Numerical methods for Partial Differential Equations

Adriano Festa
Politecnico of Turin, Italy

Undergraduate Lecture at KSU
2022

Table of Contents

THE MODEL OF ELASTIC MEMBRANE

The model for the elastic membrane extends to two space dimensions the one for the elastic string; its deduction follows similar guidelines.

Consider a thin elastic membrane, whose mean cross section occupies a bounded region Ω in the plane x_1x_2 in the rest position.

Let us assume the membrane is fixed along all its boundary $\partial\Omega$.

A (small) density of force per unit volume $\boldsymbol{f} = 0\boldsymbol{e}_1 + 0\boldsymbol{e}_2 + f_3\boldsymbol{e}_3$, normal to the median cross section, induces a (small) displacement $\boldsymbol{u} = u_1\boldsymbol{e}_1 + u_2\boldsymbol{e}_2 + u_3\boldsymbol{e}_3$ from the rest position.

As a first approximation, the components u_1 and u_2 will be considered negligible with respect to the component u_3 describing the displacement along the direction of the force.

Let us set $f = f_3$, $u = u_3$ and let $\boldsymbol{\tau} = (\tau_{31}, \tau_{32})$ be the vector collecting the vertical components of the shear stress; let $\mu > 0$ denote the shear modulus of the material.

The equilibrium condition of the membrane is expressed as

$$f + \left(\frac{\partial \tau_{31}}{\partial x} + \frac{\partial \tau_{32}}{\partial y} \right) = 0 \qquad \text{namely,} \qquad f + \nabla \cdot \boldsymbol{\tau} = 0 \ ,$$

whereas the (approssimate) constitutive equations are given by

$$\tau_{31} = \mu \frac{\partial u}{\partial x} \ , \quad \tau_{32} = \mu \frac{\partial u}{\partial y} \ , \qquad \text{namely,} \qquad \boldsymbol{\tau} = \mu \nabla u \ .$$

They hold inside Ω, wheread on the boundary $\partial\Omega$ the condition

$$u = 0$$

holds, telling that the membrane is fixed at the rim.

Thus, we obtain the system

$$\begin{cases} \nabla \cdot \boldsymbol{\tau} + f = 0 & \text{in } \Omega \ , \\ \boldsymbol{\tau} = \mu \nabla u & \text{in } \Omega \ , \\ u = 0 & \text{on } \partial\Omega \ . \end{cases}$$

Substituting the expression of $\boldsymbol{\tau}$ from the second equation into the first equation produces the boundary value problem

$$\begin{cases} -\nabla \cdot (\mu \nabla u) = f & \text{in } \Omega \,, \\ u = 0 & \text{on } \partial\Omega \,, \end{cases}$$

in the unknown u. Note that the equation in Ω is, more explicitly,

$$-\frac{\partial}{\partial x}\left(\mu \frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(\mu \frac{\partial u}{\partial y}\right) = f \,.$$

In case the coefficient μ is constant on Ω, we obtain the *Poisson equation*

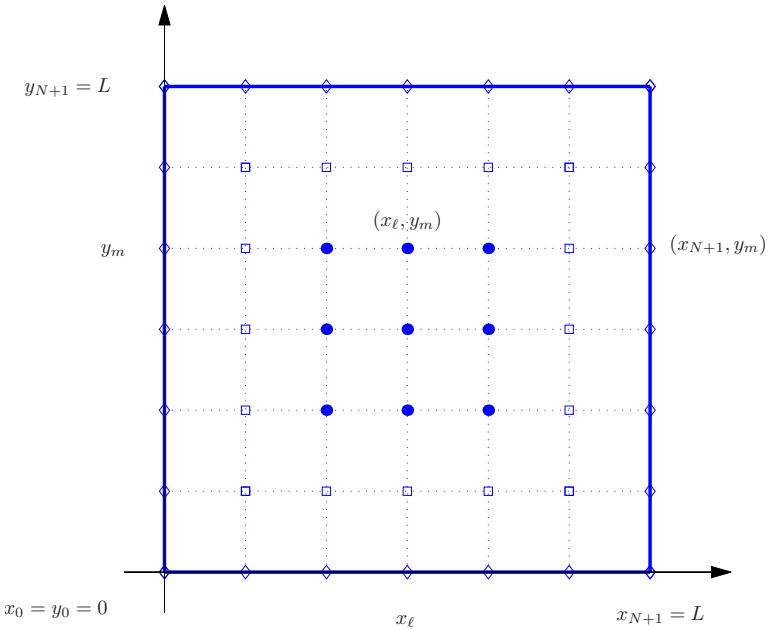
$$-\mu\, \Delta u = f \qquad \text{in } \Omega \,,$$

where the expression

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

denotes the *Laplacian* of the function u.

DISCRETIZATION BY FINITE DIFFERENCES



We assume that μ is constant and that $\Omega = (0, L) \times (0, L)$ is the square of edgelength L.

Let us fix the same discretization step $h = \dfrac{L}{N+1}$ (with $N \geq 1$) in both space directions.

Consider the equispaced grid $\mathcal{G}_h$ in $\overline{\Omega} = [0, L]^2$ given by the nodes (x_ℓ, y_m) with $x_\ell = \ell h$ for $0 \leq \ell \leq N + 1$, and $y_m = mh$ for $0 \leq m \leq N + 1$.

Note that the nodes sitting inside the square are indexed by ℓ, m such that $1 \leq \ell, m \leq N$, whereas the boundary nodes are characterised by having at least one index equal a 0 or $N + 1$.

Let us denote by $u_{\ell m} \simeq u(x_\ell, y_m)$ an approximation of the displacement u at the node (x_ℓ, y_m); moreover, let us set $f_{\ell m} = f(x_\ell, y_m)$. At boundary nodes there is no displacement, whence we define

$$u_{\ell m} = 0 \qquad \text{if} \quad \ell \in \{0, N+1\} \ \text{ or } \ m \in \{0, N+1\} \, .$$

The remaining values of $u_{\ell m}$, relative to inner nodes, are defined by imposing at each of them an approximate version of the Poisson equation, obtained by substituting the **second partial derivatives** of u in the term Δu with **centred second difference quotients**.

Precisely, noting that $x_\ell \pm h = x_{\ell \pm 1}$ and $y_m \pm h = y_{m \pm 1}$, we have

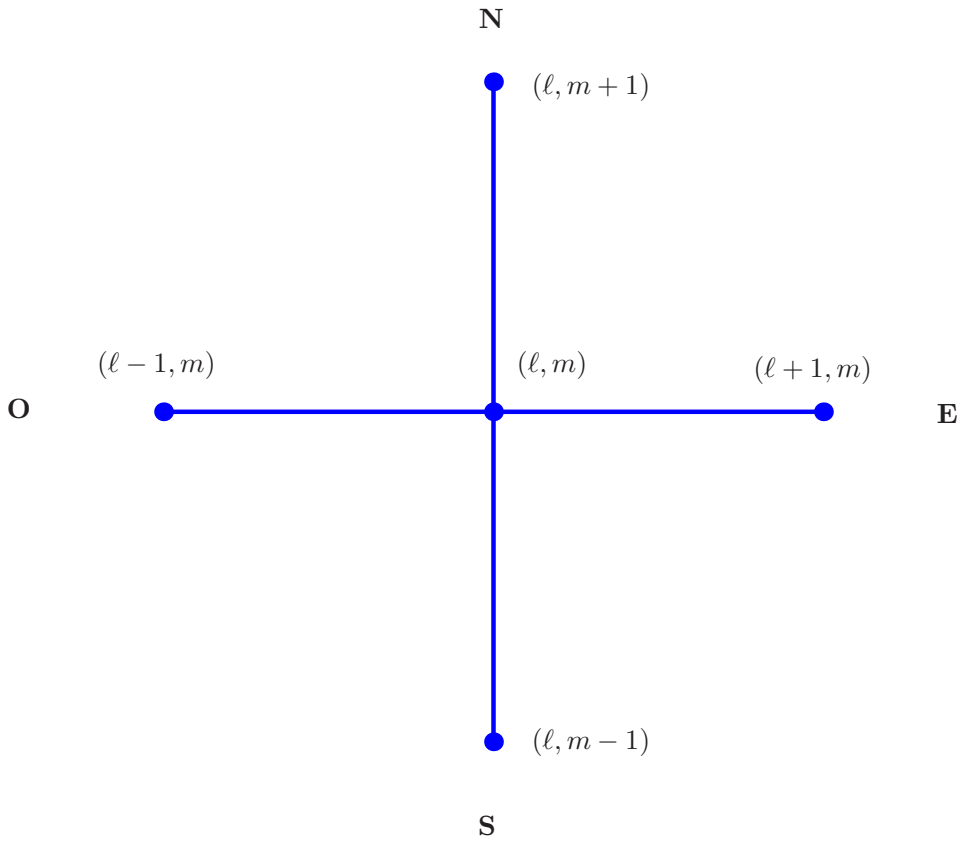
$$\frac{\partial^2 u}{\partial x^2}(x_\ell, y_m) \simeq \frac{u(x_{\ell-1}, y_m) - 2u(x_\ell, y_m) + u(x_{\ell+1}, y_m)}{h^2} \simeq \frac{u_{\ell-1,m} - 2u_{\ell m} + u_{\ell+1,m}}{h^2}$$

and

$$\frac{\partial^2 u}{\partial y^2}(x_\ell, y_m) \simeq \frac{u(x_\ell, y_{m-1}) - 2u(x_\ell, y_m) + u(x_\ell, y_{m+1})}{h^2} \simeq \frac{u_{\ell,m-1} - 2u_{\ell m} + u_{\ell,m+1}}{h^2} \, .$$

Therefore at inner nodes we impose the following equations:

$$\frac{\mu}{h^2} \left(-u_{\ell,m-1} - u_{\ell-1,m} + 4u_{\ell m} - u_{\ell+1,m} - u_{\ell,m+1} \right) = f_{\ell m}, \qquad 1 \le \ell, m \le N \, .$$

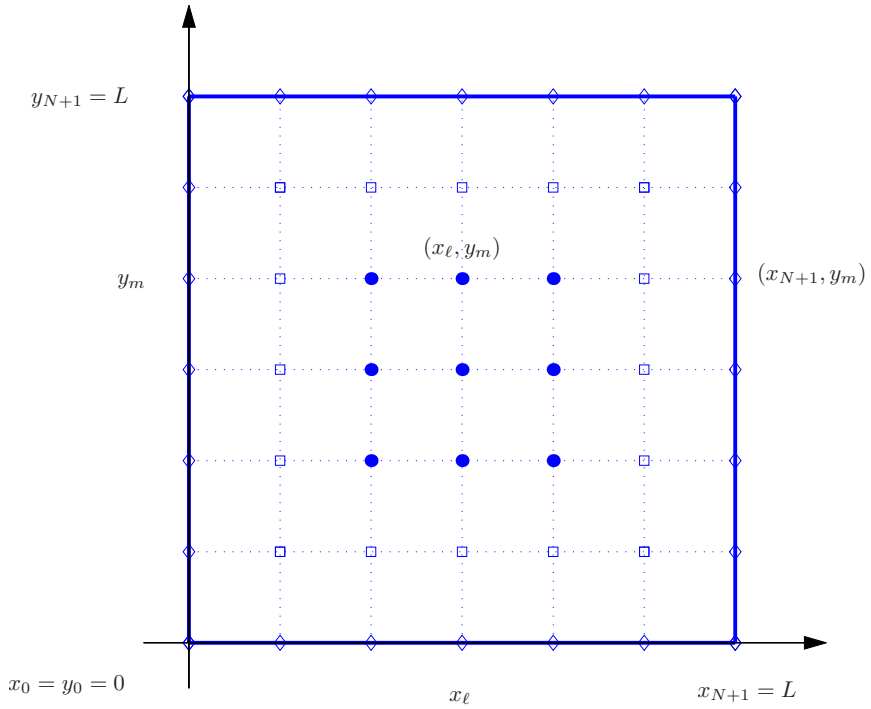


Thus, at each inner node we impose the equation

$$\frac{\mu}{h^2}\left(-u_{\ell,m-1} - u_{\ell-1,m} + 4u_{\ell m} - u_{\ell+1,m} - u_{\ell,m+1}\right) = f_{\ell m}, \qquad 1 \le \ell, m \le N\ ,$$

which involves the five nodes of the **computational molecule** centred at (x_ℓ, y_m).

Therefore, we have N^2 equations in the N^2 unknowns given by the values $u_{\ell m}$ at the internal nodes.

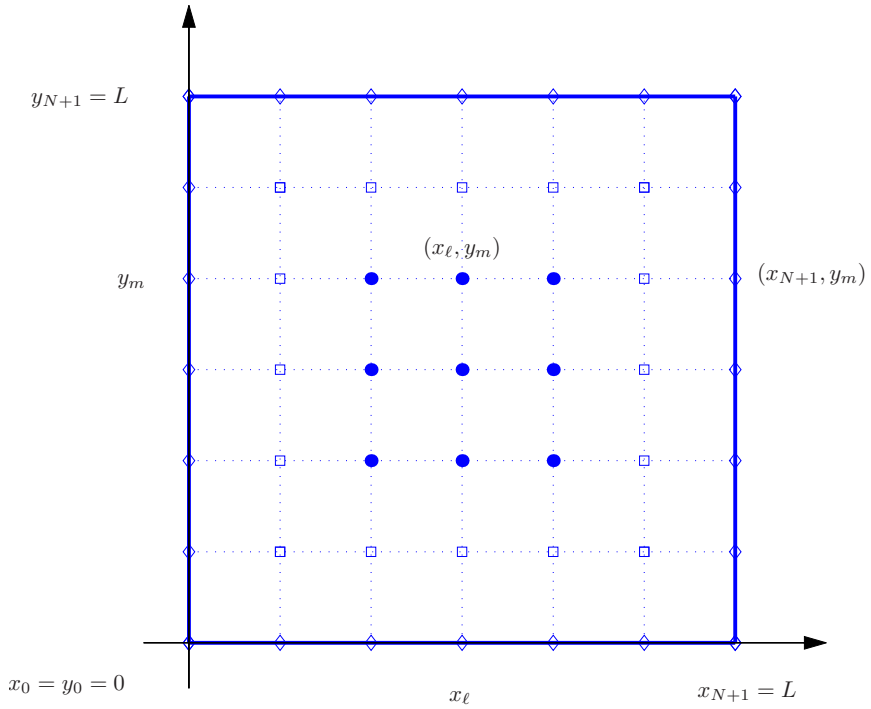


Whenever one or more nodes of the molecule sit on the boundary $\partial\Omega$, we use in the previous expression the boundary values

$$u_{\ell m} = 0 \qquad \text{if} \quad \ell \in \{0, N+1\} \ \text{ or } \ m \in \{0, N+1\} \ .$$

It is convenient to distinguish between a

- **strong** inner node: all the nodes of its computational molecule are *inner nodes*.
- **weak** inner node: one or more nodes of its molecule are *boundary nodes*.



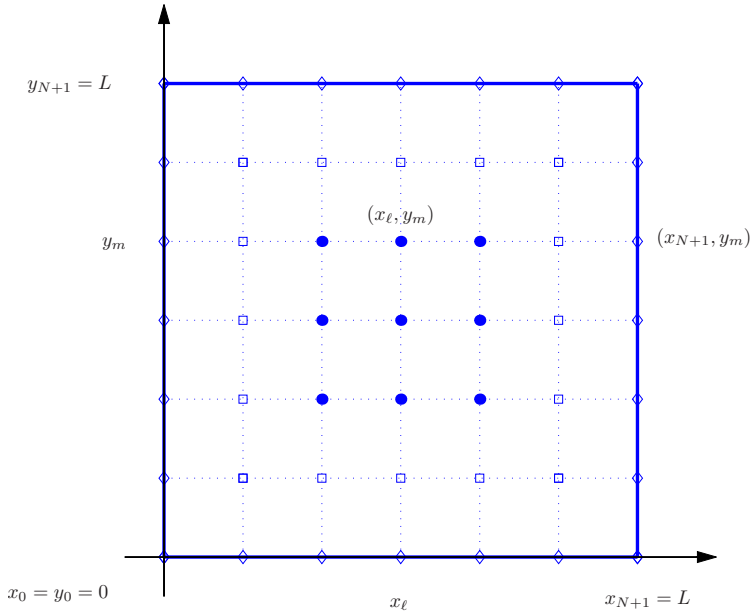
Thus, each equation

$$\frac{\mu}{h^2}\left(-u_{\ell,m-1} - u_{\ell-1,m} + 4u_{\ell m} - u_{\ell+1,m} - u_{\ell,m+1}\right) = f_{\ell m}$$

involves

- five inner unknowns, if it is associated with a **strong** inner node,
- four or three inner unknowns, if it is associated with a **weak** inner node.

Examples



- The equation at the node $(1, 1)$ takes the form

$$\frac{\mu}{h^2}\left(4u_{11} - u_{21} - u_{12}\right) = f_{11}\ ;$$

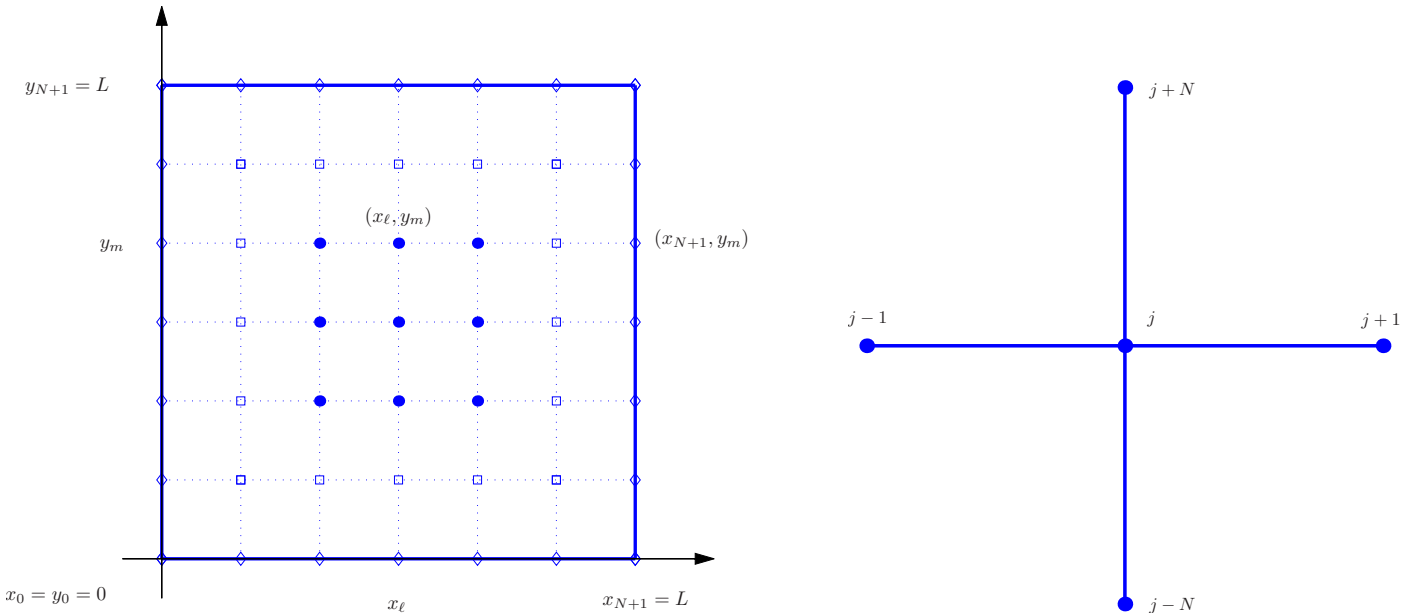
- the equation at a weak node $(\ell, 1)$ with $2 \leq \ell \leq N-1$ takes the form

$$\frac{\mu}{h^2}\left(-u_{\ell-1,1} + 4u_{\ell 1} - u_{\ell+1,1} - u_{\ell 2}\right) = f_{\ell 1}\ ;$$

- the equation at a weak node (N, m) with $2 \leq m \leq N-1$ takes the form

$$\frac{\mu}{h^2}\left(-u_{N,m-1} - u_{N-1,m} + 4u_{Nm} - u_{N,m+1}\right) = f_{Nm}\ .$$

Single-index(*lexicographic*) ordering of the unknowns



In order to write the algebraic system thus obtained in the form $\boldsymbol{Au} = \boldsymbol{f}$, we need first to shift from the two-index numbering of the unknowns to a single-index numbering:

$$(\ell, m) \qquad \longleftrightarrow \qquad j \ .$$

This is accomplieshed by resorting to the l'**lexicographical ordering**, whereby the N nodes on the first row from the bottom ($m = 1$) are ordered first, then those on the second row ($m = 2$), and so on; within each row the indexing goes from left to right.

It is not difficult to see that

$$j = \ell + (m-1)N \qquad \text{for} \quad 1 \leq \ell, m \leq N \ ,$$

and, obviously, one has

$$1 \leq j \leq N^2 \ .$$

Let us set $u_j = u_{\ell m}$ and $f_j = f_{\ell m}$. With such a notation, the equation enforced at a **strong** *inner node*, namely

$$\frac{\mu}{h^2}\left(-u_{\ell,m-1} - u_{\ell-1,m} + 4u_{\ell m} - u_{\ell+1,m} - u_{\ell,m+1}\right) = f_{\ell m} \ ,$$

becomes

$$\frac{\mu}{h^2}\left(-u_{j-N} - u_{j-1} + 4u_j - u_{j+1} - u_{j+N}\right) = f_j \ .$$

The entries of the corresponding row (the j-th one, indeed) of the matrix $\boldsymbol{A}$ are given by

$$a_{jk} = \frac{\mu}{h^2}\begin{cases} 4 & \text{if} \quad k = j \ , \\ -1 & \text{if} \quad k = j \pm 1 \text{ or } \quad k = j \pm N \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

The equation relative to a **weak** *inner nodes* involves only the unknowns (4 or 3) associated to the inner nodes of the computational molecule.

The corresponding row of the matrix differs from the one of a strong inner node by having only 3 or 2 non-null entries off the main diagonal.

- For instance, the equation relative to the weak node $(1,1)$,

$$\frac{\mu}{h^2}\left(4u_{11} - u_{21} - u_{12}\right) = f_{11} \ ,$$

  is expressed as

$$\frac{\mu}{h^2}\left(4u_{1} - u_{2} - u_{1+N}\right) = f_{1} \ ;$$

- the equation relative to the weak node $(\ell,1)$ with $2 \leq \ell \leq N-1$,

$$\frac{\mu}{h^2}\left(-u_{\ell-1,1} + 4u_{\ell 1} - u_{\ell+1,1} - u_{\ell 2}\right) = f_{\ell 1} \ ,$$

  is expressed as

$$\frac{\mu}{h^2}\left(-u_{j-1} + 4u_{j} - u_{j+1} - u_{j+N}\right) = f_{j}$$

  with $j = \ell$;

- the equation relative to the weak node (N, m) with $2 \leq m \leq N-1$,

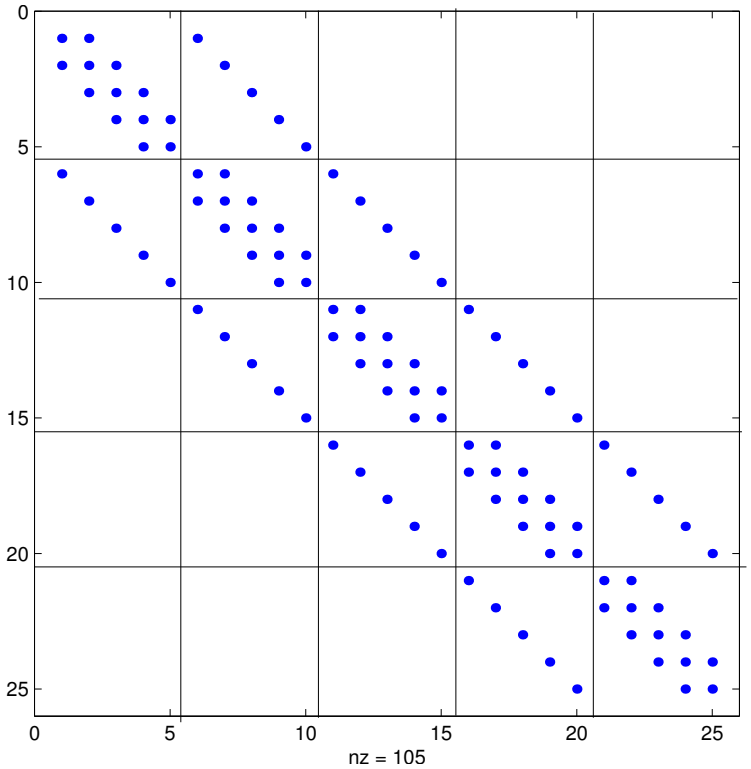
$$\frac{\mu}{h^2}\left(-u_{N,m-1} - u_{N-1,m} + 4u_{Nm} - u_{N,m+1}\right) = f_{Nm} \ ,$$

  is expressed as

$$\frac{\mu}{h^2}\left(-u_{j-N} - u_{j-1} + 4u_j - u_{j+N}\right) = f_j$$

  with $j = mN$.

Structure of the matrix A



The entries in the upper triangular part of the matrix are given by

$$a_{jk} = \frac{\mu}{h^2} \begin{cases} 4 & \text{if } \; k = j \; , \\ -1 & \text{if } \; k = j+1 \text{ with } \; j \neq pN \; , \\ -1 & \text{if } \; k = j+N \; , \\ 0 & \text{otherwise .} \end{cases}$$

The matrix is symmetric and **pentadiagonal**.

We immediately notice the **block tridiagonal** structure of the matrix, obtained by grouping the entries of the vector $\boldsymbol{u}$ of unknowns into blocks $\boldsymbol{u}_m$ of N elements, corresponding to the nodes on the same row of the grid, i.e., the nodes indexed by ℓ, m with m given.

Introducing the matrices of order N

$$\boldsymbol{D} = \frac{\mu}{h^2}\text{tridiag}\,[-1 \;\; 4 \;\; -1] \qquad \text{and} \qquad \boldsymbol{C} = -\frac{\mu}{h^2}\boldsymbol{I}\ ,$$

we have indeed

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{D} & \boldsymbol{C} & & & \\ \boldsymbol{C} & \boldsymbol{D} & \boldsymbol{C} & & \\ & \boldsymbol{C} & \boldsymbol{D} & \boldsymbol{C} & \\ & & \boldsymbol{C} & \boldsymbol{D} & \boldsymbol{C} \\ & & & \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}\ ,$$

namely,

$$\boldsymbol{A} = \text{tridiag}\,[\boldsymbol{C} \;\; \boldsymbol{D} \;\; \boldsymbol{C}]\ .$$

Properties of the discretization

- Gerschgorin's theorem shows that the eigenvalues of $\boldsymbol{A}$ are all contained in the interval

$$\left(0, \frac{8\mu}{h^2}\right)$$

hence, $\boldsymbol{A}$ is a symmetric and **positive-definite** matrix.

- The condition number worsen as the grid becomes finer, with the same behaviour as for the elastic string:

$$\mathrm{cond}_2(\boldsymbol{A}) \simeq ch^{-2} \simeq CN^2 \ .$$

- The discretization schemes yields a quadratic convergence:

$$\max_{0\leq \ell,m\leq N} |u(x_\ell, y_m) - u_{\ell,m}| \leq Ch^2 \max_{x\in\overline{\Omega}} \left(\left| \frac{\partial^4 u}{\partial x^4} \right| + \left| \frac{\partial^4 u}{\partial y^4} \right| \right) \ .$$

VARIATIONAL FORMULATION OF THE ELASTIC MEMBRANE PROBLEM

Let us recall the equilibrium equation of the membrane and the constraint of adherence to the rim:

$$\begin{cases} -\nabla \cdot (\mu \nabla u) = f & \text{in } \Omega \,, \\ u = 0 & \text{on } \partial\Omega \,. \end{cases}$$

Let us denote by V the set (vector space) of all **admissible displacements** of the membrane. An element in V is a function defined in $\overline{\Omega} = \Omega \cup \partial\Omega$, fulfilling suitable conditions of continuity and differentiability, and vanishing on $\partial\Omega$. The admissible displacements will be also termed *shape functions*, or *test functions*.
Note that the solution u is a particular admissible displacement, hence $u \in V$.

Let us multiply the equilibrium equation by the generic admissible displacement v and let us integrate on Ω:

$$-\int_\Omega \nabla \cdot (\mu \nabla u) v \, d\boldsymbol{x} = \int_\Omega f v \, d\boldsymbol{x} \,.$$

Next, we are going to "integrate by parts" on the left-hand side.

To this end, let us set $\boldsymbol{w} = \mu \nabla u$. We now recall the Divergence Theorem

$$\int_\Omega \nabla \cdot \boldsymbol{g} \, d\boldsymbol{x} = \int_{\partial\Omega} \boldsymbol{g} \cdot \boldsymbol{n} \, ds \, ,$$

where $\boldsymbol{g}$ is a vector field defined in Ω and $\boldsymbol{n}$ is the normal unit vector to the boundary $\partial\Omega$, pointing *outward* Ω.

Let us apply such result to the vector field $\boldsymbol{g} = \boldsymbol{w} v$, obtaining

$$\int_\Omega \nabla \cdot (\boldsymbol{w} v) \, d\boldsymbol{x} = \int_{\partial\Omega} (\boldsymbol{w} v) \cdot \boldsymbol{n} \, ds = \int_{\partial\Omega} (\boldsymbol{w} \cdot \boldsymbol{n}) \, v \, ds \, .$$

Next, using the differentiation rule for a product (Leibniz's rule), we have

$$\nabla \cdot (\boldsymbol{w} v) = (\nabla \cdot \boldsymbol{w}) \, v + \boldsymbol{w} \cdot \nabla v \, .$$

Hence,

$$\int_\Omega (\nabla \cdot \boldsymbol{w}) \; v \, d\boldsymbol{x} + \int_\Omega \boldsymbol{w} \cdot \nabla v \, d\boldsymbol{x} = \int_{\partial\Omega} (\boldsymbol{w} \cdot \boldsymbol{n}) \, v \, ds \, ,$$

i.e.,

$$- \int_\Omega (\nabla \cdot \boldsymbol{w}) \; v \, d\boldsymbol{x} = \int_\Omega \boldsymbol{w} \cdot \nabla v \, d\boldsymbol{x} - \int_{\partial\Omega} (\boldsymbol{w} \cdot \boldsymbol{n}) \, v \, ds \, .$$

Recalling that $\boldsymbol{w} = \mu \nabla u$, we obtain

$$-\int_\Omega \nabla\cdot(\mu\nabla u)\,v\,d\boldsymbol{x} = \int_\Omega \mu\nabla u\cdot\nabla v\,d\boldsymbol{x} - \int_{\partial\Omega}\mu\frac{\partial u}{\partial n}\,v\,ds\ ,$$

where we have indicated by

$$\frac{\partial u}{\partial n} = \nabla u\cdot\boldsymbol{n}$$

the *normal derivative* of u along $\partial\Omega$.

Finally, on the right-hand side, let us note that the boundary integral vanishes since all admissible displacements v vanish on $\partial\Omega$. In conclusion, the following identity holds:

$$-\int_\Omega \nabla\cdot(\mu\nabla u)\,v\,d\boldsymbol{x} = \int_\Omega \mu\nabla u\cdot\nabla v\,d\boldsymbol{x}\ .$$

This brings us to the following **integral formulation**, or **variational formulation**, of the elastic membrane problem:

$$\begin{cases} u\in V \text{ and satisfies} \\ \displaystyle\int_\Omega \mu\nabla u\cdot\nabla v\,d\boldsymbol{x} = \int_\Omega fv\,d\boldsymbol{x} \qquad \text{for all } \ v\in V\ . \end{cases} \tag{42}$$

From a physical point of view, it expressed the **Principle of Virtual Works**. Note the perfect analogy with the variational formulation of the 1D elastic string problem.

Discrete variational formulation

The rigorous definition of the vector space V containing all admissible displacements is such to warrant that problem (42) admits one, and one only, solution.

Among the functions of V are the continuous ones with continuous first partial derivatives on $\overline{\Omega}$. But V also contains functions that are continuous and only *piecewise differentiable* on $\overline{\Omega}$: this means there is a finite partition of $\overline{\Omega}$ into closed regions $C_1, \ldots, C_m$ such that v has continuous and bounded first partial derivatives inside each C_i.

The presence in V of the latter type of functions allows for an easy numerical approximation of the variational problem. It is enough to restrict the choice of the admissible displacements to those belonging to a subspace V_h of V; these are termed the **discrete admissible displacements**. "Piecewise polynomial" continuous functions are an important example.
The subspace has *finite dimension*, i.e., it is spanned by the linear combinations of N discrete admissible displacements $\varphi_1,\ \varphi_2, \ldots, \varphi_N$, which are linearly independent:

$$v_h(\boldsymbol{x}) = \sum_{j=1}^{N} v_j \varphi_j(\boldsymbol{x}) \qquad \text{for all} \quad v_h \in V_h \ ,$$

namely

$$V_h = \text{span}\left\{\varphi_1,\ \varphi_2, \ldots, \varphi_N\right\} .$$

The functions φ_j are the **basis functions** in V_h.

Replacing V with V_h in the formulation (42), we obtain the following **discrete variational formulation** of the elastic membrane problem:

$$\begin{cases} u_h \in V_h \text{ and satisfies} \\ \displaystyle\int_\Omega \mu \nabla u_h \cdot \nabla v_h \, d\boldsymbol{x} = \int_\Omega f v_h \, d\boldsymbol{x} \qquad \text{for all } \; v_h \in V_h \; . \end{cases} \tag{43}$$

By means of the basis $\{\varphi_1, \; \varphi_2, \ldots, \varphi_N\}$ di V_h, we can reduce this problem to an algebraic system

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$$

of N equations in N unknowns.

Indeed, choosing subsequently $v_h = \varphi_j$ for $j = 1, 2, \ldots$ in (43), we immediately obtain the N equations satisfied by u_h:

$$\int_\Omega \mu \nabla u_h \cdot \nabla \varphi_j \, d\boldsymbol{x} = \int_\Omega f \varphi_j \, d\boldsymbol{x} \; , \qquad 1 \le j \le N \; .$$

If we represent u_h with respect to the basis of V_h, as $u_h = \sum_{k=1}^{N} u_k \varphi_k$, and if we use the linearity property of differentiation and integration, we get

$$\sum_{k=1}^{N} u_k \int_\Omega \mu \nabla \varphi_k \cdot \nabla \varphi_j \, d\boldsymbol{x} = \int_\Omega f \varphi_j \, d\boldsymbol{x} \,, \qquad 1 \leq j \leq N \,.$$

Hence, the formulation (43) is translated into the following algebraic system:

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f} \,, \tag{44}$$

where we have set

$$\boldsymbol{A} = (a_{jk}) \in \mathbb{R}^{N \times N} \,, \qquad \boldsymbol{u} = (u_k) \in \mathbb{R}^N \,, \qquad \boldsymbol{f} = (f_j) \in \mathbb{R}^N \,,$$

with

$$a_{jk} = \int_\Omega \mu \nabla \varphi_k \cdot \nabla \varphi_j \, d\boldsymbol{x} \,, \qquad f_j = \int_\Omega f \varphi_j \, d\boldsymbol{x} \,. \tag{45}$$

The matrix $\boldsymbol{A}$ will be called the **stiffness matrix** of the system.

Theorem

The matrix

$$\boldsymbol{A} = (\, a_{jk} \,) \,, \qquad a_{jk} = \int_\Omega \mu \nabla \varphi_k \cdot \nabla \varphi_j \, d\boldsymbol{x} \,,$$

is symmetric and positive definite.

Proof. The matrix is obviously symmetric. Let us show that it is positive definite by checking the equivalent condition

$$\boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} > 0 \qquad \text{for each vector} \quad \boldsymbol{v} \neq \boldsymbol{0} \,.$$

The j-th entry of the vector $\boldsymbol{A}\boldsymbol{v}$ is given by

$$(\boldsymbol{A}\boldsymbol{v})_j = \sum_{k=1}^{N} a_{jk} v_k = \sum_{k=1}^{N} v_k \int_\Omega \mu \nabla \varphi_k \cdot \nabla \varphi_j \, d\boldsymbol{x} = \int_\Omega \mu \nabla \left(\sum_{k=1}^{N} v_k \varphi_k \right) \cdot \nabla \varphi_j \, d\boldsymbol{x} \,,$$

by the linearity property of differentiation and integration. Hence,

$$\begin{aligned}\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} &= \sum_{j=1}^{N} v_j(\boldsymbol{A}\boldsymbol{v})_j = \sum_{j=1}^{N} v_j \int_\Omega \mu\nabla\left(\sum_{k=1}^{N} v_k\varphi_k\right)\cdot\nabla\varphi_j\,d\boldsymbol{x}\\ &= \int_\Omega \mu\nabla\left(\sum_{k=1}^{N} v_k\varphi_k\right)\cdot\nabla\left(\sum_{j=1}^{N} v_j\varphi_j\right)\,d\boldsymbol{x}\ .\end{aligned}$$

Let us denote by $v_h \in V_h$ the function

$$v_h = \sum_{k=1}^{N} v_k\varphi_k = \sum_{j=1}^{N} v_j\varphi_j$$

(note that the two expressions coincide, since k and j are just "dummy" indices of summation).
Thus, we have

$$\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} = \int_\Omega \mu\nabla v_h\cdot\nabla v_h\,d\boldsymbol{x} = \int_\Omega \mu\|\nabla v_h\|^2\,d\boldsymbol{x} \geq 0$$

for any $\boldsymbol{v}\in\mathbb{R}^N$, since the elastic coefficient μ is strictly positive.

Our check is complete if we show that

$$\boldsymbol{v}^T \boldsymbol{A}\boldsymbol{v} = 0 \qquad \text{implies} \qquad \boldsymbol{v} = \mathbf{0}\ .$$

If $\boldsymbol{v}^T \boldsymbol{A}\boldsymbol{v} = \int_\Omega \mu \|\nabla v_h\|^2\, d\boldsymbol{x} = 0$, then necessarily

$$\mu \|\nabla v_h\|^2 = 0 \quad \text{at each point in } \Omega\ ,$$

hence,

$$\nabla v_h = \mathbf{0} \quad \text{at each point in } \Omega\ ,$$

i.e, the function

$$v_h \quad \text{is constant in } \Omega\ ;$$

on the other hand, since v_h vanishes on the boundary of Ω, necessarily one has

$$v_h = 0 \quad \text{at each point in } \Omega\ .$$

From the relation so obtained,

$$v_h = \sum_{k=1}^{N} v_k \varphi_k = 0\ ;$$

we deduce that $v_k = 0$ for each k, since the functions φ_k are linearly independent by assumption. Hence, recalling that $\boldsymbol{v} = (v_k)$, we conclude that $\boldsymbol{v} = \mathbf{0}$.

DISCRETIZATION BY FINITE ELEMENTS

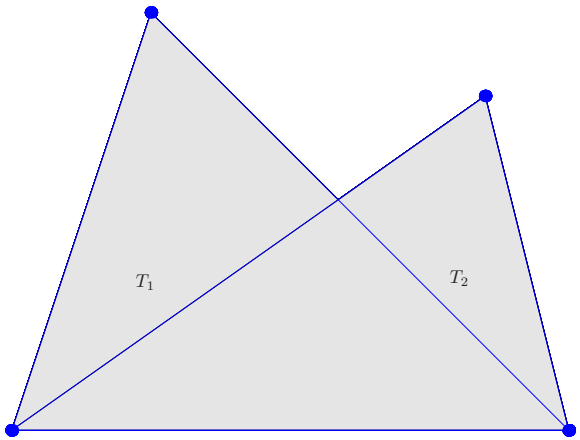
From now on, let us assume that $\overline{\Omega} = \Omega \cup \partial\Omega$ is a polygon.

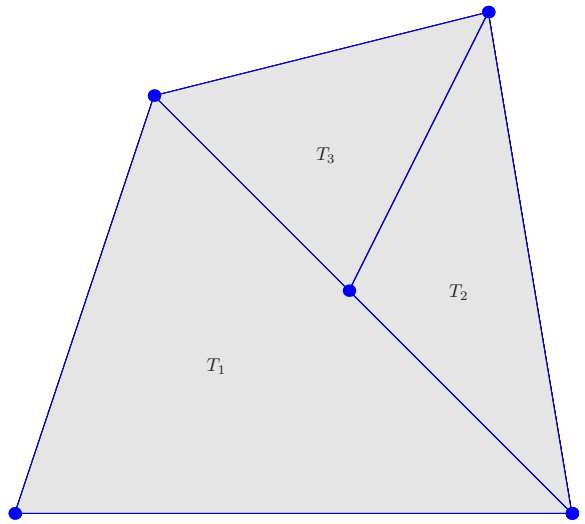
Let us decompose $\overline{\Omega}$ in the union of a finite number of non-degenerate triangles T (the *geometric elements* of the method) that satisfy the following admissibility condition:

the intersection of two distinct triangles can be only either

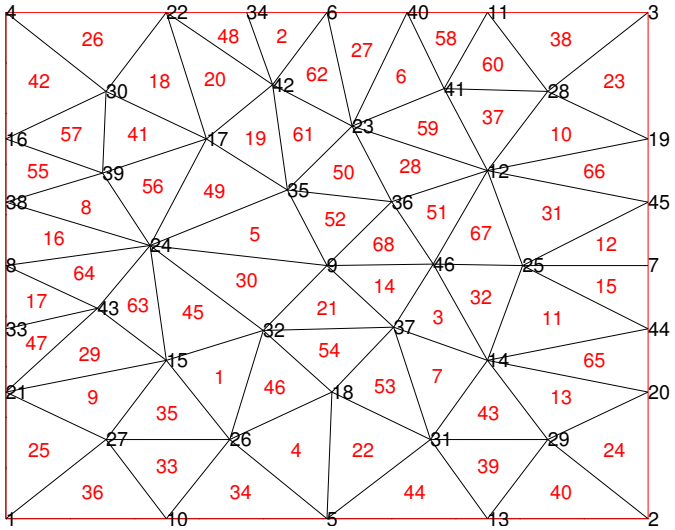
- a whole edge common to both, or
- a common vertex, or
- the empty set.

Forbidden situations:





Example of an admissible triangulation of $\overline{\Omega}$*:*



Let $\mathcal{T}$ be any admissible collection of triangles which decompose $\overline{\Omega}$; we say that $\mathcal{T}$ is a **triangulation** of Ω. Hence,

$$\overline{\Omega} = \bigcup_{T \in \mathcal{T}} T \, .$$

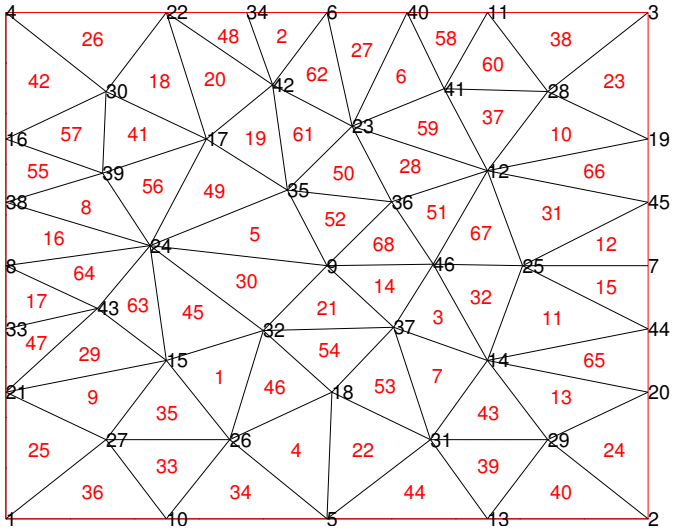
Given a triangle $T \in \mathcal{T}$, let us denote by

$$h_T = \mathrm{diam}(T)$$

its *diameter*, i.e., the maximum distance between any two points (or the length of the longest edge). Next, let us set

$$h = \max_{T \in \mathcal{T}} h_T \; ;$$

this parameter represents a measure of how *refined* the triangulation is.



The vertices $\boldsymbol{x}$ of the triangles of $\mathcal{T}$ are said **nodes** of the triangulation.

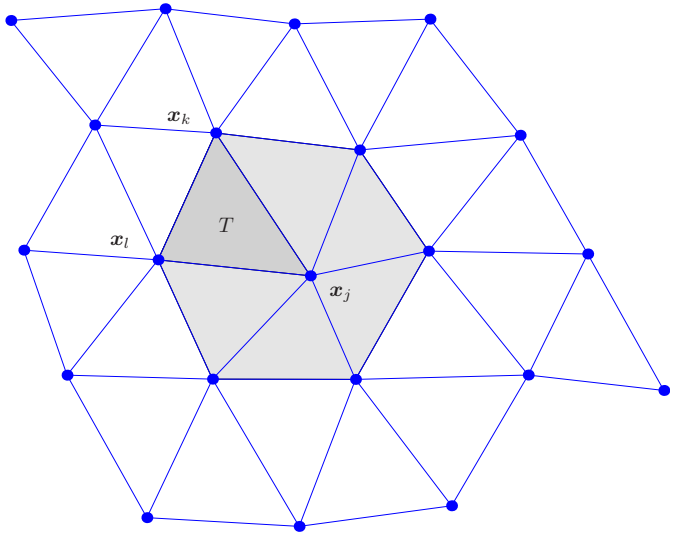
We distinguish between

- *inner nodes*, sitting in Ω, whose number will be denoted by $\mathcal{N}_h^i$,
- *boundary nodes*, sitting on $\partial\Omega$, whose number will be denoted by $\mathcal{N}_h^b$.

The total number of nodes of the triangulation is then

$$\mathcal{N}_h = \mathcal{N}_h^i + \mathcal{N}_h^b \ .$$

Let us denote by $\boldsymbol{x}_j$ the j-th node of the triangulation, whose coordinates are $(x_{j1}, x_{j2}) = (x_j, y_j)$.



Let $T \in \mathcal{T}$ be any triangle of the triangulation, and let $\boldsymbol{x}_j$, $\boldsymbol{x}_k$, $\boldsymbol{x}_l$ be its three vertices. Since triangles are supposed non-degenerate (the vertices are not collinear), the vectors $\boldsymbol{x}_j - \boldsymbol{x}_l$ and $\boldsymbol{x}_k - \boldsymbol{x}_l$ are not parallel, i.e., their cross product is not zero:

$$(\boldsymbol{x}_j - \boldsymbol{x}_l) \times (\boldsymbol{x}_k - \boldsymbol{x}_l) \neq \boldsymbol{0} \ .$$

Setting

$$\boldsymbol{G} = \begin{pmatrix} x_j - x_l & y_j - y_l \\ x_k - x_l & y_k - y_l \end{pmatrix} \ , \tag{46}$$

we have

$$(\boldsymbol{x}_j - \boldsymbol{x}_l) \times (\boldsymbol{x}_k - \boldsymbol{x}_l) = (\det \boldsymbol{G})\, \boldsymbol{e}_3 \ ,$$

hence, the non-degeneracy of T amounts to the matrix $\boldsymbol{G}$ being non-singular. In addition, it holds

$$\text{area}(T) = \frac{1}{2} |\det \boldsymbol{G}| \ .$$

The discrete admissible displacements

Given a triangle $T \in \mathcal{T}$, let

$$\mathbb{P}_1(T) = \{p : T \to \mathbb{R} \ \mid \ p(x,y) = \alpha x + \beta y + \gamma \ , \quad \text{with } \alpha, \beta, \gamma \in \mathbb{R}\} \ ,$$

denote the set of all algebraic polynomials of total degree ≤ 1, defined in T. It is a vector space of dimension 3.

Our admissible displacements v_h will be continuous functions in $\overline{\Omega}$, vanishing on $\partial\Omega$ and such that on each triangle T they belong to $\mathbb{P}_1(T)$. Thus, we are led to introduce the space of all functions which are *continuous and piecewise polynomial on the triangulation*, more precisely

$$\mathcal{V}_h = \{v_h : \overline{\Omega} \to \mathbb{R} \ : \ v_h \text{ is continuous and } \ v_{h|T} \in \mathbb{P}_1(T) \text{ for every } T \in \mathcal{T}\} \ .$$

Then, the space of all **discrete admissible displacements** will be

$$V_h = \{v_h \in \mathcal{V}_h \ : \ v_h = 0 \ \text{ on } \partial\Omega\} \ .$$

Let v_h be a function in $\mathcal{V}_h$. On the generic triangle $T \in \mathcal{T}$, the polynomial

$$p = v_{h|T} \in \mathbb{P}_1(T) \ , \qquad p(x,y) = \alpha x + \beta y + \gamma \ ,$$

is uniquely determined by 3 linearly independent conditions.

A quite natural choice consists of assigning the values of p at the three vertices of the triangle T, namely, the three conditions

$$p(\boldsymbol{x}_j) = v_j \ , \qquad p(\boldsymbol{x}_k) = v_k \ , \qquad p(\boldsymbol{x}_l) = v_l \ .$$

Indeed:

- (*geometric motivation:*) 3 non-aligned points determine a unique plane in space;
- (*algebraic motivation*) the coefficients α, β and γ have to satisfy

$$\begin{cases} \alpha x_j & + & \beta y_j & + & \gamma & = & v_j \\ \alpha x_k & + & \beta y_k & + & \gamma & = & v_k \\ \alpha x_l & + & \beta y_l & + & \gamma & = & v_l \end{cases}$$

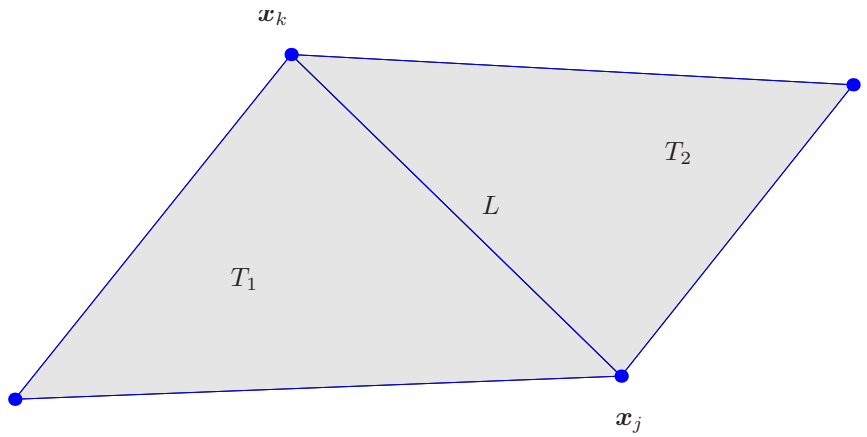
namely,

$$\begin{pmatrix} x_j & y_j & 1 \\ x_k & y_k & 1 \\ x_l & y_l & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} v_j \\ v_k \\ v_l \end{pmatrix} \ .$$

Such a system admits a unique solution, since the determinant of its matrix coincides with that of the matrix $\boldsymbol{G}$ defined in (46), which is non-singular.

Glueing polynomials together

Assigning the values of v_h at the nodes of the triangulation guarantees the continuity across the common edge of any two adjacent triangles, which in turn implies the global continuity.



Indeed, if p_1 is a polynomial of degree one defined in T_1, p_2 is a polynomial of degree one defined in T_2, and if

$$p_1(\boldsymbol{x}_j) = p_2(\boldsymbol{x}_j) \ , \qquad p_1(\boldsymbol{x}_k) = p_2(\boldsymbol{x}_k) \ ,$$

then

$$p_1 \text{ and } p_2 \text{ coincide along the edge } L = [\boldsymbol{x}_j, \boldsymbol{x}_k] \ .$$

Consequently, the function

$$v_h(\boldsymbol{x}) = \begin{cases} p_1(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in T_1 \ , \\ p_2(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in T_2 \ , \end{cases}$$

is continuous in $T_1 \cup T_2$.

By applying this line of reasoning to any pair of adjacent triangles of the triangulation $\mathcal{T}$, we see that to manifacture a function $v_h \in \mathcal{V}_h$ it suffices to prescribe its values

$$v_j = v_h(\boldsymbol{x}_j) \ , \qquad j = 1, \ldots, \mathcal{N}_h \ ,$$

at the nodes of the triangulation. The above argument guarantees that v_h will automatically be continuous across the triangulation's edges.

Thus, a function $v_h \in \mathcal{V}_h$ is uniquely determined by its values at the nodes of the triangulation. We may thus associate to it the column vector

$$\boldsymbol{v} = (v_j)_{1 \le j \le \mathcal{N}_h} \in \mathbb{R}^{\mathcal{N}_h} \ .$$

On the other hand, a function $v_h \in V_h$ (i.e., a function in $\mathcal{V}_h$ vanishing on $\partial\Omega$) is characterized by the vanishing of all its values at the boundary nodes.

Hence, such a function will be determined by its values $v_j = v_h(\boldsymbol{x}_j)$, $j = 1, \ldots, \mathcal{N}_h^i$, at the *inner* nodes of the triangulation. We can then associate to v_h the column vector

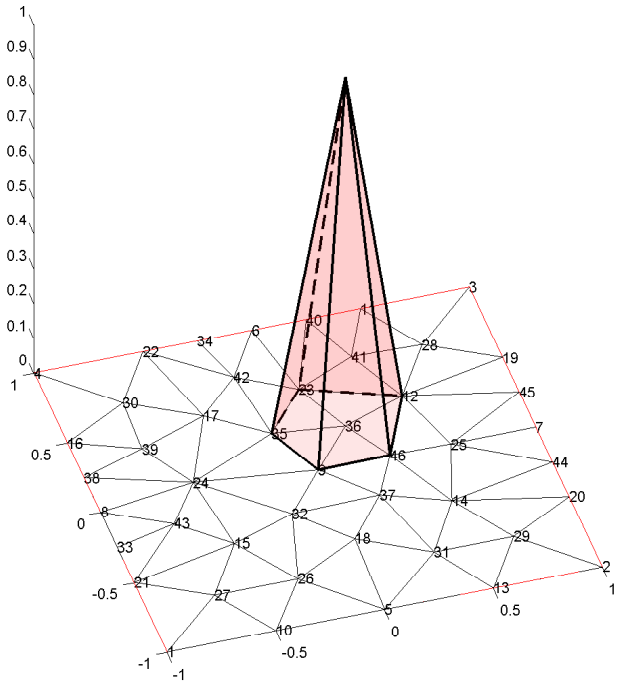
$$\boldsymbol{v} = (v_j)_{1 \le j \le \mathcal{N}_h^i} \in \mathbb{R}^{\mathcal{N}_h^i} \ .$$

(Just to keep the presentation as simple as possible, here and in the sequel we will

assume that inner nodes are invariably numbered before boundary nodes.)

The Lagrange basis

Let us now introduce bases in $\mathcal{V}_h$ and in V_h.



A natural choice is the so-called *Lagrange basis* in $\mathcal{V}_h$, which is made of the funtions of this set that equal 1 at one node or the triangulation and 0 at any other node. So let $\varphi_j \in \mathcal{V}_h$ be the function satisfying

$$\varphi_j(\boldsymbol{x}_k) = \delta_{jk} \ , \qquad 1 \leq j, k \leq \mathcal{N}_h \ .$$

Such a function is non-zero in all triangles, and only those, having $\boldsymbol{x}_j$ as vertex, i.e., the triangles of $\mathcal{T}(j)$, where

$$\mathcal{T}(j) = \{T \in \mathcal{T} \ : \ \boldsymbol{x}_j \in T\} \ .$$

The *support* of φ_j is by definition the union of these triangles, i.e.,

$$\text{supp}\,\varphi_j = \bigcup_{T\in\mathcal{T}(j)} T\,.$$

The functions φ_j form a basis of $\mathcal{V}_h$,

$$\mathcal{V}_h = \text{span}\,\{\varphi_j\ :\ 1\le j\le \mathcal{N}_h\}\,,$$

since each $v_h\in\mathcal{V}_h$ can be written as

$$v_h(\boldsymbol{x}) = \sum_{j=1}^{\mathcal{N}_h} v_j\varphi_j(\boldsymbol{x})\,, \qquad \text{with } v_j = v_h(\boldsymbol{x}_j)\,. \tag{47}$$

To produce a Lagrange basis in V_h it is enough to consider the functions φ_j with $1\le j\le \mathcal{N}_h^i$. Indeed, if $v_h\in V_h$, we have $v_j = v_h(\boldsymbol{x}_j) = 0$ for $\mathcal{N}_h^i+1\le j\le\mathcal{N}_h$, and so (47) becomes

$$v_h(\boldsymbol{x}) = \sum_{j=1}^{\mathcal{N}_h^i} v_j\varphi_j(\boldsymbol{x})\,, \tag{48}$$

i.e.,

$$V_h = \text{span}\,\{\varphi_j\ :\ 1\le j\le \mathcal{N}_h^i\}\,.$$

From now on, we set $N = \mathcal{N}_h^i$.

Structure of the stiffness matrix and the right-hand side

Let us recall the definition of the stiffness matrix $\boldsymbol{A} = (a_{jk}) \in \mathbb{R}^{N\times N}$, with

$$a_{jk} = \int_\Omega \mu \nabla\varphi_k \cdot \nabla\varphi_j \, d\boldsymbol{x} = \sum_{T\in\mathcal{T}} \int_T \mu \nabla\varphi_k \cdot \nabla\varphi_j \, d\boldsymbol{x} \ ,$$

where we have exploited the fact that Ω is the union of the triangles of $\mathcal{T}$.

Let us also recall the definitions

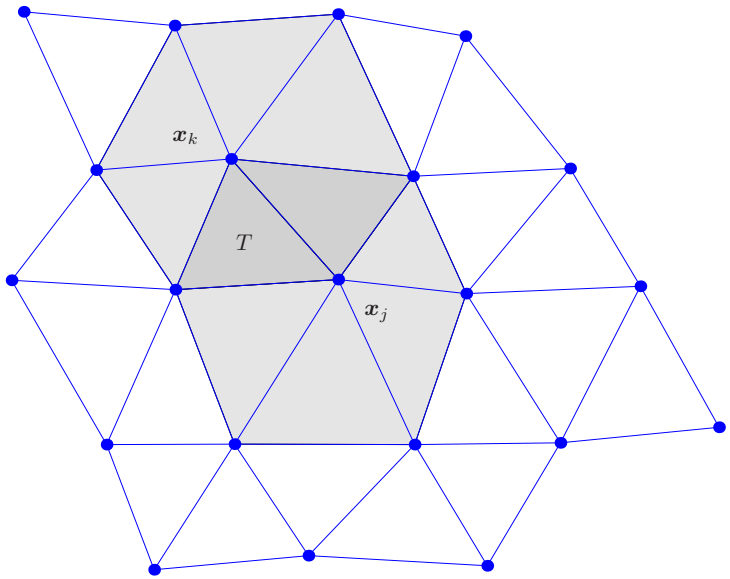
$$\mathcal{T}(j) = \{T \in \mathcal{T} \,:\, \boldsymbol{x}_j \in T\} \qquad \text{and} \qquad \operatorname{supp}\varphi_j = \bigcup_{T\in\mathcal{T}(j)} T \ .$$

Thus, if $T \notin \mathcal{T}(j)$, then $\boldsymbol{x}_j \notin T$. Hence, φ_j vanishes identically on T and therefore its gradient, too, will vanish identically therein. Consequently, we have

$$\int_T \mu \nabla\varphi_k \cdot \nabla\varphi_j \, d\boldsymbol{x} = 0 \qquad \text{if } T \notin \mathcal{T}(j) \ \text{ or if } T \notin \mathcal{T}(k) \ .$$

We conclude that

$$a_{jk} = \sum_{T\in\mathcal{T}(j)\cap\mathcal{T}(k)} \int_T \mu \nabla\varphi_k \cdot \nabla\varphi_j \, d\boldsymbol{x} \ . \tag{49}$$

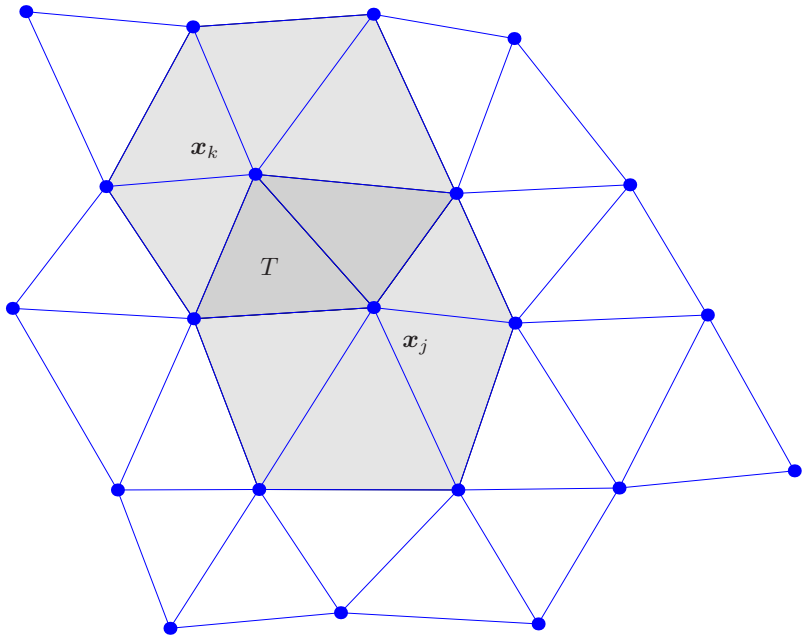


Let us first assume that $k = j$. In this case

$$a_{jj} = \sum_{T \in \mathcal{T}(j)} \int_T \mu \|\nabla \varphi_j\|^2 \, d\boldsymbol{x} \ ,$$

i.e., all diagonal elements are strictly positive and obtained by integrating over the support of the corresponding basis function.

Let now $k \neq j$. If there is a triangle T in $\mathcal{T}(j) \cap \mathcal{T}(k)$, the nodes $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ are among its vertices; clearly this can occur if and only if *the two nodes are joined by an edge of the triangulation*.



Hence, if $\boldsymbol{x}_j$, $\boldsymbol{x}_k$ belong on a same (necessarily non-boundary) edge, then $\mathcal{T}(j) \cap \mathcal{T}(k)$ is not empty: it consists, namely, of the triangles T_1, T_2 that share the edge; thus,

$$a_{jk} = \int_{T_1} \mu \nabla \varphi_k \cdot \nabla \varphi_j \, d\boldsymbol{x} + \int_{T_2} \mu \nabla \varphi_k \cdot \nabla \varphi_j \, d\boldsymbol{x} \,, \qquad \text{with } T_1, T_2 \in \mathcal{T}(j) \cap \mathcal{T}(k) \,,$$

and this matrix entry is, possibly, other than 0.

Vice versa,

$$\mathcal{T}(j) \cap \mathcal{T}(k) \text{ is empty if } \boldsymbol{x}_j \text{ and } \boldsymbol{x}_k \text{ do not belong to the same edge of } \mathcal{T} .$$

In this case, necessarily we have

$$a_{jk} = 0 .$$

For each node $\boldsymbol{x}_j$, the number of nodes $\boldsymbol{x}_k$ connected to it by an edge is equal to the number of triangles having $\boldsymbol{x}_j$ as a vertex; such a number is typically small (≤ 10), for otherwise there would exist triangles with very small angles, a feature that influences negatively the condition number of the matrix.

Thus, the number of entries a_{jk} a priori different from 0 sitting on the generic row j of the matrix $\boldsymbol{A}$ turns out to be $O(1)$.

We conclude that the total number of entries a_{jk} a priori different from 0 is $O(N)$, to be compared with the number N^2 of entries of $\boldsymbol{A}$. Therefore, $\boldsymbol{A}$ is a **sparse** matrix.

The position of the elements a_{jk} different from 0 in the matrix depends on how nodes are numbered: if two nodes connected by an edge are numbered "close to each other", then the corresponding matrix entry a_{jk} will be "close" to the main diagonal; the distance from the main diagonal is indeed given by $|j-k|$.

It is possible to number nodes in such a way that $\boldsymbol{A}$ is **banded**, with band-width $O(\sqrt{N})$.

Sophisticated techniques for **reordering** the nodes have been developed, which optimize the cost of solving the algebraic system by a direct method (such as Gauss or Choleski).

At last, as far as the right-hand side is concerned, we have

$$f_j = \int_\Omega f\varphi_j \, d\boldsymbol{x} = \sum_{T\in\mathcal{T}} \int_T f\varphi_j \, d\boldsymbol{x} = \sum_{T\in\mathcal{T}(j)} \int_T f\varphi_j \, d\boldsymbol{x} \; .$$

Assemblying the stiffness matrix and the right-hand side

We have seen that

$$a_{jk} = \sum_{T \in \mathcal{T}(j) \cap \mathcal{T}(k)} \int_T \mu \nabla \varphi_k \cdot \nabla \varphi_j \, d\boldsymbol{x} \,, \qquad f_j = \sum_{T \in \mathcal{T}(j)} \int_T f \varphi_j \, d\boldsymbol{x} \,.$$

In a finite-elements code, the construction of stiffness matrix and source term may be logically divided into two phases:

- *a cycle on the triangles* of $\mathcal{T}$, in which one computes the contribution of any triangle to the stiffness matrix and the source;
- *the assemblage* of matrix and source, where the contributions of the single triangles are suitably processed to produce the entries of the matrix and the source vector.

For efficiency reasons, though, the latter part is actually carried out simultaneously with the former: as the contributions from a triangle become available, they are added to the running value of the corresponding entries in the matrix and the source (which are initially set to 0); at the end of the cycle on the triangles every matrix and source element will have attained the correct value.

Let T be a generic triangle of the triangulation; we assume that its vertices are

$$\boldsymbol{x}_{j_1} \ , \qquad \boldsymbol{x}_{j_2} \ , \qquad \boldsymbol{x}_{j_3} \ , \qquad \text{with} \quad 1 \leq j_1, j_2, j_3 \leq \mathcal{N}_h \ .$$

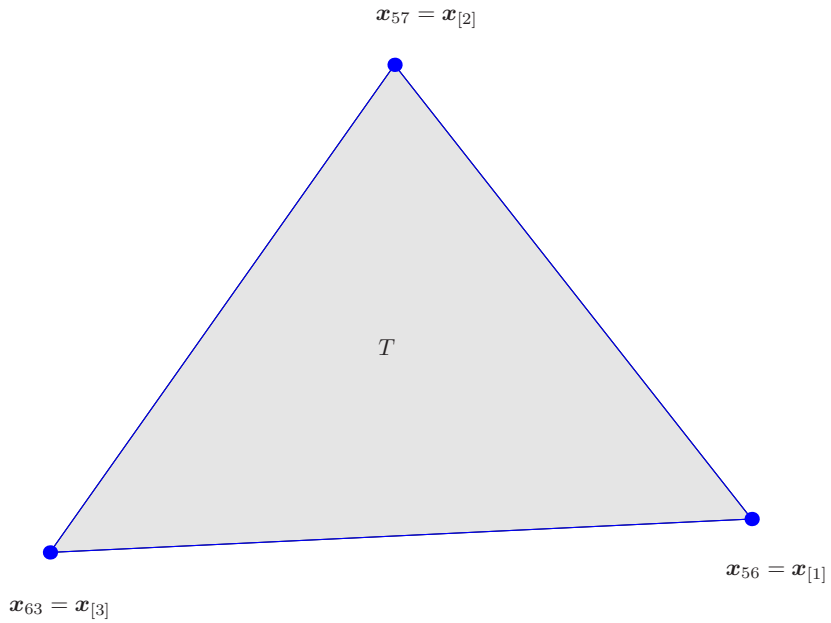
If all vertices of T are inner (i.e., if $1 \leq j_1, j_2, j_3 \leq \mathcal{N}_h^i = N$), then the triangle contributes to the matrix entries

$$a_{j_1,j_1} \ , \qquad a_{j_1,j_2} \ , \qquad a_{j_1,j_3} \ , \qquad a_{j_2,j_2} \ , \qquad a_{j_2,j_3} \ , \qquad a_{j_3,j_3}$$

to the ones obtained from these by swapping indices, and to the vector entries

$$f_{j_1} \ , \quad f_{j_2} \ , \quad f_{j_3} \ .$$

If one or two vertices of T sit on the boundary $\partial\Omega$, then the contributions to the matrix and source entries may be different, depending on the boundary conditions enforced therein. For the moment, we ignore this situation.



It is convenient to shift from the global numbering of nodes j_1, j_2, j_3 to a local numbering which uses the indices $1, 2, 3$. Hereafter, we will use greek letters $(\alpha, \beta, \dots)$ to denote local indices.

Hence,

$$
\begin{aligned}
\boldsymbol{x}_{j_\alpha} \quad &\longrightarrow \quad \boldsymbol{x}_\alpha \,, \qquad 1 \le \alpha \le 3 \,; \\
\varphi_{j_\alpha|T} \quad &\longrightarrow \quad \varphi_\alpha \in \mathbb{P}_1(T) \,, \qquad 1 \le \alpha \le 3 \,; \\
\int_T \mu \nabla \varphi_{j_\beta} \cdot \nabla \varphi_{j_\alpha} \, d\boldsymbol{x} \quad &\longrightarrow \quad \int_T \mu \nabla \varphi_\beta \cdot \nabla \varphi_\alpha \, d\boldsymbol{x} \,, \qquad 1 \le \alpha, \beta \le 3 \,; \\
\int_T f \varphi_{j_\alpha} \, d\boldsymbol{x} \quad &\longrightarrow \quad \int_T f \varphi_\alpha \, d\boldsymbol{x} \,, \qquad 1 \le \alpha \le 3 \,.
\end{aligned}
$$

Thus, we are led to introduce the **elemental stiffness matrix** of the triangle T ("elemental" means related to an element, i.e., a triangle)

$$\boldsymbol{A}^{(T)} = (a_{\alpha\beta}^{(T)})_{1\leq\alpha,\beta\leq 3} \in \mathbb{R}^{3\times 3}\,, \qquad \text{with } \; a_{\alpha\beta}^{(T)} = \int_T \mu \nabla\varphi_\beta \cdot \nabla\varphi_\alpha \, d\boldsymbol{x}\,,$$

and the **elemental source vector**

$$\boldsymbol{f}^{(T)} = (f_{\alpha}^{(T)})_{1\leq\alpha\leq 3} \in \mathbb{R}^{3}\,, \qquad \text{with } \; f_{\alpha}^{(T)} = \int_T f\varphi_\alpha \, d\boldsymbol{x}\,.$$

[Beware not to mix the superscript $^{(T)}$ with the symbol of transposition of a matrix or vector!]

Computing the elemental stiffness matrix

Observe that $\nabla\varphi_\alpha$ is a constant vector in T, since φ_α is an affine function. Hence,

$$a_{\alpha\beta}^{(T)} = \nabla\varphi_\beta \cdot \nabla\varphi_\alpha \int_T \mu \, d\boldsymbol{x} \; .$$

Let us introduce the *mean value* of μ on T

$$\mu_T = \frac{1}{\text{area}(T)} \int_T \mu \, d\boldsymbol{x} \; .$$

Note that if μ is constant on T, then $\mu_T = \mu$, otherwise μ_T can be safely approximated by the value $\mu(\boldsymbol{x}_b)$ in the baricenter $\boldsymbol{x}_b = \frac{1}{3}(\boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_3)$ of the triangle.

In any case, we obtain

$$a_{\alpha\beta}^{(T)} = \mu_T \, \text{area}(T) \; \nabla\varphi_\beta \cdot \nabla\varphi_\alpha \; .$$

We are left with the problem of computing the gradients of the basis functions.

Recalling Taylor's formula for functions of several variables, we first observe that if $\boldsymbol{x}$ and $\boldsymbol{x}_0$ are points in the triangle T, the variation of φ_α between $\boldsymbol{x}_0$ and $\boldsymbol{x}$ can be written as

$$\varphi_\alpha(\boldsymbol{x}) - \varphi_\alpha(\boldsymbol{x}_0) = \nabla\varphi_\alpha \cdot (\boldsymbol{x} - \boldsymbol{x}_0) \; ;$$

indeed, φ_α is affine and therefore all its partial derivatives of order higher than 1 vanish identically. Thus, applying the previous formula to the vertices of the triangle, we get

$$\begin{aligned} 1 &= \varphi_\alpha(\boldsymbol{x}_\alpha) - \varphi_\alpha(\boldsymbol{x}_\gamma) = \nabla\varphi_\alpha \cdot (\boldsymbol{x}_\alpha - \boldsymbol{x}_\gamma) \; , \qquad \text{with } \gamma \neq \alpha \; , \\ 0 &= \varphi_\alpha(\boldsymbol{x}_\beta) - \varphi_\alpha(\boldsymbol{x}_\gamma) = \nabla\varphi_\alpha \cdot (\boldsymbol{x}_\beta - \boldsymbol{x}_\gamma) \; , \qquad \text{with } \beta \neq \alpha, \; \beta \neq \gamma \; . \end{aligned}$$

Consequently, the components $\varphi_{\alpha,x}$ and $\varphi_{\alpha,y}$ of $\nabla\varphi_\alpha$ are the solutions of the following algebraic system

$$\begin{pmatrix} x_\alpha - x_\gamma & y_\alpha - y_\gamma \\ x_\beta - x_\gamma & y_\beta - y_\gamma \end{pmatrix} \begin{pmatrix} \varphi_{\alpha,x} \\ \varphi_{\alpha,y} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \; ,$$

where β and γ denote the two indices different from α in the set $\{1, 2, 3\}$; note that the system's matrix is indeed a particular matrix $\boldsymbol{G}$ as defined in (46). We thus have

$$\varphi_{\alpha,x} = \frac{y_\beta - y_\gamma}{\det \boldsymbol{G}} \; , \qquad \varphi_{\alpha,y} = -\frac{x_\beta - x_\gamma}{\det \boldsymbol{G}} \; . \tag{50}$$

If the nodes $\boldsymbol{x}_\alpha$, $\boldsymbol{x}_\beta$ and $\boldsymbol{x}_\gamma$ are numbered counterclockwise, we have

$$\det \boldsymbol{G} = 2\text{area}(T) > 0 \ ,$$

hence the previous formulas become

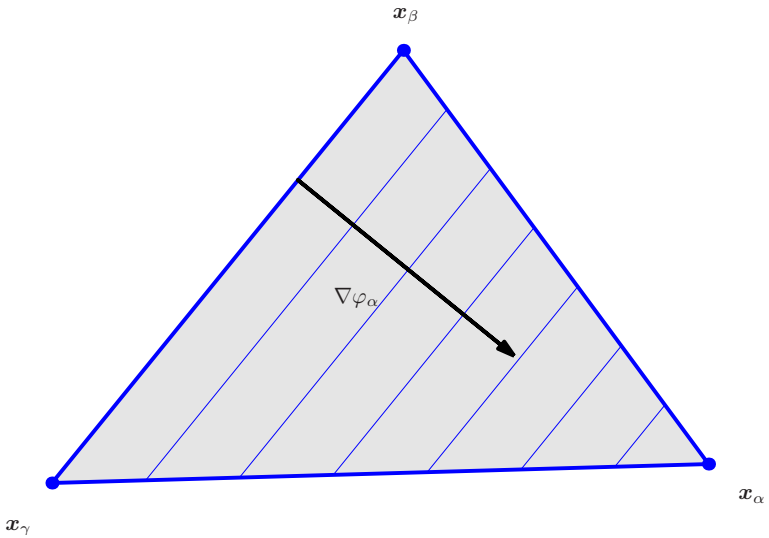
$$\varphi_{\alpha,x} = \frac{y_\beta - y_\gamma}{2\text{area}(T)} \ , \qquad \varphi_{\alpha,y} = -\frac{x_\beta - x_\gamma}{2\text{area}(T)} \ . \tag{51}$$

Making things explicit, if we consider the three cases $(\alpha, \beta, \gamma) = (1, 2, 3)$, $(\alpha, \beta, \gamma) = (2, 3, 1)$ and $(\alpha, \beta, \gamma) = (3, 1, 2)$, we get

$$\varphi_{1,x} = \frac{y_2 - y_3}{2\text{area}(T)} \ , \qquad \varphi_{1,y} = -\frac{x_2 - x_3}{2\text{area}(T)} \ ,$$

$$\varphi_{2,x} = \frac{y_3 - y_1}{2\text{area}(T)} \ , \qquad \varphi_{2,y} = -\frac{x_3 - x_1}{2\text{area}(T)} \ ,$$

$$\varphi_{3,x} = \frac{y_1 - y_2}{2\text{area}(T)} \ , \qquad \varphi_{3,y} = -\frac{x_1 - x_2}{2\text{area}(T)} \ .$$



From the geometric point of view, the previous formulas translate the following facts:

- the vector $\nabla\varphi_\alpha$ is perpendicular to the opposite edge $[\boldsymbol{x}_\gamma, \boldsymbol{x}_\beta]$,

$$\nabla\varphi_\alpha \cdot (\boldsymbol{x}_\gamma - \boldsymbol{x}_\beta) = 0\ ;$$

- the increment of φ_α in the direction of its gradient is 1 if we move from the edge $[\boldsymbol{x}_\gamma, \boldsymbol{x}_\beta]$ to the vartex $\boldsymbol{x}_\alpha$.

A useful remark. In order to compute the components of $\nabla\varphi_\alpha$, it might be easier to exploit the fact they are constant on T: thus, it suffices to find two segments, one horizontal and one vertical, at the endpoints of which we know the values φ_α. The difference quotients of φ_α on the segments will give the required components.

A useful test

It is useful to observe that the affine function $\varphi_1 + \varphi_2 + \varphi_3$ takes the value 1 at each of the vertices of T, hence, it is constant:

$$\varphi_1(\boldsymbol{x}) + \varphi_2(\boldsymbol{x}) + \varphi_3(\boldsymbol{x}) = 1 \qquad \text{for each } \boldsymbol{x} \in T \ .$$

Applying the gradient, we get the identity

$$\nabla\varphi_1 + \nabla\varphi_2 + \nabla\varphi_3 = \mathbf{0} \ ,$$

which – for instance – allows us to compute the gradient of a basis function, if the other two are known.

A consequence of the last identity is that *the sum of the elements on each row (or column) of the matrix* $\boldsymbol{A}^{(T)}$ *is zero*. Indeed, for $\alpha = 1, 2, 3$, we have

$$\begin{aligned} \sum_{\beta=1}^{3} a_{\alpha,\beta}^{(T)} &= \sum_{\beta=1}^{3} \int_T \mu \nabla\varphi_\beta \cdot \nabla\varphi_\alpha \, d\boldsymbol{x} = \int_T \mu \left(\sum_{\beta=1}^{3} \nabla\varphi_\beta \right) \cdot \nabla\varphi_\alpha \, d\boldsymbol{x} \\ &= \int_T \mu \, \mathbf{0} \cdot \nabla\varphi_\alpha \, d\boldsymbol{x} = 0 \ . \end{aligned}$$

This property gives a practical way to double-check whether the elements of the stiffness matrix have been computed correctly.

Computing the elemental source vector

Let us use the following quadrature rule

$$\int_T g\, d\boldsymbol{x} \simeq \frac{1}{3}\big(g(\boldsymbol{x}_1) + g(\boldsymbol{x}_2) + g(\boldsymbol{x}_3)\big)\,\mathsf{area}(T)\ , \tag{52}$$

which is exact for affine functions and extends to the two-dimensional situation the known trapezoidal rule on an interval of the real line.
Choosing $g = f\varphi_\alpha$ and recalling the values taken by φ_α at the vertices of the triangle, we can define the elemental source vector as the vector $\boldsymbol{f}^{(T)}$ whose entries are

$$f_\alpha = \frac{1}{3} f(\boldsymbol{x}_\alpha)\mathsf{area}(T)\ , \qquad \alpha = 1, 2, 3\ . \tag{53}$$

If f is constant on T, such expression coincides with the exact value $\int_T f\varphi_\alpha d\boldsymbol{x}$.

At last, let us note that, after assemblying the entries of the elemental source vectors coming from all triangles, the previous definition leads to the following expression for the j-th entry of the right-hand side vector $\boldsymbol{f}$:

$$f_j = \frac{1}{3} f(\boldsymbol{x}_j) \sum_{T \in \mathcal{T}(j)} \mathsf{area}(T) = \frac{1}{3} f(\boldsymbol{x}_j)\,\mathsf{area}(\mathsf{supp}\,\varphi_j)\ . \tag{54}$$

- **Condition number of the stiffness matrix** One has

$$\mathrm{cond}_2(\boldsymbol{A}) \simeq c h_{\mathsf{min}}^{-2} , \qquad \text{where} \quad h_{\mathsf{min}} = \min_{T \in \mathcal{T}} h_T .$$

- **Error behaviour** Let us measure the discretization error $u - u_h$ in one of the following norms:

$$\|v\|_2 = \left(\int_\Omega v^2 \, d\boldsymbol{x} \right)^{1/2} \qquad (\textit{quadratic norm})$$

$$\|v\|_E = \left(\int_\Omega \mu \|\nabla v\|^2 \, d\boldsymbol{x} \right)^{1/2} \qquad (\textit{energy norm})$$

$$\|v\|_\infty = \max_{\boldsymbol{x} \in \overline{\Omega}} |v(\boldsymbol{x})| \qquad (\textit{maximum norm}) .$$

Let us assume that the exact solution u has second-order partial derivatives with bounded quadratic norm, namely,

$$\|u\|_{H,2} = \sum_{i,j=1}^{2} \left\| \frac{\partial^2 u}{\partial x_i \partial x_j} \right\|_2 < +\infty .$$

Under this assumption, for a regular triangulation we can prove that the error $u - u_h$ tends to zero

- *quadratically* in h, if it is measured in the quadratic norm, namely,

$$\|u - u_h\|_2 \leq Ch^2 \|u\|_{H,2} \ ,$$

- *linearly* in h, if it is measured in the energy norm, namely,

$$\|u - u_h\|_E \leq Ch \|u\|_{H,2} \ .$$

If in addition u has second-order partial derivatives with bounded maximum norm, i.e., if

$$\|u\|_{H,\infty} = \max_{1 \leq i,j \leq 2} \left\| \frac{\partial^2 u}{\partial x_i \partial x_j} \right\|_\infty < +\infty \ ,$$

then the error tends to zero

- *"almost" quadratically* in h, if it is measured in the maximum norm, namely,

$$\|u - u_h\|_\infty \leq Ch^2 |\log h|^{3/2} \|u\|_{H,\infty} \ .$$

Further boundary conditions

Let us assume that the boundary $\partial\Omega$ is divided in a non-empty part Γ_D, on which we enforce a (non-homogeneous) *Dirichlet condition*, and its complementary part Γ_N, on which we enforce a (non-homogeneous) *Neumann condition*. Thus, the problem is

$$\begin{cases} -\nabla \cdot (\mu \nabla u) = f & \text{in } \Omega \,, \\ u = g & \text{on } \Gamma_D \,, \\ \mu \dfrac{\partial u}{\partial n} = \psi & \text{on } \Gamma_N \,, \end{cases}$$

with g and ψ given functions.

The Neumann condition means that on Γ_N we assign the normal component of the shear stress (in the elastic model), or the normal component of the heat flux (in the thermal model).

If we keep in mind the elastic model, the set $V(g)$ of the *admissible displacements*, or shape functions, is now formed by functions that take the value g on Γ_D, whereas they take arbitrary values on Γ_N.

The set $V(0)$ of the functions vanishing on Γ_D will be the set of the *admissible variations*, or test functions.

Why a function v in $V(0)$ is an admissible *variation*?

Suppose we fix an arbitrary admissible displacement $u_g \in V(g)$. The solution u of our problem is itself a function in $V(g)$, hence the difference $u - u_g$ will vanish on Γ_D, i.e.,

$$u - u_g \in V(0).$$

If we denote this difference by $u_0 = u - u_g$, we can write

$$u = u_g + u_0.$$

This shows that u_0 is the *variation* that we have to apply to u_g in order to get the solution u.

Therefore, if we know a particular admissible displacement u_g, the problem of seeking the solution u in $V(g)$ is equivalent to the problem of seeking u_0 in $V(0)$.

The variational formulation of the problem with the new boundary conditions is based on the "integration by part" formula, that we already know:

$$-\int_\Omega \nabla\cdot(\mu\nabla u)\, v\, d\boldsymbol{x} = \int_\Omega \mu\nabla u\cdot\nabla v\, d\boldsymbol{x} - \int_{\partial\Omega} \mu\frac{\partial u}{\partial n}\, v\, ds\ ,$$

where $u \in V(g)$ is now the solution of our problem, whereas $v \in V(0)$ is any test function. The boundary integral on the right-hand side can be written as

$$\int_{\partial\Omega} \mu\frac{\partial u}{\partial n}\, v\, ds = \int_{\Gamma_D} \mu\frac{\partial u}{\partial n}\, v\, ds + \int_{\Gamma_N} \mu\frac{\partial u}{\partial n}\, v\, ds = 0 + \int_{\Gamma_N} \psi\, v\, ds\ ,$$

where we have kept into account that v vanishes on Γ_D and u satisfies the Neumann condition on Γ_N.

Therefore, the variational formulation of the problem is as follows:

$$\begin{cases} u \in V(g) \text{ and satisfies} \\ \displaystyle\int_\Omega \mu\nabla u\cdot\nabla v\, d\boldsymbol{x} = \int_\Omega f v\, d\boldsymbol{x} + \int_{\Gamma_N} \psi\, v\, ds \qquad \text{for any } \ v \in V(0)\ . \end{cases} \tag{55}$$

Discretization

Let us suppose that Γ_D is a union of edges of the triangulation. In addition, let us assume (just to simplfy the presentation) that the boundary nodes on Γ_N have been numbered before those on Γ_D.

The set of discrete admissible displacements, or discrete shape functions, is defined as

$$V_h(g) = \{v_h \in \mathcal{V}_h \ : \ v_h(\boldsymbol{x}_j) = g(\boldsymbol{x}_j) \text{ for each } \boldsymbol{x}_j \in \Gamma_D\} \ ;$$

the set of discrete admissible variations, or discrete test functions, will thus be $V_h(0)$.

The functions of $V_h(g)$ and $V_h(0)$ are uniquely determined by their values at all *inner nodes* and at the *boundary nodes that belong to* Γ_N. These are the nodes carrying the degrees of freedom; let us still indicate with N their number ($\mathcal{N}_h^i \leq N < \mathcal{N}_h$).
We thus have

$$V_h(0) = \text{span}\,\{\varphi_j \ : \ 1 \leq j \leq N\} \ ,$$

and a function $u_h \in V_h(g)$ will be written as

$$u_h(\boldsymbol{x}) = \sum_{k=1}^{N} u_k \varphi_k(\boldsymbol{x}) + \sum_{k=N+1}^{\mathcal{N}_h} g_k \varphi_k(\boldsymbol{x}) \ ,$$

with $g_k = g(\boldsymbol{x}_k)$.

The discrete variational formulation is as follows:

$$\begin{cases} u_h \in V_h(g) \text{ and satisfies} \\ \displaystyle\int_\Omega \mu \nabla u_h \cdot \nabla v_h \, d\boldsymbol{x} = \int_\Omega f v_h \, d\boldsymbol{x} + \int_{\Gamma_N} \psi \, v_h \, ds \qquad \text{for each} \quad v_h \in V_h(0) \, . \end{cases}$$

Choosing subsequently $v_h = \varphi_j$ for $j = 1, \ldots, N$, we obtain the equations satisfied by u_h:

$$\int_\Omega \mu \nabla u_h \cdot \nabla \varphi_j \, d\boldsymbol{x} = \int_\Omega f \varphi_j \, d\boldsymbol{x} + \int_{\Gamma_N} \psi \, \varphi_j \, ds \, , \qquad 1 \le j \le N \, .$$

Let us replace u_h by its expansion given in the previous slide, and let us move to the right-hand side whatever depends on the datum g. Setting, as usual,

$$a_{jk} = \int_\Omega \mu \nabla \varphi_k \cdot \nabla \varphi_j \, d\boldsymbol{x} \, ,$$

we obtain the algebraic system

$$\sum_{k=1}^{N} a_{jk} u_k = \int_\Omega f \varphi_j \, d\boldsymbol{x} + \int_{\Gamma_N} \psi \, \varphi_j \, ds - \sum_{k=N+1}^{\mathcal{N}_h} a_{jk} g_k \, , \qquad 1 \le j \le N \, ,$$

which we still write in the form

$$\boldsymbol{A u} = \boldsymbol{f} \, .$$

Robin boundary condition

The Neumann boundary condition can be made more general by considering the so-called *Robin condition*

$$\mu \frac{\partial u}{\partial n} + \alpha u = \psi \qquad \text{on} \quad \Gamma_N \ ,$$

where $\alpha \geq 0$ is a given coefficient.
For instance, we obtain such a condition by imposing the heat flux to be proportional to the difference between a given temperature $\bar{u}$ and the current temperature u of the plate, i.e.,

$$\mu \frac{\partial u}{\partial n} = \alpha(\bar{u} - u) \qquad \text{on} \quad \Gamma_N \ ;$$

in this case, one has $\psi = \alpha \bar{u}$.

The Robin condition modifies the expression of the boundary term:

$$\int_{\partial\Omega} \mu \frac{\partial u}{\partial n} \, v \, ds = \int_{\Gamma_N} \mu \frac{\partial u}{\partial n} \, v \, ds = \int_{\Gamma_N} (-\alpha u + \psi) \, v \, ds = - \int_{\Gamma_N} \alpha u \, v \, ds + \int_{\Gamma_N} \psi \, v \, ds \ .$$

The new discrete variational formulation of the problem becomes:

$$\begin{cases} u_h \in V_h(g) \text{ and satisfies} \\ \displaystyle\int_\Omega \mu \nabla u_h \cdot \nabla v_h \, d\boldsymbol{x} + \int_{\Gamma_N} \alpha u_h \, v_h \, ds = \int_\Omega f v_h \, d\boldsymbol{x} + \int_{\Gamma_N} \psi \, v_h \, ds \quad \text{for each} \quad v_h \in V_h(0) \, . \end{cases}$$

The effect is a modification of the stiffness matrix $\boldsymbol{A}$, whose entries are now given by

$$a_{jk} = \int_\Omega \mu \nabla \varphi_k \cdot \nabla \varphi_j \, d\boldsymbol{x} + \int_{\Gamma_N} \alpha \, \varphi_k \, \varphi_j \, ds \, .$$

Note, however, that the integral over Γ_N is zero for all basis functions associated with the inner nodes.

Since $\alpha \geq 0$ by assumption, it is not difficult to check that the symmetric matrix $\boldsymbol{A}$ is still positive definite.