Numerical methods for Partial Differential Equations

Adriano Festa
Politecnico of Turin, Italy

Undergraduate Lecture at KSU
2022

Table of Contents

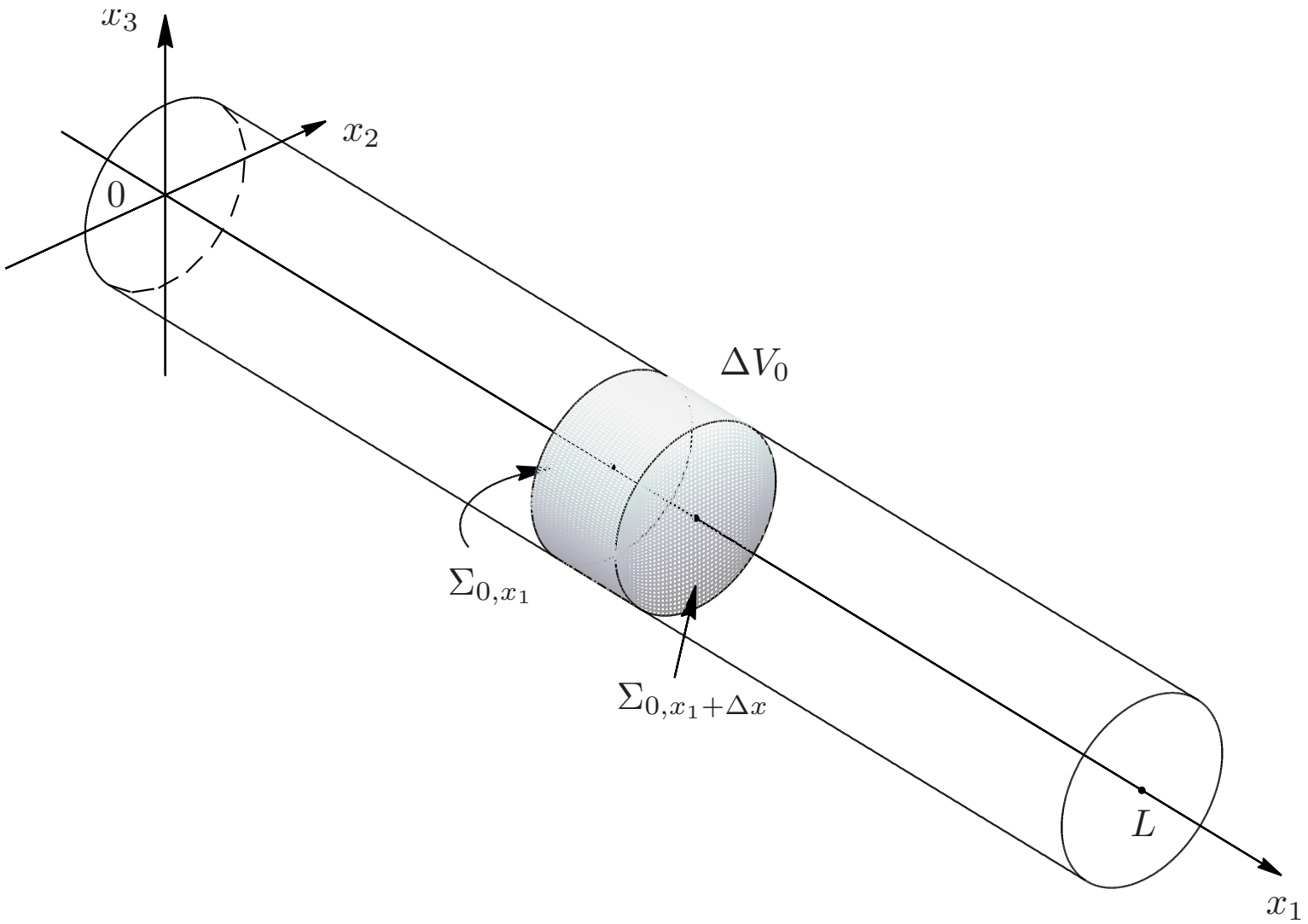
THE MODEL OF ELASTIC STRING

Consider a thin elastic string with constant round cross section S. In absence of external forces, the axis is aligned along the interval $[0, L]$ on the coordinate axis x_1. The elastic is fixed at the endpoints of the interval.

Let us apply a (small) density of force $\boldsymbol{f} = 0\boldsymbol{e}_1 + 0\boldsymbol{e}_2 + f_3\boldsymbol{e}_3$ per unit of volume lying on the plane x_1x_3 and normal to the string axis. This induces a (small) dispacement $\boldsymbol{u} = u_1\boldsymbol{e}_1 + u_2\boldsymbol{e}_2 + u_3\boldsymbol{e}_3$ of the string from the reference position; more precisely, $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$ denotes the displacement of the point particle at the point $\boldsymbol{x}$ at reference position.

The dispacement will be coplanar with the force, so that $u_2 = 0$; at first approximation, moreover, the component u_1 is negligible with respect to the component u_3 that describes the displacement along the force.

Take $x_1 \in (0, L)$ and let $\Delta V_0 = [x_1, x_1 + \Delta x] \times S$ be an element of the string of length Δx at reference; denote by $\Sigma_{0,x_1} = \{x_1\} \times S$ and $\Sigma_{0,x_1+\Delta x} = \{x_1 + \Delta x\} \times S$ the cross sections delimiting the element.



Under the applied force the element transforms into the element ΔV; let Σ_{x_1} and $\Sigma_{x_1+\Delta x}$ be the transformed cross sections. Indicate by

$$\underline{\boldsymbol{\sigma}} = \begin{pmatrix} \sigma_1 & \tau_{12} & \tau_{13} \\ \tau_{21} & \sigma_2 & \tau_{23} \\ \tau_{31} & \tau_{23} & \sigma_3 \end{pmatrix} \tag{1}$$

the stress tensor of the string.

Now let $\boldsymbol{n}$ be the normal to the transformed section, oriented from the interior to the exterior of ΔV. The equilibrium equation reads

$$\int_{\Delta V} \boldsymbol{f}\, dV + \int_{\Sigma_{x_1+\Delta x}} \underline{\boldsymbol{\sigma}}\, \boldsymbol{n}\, d\Sigma + \int_{\Sigma_{x_1}} \underline{\boldsymbol{\sigma}}\, \boldsymbol{n}\, d\Sigma = \mathbf{0}\,. \tag{2}$$

Having assumed small dispacements we may, as first approximation, identify ΔV with ΔV_0, as well as the transformed cross sections with the reference one. This and the assumption that the string is thin allows us to switch to a one-dimensional model. In fact,

$$\int_{\Delta V} \boldsymbol{f}\, dV \simeq \int_{\Delta V_0} \boldsymbol{f}\, dV = \int_{x_1}^{x_1+\Delta x} \left(\int_S \boldsymbol{f}(x, x_2, x_3)\, dx_2 dx_3 \right) dx = |S| \int_{x_1}^{x_1+\Delta x} \tilde{\boldsymbol{f}}(x)\, dx \ ,$$

where $|S|$ is the area of the section S and $\tilde{\boldsymbol{f}}(x)$ is the mean value of $\boldsymbol{f}$ over the section S at $x \in (0, L)$:

$$\tilde{\boldsymbol{f}}(x) = \frac{1}{|S|} \int_S \boldsymbol{f}(x, x_2, x_3)\, dx_2 dx_3 \ .$$

Proceeding in a similar manner with surface integrals we obtain the approximate equations

$$|S| \int_{x_1}^{x_1+\Delta x} \tilde{\boldsymbol{f}}(x)\, dx + |S|\, \underline{\tilde{\boldsymbol{\sigma}}}\, \tilde{\boldsymbol{n}}_{|x_1+\Delta x} + |S|\, \underline{\tilde{\boldsymbol{\sigma}}}\, \tilde{\boldsymbol{n}}_{|x_1} = \boldsymbol{0} \ . \tag{3}$$

Let us divide by $|S|$ and note that by assumption, $\tilde{\boldsymbol{n}}_{|x_1+\Delta x} \simeq \boldsymbol{e}_1$ and $\tilde{\boldsymbol{n}}_{|x_1} \simeq -\boldsymbol{e}_1$. Furthermore, let us omit the symbol $\tilde{}$ to make the notation simpler. Therefore

$$\int_{x_1}^{x_1+\Delta x} \boldsymbol{f}(x)\, dx + \underline{\boldsymbol{\sigma}}\, \boldsymbol{e}_{1\,|x_1+\Delta x} - \underline{\boldsymbol{\sigma}}\, \boldsymbol{e}_{1\,|x_1} = \boldsymbol{0}\ . \tag{4}$$

Now we take the component along x_3 of this vector equation, i.e., we apply the dot product with $\boldsymbol{e}_3$ to obtain

$$\int_{x_1}^{x_1+\Delta x} f_3(x)\, dx + \tau_{31\,|x_1+\Delta x} - \tau_{31\,|x_1} = 0\ . \tag{5}$$

Dividing by Δx and taking the limit as $\Delta x \to 0$ brings us to the differential relation

$$f_3(x_1) + \frac{d\tau_{31}}{dx}(x_1) = 0\ , \qquad x_1 \in (0, L)\ , \tag{6}$$

expressing the string's equilibrium state. Here

$$\tau_{31}$$

is the vertical component of the **shear stress** acting on the cross section.

Now, we invoke *Hooke's law*, that relates the stress tensor $\underline{\boldsymbol{\sigma}}$ to the deformation, or strain, tensor $\underline{\boldsymbol{\varepsilon}}$:

$$\underline{\boldsymbol{\sigma}} = 2\mu\underline{\boldsymbol{\varepsilon}} + \lambda \mathrm{tr}(\underline{\boldsymbol{\varepsilon}})\underline{\boldsymbol{I}} \ , \tag{7}$$

where $\lambda > 0$, $\mu > 0$ are known as Lamé coefficients of the elastic material, $\underline{\boldsymbol{\varepsilon}}$ is the strain tensor

$$\underline{\boldsymbol{\varepsilon}} = (\varepsilon_{ij})_{1\leq i,j\leq 3} \ , \qquad \varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \ , \tag{8}$$

$\mathrm{tr}(\underline{\boldsymbol{\varepsilon}}) = \varepsilon_{11} + \varepsilon_{22} + \varepsilon_{33} = \nabla \cdot \boldsymbol{u}$ is its trace and $\underline{\boldsymbol{I}} = (\delta_{ij})_{1\leq i,j\leq 3}$ is the identity tensor. Taking the component $3,1$ of equation (7) and recalling that u_1 is negligible with respect to u_3, gives the approximate constitutive equation

$$\tau_{31} = \mu \frac{\partial u_3}{\partial x_1} \ . \tag{9}$$

Eventually, we omit indices in u_3, τ_{31} f_3 just to simplify the notation. The coefficient μ, called *shear modulus*, can be expressed using *Young's module* E and the *Poisson coefficient* ν, as follows

$$\mu = \frac{E}{2(1+\nu)} \ . \tag{10}$$

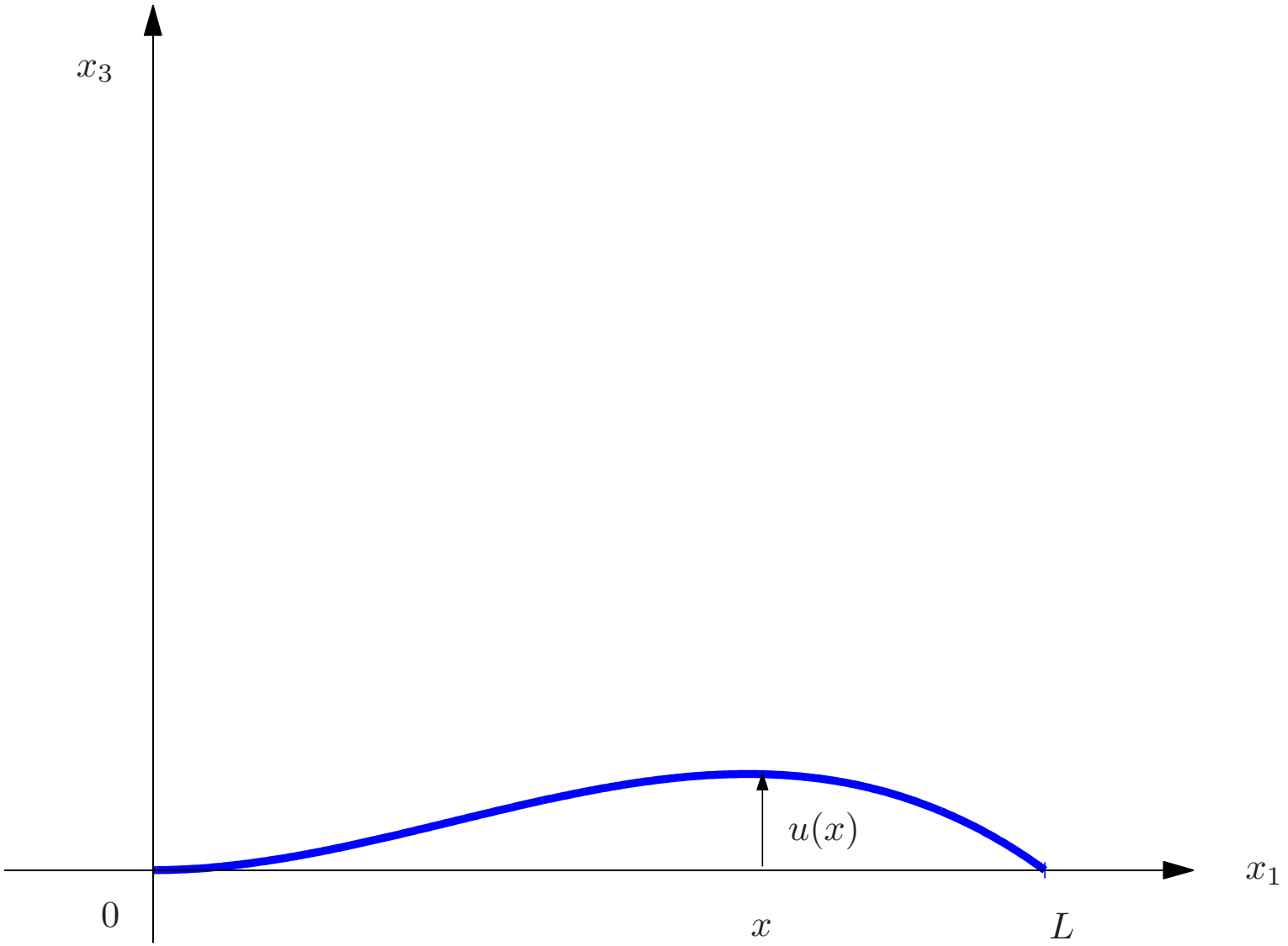
Thus, we have obtained the following system of equations:

$$\begin{cases} \dfrac{d\tau}{dx} + f = 0 & \text{in } (0, L) \ , \\ \tau = \mu \dfrac{du}{dx} & \text{in } (0, L) \ , \\ u(0) = u(L) = 0 \ , \end{cases} \tag{11}$$

where the latter tells that the string is fixed at the endpoints of the interval $[0, L]$. Substituting the expression for τ in the former equation produces

$$\begin{cases} -\dfrac{d}{dx}\left(\mu \dfrac{du}{dx}\right) = f & \text{in } (0, L) \ , \\ u(0) = u(L) = 0 \ . \end{cases} \tag{12}$$

Hence the string's displacement solves a *boundary value problem* for a linear differential equation of order two. The problem admits one, and one only, solution if, for instance, μ and f are continuous functions (or piecewise continuous) on $[0, L]$.


x_3
$u(x)$
x_1
0
x
L

It is easy to write u in terms of f by integration. Integrating the first equation of (11), in fact, gives

$$\tau(x) = C_1 - \int_0^x f(s)\, ds \tag{13}$$

while integrating the second of (11) produces

$$u(x) = C_2 + \int_0^x \frac{\tau(s)}{\mu(s)}\, ds\ ; \tag{14}$$

substituting (13) in (14) leads to the required expression for u. The constraint $u(0) = 0$ implies straightforward $C_2 = 0$, while the constant C_1 is determined by imposing $u(L) = 0$, that is $\displaystyle\int_0^L \frac{\tau(s)}{\mu(s)}\, ds = 0$.

However, the procedure to find the analytical solution just described may be rather involved due to the computation of the integrals. In addition, it cannot be generalised to bidimensional models.

For these reasons we opt for another way that consists in *discretising* problem (12) and reducing it to a linear algebraic system.

THE MODEL OF HEATED BEAM

Let us consider a thin metallic beam whose axis occupies the position of the interval $[0, L]$ on the x-axis. If we identify the beam with its axis, we are allowed to describe all physical quantity of interest as functions of the abscissa x and time t.

In particular, we will see that the temperature $u = u(x, t)$ of the material point of abscissa x at time t satisfies the equation

$$c\,\rho\,\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(\kappa\frac{\partial u}{\partial x}\right) = \rho\,q\ ,$$

where

- ρ is the mass density per unit of length,
- c is the specific heat of the beam,
- κ is the coefficient of termal conductivity of the beam,
- q is the external heat contribution per unit of mass and length.

In the steady-state case, i.e., when all variables do not depend upon time, the equation takes the simplified form

$$-\frac{d}{dx}\left(\kappa\frac{du}{dx}\right) = \rho\,q\ .$$

Setting $\mu = \kappa$ and $f = \rho\, q$, we thus obtain the same mathematical equation which describes the equilibrium of the elastic string, namely,

$$-\frac{d}{dx}\left(\mu \frac{du}{dx}\right) = f \qquad \text{in} \quad (0, L) \ .$$

The equation has to be supplemented by one condition at each extremum of the interval (boundary conditions). For instance, we can prescribe the temperature values g_0 and g_L at the extrema:

$$u(0) = g_0 \ , \qquad u(L) = g_L \ .$$

As an alternative, in addition to the temperature at one extremum we can prescribe the heat flux at the opposite extremum, e.g.,

$$u(0) = g_0 \ , \qquad \mu \frac{du}{dx}(L) = \psi_L \ ,$$

for a given ψ_L.

Let N be any integer ≥ 1; let us set $h = \dfrac{L}{N+1}$. In the interval $[0, L]$ let us introduce the equally-spaced *nodes* $x_j = hj$, with $j = 0, 1, \ldots, N+1$. We have

$$0 = x_0 < x_1 < \ \ldots \ < x_{j-1} < x_j < x_{j+1} < \ \ldots \ < x_N < x_{N+1} = L \, .$$

Such nodes form our *computational grid*.

Let us associate to each node x_j a value u_j, that we think as an approximation of the displacement u at this node, i.e., $u_j \simeq u(x_j)$.

The prescribed boundary conditions immediately yield the values $u_0 = u_{N+1} = 0$.

Hence, we have to determine the values u_j at the internal nodes, whose number is N. These values will be our *discrete unknowns*. In order to accomplish this task, we use the differential equations at suitable internal points of the interval $[0, L]$.

A *finite difference method* is based on the two following fundamental ingredients:

1. the approximation of the derivatives that appear in the equations, by means of suitable numerical differentiation formulas (such as incremental quotients); these involve continguous nodes of the grid;
2. the requirement that the resulting equations be satisfied at the internal nodes.

In order to accomplish Step 1 above, a simple and natural choice consists of approximating any derivative that appears in the equations by means of a *centered* incremental quotient, based on two points symmetrically placed with respect to the point at which we want to approximate the derivative; such an approximation turns out to be more accurate than the one given by a *backward* or *forward* incremental quotient, again based on two points.

Let us first consider the remarkable particular case when the elastic coefficient μ is constant in $[0, L]$. In this situation, the differential equation (12) which defines the displacement becomes

$$-\mu \frac{d^2 u}{dx^2} = f \ .$$

Let us approximate the second derivative by the centered second incremental quotient

$$\frac{d^2 u}{dx^2}(x_j) \simeq \frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2} \simeq \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} \ .$$

Step 2 consists in enforcing the equations

$$-\mu \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} = f_j \ , \qquad j = 1, \ldots, N \ , \tag{15}$$

namely

$$\frac{\mu}{h^2}\big(- u_{j-1} + 2u_j - u_{j+1}\big) = f_j \ , \qquad j = 1, \ldots, N \ ,$$

where we have set $f_j = f(x_j)$ (we assume here that f is a continuous function in $[0, L]$).

The effect of the boundary conditions

- Note that, taking into account the boundary condition $u_0 = 0$, the first equation $(j = 1)$ involves only two unknowns:

$$\frac{\mu}{h^2}\left(2u_1 - u_2\right) = f_1 \ .$$

- Similarly, taking into account the boundary condition $u_{N+1} = 0$, the last equation $(j = N)$ involves only two unknowns:

$$\frac{\mu}{h^2}\left(- u_{N-1} + 2u_N\right) = f_N \ .$$

- On the other hand, all other equations $(2 \leq j \leq N - 1)$ involve three consecutive unknowns:

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

Hence, we have got a system of N linear equations in the N internal unknowns u_j. It can be written in matrix form as

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f} \ , \tag{16}$$

with column vectors in $\mathbb{R}^N$

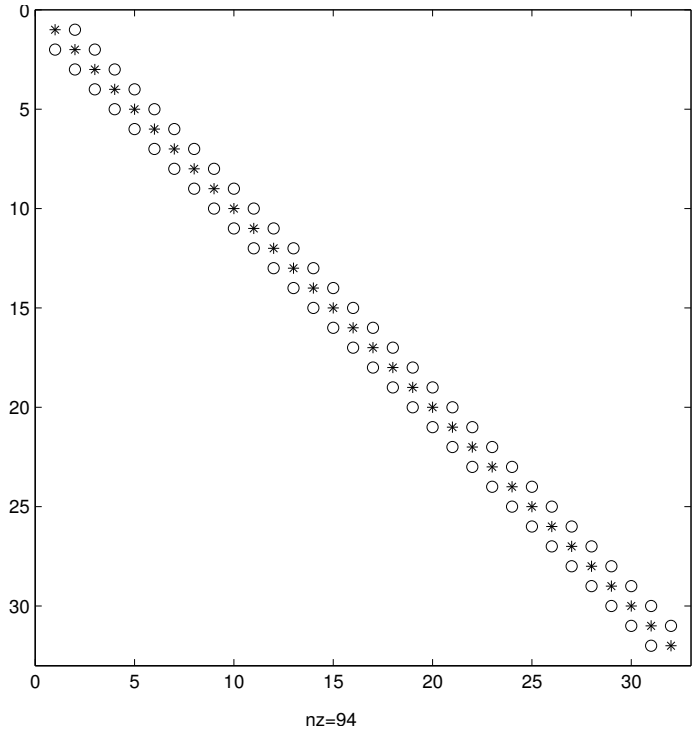
$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} \ , \qquad \boldsymbol{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix} \ , \tag{17}$$

and square matrix $\boldsymbol{A}$ of order N, whose elements a_{jk} are given by

$$a_{jk} = \frac{\mu}{h^2} \begin{cases} 2 & \text{if} \quad k = j \ , \\ -1 & \text{if} \quad k = j \pm 1 \ , \\ 0 & \text{otherwise} \ , \end{cases} \tag{18}$$

which we write as

$$\boldsymbol{A} = \frac{\mu}{h^2} \, \text{tridiag} \, [-1 \quad 2 \quad -1] \ .$$



Note that the matrix $\boldsymbol{A}$ is *tridiagonal*, i.e., the non-zero elements appear only in the main diagonal and in the first upper- and lower-diagonal.
This is a particular instance of a *banded* matrix, namely a matrix whose non-zero elements are contained in a band made of $2m + 1$ diagonals symmetrically placed around the main diagonal (the integer m is called the *band half-width*); in our case, we have $m = 1$.
In turns, a banded matrix with m much smaller than N is a particular instance of a *sparse* matrix, i.e., a matrix such that the number of its non-zero elements is small compared to the total number of elements; in our case, the non-zero elements are $3N - 2$ out of a total of N^2 elements.
At last, our matrix is symmetric, since $a_{j,j+1} = a_{j+1,j}$ for any j.

Notation It is convenient to introduce a specific notation to indicate those tridiagonal matrices having equal elements in each diagonal (with the possible exception of those in the first and/or last row). Precisely, we will set

$$\text{tridiag}\,[a \quad b \quad c] = \begin{pmatrix} b & c & & & & & & \\ a & b & c & & & & & \\ & a & b & c & & & & \\ & & \cdot & \cdot & \cdot & & & \\ & & & \cdot & \cdot & \cdot & & \\ & & & & a & b & c & \\ & & & & & a & b & c \\ & & & & & & a & b \end{pmatrix} \tag{19}$$

and

$$\text{tridiag}\,[b' \quad c';\quad a \quad b \quad c;\quad a'' \quad b''] = \begin{pmatrix} b' & c' & & & & & & \\ a & b & c & & & & & \\ & a & b & c & & & & \\ & & \cdot & \cdot & \cdot & & & \\ & & & \cdot & \cdot & \cdot & & \\ & & & & a & b & c & \\ & & & & & a & b & c \\ & & & & & & a'' & b'' \end{pmatrix}. \tag{20}$$

Let us now consider the general case of a variable coefficient μ in $[0, L]$, namely equations (11).

In order to realize the centered incremental quotients, it is convenient to enrich the computational grid by introducing new nodes having a semi-integer index, namely $x_{j+1/2} = h(j + 1/2)$, with $j = 0, \ldots, N$; each of them is placed in between two continguous nodes with integer indices. We associate to them the approximate values $\tau_{j+1/2} \simeq \tau(x_{j+1/2})$ of the shear stress.

Based on these ideas, let us introduce the following approximations of the first derivative of τ and of u:

$$\frac{d\tau}{dx}(x_j) \simeq \frac{\tau_{j+1/2} - \tau_{j-1/2}}{h} \qquad \text{and} \qquad \frac{du}{dx}(x_{j+1/2}) \simeq \frac{u_{j+1} - u_j}{h} \ . \tag{21}$$

Step 2 above is accomplished by enforcing the approximate form of the equilibrium equation at the nodes with integer indices:

$$\frac{\tau_{j+1/2} - \tau_{j-1/2}}{h} + f_j = 0 \,, \qquad j = 1, \ldots, N \,, \tag{22}$$

as well as the approximate form of the constitutive equation at the nodes with semi-integer indices:

$$\tau_{j+1/2} = \mu_{j+1/2} \frac{u_{j+1} - u_j}{h} \,, \qquad j = 0, \ldots, N \,, \tag{23}$$

where we have set $f_j = f(x_j)$ and $\mu_{j+1/2} = \mu(x_{j+1/2})$.

(We assume throughout this Section that f and μ are continuous functions in $[0, L]$).

Substituting (23) into (22), we obtain the equations

$$-\frac{1}{h}\left(\mu_{j+1/2}\frac{u_{j+1}-u_j}{h}-\mu_{j-1/2}\frac{u_j-u_{j-1}}{h}\right)=f_j\ ,\qquad j=1,\ldots,N\ ,$$

which can be written as

$$\frac{1}{h^2}\left(-\mu_{j-1/2}u_{j-1}+(\mu_{j-1/2}+\mu_{j+1/2})u_j-\mu_{j+1/2}u_{j+1}\right)=f_j\ ,\qquad j=1,\ldots,N\ . \tag{24}$$

Hence, we obtain again a system of N linear equations in the N internal unknowns u_j. We write this system in matrix form as

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

where $\boldsymbol{A}$ is the square tridiagonal symmetric matrix of order N whose elements are given by

$$a_{jk}=\frac{1}{h^2}\begin{cases}\mu_{j-1/2}+\mu_{j+1/2} & \text{if } k=j\ , \quad \text{with } 1\le j\le N\ ,\\ -\mu_{j-1/2} & \text{if } k=j-1\ , \quad \text{with } 2\le j\le N\ ,\\ -\mu_{j+1/2} & \text{if } k=j+1\ , \quad \text{with } 1\le j\le N-1\ ,\\ 0 & \text{otherwise .}\end{cases} \tag{25}$$

To the equation

$$-\frac{d}{dx}\left(\mu\frac{du}{dx}\right) = f \qquad \text{in } (0, L) \ ,$$

we may associate the so-called **Dirichlet boundary conditions**

$$u(0) = g_0 \ , \qquad u(L) = g_L \ ;$$

they are said to be *non-homogeneous* when the assigned values are $\neq 0$.

As an alternative, one of the so-called **Neumann boundary conditions**

$$\mu\frac{du}{dx}(0) = \psi_0 \ , \qquad \text{or} \qquad \mu\frac{du}{dx}(L) = \psi_L \ ,$$

may be enforced at the corresponding endpoint of the interval; physically, any such condition amounts to assigning the value of the shear stress (in the elastic model) or the heat flux (in the thermal model).

Non-homogeneous Dirichlet boundary conditions

Let us suppose that we have to enforce

$$u(0) = g_0 \ .$$

In the first equation (relative to the first internal node x_1)

$$\frac{1}{h^2}\left(-\mu_{1/2}u_0 + (\mu_{1/2} + \mu_{3/2})u_1 - \mu_{3/2}u_2\right) = f_1 \ ,$$

let us replace u_0 by the value g_0, and let us move the corresponding term to the right-hand side

$$\frac{1}{h^2}\left((\mu_{1/2} + \mu_{3/2})u_1 - \mu_{3/2}u_2\right) = f_1 + \frac{\mu_{1/2}}{h^2}g_0 \ .$$

If we have to enforce

$$u(L) = g_L \ ,$$

we manipulate in a similar manner the last equation (relative to the last internal node x_N), and we get

$$\frac{1}{h^2}\left(-\mu_{N-1/2}u_{N-1} + (\mu_{N-1/2} + \mu_{N+1/2})u_N \right) = f_N + \frac{\mu_{N+1/2}}{h^2}g_L \ .$$

In conclusion, it is enough **to modify the first and last entry of the right-hand side of the algebraic system**.

Neumann boundary conditions

Let us suppose that we have to enforce

$$\mu \frac{du}{dx}(L) = \psi_L \ .$$

Since the value of u in $x_{N+1} = L$ is not prescribed, we have an additional unknown, $u_{N+1} \simeq u(x_{N+1}) = u(L)$.

Thus, we need to add a new equation, at $x = x_{N+1} = L$. Assuming for simplicity that μ is constant, let us add

$$-\mu \frac{u_N - 2u_{N+1} + u_{N+2}}{h^2} = f(L) \ ;$$

however, a further unknown u_{N+2}, associated with the *faked node* $x_{N+2} = L + h$ outside the interval $[0, L]$, is introduced.

This unknown, though, is swiftly eliminated by imposing at the node x_{N+1} the approximate Neumann condition given by the centred difference quotient, i.e.,

$$\mu \frac{u_{N+2} - u_N}{2h} = \psi_L \ .$$

In such a way, the **second-order accuracy** of the discretization is preserved.

From this equation, we get

$$u_{N+2} = u_N + \frac{2h}{\mu}\psi_L \ .$$

Substituting this value in the preceding equation and dividing by 2, we thus obtain the $(N+1)$-th equation of our algebraic system:

$$\frac{\mu}{h^2}\big(-u_N + u_{N+1}\big) = \frac{1}{2}f(L) + \frac{1}{h}\psi_L = f_{N+1} \ .$$

Hence, the matrix of the algebraic system

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

obtained in this way, now of order $N+1$, is given by

$$\boldsymbol{A} = \frac{\mu}{h^2}\,\text{tridiag}\,[-1 \quad 2 \quad -1;\quad -1 \quad 1]\ ,$$

whereas the value ψ_L of the Neumann condition appears in the last entry of the vector $\boldsymbol{f}$.

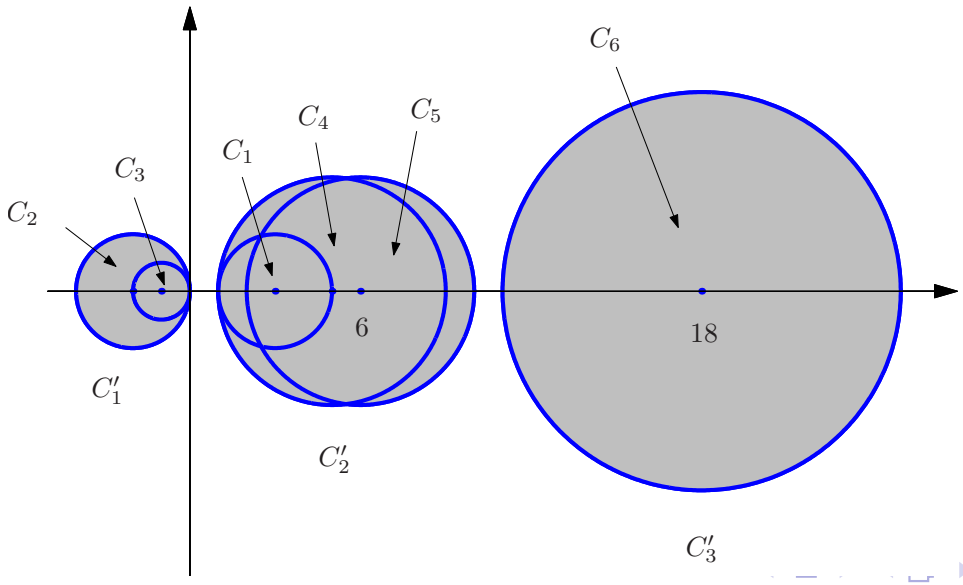
Gerschgorin's Theorem

Gerschgorin's Theorem provides some easy-to-check criteria in order to localize the eigenvalues of a square matrix in the complex plane.

Given a real square matrix $\boldsymbol{A}$ of order n, let us define the *Gerschgorin circles*

$$C_i = \{z \in \mathbb{C} \ : \ |z - a_{ii}| \leq r_i = \sum_{j=1, j\neq i}^{n} |a_{ij}| \} \ , \qquad i = 1, 2, \ldots, n \ . \tag{26}$$

Note that the circle C_i has center on the real axis at the point of abscissa a_{ii}, and radius equal to the sum of the moduli of the elements belonging to the row i, outside the diagonal.



Theorem

Let $\boldsymbol{A}$ be a real square matrix of order n and $C_i,\ i = 1, 2, \ldots, n$, its Gerschgorin discs. Then:

1. *each eigenvaue λ di $\boldsymbol{A}$ belongs to the union $C = \bigcup_{i=1}^{n} C_i$ of all Gerschgorin discs.*
2. *If the union $C' = \bigcup_{k=1}^{m} C_{i_k}$ of m Gerschgorin discs is disjoint from the union of the remaining $n - m$ discs (we say C' is a* connected component *of C), then exactly m eigenvalues of $\boldsymbol{A}$ belong in C'.*
3. *Let $\boldsymbol{A}$ be irreducible, meaning there exists no row or column permutation making $\boldsymbol{A}$ block diagonal, i.e., of the form*

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{O} \\ \boldsymbol{O}^T & \boldsymbol{A}_{22} \end{pmatrix}$$

with $\boldsymbol{A}_{11},\ \boldsymbol{A}_{22}$ square of order $< n$. If one eigenvalue λ lies on the boundary of C, then λ belongs to every Gerschgorin disc of $\boldsymbol{A}$.

Since the eigenvalues of the transpose matrix $\boldsymbol{A}^T$ coincide with those of $\boldsymbol{A}$, we can apply Gerschgorin's theorem to the transpose, and locate eigenvalues with more accuracy; the centres of the discs are unchanged, whereas radia are the sums of the absolute values of off-diagonal column elements of $\boldsymbol{A}$.

If $\boldsymbol{A}$ is symmetric its eigenvalues are real, so it is enough to analyse *Gerschgorin intervals* $\widehat{C}_i$, given by the intersections of the discs C_i with the real axis.

Example

Let $\boldsymbol{A}$ be the irreducible, symmetric 6×6 matrix

$$\boldsymbol{A} = \begin{pmatrix} 3 & 0 & 1 & 0 & 0 & -1 \\ 0 & -2 & 0 & 1 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 5 & 0 & 3 \\ 0 & 1 & 0 & 0 & 6 & -3 \\ -1 & 0 & 0 & 3 & -3 & 18 \end{pmatrix} .$$

Gershgorin's intervals are

$$\begin{aligned} \widehat{C}_1 &= \{x \in \mathbb{R} \ : \ |x-3| \leq 2\} = [1,5] \, , \\ \widehat{C}_2 &= \{x \in \mathbb{R} \ : \ |x+2| \leq 2\} = [-4,0] \, , \\ \widehat{C}_3 &= \{x \in \mathbb{R} \ : \ |x+1| \leq 1\} = [-2,0] \, , \\ \widehat{C}_4 &= \{x \in \mathbb{R} \ : \ |x-5| \leq 4\} = [1,9] \, , \\ \widehat{C}_5 &= \{x \in \mathbb{R} \ : \ |x-6| \leq 4\} = [2,10] \, , \\ \widehat{C}_6 &= \{x \in \mathbb{R} \ : \ |x-18| \leq 7\} = [11,25] \, , \end{aligned}$$

whereas Gershgorin's disks have been shown three slides back.

Example (continued)

The connected components are

$$\widehat{C}_1' = \widehat{C}_2 \cup \widehat{C}_3 = [-4, 0], \qquad \widehat{C}_2' = \widehat{C}_1 \cup \widehat{C}_4 \cup \widehat{C}_5 = [1, 10], \qquad \widehat{C}_3' = \widehat{C}_6 = [11, 25] \, .$$

By Gershgorin's theorem two eigenvalues of $\boldsymbol{A}$ belong to the open interval $(-4, 0)$, three are in the open interval $(1, 10)$ and one in the open interval $(11, 25)$. The eigenvalues of $\boldsymbol{A}$, computed in MATLAB, read:

$$\begin{aligned} \lambda_1 &= -2.2590... \\ \lambda_2 &= -1.2393... \\ \lambda_3 &= 3.1028... \\ \lambda_4 &= 4.1440... \\ \lambda_5 &= 5.8899... \\ \lambda_6 &= 19.3616... \end{aligned}$$

Properties of the finite difference matrix

Let us go back to the matrix $\boldsymbol{A}$ obtained from the discretization of the elastic string problem by finite differences (defined in (25)).

Theorem

The matrix $\boldsymbol{A}$ is symmetric and positive definite, hence in particular it is non-singular.

Let us remember that a symmetric matrix $\boldsymbol{A}$ is **positive-definite** if

$$\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} > 0 \qquad \text{for any vector } \boldsymbol{x} \neq \boldsymbol{0}\ ,$$

or, equivalently, if

$$\text{all the eigenvalues of } \boldsymbol{A} \text{ are } > 0\ .$$

The theorem is a consequence of Gerschgorin's theorem.

Let us check this statement in the particular case in which μ is constant, i.e., when

$$\boldsymbol{A} = \frac{\mu}{h^2} \text{ tridiag} \begin{bmatrix} -1 & 2 & -1 \end{bmatrix} .$$

The Gerschgorin intervals $\boldsymbol{A}$ are

$$\widehat{C}_1 = \widehat{C}_N = \frac{\mu}{h^2} \{x \in \mathbb{R} \ : \ |x-2| \leq 1\} = \frac{\mu}{h^2} [1,3] \ ,$$

$$\widehat{C}_j = \frac{\mu}{h^2} \{x \in \mathbb{R} \ : \ |x-2| \leq 2\} = \frac{\mu}{h^2} [0,4] \ , \quad j = 2, \ldots, N-1 \ .$$

Hence all the eigenvalues of $\boldsymbol{A}$ are strictly positive: in fact, every Gerschgorin interval is contained in the positive x-semiaxis, and the origin does not belong to all Gerschgorin intervals; therefore 0 cannot be an eigenvalue.

Let $\boldsymbol{x} = (x_i)_{1 \leq i \leq n} \in \mathbb{R}^n$ be a column vector with n real entries. If p is an arbitrary real number ≥ 1, one calls p-*norm* of $\boldsymbol{x}$ the quantity

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} .$$

Especially important are the *norms*

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i| , \qquad \|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2} , \qquad \|\boldsymbol{x}\|_\infty = \max_{1 \leq i \leq n} |x_i| .$$

Let then $\boldsymbol{A} = (a_{ij})_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$ be a square matrix of size n. To each vector norm $\|\boldsymbol{x}\|$ is associated a *matrix norm* $\|\boldsymbol{A}\|$, defined by

$$\|\boldsymbol{A}\| = \max_{\boldsymbol{x} \in \mathbb{R}^n \ \boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{x} \in \mathbb{R}^n \ \|\boldsymbol{x}\| = 1} \|\boldsymbol{A}\boldsymbol{x}\| .$$

From the definition, it easily follows

$$\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\| \, \|\boldsymbol{x}\| \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^n ,$$

and

$$\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\| \, \|\boldsymbol{B}\| , \qquad \|\boldsymbol{I}\| = 1 .$$

In particular, one has

$$\|\boldsymbol{A}\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| = \|\boldsymbol{A}^T\|_1$$

and

$$\|\boldsymbol{A}\|_2 = \sqrt{\rho(\boldsymbol{A}^T\boldsymbol{A})}\ ,$$

where $\rho(\boldsymbol{B})$ denotes the *spectral radius* of a matrix $\boldsymbol{B}$, namely

$$\rho(\boldsymbol{B}) = \max\{\,|\lambda| \ :\ \lambda \text{ is an eigenvalue of } \boldsymbol{B}\,\}\ .$$

If $\boldsymbol{A}$ is a symmetric matrix (hence, it has all real eigenvalues), it holds

$$\|\boldsymbol{A}\|_2 = \max\{\,|\lambda| \ :\ \lambda \text{ is an eigenvalue of } \boldsymbol{A}\,\}\ .$$

If, in addition, $\boldsymbol{A}$ is positive definite, with eigenvalues that satisfy

$$0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n\ ,$$

then, setting $\lambda_{\mathsf{min}} = \lambda_1$ and $\lambda_{\mathsf{max}} = \lambda_n$, we have

$$\|\boldsymbol{A}\|_2 = \lambda_{\mathsf{max}}\ .$$

Recalling that

$$\boldsymbol{A}\boldsymbol{w} = \lambda\boldsymbol{w} \quad \Leftrightarrow \quad \boldsymbol{A}^{-1}\boldsymbol{w} = \lambda^{-1}\boldsymbol{w}\ ,$$

we immediately obtain

$$\|\boldsymbol{A}^{-1}\|_2 = \frac{1}{\lambda_{\mathsf{min}}}\ .$$

The condition number of a matrix

Let $\boldsymbol{A}$ be a nonsingular, square matrix. The number

$$\text{cond}_p(\boldsymbol{A}) = \|\boldsymbol{A}\|_p \, \|\boldsymbol{A}^{-1}\|_p$$

is called the *condition number* of $\boldsymbol{A}$ (with respect to the p-norm).
We always have

$$\text{cond}_p(\boldsymbol{A}) \geq 1 \, .$$

A matrix $\boldsymbol{A}$ is called *well-conditioned* if $\text{cond}_p(\boldsymbol{A}) \simeq 1$, *ill-conditioned* if $\text{cond}_p(\boldsymbol{A}) >> 1$.

Let $\boldsymbol{b} \in \mathbb{R}^n$ be a non-zero vector (representing the "input data" of a certain problem, or the "initial state" of a physical system), and let $\boldsymbol{x} \in \mathbb{R}^n$ be the solution to the linear system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \, ,$$

(representing the "solution" to the problem, or the "exit state" of the physical system). Now suppose to know not $\boldsymbol{b}$, but rather only an approximation $\tilde{\boldsymbol{b}}$ of it, by a series of reasons (measuring errors, errors in numerical representation, et c.); correspondingly, we have a solution $\tilde{\boldsymbol{x}}$ defined by

$$\boldsymbol{A}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}} \, ;$$

we can reasonably expect $\tilde{\boldsymbol{x}}$ to approximate $\boldsymbol{x}$. Then, one has:

$$\frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_p}{\|\boldsymbol{x}\|_p} \leq \text{cond}_p(\boldsymbol{A}) \, \frac{\|\boldsymbol{b} - \tilde{\boldsymbol{b}}\|_p}{\|\boldsymbol{b}\|_p} \, .$$

For a positive-definite, symmetric matrix, the condition number in Euclidean norm is given by

$$\mathrm{cond}_2(\boldsymbol{A}) = \frac{\lambda_{\max}}{\lambda_{\min}} \, .$$

Hence, *a positive-definite, symmetric matrix is ill-conditioned whenever the orders of magnitude of its eigenvalues are considerably different*.

Example

A typical example of (positive-definite, symmetric) matrices that are very ill-conditioned is provided by the class of *Hilbert matrices* $\boldsymbol{H}_n$ $(n \geq 1)$, whose entries are

$$h_{ij} = \frac{1}{i+j-1} \, , \qquad 1 \leq i, j \leq n \, ,$$

(such matrices are defined by the MATLAB instruction hilb).
The condition numbers $\mathrm{cond}_2(\boldsymbol{H}_n)$ (estimated by the MATLAB command cond) grow exponentially fast as n grows!

Going back to the matrix $\boldsymbol{A}$ of the finite difference discretization, if the physical coefficient μ is constant one can explicitly compute its eigenvalues, that we denote by

$$\lambda_{h,1} < \lambda_{h,2} < \cdots < \lambda_{h,p} < \cdots < \lambda_{h,N-1} < \lambda_{h,N} \; ;$$

indeed, one has

$$\lambda_{h,p} = \frac{2\mu}{h^2} \left(1 - \cos\left(p \, \frac{\pi h}{L} \right) \right) \; , \qquad p = 1, \ldots, N \; .$$

Recalling that $1 - \cos t \sim \frac{1}{2} t^2$ as $t \to 0$, and that $\cos t \to -1$ as $t \to \pi$, we get the asymptotic behaviour of the minimum aigenvalue $\lambda_{h,1}$ and maximum eigenvalue $\lambda_{h,N}$ when h tends 0 (or, equivalently, when N tends to $+\infty$):

$$\lambda_{h,1} \sim \lambda_1 := \mu \, \frac{\pi^2}{L^2} \; , \qquad \lambda_{h,N} \sim \frac{4\mu}{h^2} \; .$$

In particular, we derive that the **condition number** of $\boldsymbol{A}$ in Euclidean norm satisfies

$$\mathrm{cond}_2(\boldsymbol{A}) = \frac{\lambda_{h,N}}{\lambda_{h,1}} \sim \frac{4L^2}{\pi^2} \, h^{-2} \; ;$$

Hence, the matrix becomes **more and more ill-conditioned** as the discretisation step h decreases.

Hereafter, let us measure the magnitude of a vector using the mean-square norm

$$\|\boldsymbol{v}\|_{2,m} = \frac{1}{\sqrt{N}}\|\boldsymbol{v}\|_2 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|v_i|^2}\ ,$$

or the maximum norm $\|\boldsymbol{v}\|_\infty$.
The vector $\boldsymbol{u}$ that solves the algebraic system generated by the finite difference method satisfies

$$\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f} = \boldsymbol{0}\ .$$

Let $\boldsymbol{u}^e = (u(x_j))_{1\le j\le N} \in \mathbb{R}^N$ be the vector whose entries are the values of the exact solution of the equation at the inner nodes. In general, the *residual vector*, or *truncation error*,

$$\boldsymbol{A}\boldsymbol{u}^e - \boldsymbol{f} = \boldsymbol{r}$$

does not vanish. However, since we have used *second order* numerical differentiation formulas, one can prove that

$$\|\boldsymbol{r}\|_{2,m} \le \frac{\mu}{12}h^2 \max_{x\in[0,L]}\left|\frac{d^4u}{dx^4}(x)\right| = \frac{1}{12}h^2 \max_{x\in[0,L]}\left|\frac{d^2f}{dx^2}(x)\right|\ .$$

Hence, we deduce that

$$\boldsymbol{r} \to \boldsymbol{0} \qquad \text{as } h \to 0\ .$$

Such a property is termed **consistency** of the numerical method.

By substracting equations $\boldsymbol{A}\boldsymbol{u}^e - \boldsymbol{f} = \boldsymbol{r}$ and $\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f} = \boldsymbol{0}$, we get

$$\boldsymbol{A}(\boldsymbol{u}^e - \boldsymbol{u}) = \boldsymbol{r} \ , \qquad \text{i.e.,} \qquad \boldsymbol{u}^e - \boldsymbol{u} = \boldsymbol{A}^{-1}\boldsymbol{r} \ ,$$

whence

$$\|\boldsymbol{u}^e - \boldsymbol{u}\|_{2,m} \leq \|\boldsymbol{A}^{-1}\|_2 \|\boldsymbol{r}\|_{2,m} \ .$$

But

$$\|\boldsymbol{A}^{-1}\|_2 = \frac{1}{\lambda_{h,1}} \leq C\frac{1}{\lambda_1} \qquad \text{(independent of } h) \ ;$$

such an inequality is referred to as the **stability** of the numerical scheme.

We conclude that

$$\|\boldsymbol{u}^e - \boldsymbol{u}\|_{2,m} \leq Ch^2 \max_{x\in[0,L]} \left| \frac{d^2 f}{dx^2}(x) \right| \ .$$

This shows that the numerical method is **convergent**.
More precisely, if the data f is smooth enough, *the mean-square norm of the difference between the exact and numerical solutions at the internal nodes tends to* 0 *as* $h \to 0$*; in addition, convergence is second-order, namely quadratic in* h.

A similar result holds for the maximum norm $\|\boldsymbol{u}^e - \boldsymbol{u}\|_\infty$.

Finite element methods are based on an *integral formulation*, or *variational formulation*, of the boundary-value problem to be approximated.

Consider the elastic string problem (12), in which we assume the density of force f to be a piecewise-continuous function on $[0, L]$.

Let us introduce a generic function v, defined on $[0, L]$, representing the string's generic displacement from the reference position under external forces.
Using the physics' language, we shall say v is an **ammissible displacement**; in mathematical language, we will call v a **shape function**.
It is absolutely natural for v to be a continuous map (the elastic string should not break), and to vanishes at the interval's endpoints (as the string is fixed there).

Let us multiply the differential equation by an admissible displacement v (called **test function**, in this situation) and then integrate over $[0, L]$; this gives

$$-\int_0^L \frac{d}{dx}\left(\mu \frac{du}{dx}\right) v\, dx = \int_0^L f v\, dx\ . \tag{27}$$

Now we can integrate by parts the left hand side, thus supposing v is piecewiese differentiable (at least), with continuous derivative. The equation becomes

$$\int_0^L \mu \frac{du}{dx}\frac{dv}{dx}\,dx - \left[\mu \frac{du}{dx} v\right]_0^L = \int_0^L fv\,dx\ . \tag{28}$$

However, recalling that we had assumed $v(0) = v(L) = 0$, the boundary terms at $x = 0$ and $x = L$ are actually zero.

Let us denote by V the set of admissible displacements, i.e., let us define

$$V = \{v : [0, L] \to \mathbb{R} \mid v \text{ is continuous on } [0, L],\ \text{piecewise differentiable with continuous derivative, and such that } v(0) = v(L) = 0\}\ .$$

Note that the solution itself to our problem u is an admissible displacement (it describes the string's displacement exactly in correspondence to the load f), hence $u \in V$; this condition incorporates the vanishing of u at the endpoints.

So, we can formulate problem (12) in the following integral manner:

$$\begin{cases} u \in V \text{ and satisfies} \\ \displaystyle\int_0^L \mu \frac{du}{dx}\frac{dv}{dx}\,dx = \int_0^L f v\,dx \qquad \text{for all } \; v \in V\ . \end{cases} \tag{29}$$

We shall call the above the **variational**, or **weak**, formulation of the elastic string problem.

It translates in mathematical terms what in Mechanics is known as the **Principle of Virtual Work**: the work of an external force under an admissible displacement (given by the right-hand-side integral of (29)) equals the work of all elastic reactions of the material (left-hand-side integral).

The differential formulation (12) and the variational one (29) are equivalent if the problem's data μ and f (hence the solution u) are regular enough.

The integral formulation allows to treat more general situations
(a piecewiese-continuous elastic coefficient μ, or a concentrated weight f).

The set V of all admissible displacements is a *vector space*: if v_1 and v_2 are two admissible displacements, then any *linear combination* of them, $\alpha v_1 + \beta v_2$ with arbitrary $\alpha, \beta \in \mathbb{R}$, will be an admissible displacement as well.

Furthermore, if both v_1 and v_2 satisfy the variational equation

$$\int_0^L \mu \frac{du}{dx} \frac{dv}{dx} \, dx = \int_0^L f v \, dx \ ,$$

then also $\alpha v_1 + \beta v_2$ will satisfy it automatically, due to the linearity of definite integrals and derivatives.

We can define a discretization method starting from the variational formulation (29), by considering only *a finite number of independent admissible displacements*.

Their linear combinations will give rise to a vector space, that we denote by V_h and is a subspace of V.

The space V_h is finite-dimensional: the admissible displacements of V_h, called *discrete displacements*, will be determined by a finite set of parameters, known as *degrees of freedom* of the displacement.

We are thus led to consider the following *discrete variational formulation*:

$$\begin{cases} u_h \in V_h \text{ and satisfies} \\ \displaystyle\int_0^L \mu \frac{du_h}{dx}\frac{dv_h}{dx}\,dx = \int_0^L f v_h\,dx \qquad \text{for all } \; v_h \in V_h \; . \end{cases} \tag{30}$$

The *finite element method* represents a simple yet effective way to define spaces of discrete displacements V_h to be employed in the discrete variational formulation just defined.

Once more, we take $N+2$ nodes x_j in $[0, L]$, satisfying

$$0 = x_0 < x_1 < \ \ldots \ < x_{j-1} < x_j < x_{j+1} < \ \ldots \ < x_N < x_{N+1} = L \, ;$$

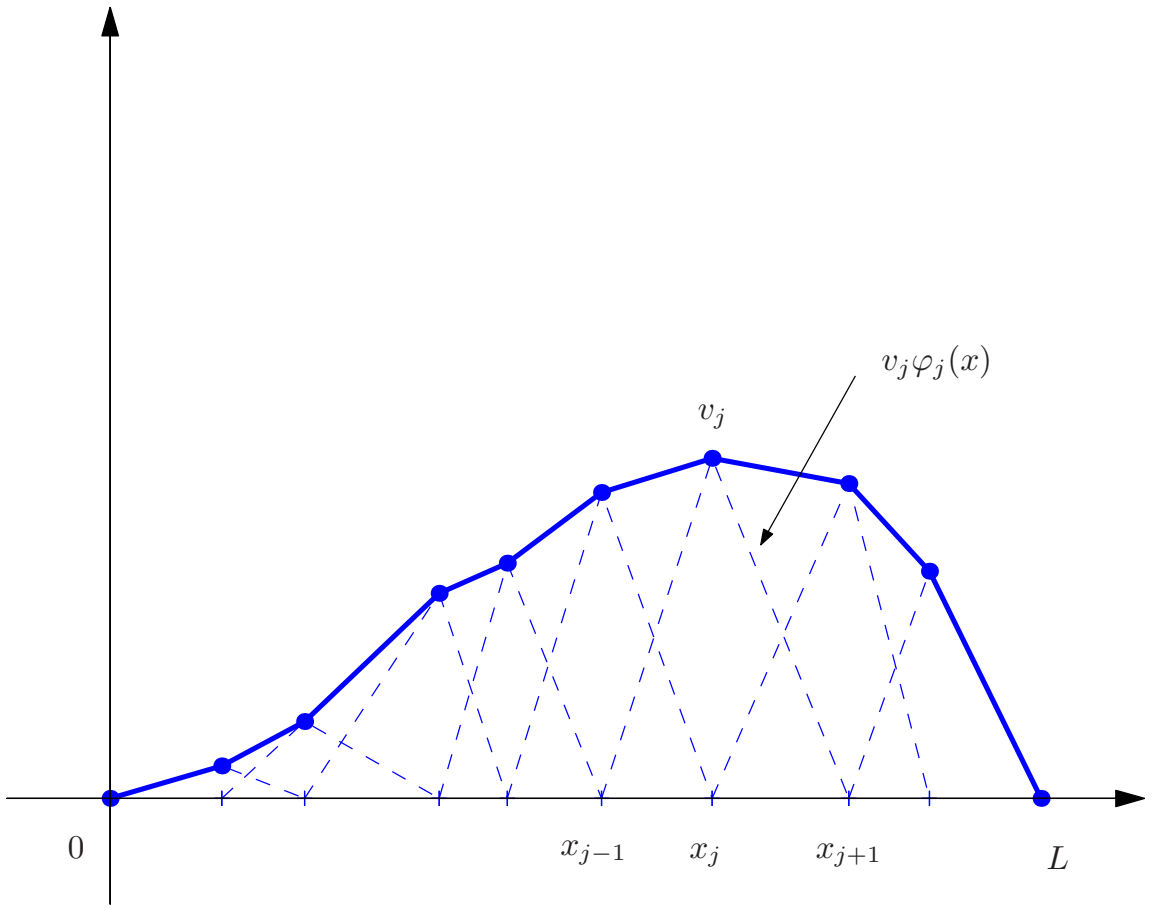
in order to warrant the method a broader generality, we will assume the nodes are not necessarily equidistant. These nodes define a partition of $[0, L]$ into subintervals $I_j = [x_{j-1}, x_j]$, $j = 1, \ldots, N+1$, of length $h_j = x_j - x_{j-1}$; let also conveniently set $h = \max\limits_j h_j$.

The easiest choice for discrete displacements consists in looking at the admissible displacements that, on each interval I_j, are polynomials of degree 1 at most. So let us set

$$V_h = \{v_h \in V \mid v_{h\,|I_j} \in \mathbb{P}_1 \text{ for } j = 1, \ldots, N+1\} \, , \tag{31}$$

where $\mathbb{P}_1$ denotes the set of polynomials on I_j of degree less than or equal to 1.

This choice generates the so-called **linear finite elements**.

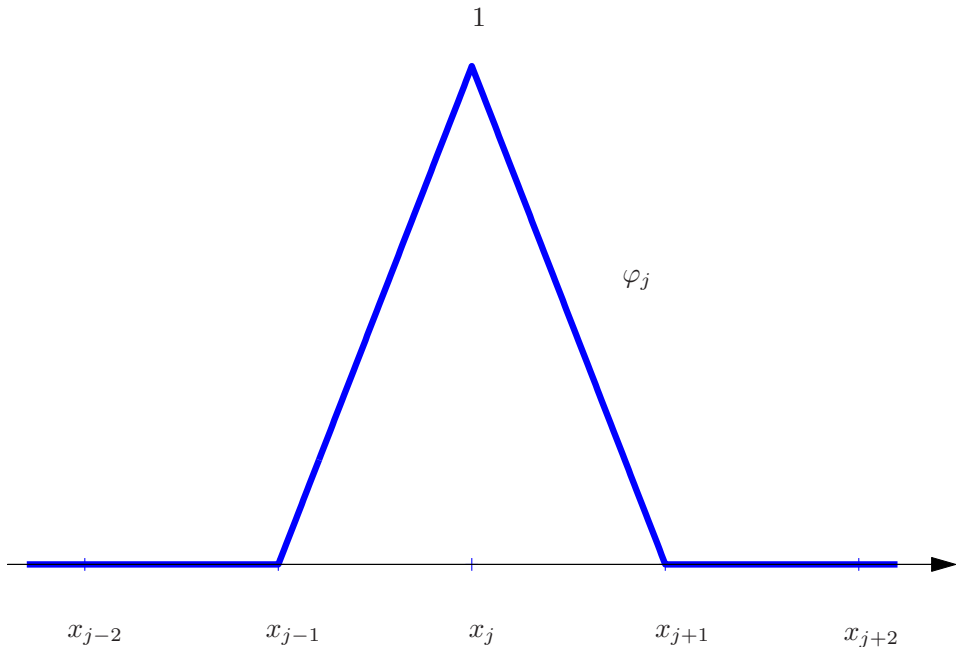


The discrete displacement v_h is uniquely, and comfortably, determined by its values $v_j = v_h(x_j)$ at the internal nodes ($j = 1, \ldots, N$). On each interval I_j, in fact, we can write it as

$$v_h(x) = v_{j-1}\frac{x_j - x}{h_j} + v_j\frac{x - x_{j-1}}{h_j}\,.$$

Therefore we can indentify v_h with the column vector

$$\mathbf{v} = (v_1, v_2, \ldots, v_{N-1}, v_N)^T \in \mathbb{R}^N\,.$$



This fact naturally leads to defining a *basis* in V_h, which allows us to write discrete displacement as a linear combination of the basis functions. We remind that every $\mathbf{v} \in R^N$ can be expressed as

$$\mathbf{v} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + \cdots + v_{N-1}\mathbf{e}_{N-1} + v_N\mathbf{e}_N \ , \tag{32}$$

where $\mathbf{e}_j = (\delta_{jk})$ is the column vector, in the canonical basis, whose components are all 0, except the j-th one, which equals 1.

The vector $\mathbf{e}_j$ defines the discrete displacement $\varphi_j \in V_h$ that is 0 at all nodes except x_j, where is equals 1. Such a function is termed *hat function*, or *tent function*.

The hat function φ_j is written as

$$\varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{h_j} & \text{if } x \in I_j \ , \\ \dfrac{x_{j+1} - x}{h_{j+1}} & \text{if } x \in I_{j+1} \ , \\ 0 & \text{otherwise .} \end{cases} \tag{33}$$

Thus, we can represent each $v_h \in V_h$ as

$$v_h(x) = v_1\varphi_1(x) + v_2\varphi_2(x) + \cdots + v_{N-1}\varphi_{N-1}(x) + v_N\varphi_N(x) = \sum_{j=1}^{N} v_j\varphi_j(x) \ . \tag{34}$$

The functions φ_j, $j = 1, \ldots, N$, form the so-called **Lagrange basis** in V_h.

Reduction to an algebraic system

Let us recall that the approximate solution $u_h \in V_h$ satisfies the discrete variational equations

$$\int_0^L \mu \frac{du_h}{dx} \frac{dv_h}{dx}\, dx = \int_0^L f v_h\, dx \qquad \text{for all} \quad v_h \in V_h\ . \tag{35}$$

Choosing as v_h the basis functions φ_j one at a time, we immediately see that u_h satisfies, in particular, the N equations

$$\int_0^L \mu \frac{du_h}{dx} \frac{d\varphi_j}{dx}\, dx = \int_0^L f \varphi_j\, dx \qquad \text{for} \quad j = 1, \ldots, N\ . \tag{36}$$

This system is indeed *equivalent* to the system of infinitely many equations (35). In fact, each such equation is a linear combination of the equations (36), due to the linearity of definite integrals and derivatives.

Next, let us represent u_h in the Lagrange basis, as

$$u_h = \sum_{k=1}^{N} u_k \varphi_k .$$

Substituting in (36), we get:

$$\int_0^L \mu \frac{d}{dx}\left(\sum_{k=1}^N u_k \varphi_k\right) \frac{d\varphi_j}{dx}\, dx = \int_0^L f\varphi_j\, dx \qquad \text{for} \quad j = 1, \ldots, N\ .$$

Using once more the linearity of derivatives and definite integrals, we arrive at the system of algebraic equations

$$\sum_{k=1}^N u_k \int_0^L \mu \frac{d\varphi_k}{dx}\frac{d\varphi_j}{dx}\, dx = \int_0^L f\varphi_j\, dx \qquad \text{for} \quad j = 1, \ldots, N\ . \tag{37}$$

Setting

$$a_{jk} = \int_0^L \mu \frac{d\varphi_k}{dx}\frac{d\varphi_j}{dx}\, dx\ , \qquad f_j = \int_0^L f\varphi_j\, dx\ , \tag{38}$$

the system (37) is written as

$$\sum_{k=1}^N a_{jk} u_k = f_j \qquad \text{for} \quad j = 1, \ldots, N\ .$$

Repeating,

$$\sum_{k=1}^{N} a_{jk} u_k = f_j \qquad \text{for} \quad j = 1, \ldots, N \ .$$

It is convenient to write such a system in matrix form, as

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f} \ , \tag{39}$$

after setting

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

The matrix $\boldsymbol{A}$ is referred to as the **stiffness matrix**.

Computation of the stiffness matrix

Let us recall that

$$a_{jk} = \int_0^L \mu \frac{d\varphi_k}{dx} \frac{d\varphi_j}{dx}\, dx = \sum_{m=1}^{N+1} \int_{I_m} \mu \frac{d\varphi_k}{dx} \frac{d\varphi_j}{dx}\, dx \ .$$

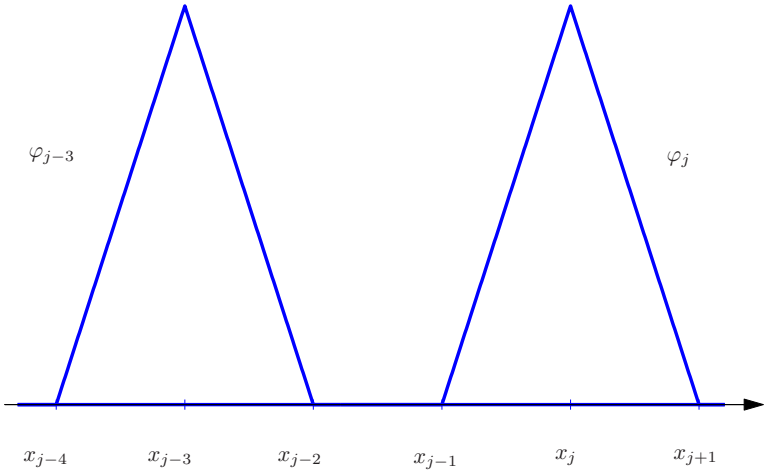
Thus, we note that

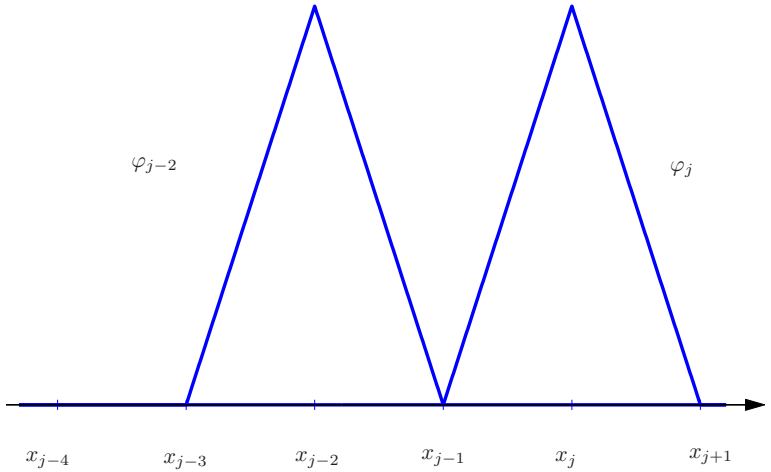
- the matrix $\boldsymbol{A}$ is symmetric,
- each entry of the matrix can be obtained by adding the contribution of each interval of the partition.

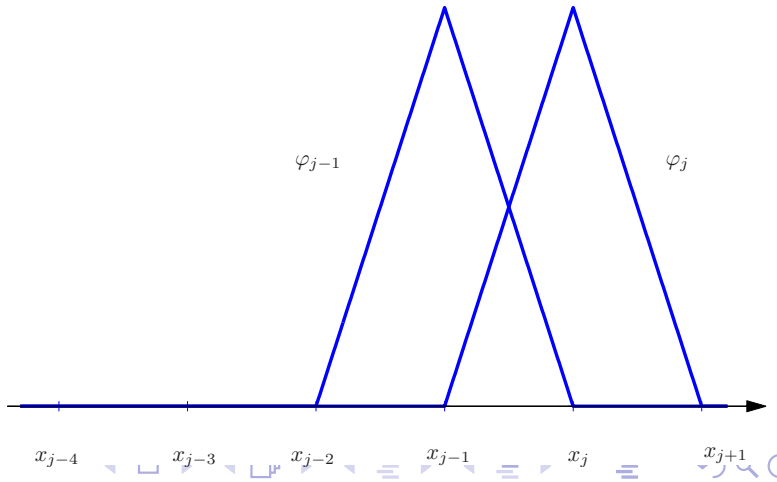
A crucial observation is that the basis function φ_j vanishes identically outside the interval $[x_{j-1}, x_{j+1}]$, that we call the *support* of φ_j. Hence, the product

$$\frac{d\varphi_k}{dx} \frac{d\varphi_j}{dx}$$

vanishes identically if the supports $[x_{k-1}, x_{k+1}]$ and $[x_{j-1}, x_{j+1}]$ do not intersect, i.e., if $|j-k| > 2$.
If $|j-k| = 2$, instead, the supports meet at one point only, yet the product is still 0 on each I_m.
In either case, i.e. for $|j-k| \geq 2$, the entry a_{jk} is zero.


φ_{j-3}
φ_j
x_{j-4}
x_{j-3}
x_{j-2}
x_{j-1}
x_j
x_{j+1}


φ_{j-2}
φ_j
x_{j-4}
x_{j-3}
x_{j-2}
x_{j-1}
x_j
x_{j+1}


φ_{j-1}
φ_j
x_{j-4}
x_{j-3}
x_{j-2}
x_{j-1}
x_j
x_{j+1}

Now let us suppose $k = j - 1$. The functions φ_{j-1} and φ_j have supports intersecting in $[x_{j-1}, x_j] = I_j$; the product of their derivatives is therefore null outside that interval, so

$$a_{j,j-1} = \int_{I_j} \mu \frac{d\varphi_{j-1}}{dx} \frac{d\varphi_j}{dx}\, dx \ .$$

On I_j basis functions are linear, so their derivatives are constant; to be precise, recalling that

$$\varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{h_j} & \text{if } x \in I_j \ , \\ \dfrac{x_{j+1} - x}{h_{j+1}} & \text{if } x \in I_{j+1} \ , \\ 0 & \text{otherwise} \ , \end{cases}$$

we have

$$\frac{d\varphi_{j-1}}{dx} = -\frac{1}{h_j} \ , \qquad \frac{d\varphi_j}{dx} = \frac{1}{h_j} \ ;$$

hence,

$$a_{j,j-1} = -\frac{1}{h_j^2} \int_{I_j} \mu \, dx = -\frac{h_j}{h_j^2} \left(\frac{1}{h_j} \int_{I_j} \mu \, dx \right) = -\frac{1}{h_j} \bar{\mu}_{j-1/2} \ ,$$

where

$$\bar{\mu}_{j-1/2} = \frac{1}{h_j} \int_{I_j} \mu \, dx$$

denotes the mean value over I_j of the elastic coefficient μ. Often this number cannot be computed exactly, rather we can approximate it, for instance, by the coefficient's value $\mu_{j-1/2} = \mu(x_{j-1/2})$ at the interval's middle point $x_{j-1/2} = (x_{j-1} + x_j)/2$.

In conclusion, we set

$$a_{j,j-1} = -\frac{1}{h_j} \mu_{j-1/2} \ .$$

At last, let us consider the diagonal entry a_{jj}. We have

$$a_{jj} = \int_{I_j} \mu \left(\frac{d\varphi_j}{dx} \right)^2 dx + \int_{I_{j+1}} \mu \left(\frac{d\varphi_j}{dx} \right)^2 dx$$

with

$$\frac{d\varphi_j}{dx} = \frac{1}{h_j} \quad \text{on } I_j \ , \qquad \frac{d\varphi_j}{dx} = -\frac{1}{h_{j+1}} \quad \text{on } I_{j+1} \ .$$

Hence, possibly with the same approximation of the elastic coefficient as above, we get

$$a_{jj} = \frac{1}{h_j} \mu_{j-1/2} + \frac{1}{h_{j+1}} \mu_{j+1/2} \ .$$

To sum up, the stiffness matrix is symmetric and tridiagonal, with entries:

$$a_{jk} = \begin{cases} \dfrac{\mu_{j-1/2}}{h_j} + \dfrac{\mu_{j+1/2}}{h_{j+1}} & \text{if } \; k = j \; , \\ -\dfrac{\mu_{j-1/2}}{h_j} & \text{if } \; k = j-1 \; , \\ -\dfrac{\mu_{j+1/2}}{h_{j+1}} & \text{if } \; k = j+1 \; , \\ 0 & \text{otherwise .} \end{cases} \tag{40}$$

In the particular case where μ is constant and the mesh is equally spaced (i.e., $h_j = h$ for all j), one has

$$a_{jk} = \frac{\mu}{h} \begin{cases} 2 & \text{if } \; k = j \; , \\ -1 & \text{if } \; k = j-1 \text{ or } \; k = j+1 \; , \\ 0 & \text{otherwise ,} \end{cases}$$

namely,

$$\boldsymbol{A} = \frac{\mu}{h} \, \text{tridiag} \, [-1 \quad 2 \quad -1] \, .$$

Computation of the forcing term

We have

$$f_j = \int_{I_j} f\varphi_j \, dx + \int_{I_{j+1}} f\varphi_j \, dx \; .$$

For a generic force density we cannot calculate the integrals exactly; each one must therefore be approximated by a numerical integration formula. Using the *trapezoidal rule*

$$\int_a^b g \, dx \simeq (g(a) + g(b)) \frac{b-a}{2}$$

(which is exact if g is linear in $[a, b]$) and recalling the φ_j vanishes at x_{j-1} and x_{j+1}, and equals 1 at x_j, we obtain the approximate value

$$\int_{I_j} f\varphi_j \, dx \simeq f(x_j) \frac{h_j}{2} \; , \qquad \int_{I_{j+1}} f\varphi_j \, dx \simeq f(x_j) \frac{h_{j+1}}{2} \; .$$

Thus, in the actual computations we set

$$f_j = f(x_j) \frac{h_j + h_{j+1}}{2} \; . \tag{41}$$

It is interesting to compare the structure of the algebraic system obtained by the finite difference discretisation,

$$\boldsymbol{A}^{DF}\boldsymbol{u}^{DF} = \boldsymbol{f}^{DF} \ ,$$

with that of the system obtained by the finite element discretisation,

$$\boldsymbol{A}^{EF}\boldsymbol{u}^{EF} = \boldsymbol{f}^{EF} \ .$$

Each entry of either the matrix $\boldsymbol{A}^{EF}$ or the source $\boldsymbol{f}^{EF}$ is of the order of h times the corresponding entry of $\boldsymbol{A}^{DF}$ or $\boldsymbol{f}^{DF}$, where h is the local discretisation spacing. Dimensionally, this is absolutely consistent if we recall that the variational formulation of the elastic string problem, generating the finite element discretisation, arises by *integrating* over the spatial interval the differential formulation, at the base of the discretisation by finite differences.

For equidistant subdivisions of $[0, L]$ (h constant) and source terms computed via the trapezoidal rule, we even have

$$\boldsymbol{A}^{EF} = h\boldsymbol{A}^{DF} \qquad \text{and} \qquad \boldsymbol{f}^{EF} = h\boldsymbol{f}^{DF} \ ,$$

whence

$$\boldsymbol{u}^{EF} = \boldsymbol{u}^{DF} \ .$$

Thus, the two methods provide the same approximation for the displacement u.

Properties of the finite element discretization

- The matrix $\boldsymbol{A}$ is symmetric and positive definite (again by Gerschgorin Theorem).
- The condition number of $\boldsymbol{A}$ satisfies:

$$\lambda_{h,\min} \sim c\,h\,, \qquad \lambda_{h,\max} \sim c\,h^{-1} \qquad \Rightarrow \qquad \mathrm{cond}_2(\boldsymbol{A}) = \frac{\lambda_{h,\max}}{\lambda_{h,\min}} \sim c\,h^{-2}\,.$$

- The error between the exact solution u and the discrete solution u_h satifies:

$$\max_{x\in[0,L]} |u(x) - u_h(x)| \leq Ch^2 \max_{x\in[0,L]} \left| \frac{d^2u}{dx^2}(x) \right| \,.$$

Non-homogeneous Dirichlet boundary conditions

Let us assume that we want to satisfy

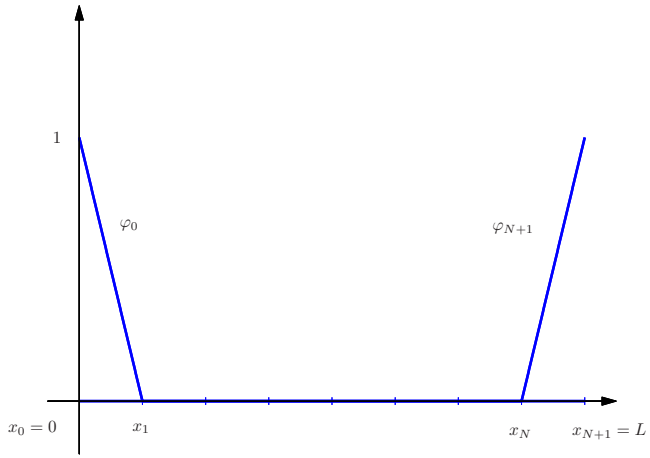
$$u(0) = g_0 \ , \quad u(L) = g_L \ .$$

Two new basis functions are needed, i.e., φ_0 associated with node $x_0 = 0$ and φ_{N+1} associated with node $x_{N+1} = L$.

The discrete solution u_h is now expressed as

$$u_h(x) = g_0\varphi_0(x) + \sum_{k=1}^{N} u_k\varphi_k(x) + g_L\varphi_{N+1}(x)$$

namely $u_h \in V_h(g_0, g_L)$
(space of *discrete admissible displacemets*)



The *discrete variational formulation* is now:

$$\begin{cases} u_h \in V_h(g_0, g_L) \text{ and satisfies} \\ \displaystyle\int_0^L \mu \frac{du_h}{dx}\frac{dv_h}{dx}\,dx = \int_0^L f v_h\,dx \end{cases} \qquad \text{for all } \ v_h \in V_h(0,0) \ .$$

(space of *discrete test displacements*)

The first equation becomes

$$a_{10}g_0 + a_{11}u_1 + a_{12}u_2 = f(x_1)\frac{h_1 + h_2}{2}$$

with

$$a_{10} = \int_{I_1} \mu \frac{d\varphi_0}{dx}\frac{d\varphi_1}{dx}\,dx \simeq -\frac{\mu_{1/2}}{h_1}\,.$$

Hence,

$$\left(\frac{\mu_{1/2}}{h_1} + \frac{\mu_{3/2}}{h_2}\right)u_1 - \frac{\mu_{3/2}}{h_2}u_2 = f(x_1)\frac{h_1 + h_2}{2} + \frac{\mu_{1/2}}{h_1}g_0\,.$$

In an analogous manner we enforce the condition in $x = L$.

In conclusion, for finite elements too, it is enough to **modify the first and last entries of the right-hand side**.

Neumann boundary conditions

Let us assume that we have to enforce

$$u(0) = 0 \ , \qquad \mu \frac{du}{dx}(L) = \psi_L \ .$$

Now, any admissible displacement v is required to vanish only at $x = 0$:

$$V = \{v : [0, L] \to \mathbb{R} \mid v \text{ is continuous on } [0, L], \text{ piecewise differentiable with continuous derivative, and such that } v(0) = 0\} \ ,$$

whereas the discrete admissible displacements

$$V_h = \{v_h \in V \mid v_{h|I_j} \in \mathbb{P}_1 \text{ for } j = 1, \ldots, N+1\}$$

are generated by the basis functions

$$\varphi_1(x), \quad \ldots \quad , \varphi_N \quad \text{and} \quad \varphi_{N+1} \ ,$$

i.e.,

$$v_h(x) = \sum_{j=1}^{N} v_j \varphi_j(x) + v_{N+1} \varphi_{N+1}(x) \ .$$

The variational formulation gets modified: indeed, starting from

$$\int_0^L \mu \frac{du}{dx}\frac{dv}{dx}\,dx - \left[\mu \frac{du}{dx} v\right]_0^L = \int_0^L f v\,dx \qquad \text{for any } v \in V\ ,$$

we now use the Dirichlet condition $v(0) = 0$ and we enforce the Neumann boundary condition at $x = L$, to get

$$\left[\mu \frac{du}{dx} v\right]_0^L = \mu \frac{du}{dx}(L)v(L) - \mu \frac{du}{dx}(0)v(0) = \psi_L v(L)\ ,$$

Hence,

$$\begin{cases} u \in V \text{ and satisfies} \\ \displaystyle\int_0^L \mu \frac{du}{dx}\frac{dv}{dx}\,dx = \int_0^L f v\,dx + \psi_L v(L) \qquad \text{for all } \ v \in V\ , \end{cases}$$

whereas

$$\begin{cases} u_h \in V_h \text{ and satisfies} \\ \displaystyle\int_0^L \mu \frac{du_h}{dx}\frac{dv_h}{dx}\,dx = \int_0^L f v_h\,dx + \psi_L v_h(L) \qquad \text{for all } \ v_h \in V_h\ . \end{cases}$$

The Neumann condition influences only the last equation of the algebraic system $\boldsymbol{Au} = \boldsymbol{f}$, whose size is now $N + 1$.

The entries a_{jk} of $\boldsymbol{A}$ are defined as for the case of Dirichlet boundary conditions, except for the element $a_{N+1,N+1}$, whose value is

$$a_{N+1,N+1} = \int_{I_{N+1}} \mu \left(\frac{d\varphi_{N+1}}{dx} \right)^2 dx = \frac{1}{h_{N+1}} \mu_{N+1/2} \ ,$$

since the last basis function, φ_{N+1}, has support only on the interval I_{N+1}.

Similarly, for the last entry f_{N+1} of the right-hand side $\boldsymbol{f}$, we have

$$\int_0^L f\varphi_{N+1}\, dx = \int_{I_{N+1}} f\varphi_{N+1}\, dx \simeq f(x_{N+1}) \frac{h_{N+1}}{2} \qquad \text{(by the trapezoidal rule)} \ ,$$

hence, in conclusion, we set

$$f_{N+1} = f(L) \frac{h_{N+1}}{2} + \psi_L \ .$$

An elastic model with restoring

Let us now consider a slightly more complicated model of elastic string. To be precise, we shall assume that on the string acts, in addition to the volume density of force $\boldsymbol{f}$, also a (density of) restoring force $\boldsymbol{r}$ proportional to the displacement $\boldsymbol{u}$ and oppositely oriented:

$$\boldsymbol{r} = -\gamma \boldsymbol{u} \ ,$$

with $\gamma \geq 0$ being the proportionality factor.

The mathematical model becomes

$$\begin{cases} \dfrac{d\tau}{dx} + f - \gamma u = 0 & \text{in } (0, L) \ , \\ \tau = \mu \dfrac{du}{dx} & \text{in } (0, L) \ , \\ u(0) = u(L) = 0 \ , \end{cases}$$

i.e.,

$$\begin{cases} -\dfrac{d}{dx}\left(\mu \dfrac{du}{dx}\right) + \gamma u = f & \text{in } (0, L) \ , \\ u(0) = u(L) = 0 \ . \end{cases}$$

The problem's discretisation by finite differences or finite elements still leads to an algebraic system like

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

where now the stiffness matrix $\boldsymbol{A}$ can be written as sum of two matrices,

$$\boldsymbol{A} = \boldsymbol{A}^{(\mu)} + \boldsymbol{A}^{(\gamma)} \ ,$$

the former (already known) accounting for the shear effects, the latter being a consequence of elastic restoring.

- *Discretization by finite differences*

  Equations now become

  $$\frac{1}{h^2}\left(-\mu_{j-1/2}u_{j-1} + (\mu_{j-1/2} + \mu_{j+1/2})u_j - \mu_{j+1/2}u_{j+1}\right) + \gamma_j u_j = f_j \ ,$$

  where we have set $\gamma_j = \gamma(x_j)$.

  Hence,

  $$\boldsymbol{A}^{(\gamma)} = \mathrm{diag}\left((\gamma_j)_{1\leq j\leq N}\right) .$$

- *Discretization by finite elements*

The discrete variational formulation now becomes

$$\begin{cases} u_h \in V_h \text{ and satisfies} \\ \displaystyle\int_0^L \left(\mu \frac{du_h}{dx}\frac{dv_h}{dx} + \gamma u_h v_h \right) dx = \int_0^L f v_h \, dx \qquad \text{for all} \quad v_h \in V_h \ ; \end{cases}$$

after introducing the Lagrange basis, we obtain from it the equations

$$\int_0^L \left(\mu \frac{du_h}{dx}\frac{d\varphi_j}{dx} + \gamma u_h \varphi_j \right) dx = \int_0^L f \varphi_j \, dx \qquad \text{for} \quad j = 1, \ldots, N \ ,$$

i.e.,

$$\sum_{k=1}^{N} u_k \left(\int_0^L \mu \frac{d\varphi_k}{dx}\frac{d\varphi_j}{dx} \, dx + \int_0^L \gamma \varphi_k \varphi_j \, dx \right) = \int_0^L f \varphi_j \, dx \qquad \text{for} \quad j = 1, \ldots, N \ .$$

Thus,

$$\boldsymbol{A}^{(\gamma)} = \boldsymbol{B} = \{b_{jk}\}_{1 \le j,k \le N} \qquad \text{with} \quad b_{jk} = \int_0^L \gamma \varphi_k \varphi_j \, dx \ .$$

The matrix $\boldsymbol{B}$ is **tridiagonal**, **symmetric** and **positive semi-definite**. Precisely, one has

$$b_{jk} = \begin{cases} \frac{1}{3}(\gamma_{j-1/2} h_j + \gamma_{j+1/2} h_{j+1}) & \text{if } \; k = j \; , \\ \frac{1}{6}\gamma_{j-1/2} h_j & \text{if } \; k = j-1 \; , \\ \frac{1}{6}\gamma_{j+1/2} h_{j+1} & \text{if } \; k = j+1 \; , \\ 0 & \text{otherwise} \; , \end{cases}$$

with

$$\gamma_{j-1/2} \sim \frac{1}{h_j} \int_{I_j} \gamma(x)\, dx \; , \qquad \gamma_{j+1/2} \sim \frac{1}{h_{j+1}} \int_{I_{j+1}} \gamma(x)\, dx \; .$$

In the particular case in which γ is constant on $[0, L]$ and the partition of the interval is equally spaced with step h, the previous expression takes the simplified form

$$b_{jk} = \gamma h \begin{cases} \frac{2}{3} & \text{if } \; k = j \; , \\ \frac{1}{6} & \text{if } \; k = j \pm 1 \; , \\ 0 & \text{otherwise} \; , \end{cases}$$

i.e., one has

$$\boldsymbol{B} = \gamma h \, \text{tridiag} \begin{bmatrix} \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{bmatrix} \; .$$