

# Introduction to the Finite Element Method

Marek K. Misztal  
Niels Bohr Institute  
*misztal@nbi.ku.dk*

March 28, 2016

## 1 Introduction

The finite element methods (FEM) is a family of numerical methods designed to find approximate solutions to linear, partial differential equations (PDEs) on a variety of domains. It relies on domain decomposition into simpler *elements* (triangles, tetrahedra, quadrilaterals, etc.) and interpolation of the solution based on a discrete set of values defined at specified domain sites. By applying the Galerkin method, the original PDE is converted into a *weak formulation*, which can be written as a linear system of equations. The Galerkin method ensures a solution that minimizes (although not strictly) the residual between the actual solution and the functions from the space of approximate solutions.

This note is meant as a practical guide to obtaining the linear, finite element formulations of Poisson-like PDEs, with emphasis on deriving equations suitable for numeric implementation. Because of that, this note makes use of the matrix notation, rather than tensor notation.

### 1.1 Focus problems

**Problem A (Poisson's equation)** Let  $\Omega$  be a bounded, compact subset of  $\mathbb{R}^2$ . Find such  $u : \Omega \rightarrow \mathbb{R}$ , that

$$\nabla^2 u = 0 \tag{1}$$

$$u(x)|_{\partial\Omega} = u_0(x) \tag{2}$$

**Problem B (Stokes' equation)\*** Compute the laminar, steady state flow through a finite pipe (in 2D), due to applied pressure difference between the

*inlet and the outlet.*

Here  $\Omega \subset \mathbb{R}^2$  is the geometric representation of the pipe. We introduce  $\partial\Omega_{in}$  to denote the inlet surface,  $\partial\Omega_{out}$  to denote the outlet surface,  $\partial\Omega_{in} \cap \partial\Omega_{out} = \emptyset$ . Further,  $\partial\Omega_s = \partial\Omega - (\partial\Omega_{in} \cup \partial\Omega_{out})$  denotes the solid wall. Formally, we seek such  $u : \Omega \rightarrow \mathbb{R}^2$  and  $p : \Omega \rightarrow \mathbb{R}$ , that

$$\nabla^2 u - \nabla p + f = 0, \quad (\text{Stokes' equation}) \quad (3)$$

$$\nabla \cdot u = 0, \quad (\text{continuity equation}) \quad (4)$$

subject to the Dirichlet boundary conditions

$$u(x)|_{\partial\Omega_s} = 0, \quad (5)$$

$$p(x)|_{\partial\Omega_{in}} = p_{in}, \quad (6)$$

$$p(x)|_{\partial\Omega_{out}} = p_{out}, \quad (7)$$

where  $p_{in}$  and  $p_{out}$  are constant values.

## 2 Discretization

### 2.1 Domain discretization

In order to numerically solve the PDE, we first have to approximate the domain with a finite number of simpler, geometric objects (referred to as *elements*). Popular types of elements include

- *simplicial*: segments in 1D, triangles in 2D, tetrahedra in 3D,
- quadrilateral (in 2D),
- hexahedral (in 3D).

From now we focus on simplicial domains in 2D, which are known as *triangle meshes*. Most of the following equations significantly simplify in the 1D case, and the overall procedure is very similar in the 3D case. We will denote the *vertices* (or *nodes*) of the mesh  $x_i$ ,  $i = 1, 2, \dots, N$ , and the triangle elements  $\Omega^e$ ,  $e = 1, 2, \dots, M$ . Typically, the number of elements in the mesh  $M$  scales linearly with the number of vertices  $N$ . Elements are only allowed to intersect along the common edge (no T-junctions allowed). We define the entire discrete domain as

$$\Omega = \bigcup_{e=1}^M \Omega^e. \quad (8)$$

Once we have decided on the domain type, we have to decide on how to discretize the solution. In this note, we focus on linear schemes, i.e. ones where the approximate solution is given as a piece-wise linear function. Higher-order schemes are often more robust, however, they require maintaining the values of the solution's derivatives. Having a piecewise linear approximation of the solution  $\tilde{u}(x)$  requires us to store only a finite number of its values, typically at the vertices or elements' centres. For now we will focus on vertex-centred schemes, i.e. where we specify  $u_i = \tilde{u}(x_i)$ , for  $i = 1, 2, \dots, N$ .

## 2.2 Solution space

The decisions we have made about the discretization determines the *solution space*, i.e. the set of admissible solutions. These are constructed through interpolation, based on the nodal values, using *interpolant* functions  $\varphi_i$  for  $i = 1, 2, \dots, N$ , such that

$$\varphi_i(x_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}, \quad \sum_{i=1}^N \varphi_i(x) = 1. \quad (9)$$

Then the interpolated (approximate) solution is defined as

$$\tilde{u}(x) = \sum_{i=1}^N u_i \varphi_i(x) = \hat{\varphi}(x)^T \hat{\mathbf{u}}, \quad (10)$$

where  $u_i = u(x_i)$  for  $i = 1, 2, \dots, N$  are the (stored) nodal values of the solution,  $\hat{\varphi}(x) = [\varphi_1(x), \varphi_2(x), \dots, \varphi_N(x)]^T$  and  $\hat{\mathbf{u}} = [u_1, u_2, \dots, u_N]^T$ . Formally, the solution space is defined as:

$$\mathcal{S} = \{f : \Omega \rightarrow T; \exists (\hat{\boldsymbol{\alpha}} \in T^N) f(x) = \hat{\varphi}(x)^T \hat{\boldsymbol{\alpha}}\}. \quad (11)$$

Here we focus on linear elements with linear shape functions. In this case  $\nabla \varphi_i(x)|_{\Omega_e} = \text{const}$ . The gradient of the interpolated function  $\tilde{u}$  is well-defined in the interior of each element  $\Omega_e$ , and is given by

$$\nabla \tilde{u}(x) = \sum_{i=1}^N u_i \nabla \varphi_i(x) = (\nabla \hat{\varphi}(x))^T \hat{\mathbf{u}}, \quad (12)$$

where

$$\mathbb{R}^{N \times 2} \ni \nabla \hat{\varphi}(x) = \begin{bmatrix} \partial_x \varphi_1(x) & \partial_x \varphi_2(x) & \dots & \partial_x \varphi_N(x) \\ \partial_y \varphi_1(x) & \partial_y \varphi_2(x) & \dots & \partial_y \varphi_N(x) \end{bmatrix}^T. \quad (13)$$

## 2.3 Galerkin method

The Galerkin method allows us to find the quasi-best approximation of the real solution to the PDE among the functions from the solution space  $\mathcal{S}$ . The specific details of the mathematics behind the Galerkin method are out of scope of this note, and are available in specialized literature. In short, the Galerkin solution  $u_g \in \mathcal{S}$  is found by transforming the PDE

$$\mathcal{L}u = 0, \quad (14)$$

where  $\mathcal{L}$  is a linear operator (for example  $\mathcal{L} = \nabla^2$  for the Poisson problem) into its *weak formulation*

$$\forall (\tilde{w} \in \mathcal{T}) \langle \tilde{w}, \mathcal{L}\tilde{u} \rangle = 0, \quad (15)$$

where  $\langle \cdot \rangle$  is the inner product in the space of the solution functions, and  $\mathcal{T}$  is the set of admissible *test functions* ( $\tilde{w}$ ),  $\mathcal{T} = \{v \in \mathcal{S} : v(x)|_{\partial\Omega} = 0\}$ . Under the  $L^2$ -norm, this becomes

$$\forall (\tilde{w} \in \mathcal{T}) \int_{\Omega} \tilde{w}(x)^T \mathcal{L}\tilde{u}(x) \, d\Omega = 0. \quad (16)$$

## 3 Poisson's equation

### 3.1 Galerkin method in practice

The weak formulation of the Problem A reads

$$\forall (\tilde{w} \in \mathcal{T}) \int_{\Omega} \tilde{w}^T \nabla^2 \tilde{u} \, d\Omega = 0, \quad (17)$$

where  $\tilde{u}, \tilde{w} : \mathbb{R}^2 \rightarrow \mathbb{R}$ . By applying the divergence theorem, we obtain

$$\int_{\Omega} \tilde{w}^T \nabla^2 \tilde{u} \, d\Omega = \oint_{\partial\Omega} \tilde{w}^T (\nabla \tilde{u})^T \bar{n} \, dS - \int_{\Omega} (\nabla \tilde{w})^T \nabla \tilde{u} \, d\Omega, \quad (18)$$

where  $\bar{n}$  is the normal vector to  $\partial\Omega$ . Note that because  $w(x)|_{\partial\Omega} = 0$ , the first term on the right hand side vanishes. Now recall that  $\nabla \tilde{u}(x) = (\nabla \hat{\varphi}(x))^T \hat{\mathbf{u}}$  and  $\nabla \tilde{w}(x) = (\nabla \hat{\varphi}(x))^T \hat{\mathbf{w}}$ , which allows us to rewrite Eq. (18) into a linear equation

$$\int_{\Omega} \nabla \tilde{w}(x)^T \nabla \tilde{u}(x) \, d\Omega = \int_{\Omega} \hat{\mathbf{w}}^T \nabla \hat{\varphi}(x) (\nabla \hat{\varphi}(x))^T \hat{\mathbf{u}} \, d\Omega \quad (19)$$

$$= \hat{\mathbf{w}}^T \left\{ \int_{\Omega} \nabla \hat{\varphi}(x) (\nabla \hat{\varphi}(x))^T \, d\Omega \right\} \hat{\mathbf{u}} \quad (20)$$

$$= \hat{\mathbf{w}}^T \mathbf{K} \hat{\mathbf{u}}, \quad (21)$$

where  $\mathbf{K}$  is an  $N \times N$  real matrix (called the *stiffness matrix* of the PDE), with coefficients defined as  $K_{ij} = \int_{\Omega} \nabla \varphi_i(x) \cdot \nabla \varphi_j(x) d\Omega$ . We have arrived at an equivalent form of Eq. (17)

$$\forall (\hat{\mathbf{w}} \in \mathbb{R}^N) \hat{\mathbf{w}}^T \mathbf{K} \hat{\mathbf{u}} = 0, \quad (22)$$

which is fulfilled by  $\hat{\mathbf{u}}$  such that

$$\mathbf{K} \hat{\mathbf{u}} = \mathbf{0}. \quad (23)$$

### 3.2 Matrix assembly\*

We can learn more about the structure of matrix  $\mathbf{K}$  by rewriting the integral over  $\Omega$  as a sum of integrals over each element  $\Omega^e$

$$\mathbf{K} = \int_{\Omega} \nabla \hat{\varphi}(x) (\nabla \hat{\varphi}(x))^T d\Omega = \sum_{e=1}^M \int_{\Omega^e} \nabla \hat{\varphi}^e (\nabla \hat{\varphi}^e)^T d\Omega = \sum_{e=1}^M \mathbf{K}^e, \quad (24)$$

where  $\nabla \hat{\varphi}^e = \nabla \hat{\varphi}(x)|_{\Omega^e} = \text{const}$ , and  $\mathbf{K}^e$  is the *element stiffness matrix*. In the 2D case with triangle elements, only three shape functions are non-zero over a given element  $\Omega^e$  (the three shape functions associated with the triangle's vertices). That means, only three out of  $N$  rows of  $\nabla \hat{\varphi}^e$  are non-zero. That also means that  $\mathbf{K}^e$  has at most nine non-zero coefficients, forming a  $3 \times 3$  submatrix

$$\mathbf{K}_3^e = |\Omega^e| \begin{bmatrix} \nabla \varphi_i^e \cdot \nabla \varphi_i^e & \nabla \varphi_i^e \cdot \nabla \varphi_j^e & \nabla \varphi_i^e \cdot \nabla \varphi_k^e \\ \nabla \varphi_j^e \cdot \nabla \varphi_i^e & \nabla \varphi_j^e \cdot \nabla \varphi_j^e & \nabla \varphi_j^e \cdot \nabla \varphi_k^e \\ \nabla \varphi_k^e \cdot \nabla \varphi_i^e & \nabla \varphi_k^e \cdot \nabla \varphi_j^e & \nabla \varphi_k^e \cdot \nabla \varphi_k^e \end{bmatrix}, \quad (25)$$

where  $|\Omega^e|$  is the volume (area) of the element  $\Omega^e$ . Note that matrix  $\mathbf{K}$  has at most  $9M$  non-zero coefficients, which is one order of magnitude lower than its size  $N^2$ . This *sparsity* property is important from the point of view of numeric performance, allowing to solve the linear system using fast, iterative solvers (such as the conjugate gradient, GMRES, etc.).

### 3.3 Dirichlet boundary conditions

In order to complete solving Problem A we have to enforce the Dirichlet boundary conditions

$$u(x)|_{\partial\Omega} = u_0(x). \quad (26)$$

In terms of the piecewise-linear function  $\tilde{u}(x)$  represented by a discrete vector  $\hat{\mathbf{u}}(x)$ , such boundary conditions now read

$$u_i = \tilde{u}(x_i) = \tilde{u}_0(x_i) = u_{0i} \quad \text{for } i \in I_D, \quad (27)$$

where  $I_D$  is the set of indices of the nodes lying on the Dirichlet boundary, and  $u_{0i}$  are the prescribed nodal values. There are two ways of enforcing such boundary conditions: the first one reduces to elimination of the unknown values corresponding to the indices from  $I_D$  from the linearised form of the PDE (and appropriately modifying the stiffness matrix  $\mathbf{K}$ ). The second one is the *Lagrange multipliers* method. It requires stating the Dirichlet conditions in a matrix form

$$\mathbf{B}\hat{\mathbf{u}} = \hat{\mathbf{u}}_0, \quad (28)$$

where  $\hat{\mathbf{u}}, \hat{\mathbf{u}}_0 \in \mathbb{R}^{N_D}$ ,  $N_D$  is the number of the boundary nodes, and the matrix  $\mathbf{B} \in \mathbb{R}^{N_D \times N}$  is constructed so that each row corresponds a single index from  $I_D$ , and has a single non-zero value (typically 1) at the column corresponding to the index of the boundary vertex. In order to incorporate the boundary conditions into the Galerkin system of equations, a new, larger system is solved

$$\begin{bmatrix} \mathbf{K} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \hat{\mathbf{u}}_0 \end{bmatrix}, \quad (29)$$

where  $\boldsymbol{\lambda} \in \mathbb{R}^{N_D}$  is the vector of Lagrange multipliers corresponding to the Dirichlet boundary conditions.

### 3.4 Vector functions\*

All the formulas above remain valid if we operate with vector-valued functions, e.g. when  $u : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  in the Poisson problem. However, such representation is not best suited for numerical implementation. In practice, we often expand the discrete solution vector

$$\mathbb{R}^{2N} \ni \hat{\mathbf{u}} = [u_1, u_2, \dots, u_N]^T \quad (30)$$

into the “long” form

$$\mathbb{R}^{2N} \ni \tilde{\mathbf{u}} = [u_{1x}, u_{1y}, u_{2x}, u_{2y}, \dots, u_{Nx}, u_{Ny}]^T. \quad (31)$$

For the relation  $\tilde{u}(x) = \hat{\boldsymbol{\varphi}}(x)^T \hat{\mathbf{u}} = \check{\boldsymbol{\varphi}}(x)^T \tilde{\mathbf{u}}$  to hold, we have to replace

$$\mathbb{R}^N \ni \hat{\boldsymbol{\varphi}}(x) = [\varphi_1(x), \varphi_2(x), \dots, \varphi_N(x)]^T \quad (32)$$

with

$$\mathbb{R}^{2N \times 2} \ni \check{\boldsymbol{\varphi}}(x) = [\mathbf{I}_2 \varphi_1(x), \mathbf{I}_2 \varphi_2(x), \dots, \mathbf{I}_2 \varphi_N(x)]^T, \quad (33)$$

where  $\mathbf{I}_2$  is the  $2 \times 2$  identity matrix. Finally, in order to avoid introducing 3-tensors, we define

$$\nabla \check{\boldsymbol{\varphi}}(x) = \begin{bmatrix} \partial_x \varphi_1(x) & 0 & \partial_x \varphi_2(x) & \cdots & \partial_x \varphi_N(x) & 0 \\ 0 & \partial_x \varphi_1(x) & 0 & \cdots & 0 & \partial_x \varphi_N(x) \\ \partial_y \varphi_1(x) & 0 & \partial_y \varphi_2(x) & \cdots & \partial_y \varphi_N(x) & 0 \\ 0 & \partial_y \varphi_1(x) & 0 & \cdots & 0 & \partial_y \varphi_N(x) \end{bmatrix}^T. \quad (34)$$

This way

$$(\nabla \check{\boldsymbol{\varphi}}(x))^T \check{\mathbf{u}} = \begin{bmatrix} \sum_i u_{ix} \partial_x \varphi_i(x) \\ \sum_i u_{iy} \partial_x \varphi_i(x) \\ \sum_i u_{ix} \partial_y \varphi_i(x) \\ \sum_i u_{iy} \partial_y \varphi_i(x) \end{bmatrix} = \begin{bmatrix} \partial_x \tilde{u}_x(x) \\ \partial_x \tilde{u}_y(x) \\ \partial_y \tilde{u}_x(x) \\ \partial_y \tilde{u}_y(x) \end{bmatrix} \equiv \nabla \tilde{u}(x). \quad (35)$$

Note that under this notation  $(\nabla \tilde{w})^T \nabla \tilde{u}$ , as seen in Eq. (18) corresponds to the full contraction between the Jacobians of  $\tilde{w}$  and  $\tilde{u}$ . After establishing those relations we can repeat the procedure from Section 3.1, obtaining the new stiffness matrix  $\mathbb{R}^{2N \times 2N} \ni \mathbf{K}_2 = \int_{\Omega} \nabla \check{\boldsymbol{\varphi}}(x) (\nabla \check{\boldsymbol{\varphi}}(x))^T d\Omega$ .

## 4 Stokes' equation\*

We now focus on Problem B.

$$\mu \nabla^2 u - \nabla p + f = 0, \quad (36)$$

$$\nabla \cdot u = 0. \quad (37)$$

Notice that this time we are seeking two functions  $u : \Omega \rightarrow \mathbb{R}^2$  and  $p : \Omega \rightarrow \mathbb{R}$ . There are many approaches to discretizing those two fields, however here we will only focus on the (minimal) linear, *staggered* approach, where  $u$  is sampled at mesh nodes and interpolated using linear shape functions (as in the previous example), however  $p$  is sampled at the centers of the elements and is constant over each element;  $p(x) = \check{\boldsymbol{\xi}}(x)^T \check{\mathbf{p}}$ , where  $\check{\mathbf{p}} \in \mathbb{R}^M$  is the vector of discrete pressure values and  $\check{\boldsymbol{\xi}}(x) \in \mathbb{R}^M$  is the vector of element-wise constant shape functions. The weak formulation Eq. (36) reads

$$\forall (\tilde{w} \in \mathcal{T}) \int_{\Omega} \tilde{w}^T [\mu \nabla^2 \tilde{u} - \nabla \tilde{p} + \tilde{f}] d\Omega = 0, \quad \tilde{u} \in \mathcal{S}, \tilde{p} \in \mathcal{P}, \quad (38)$$

where  $\mathcal{S}$  and  $\mathcal{T}$  are defined as in the previous section, and  $\mathcal{P}$  is the pressure solution space, spanned by  $\check{\boldsymbol{\xi}}(x)$ . The first term  $\int_{\Omega} \mu \tilde{w}^T \nabla^2 \tilde{u} d\Omega$  we have already linearised in the previous section. The third term  $\int_{\Omega} \tilde{w}^T \tilde{f} d\Omega$  is left

to the reader as an exercise, and will be omitted in the remainder of this note. What remains is the pressure term, to which we apply the divergence theorem

$$\int_{\Omega} \tilde{w}^T \nabla \tilde{p} \, d\Omega = \int_{\partial\Omega} \tilde{p} \tilde{w}^T \bar{n} \, dS - \int_{\Omega} (\nabla \cdot \tilde{w})^T \tilde{p} \, d\Omega. \quad (39)$$

Again, the first term on the right hand side vanishes, and we rewrite the second term as

$$\int_{\Omega} (\nabla \cdot \tilde{w})^T \tilde{p} \, d\Omega = \int_{\Omega} \tilde{\mathbf{w}}^T (\nabla \cdot \tilde{\boldsymbol{\varphi}}) \tilde{p} \, d\Omega = \tilde{\mathbf{w}}^T \int_{\Omega} (\nabla \cdot \tilde{\boldsymbol{\varphi}}) \tilde{p} \, d\Omega. \quad (40)$$

where  $\nabla \cdot \tilde{\boldsymbol{\varphi}}(x) = [\partial_x \varphi_1(x), \partial_y \varphi_1(x), \partial_x \varphi_2(x), \dots, \partial_x \varphi_N(x), \partial_y \varphi_N(x)]^T$ . Now, rather than expanding  $\tilde{p}(x) = \boldsymbol{\xi}(x)^T \tilde{\mathbf{p}}$ , we skip right to the decomposition into sum of integrals, while remembering that  $\tilde{p}$  is constant in each element,  $p(x)|_{\Omega^e} = p_e$ .

$$\tilde{\mathbf{w}}^T \int_{\Omega} (\nabla \cdot \tilde{\boldsymbol{\varphi}}) \tilde{p} \, d\Omega = \tilde{\mathbf{w}}^T \sum_{e=1}^M \left\{ \int_{\Omega^e} (\nabla \cdot \tilde{\boldsymbol{\varphi}}) \, d\Omega \right\} p_e = \tilde{\mathbf{w}}^T \mathbf{P} \tilde{\mathbf{p}}, \quad (41)$$

where  $\mathbf{P} \in \mathbb{R}^{2N \times M}$ , and the  $e$ -th column  $\mathbf{P}^e$  of  $\mathbf{P}$  is given as

$$\mathbf{P}^e = |\Omega^e| [\partial_x \varphi_1|_{\Omega^e}, \partial_y \varphi_1|_{\Omega^e}, \dots, \partial_x \varphi_N|_{\Omega^e}, \partial_y \varphi_N|_{\Omega^e}]^T. \quad (42)$$

Once again, due to the properties of the linear shape functions, at most six out of  $2N$  values in each column are non-zero. Finally, Eq. (38) becomes (save for the body force term)

$$\forall (\tilde{\mathbf{w}} \in \mathbb{R}^{2N}) \quad \tilde{\mathbf{w}}^T (\mu \mathbf{K} \tilde{\mathbf{u}} + \mathbf{P} \tilde{\mathbf{p}}) = 0, \quad (43)$$

which is fulfilled as long as

$$\mu \mathbf{K} \tilde{\mathbf{u}} + \mathbf{P} \tilde{\mathbf{p}} = \mathbf{0}. \quad (44)$$

## 4.1 Continuity equation

The weak formulation of the continuity equation  $\nabla \cdot u = 0$  can be easily obtained by substituting  $\tilde{u} = (\tilde{\boldsymbol{\varphi}}(x))^T \tilde{\mathbf{u}}$  into the volume integral of  $\nabla \cdot u$

$$\int_{\Omega} \nabla \cdot \tilde{u} \, d\Omega = \int_{\Omega} (\nabla \cdot \tilde{\boldsymbol{\varphi}}(x))^T \tilde{\mathbf{u}} \, d\Omega = \sum_{e=1}^M \left\{ \int_{\Omega^e} (\nabla \cdot \tilde{\boldsymbol{\varphi}})^T \, d\Omega \right\} \tilde{\mathbf{u}} \quad (45)$$

$$= \mathbf{P}^T \tilde{\mathbf{u}}. \quad (46)$$



Finally, the full system of linear equations of the steady state flow (without the boundary conditions) reads

$$\begin{bmatrix} \mu\mathbf{K} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \check{\mathbf{u}} \\ \check{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \quad (47)$$

Note that in the absence of boundary conditions, the solution to this system is trivially  $\check{\mathbf{u}} = \mathbf{0}$ ,  $\check{\mathbf{p}} = \mathbf{0}$ .