Omega) \, , \qquad (\bullet)$$

where n_i^K is the ith component of the unit outward normal to ∂K. Let $z^i \in L^2(\Omega)$ be defined through the relation

$$z^i|_K = \frac{\partial v_h}{\partial x_i}, \ K \in \mathcal{T}_h \ .$$

Summing $(\bullet)$ over all the elements $K \in \mathcal{T}_h$, we obtain

$$\int_\Omega z^i w \, dx + \int_\Omega v_h \frac{\partial w}{\partial x_i} \, dx = \sum_{K \in \mathcal{T}_h} \int_{\partial K} v_h|_K w n_i^K \, ds \ .$$

We see, however, that the sum vanishes. Either a portion of ∂K is a portion of the boundary $\partial\Omega$, where $w = 0$, or the contribution of any two adjacent elements $K, K' \in \mathcal{T}_h$ is zero as $n^K + n^{K'} = 0$. Hence,

$$\int_\Omega z^i w \, dx = -\int_\Omega v_h \frac{\partial w}{\partial x_i} \, dx \quad \forall w \in C_0^\infty(\Omega)$$

and the functions $z^i,\ i = 1, \ldots, d$, are the first generalized derivatives of v_h. Thus we may write $z^i = \partial v_h / \partial x_i$ in the whole Ω. ■

The theorem can be also modified to the case when P_K contains generally nonpolynomial functions from $C(K) \cap H^1(K)$ and for nonconvex elements. This may be useful to construct some more complicated spaces of finite elements.

Suppose further that in the case $d \geq 2$ elements of $\mathcal{T}_h$ are convex polygons or polyhedra, which is fulfilled quite often. Then $\overline{\Omega}$ is, of course, a polygon or polyhedron. We add two more assumptions upon $\mathcal{T}_h$:

(5) any face of any $K \in \mathcal{T}_h$ is either a subset of the boundary $\partial\Omega$, or a face of another element $K' \in \mathcal{T}_h$,

(6) the interior of any face of any $K \in \mathcal{T}_h$ is disjoint with Γ_0, where Γ_0 is a set of those points where one type of boundary condition changes into another.

Remark: For any generally nonconvex polygon (polyhedron) $\overline{\Omega}$ there exists a triangulation $\mathcal{T}_h$ satisfying (1)–(5) such that all the elements $K \in \mathcal{T}_h$ are triangles (tetrahedra). Construction can be done as follows. Let $S^1, \ldots, S^q$ be faces of $\overline{\Omega}$ and let $L^1, \ldots, L^q$ be straight lines (planes) such that $S^j \subset L^j$. First show that components of the set $\overline{\Omega} \setminus \bigcup_{j=1}^{q} L^j$ are convex. Then cut these components into simplexes so that (5) holds.

Finite Elements

The following general definition of the finite element will be used to construct finite element spaces V_h.

Definition: The *finite element* is a triple (K, P, Σ), where:

(I) K is a closed subset of R^d with a non-empty interior and a Lipschitz boundary,

(II) P is a space of real-valued functions defined over the set K,

(III) Σ is a finite set of linearly independent linear forms $\Phi_i,\ 1 \leq i \leq N$, defined over the space P (or over a space which contains P).

The set Σ is said to be *P-unisolvent* if for any real scalars $\alpha_i,\ 1 \le i \le N$, there exists a unique function $p \in P$ which satisfies

$$\Phi_i(p) = \alpha_i \ , \qquad 1 \le i \le N \ .$$

Consequently, if Σ is P-unisolvent then there exist functions $p_i \in P$, $1 \le i \le N$, which satisfy

$$\Phi_j(p_i) = \delta_{ij} \ , \qquad 1 \le j \le N \ ,$$

where δ_{ij} is Kronecker's symbol. Since

$$\forall p \in P \qquad p = \sum_{i=1}^{N} \Phi_i(p) p_i \ ,$$

we have $\dim P = N$.

The linear forms $\Phi_i,\ 1 \le i \le N$, are called the *degrees of freedom of the finite element,* and the functions $p_i,\ 1 \le i \le N$, are called the *basis functions of the finite element.*

In what follows we introduce several examples of the most used finite elements, where K will be a convex set of a simple form and P a space of polynomials.

Functions from P are sometimes called *ansatz-functions.*

The space of all polynomials of degree at most k defined on K is denoted by $P_k(K)$.

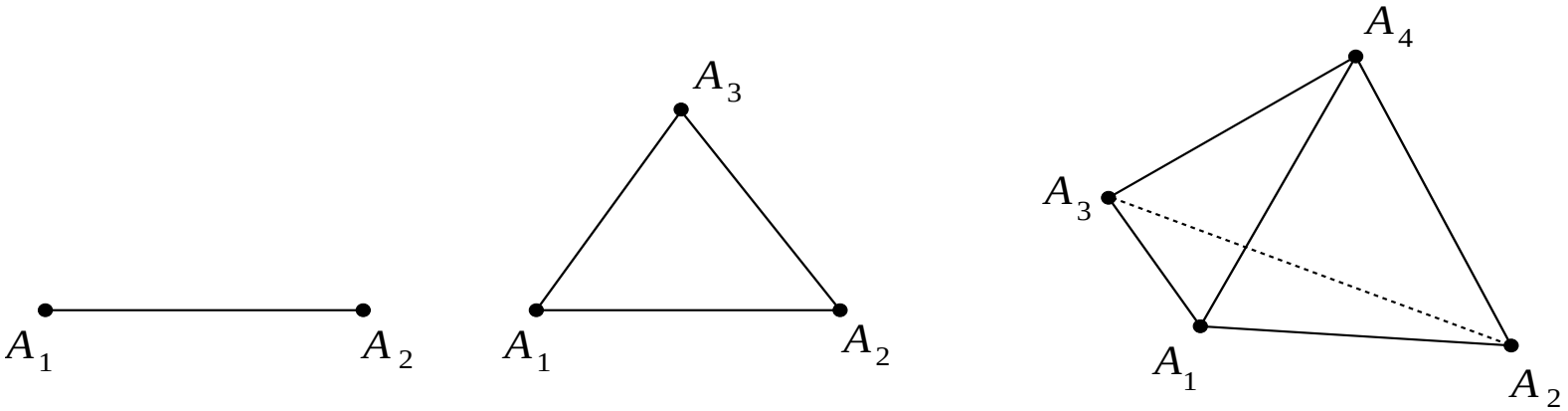
We introduce only some lower order elements, since the use of higher order elements requires certain higher smoothness of the true solution, which is usually not present in practical problems.

Linear simplicial element

The set K is a d-simplex, the space $P = P_K$ is the space $P_1(K)$ of linear functions

$$p(x_1, \ldots, x_d) = \gamma_0 + \gamma_1 x_1 + \ldots + \gamma_d x_d \ , \qquad \gamma_i \in R^1 \ ,$$

$$\dim P_1(K) = d + 1 \ .$$

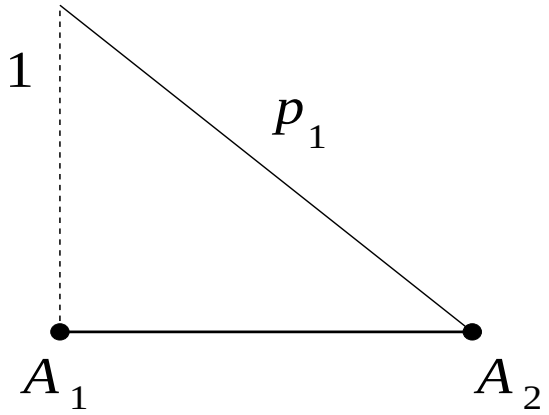


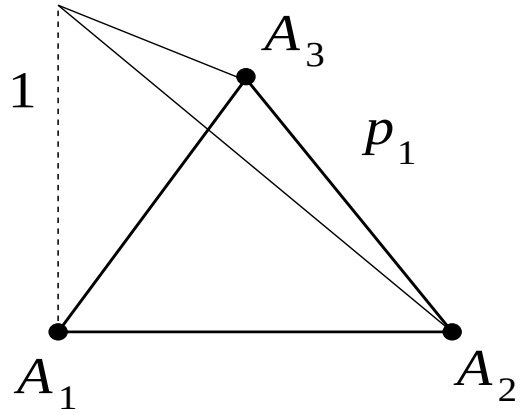
The set of degrees of freedom $\Sigma = \Sigma_K$ consists of the forms

$$\Phi_i(p) = p(A_i), \qquad 1 \le i \le d+1,\ p \in P_1(K)\ ,$$

where $A_1, \ldots, A_{d+1}$ are the vertices of K. For the sake of simplicity we shall write only symbolically

$$\Sigma_K = \{p(A_i),\ 1 \le i \le d+1\}\ .$$





In the case $d = 2$, this element is also known as Courant's triangle (element). The first basis function p_1 is sketched above for $d = 1, 2$.

Quadratic simplicial element

K is again a d-simplex, $P_K = P_2(K)$ is the space of quadratic functions

$$p(x_1, \ldots, x_d) = \gamma_0 + \sum_{i=1}^{d} \gamma_i x_i + \sum_{1 \le i \le j \le d} \gamma_{ij} x_i x_j \ , \qquad \gamma_i, \gamma_{ij} \in R^1 \ ,$$

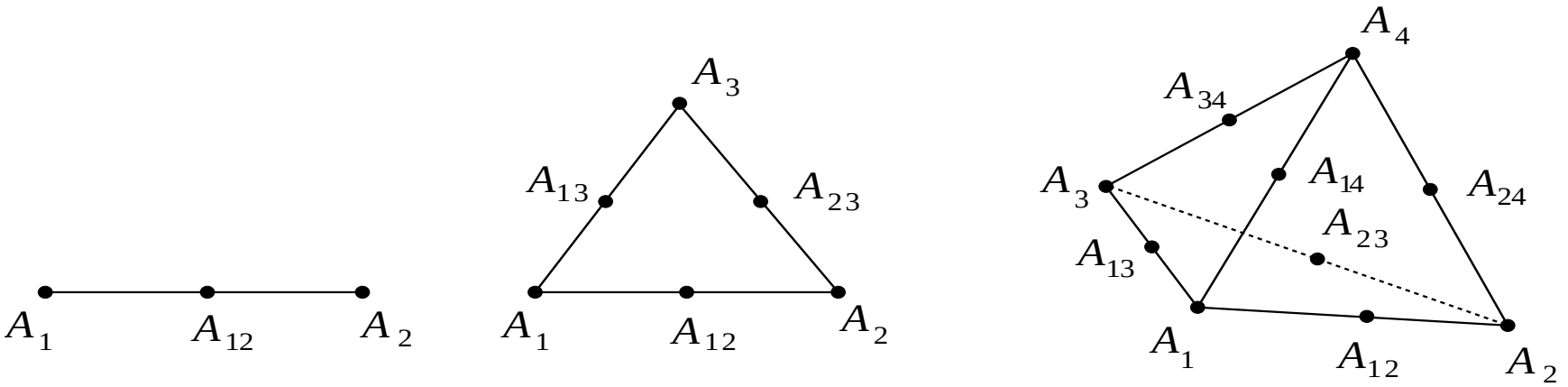
and

$$\Sigma_K = \{p(A_i),\ 1 \le i \le d+1;\ p(A_{ij}),\ 1 \le i < j \le d+1\} \ ,$$

where A_i are the vertices of K and

$$A_{ij} = \frac{1}{2}(A_i + A_j), \qquad 1 \le i < j \le d+1 \ ,$$

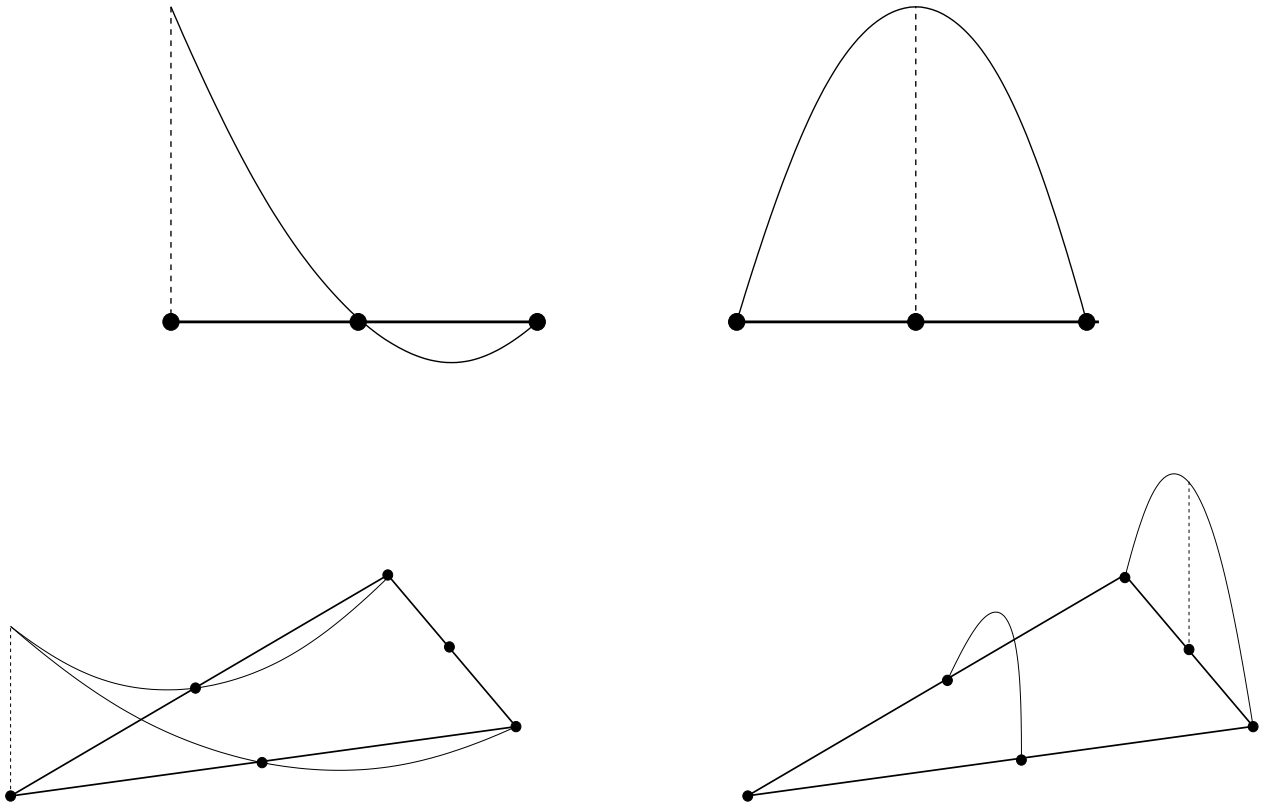
are the midpoints of the edges of the d-simplex K:



We may easily check that

$$\dim P_2(K) = \frac{1}{2}(d+1)(d+2) \ .$$

Two figures in below show two qualitatively different types of basis functions of the quadratic element for $d = 1$ and $d = 2$, respectively:



Bilinear and trilinear rectangular elements

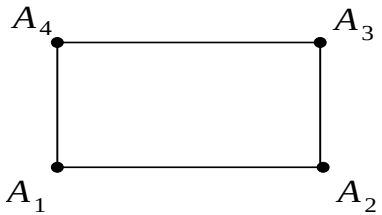
Here K is a d-rectangle, $P_K = Q_1(K)$ is the space of d-linear functions

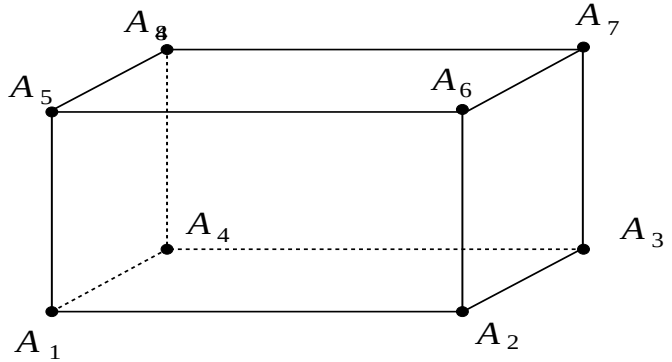
$$p(x_1, \ldots, x_d) = \sum_{0 \le q_i \le 1,\ 1 \le i \le d} \gamma_{q_1 \ldots q_d}\ x_1^{q_1} \ldots x_d^{q_d}\ , \qquad \gamma_{q_1 \ldots q_d} \in R^1$$

(for $d = 2$ and $d = 3$ these functions are called bilinear and trilinear polynomials, respectively), $\dim Q_1(K) = 2^d$ and

$$\Sigma_K = \left\{ p(A_i),\ 1 \le i \le 2^d \right\}\ ,$$

where A_i are the vertices of K.





Prismatic element

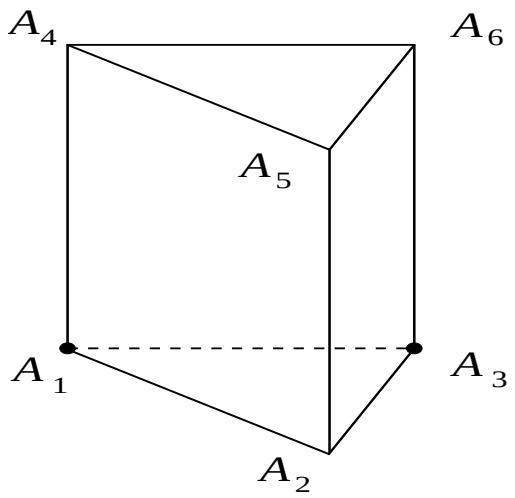
The set K is a prismatic domain. The space P_K is the tensor product of the space P_1 in variables x_1, x_2 by the space Q_1 in the variable x_3, i.e. dim $P_K = 6$ and P_K consists of polynomials of the form

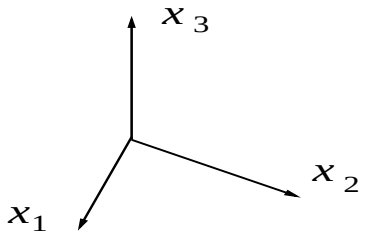
$$p(x_1, x_2, x_3) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_1 x_3 + \gamma_5 x_2 x_3 \ , \ \gamma_i \in R^1 \ ,$$

and

$$\Sigma_K = \{p(A_i) \ , \ 1 \leq i \leq 6\} \ ,$$

where A_i are again the vertices of K.





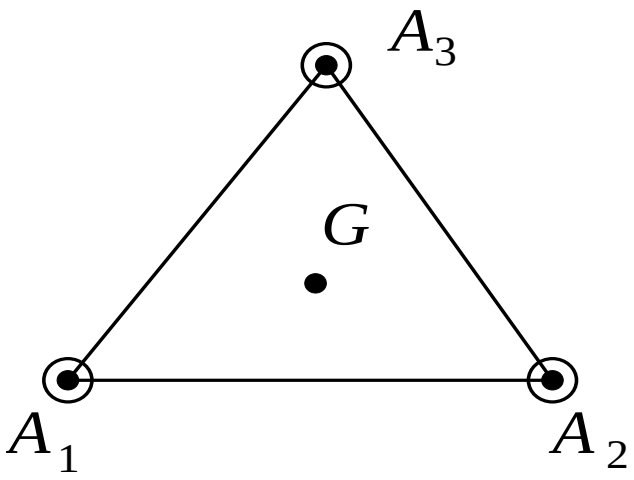
All the previous elements are called *Lagrange* finite elements since the degrees of freedom are of the form $p \mapsto p(A)$, $A \in K$. The point A is then called the node. These elements are P_K-unisolvent, i.e; given any $(\alpha_1, \ldots, \alpha_N)^T \in R^N$ there exists just one polynomial $p \in P_K$ such that, for all the nodes $A_1, \ldots, A_N$, $p(A_i) = \alpha_i$ holds.

If at least one directional derivative occurs as a degree of freedom, the associated finite element is said to be a *Hermite* finite element.

As an example we introduce the Hermite cubic element which is also P_K-unisolvent. Here K is a triangle, $P_K = P_3(K)$ is the space of cubic polynomials, and

$$\Sigma_K = \left\{ p(A_i),\ \frac{\partial p}{\partial x_1}(A_i),\ \frac{\partial p}{\partial x_2}(A_i),\ i = 1, 2, 3\ ;\ p(G) \right\}\ ,$$

where A_i are the vertices of K and G is its centre of gravity.



Some other types of degrees of freedom often used in FEM:

$$\Phi(p) = \frac{\partial p}{\partial n}(A) \text{ – normal derivative}$$

$$\Phi(p) = (D^i p)(A),\ |i| \geq 1 \text{ – higher order derivative}$$

$$\Phi(p) = \int_K p(x)\,dx \text{ – integral over the element } K$$

$\Phi(v) = n_1 v_1(A) + \cdots + n_d v_d(A)$ – normal component of vector-function $(v_1, \ldots, v_d)^T$

Note that, in engineering literature, the degrees of freedom of a finite element are sometimes called *parameters.*

Finite Element Spaces

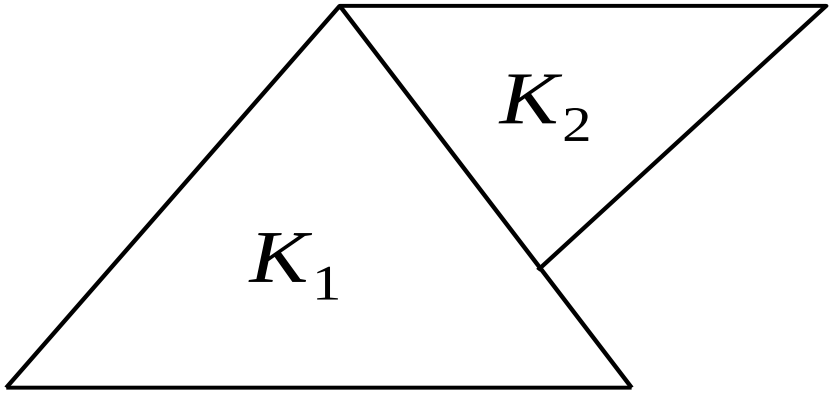
We shall give now a description of a *finite element space* generated by Lagrange elements (K, P_K, Σ_K), $K \in \mathcal{T}_h$.

The sets of degrees of freedom of adjacent finite elements will be related as follows: Whenever $(K_\ell, P_{K_\ell}, \Sigma_{K_\ell})$ with $\Sigma_{K_\ell} = \{p(A_i^\ell),\ 1 \leq i \leq N_\ell\}$, $\ell = 1, 2$, are two adjacent finite elements, then

$$\left(\bigcup_{i=1}^{N_1} \{A_i^1\}\right) \cap K_2 = \left(\bigcup_{i=1}^{N_2} \{A_i^2\}\right) \cap K_1 \ .$$

Let us denote this set by $\mathcal{N}_S$, where $S = K_1 \cap K_2$.

Note that the situation in below cannot occur due to property (5).



When we want to joint adjacent finite elements, we moreover need that

$$\{p^1|_S \mid \ \ p^1 \in P_{K_1}\} = \{p^2|_S \mid \ \ p^2 \in P_{K_2}\} \ .$$

We denote this space by P_S. Finally suppose that if $p \in P_S$ and $p(A) = 0$ for all $A \in \mathcal{N}_S$, then $p = 0$ on S.

Let us define the set

$$\mathcal{N}_h = \bigcup_{K \in \mathcal{T}_h} \mathcal{N}_K \ ,$$

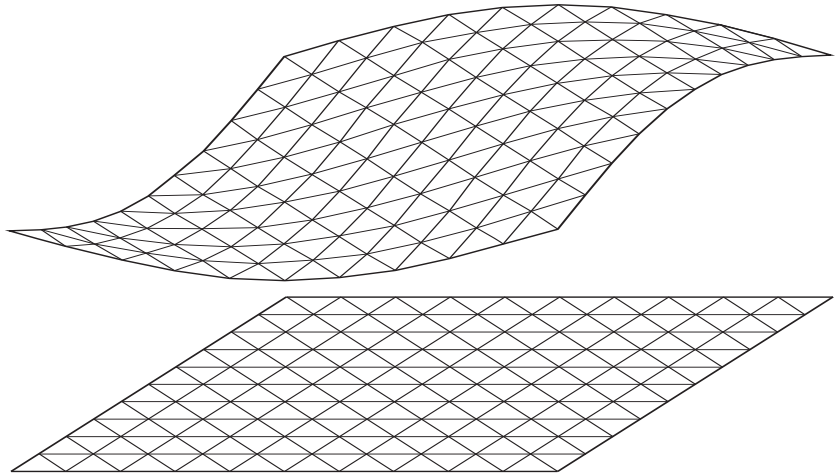
where, for each finite element (K, P_K, Σ_K), the symbol $\mathcal{N}_K$ denotes the set of nodes. The associated finite element space X_h is then given by

$$X_h = \{v_h \in C(\overline{\Omega}) \mid \ v_h|_K \in P_K \quad \forall K \in \mathcal{T}_h\} \ .$$

Therefore a function in the space X_h is uniquely determined by the set

$$\Sigma_h = \{v(A), \ A \in \mathcal{N}_h\} \ ,$$

which is called the set of degrees of freedom of the finite element space X_h. Given values at nodal points, we always get a continuous function.



An example of function from X_h defined by the linear finite elements for $d = 2$.

According to Theorem 4 and the consistency condition (6) we define finite element subspaces of the space

$$V = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_1\} \ .$$

by

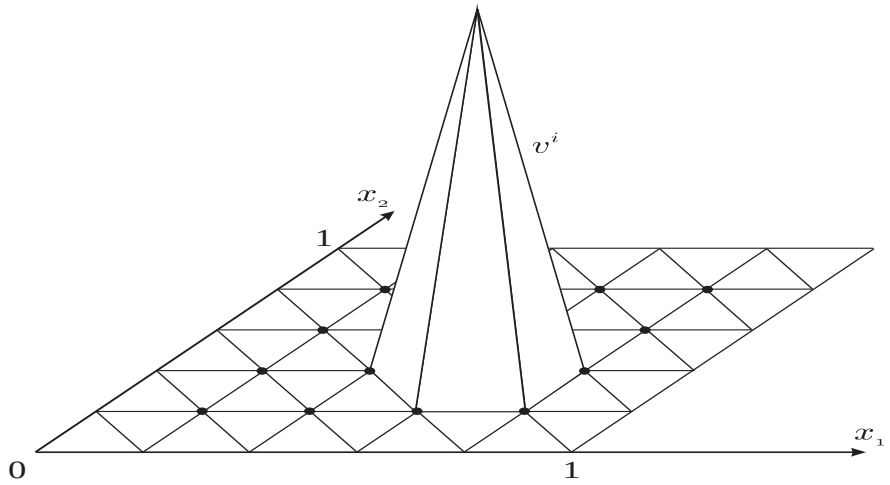
$$V_h = \{v_h \in X_h \mid v_h(A) = 0 \text{ for } A \in \mathcal{N}_h \cap \overline{\Gamma}_1\} \subset V \ .$$

It is really a subspace of V, as $v_h = 0$ on Γ_1, by our assumptions.

To guarantee the requirement (iii) we choose the basis $\{v^i\}_{i=1}^m \subset V_h$

$$v^i(A_j) = \delta_{ij} \ ,$$

where $m = \dim V_h$, $i, j = 1, \ldots, m$, $A_i \in \mathcal{N}_h$, and $A_i \notin \overline{\Gamma}_1$.



Let for example $d = 1$, $\Omega = (0, 1)$,

$$a(v, w) = \left(\frac{\partial v}{\partial x}, \frac{\partial w}{\partial x} \right)_{0,\Omega}, \quad v, w \in H_0^1(\Omega),$$

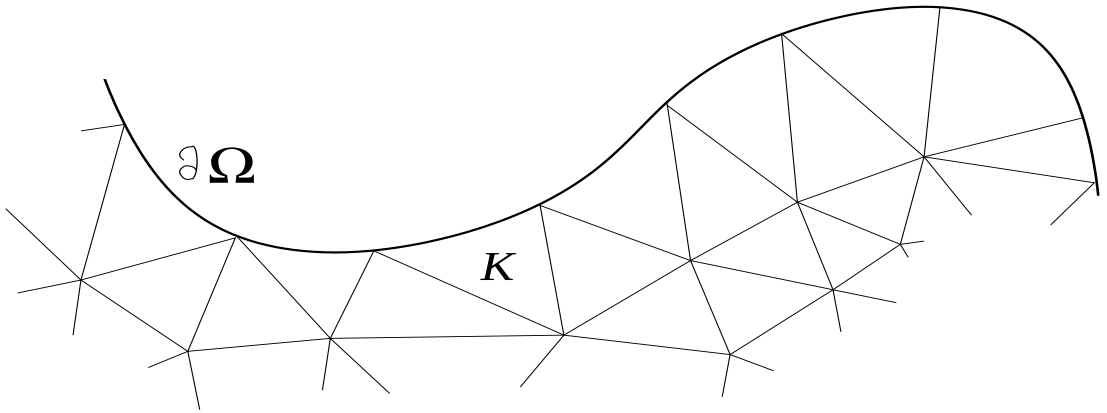
and $0 < A_1 < A_2 < \cdots < A_m < 1$. Then

$$a(v^i, v^j) = \left(\frac{\partial v^i}{\partial x}, \frac{\partial v^j}{\partial x} \right)_{0,\, \mathrm{supp} v^i \cap \mathrm{supp} v^j}$$

and the standard Courant basis functions yield $a(v^i, v^j) = 0$ whenever $|i - j| > 1$. Thus we see that the matrix $A = (a(v^i, v^j))_{i,j=1}^m$ is only tridiagonal.

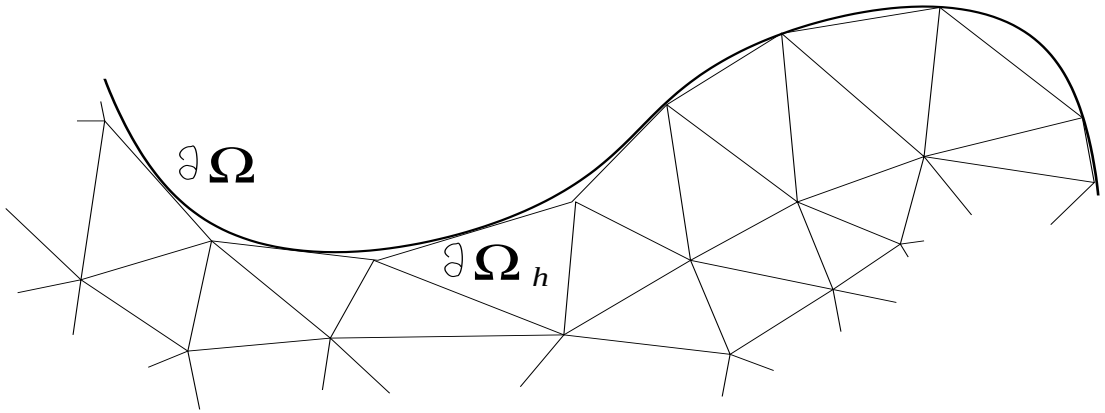
For $d > 1$ the nodes can be numbered so that A is a band matrix.

When $\partial\Omega$ is piecewise curved, there are several ways of constructing finite element spaces. One way is to generate them by the so-called isoparametric (curved) elements. Note that a curved element, K need not be convex and P_K may be formed e.g. by rational functions.



Another way is to approximate Ω by a polygonal (polyhedral) domain $\Omega_h \subset \Omega$ and then to extend finite element functions from Ω_h to the whole Ω in an appropriate manner.

Consider for instance a bounded plane domain with a Lipschitz boundary which consists of a finite number of smooth convex and concave arcs.



The sides of $\partial\Omega_h$ are chords or tangents of convex or concave arcs, respectively. The points of inflexion of $\partial\Omega$ coincide with some vertices of $\partial\Omega_h$.

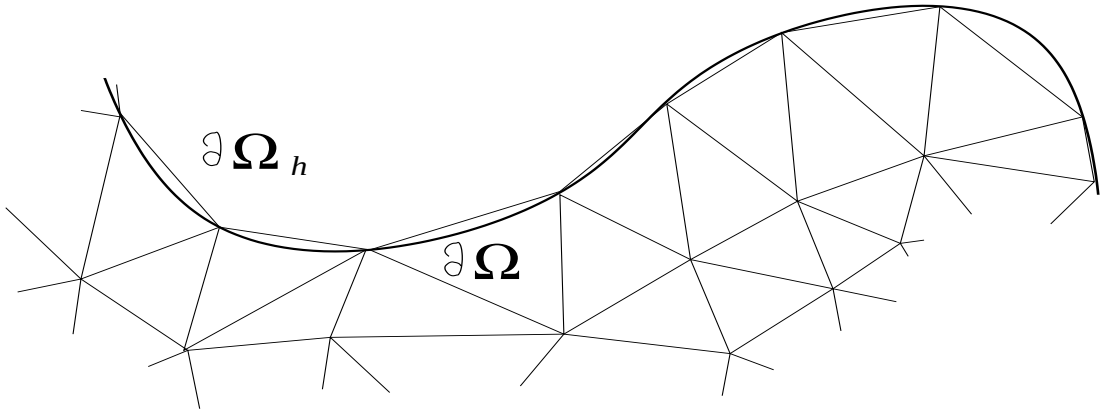
For such a polygon Ω_h there is a triangulation $\mathcal{T}_h$ consisting of triangles such that the maximum length of all their sides equals h. Let us define the space of continuous piecewise linear functions as follows

$$V_h = \{v_h \in C(\overline{\Omega}) \mid \quad v_h|_K \in P_1(K) \quad \forall K \in \mathcal{T}_h \ , \ v_h|_{\overline{\Omega}\setminus\Omega_h} = 0\} \ .$$

We see now that V_h is entirely contained in the space of test functions $V = H_0^1(\Omega)$ for the homogeneous Dirichlet problem.

Note that in the case of the mixed boundary conditions, the parts Γ_2 and Γ_3 need not be approximated by "polygonal" curves and we still have $V_h \subset V$.

When $V_h \subset V$, and when the bilinear and linear form of the discrete problem are identical to the original ones, the finite element method is said to be *conforming.* That was the case before.



A nonconforming method arises when $V_h \not\subset V$ or when e.g. some numerical integration is used. Figure in above shows another manner of boundary approximation which is often used in practice and which also yields $V_h \not\subset V$.

A nonconforming method is also obtained when the function $\overline{u} \in H^1(\Omega)$ from the Dirichlet boundary condition is approximized by a piecewise polynomial continuous function.