Galerkine methods for the 1D Helmholtz equation

An introduction to PDEs and finite element methods through the example of wave propagation

V. de Nodrest[1]

[1]CentraleSupélec, Université Paris Saclay

October 15, 2025

CentraleSupélec

Table of Contents

UNIVERSITY OF GREENWICH

Table of Contents

The wave equation

Many physical problems invole the wave equation, which describes wave propagation in media that are homogeneous, isotropic, linear et non dispersive:

The wave equation

$$\frac{\partial^2 p}{\partial t^2} = c^2 \Delta p$$

$\Delta = \nabla^2$ is the Laplacian, a differential operator.
p is a physical phenomenon propagated through one of the aforementioned media at a speed c.
It depends on both space $\mathbf{r}$ and time t.

Variable seperation

Assuming the solution is separable in time and space ($p(\mathbf{r}, t) = u(\mathbf{r})T(t)$), the wave equation can be rewritten as such:

$$\frac{\Delta u}{u} = \frac{1}{c^2}\frac{\mathrm{d}^2 T}{\mathrm{d}t^2}$$

The left-hand side only depends on space and the right-hand side only depends on time. In order to be equal in any situation, both members need to be equal to the same constant. This constant is set to $-k^2$ for convenience:

$$\frac{\Delta u}{u} = -k^2$$

$$\frac{1}{c^2 T}\frac{\mathrm{d}^2 T}{\mathrm{d}t^2} = -k^2$$

The Helmholtz equation

Rearranging the first equation (space-dependent) yields:

The homogeneous Helmholtz equation

$$\Delta u + k^2 u = 0$$

It is possible to account for sources using f, a function with compact support, thus yielding:

The inhomogeneous Helmholtz equation

$$\Delta u + k^2 u = f$$

Our 1D Helmholtz problem

We consider the following problem, a simple yet telling example of a situation involving the Helmholtz equation:

Our 1D Helmholtz problem

$$u'' + k^2 u = 0 \text{ in }]0,1[\tag{1}$$

$$u'(0) = ik \tag{2}$$

$$u'(1) = iku(1) \tag{3}$$

About our problem

Our 1D Helmholtz problem

$$u'' + k^2 u = 0 \text{ in }]0,1[\tag{1}$$

$$u'(0) = ik \tag{2}$$

$$u'(1) = iku(1) \tag{3}$$

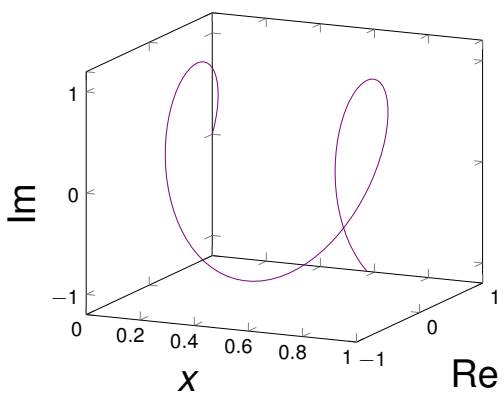
- u is a 1D complex-valued function, at least two times derivable
- $f = 0$, so this problem ”sourceless”/homogeneous
- (2), assigning a value to the derivative, is called a Neumann ”flux” boundary condition.
- (3), establishing a linear relationship between the value and the derivative, is called a Robin ”impedance” boundary condition.

The exact solution

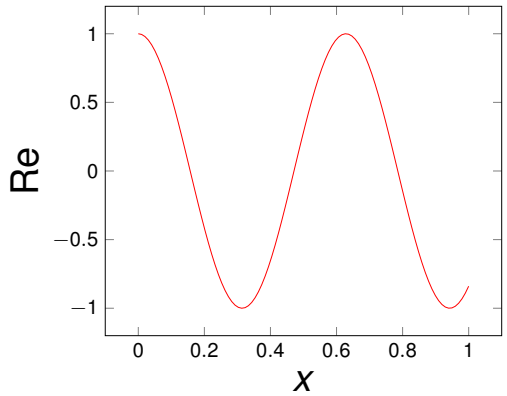
The problem can be solved with linear PDE tools, yielding an unique solution:

The ”Euler wave” exact solution

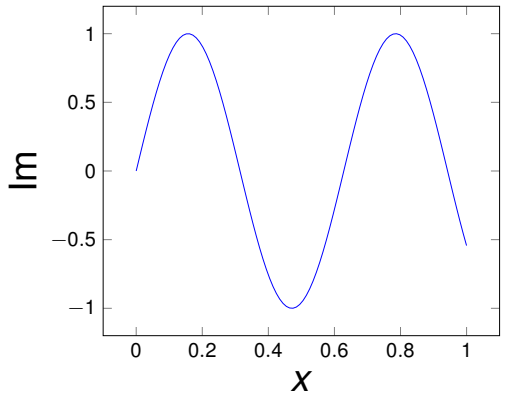
$$\forall x \in [0,1],\ u(x) = e^{ikx}$$



(a) 3D complex plot



(b) Real part



(c) Imaginary part

Figure: Exact solution with $k = 10$

Weak formulation (1)

Let's forget about our exact solution and apply conventional PDE tools, which are mandatory for harder problems and numerical methods.

We will assume that u is a solution and is in $H^2(]0,1[,\mathbb{C})$ (the $\mathbb{C}$ might be omitted in the next slides).

Then, for every v in $H^2(0,1)$ we multiply the equation in our domain by $\overline{v}$ and integrate over the domain, yielding:

$$\int_{]0,1[} u''\overline{v} + k^2 \int_{]0,1[} u\overline{v} = 0$$

Weak formulation (2)

Integrating by parts, noting:
$\langle u, v \rangle_{L^2} = \int_{]0,1[} u\overline{v}$ (scalar product for the energy norm over our domain)
and using the boundary conditions, we get:

Our weak formulation

$$\forall v \in H^2(0,1),\ k^2 \langle u, v \rangle_{L^2} - \langle u', v' \rangle_{L^2} + ik \left[u(1)\overline{v(1)} - \overline{v(0)} \right] = 0$$

This is only one of the many possible weak formulations we could have obtained. Other choices could have been made regarding:

- The space to which u belongs (test functions)
- The space to which v belongs (weighting functions)
- The norm and the scalar product

Variational formulation (1)

We need to rewrite our weak formulation in the standardized form:

A variational formulation

Find $u \in V_1$ such that

$$\forall v \in V_2,\ a(u, v) = l(v)$$

Where:

- V_1 and V_2 are Hilbert spaces (with their respective scalar products)
- a is sesquilinear
- l is antilinear or linear

Other properties can work, but this covers most problems. Real-valued problems can be seen a trivial particular case.

Variational formulation (2)

From the weak formulation, we deduce:

Our sesquilinear form

$$a : (H^1(0,1))^2 \longrightarrow \mathbb{C}$$
$$(u,v) \longmapsto k^2 \langle u, v \rangle_{L^2} - \langle u', v' \rangle_{L^2} + ik \left[u(1)\overline{v(1)} \right]$$

Our antilinear form

$$l : H^1(0,1) \longrightarrow \mathbb{C}$$
$$v \longmapsto ik \left[\overline{v(0)} \right]$$

We had to use H^2 spaces to obtain the weak formulation but our forms can be expressed in H^1 spaces (no second order derivatives).

Continuity of the sesquilinear form (1)

A nice property a sesquilinear form $a : V^2 \to \mathbb{C}$ can have is continuity:

Continuity of a sesquilinear form

$$\exists C_a \in \mathbb{R},\ \ \forall u, v \in V,\ |a(u,v)| \leq C_a ||u||_V ||v||_V$$

Given our sesquilinear form, proving its continuity over the Hilbert space $(H^1(0,1), \langle .,. \rangle_{L^2})$ would require a constant $\lambda \in \mathbb{R}$ for which:

$$\forall u \in H^1(0,1), ||u'||_{L^2} \leq \lambda ||u||_V$$

This is not the case, and thus a is unlikely to be continuous in this case. However, continuity is posible with other Hilbert spaces.

Continuity of the sesquilinear form (2)

Continuity of the antilinear form

A nice property an antilinear form $l : V \to \mathbb{C}$ can have is continuity:

Continuity of an antilinear form

$$\exists C_l \in \mathbb{R}, \ \ \forall v \in V, \ |l(v)| \leq C_l ||v||_V$$

Coercivity of the sesquilinear form

Lax-Milgram theory

After proving all of those properties and if a was coercive, we could have used:

The adequate complex-valued version of the Lax-Milgram theorem

If:

- $(V, \langle ., . \rangle_V)$ is a Hilbert space with any valid scalar produt
- $a : V^2 \to \mathbb{C}$ is sesquilinear (with antilinearity on the second argument)
- $l : V \to \mathbb{C}$ is antilinear
- a and l are continuous (let's name the constants C_a and C_l)
- $\Re(a)$ is coercive (let's name the constant α)

Then the problem "find $u \in V$ such that for all $v \in V$, $a(u, v) = l(v)$":

- has un unique solution u
- $||u||_V \leq \frac{1}{\alpha}||l||_{V'} = \frac{C_l}{\alpha}$

Note that other versions of the theorem exist and that some properties were only chosen as conventions.

However, a is not coercive

Well-posedness of our problem

Galerkin equation

Straightforward numerical solving of the variational formulation is not possible, as there are infinitely many possible trial functions and weighting functions.
We discretize those spaces, thus yielding:

A Galerkin equation

Find $u^h \in V_1^h$ such that

$$\forall v^h \in V_2^h,\ a(u^h, v^h) = l(v^h)$$

Where $V_1^h \subset V_1$ and $V_2^h \subset V_2$ are finite.
Be reminded that in our case, $V_1 = V_2 = H^1(0,1)$, with the L^2 scalar product.
One might notice that the error is orthogonal to the subspaces:

$$a(u - u^h, v^h) = a(u, v^h) - a(u^h, v^h) = f(v^h) - f(v^h) = 0$$

Lax-Milgram Galerkin theory

Fredholm Galerkin theory

Matrix form of the Galerkin equation (1)

Trivial extension of Galerkine methods to complex-valued problems would be conducted by expressing functions of V_1^h and V_2^h as linear combinations of complex functions (base of the discrete Hilbert spaces) with real coefficients.

However, for convenience, we will rather use complex coefficients and real functions. This approach is possible for complex Hilbert spaces that are complexifications of real Hilbert spaces, which is true in our case.

The only consequence of such a choice is that we must choose the same dicretization for the real and imaginary parts of our trial and weighting functions.

Let's note $(e_{1,i})_{i\in[\![1,n]\!]}$ and $(e_{2,i})_{i\in[\![1,m]\!]}$ our bases of $\Re V_1^h$ and $\Re V_2^h$.

Matrix form of the Galerkin equation (2)

It can easily be shown that to solve the Galerkin problem, it is sufficient to test the solution over every single weighting (test) function.

Thus, the Galerkin problem is equivalent to:
Find $(u_i)_i \in \mathbb{C}^n$ such that:

$$\forall j \in [\![1, m]\!],\ a\left(\sum_{i=1}^{n} u_i e_{1,i}\ ,\ e_{2,j}\right) = l(e_{2,j})$$

Note: we wrote u^h in its base, i.e $u^h = \sum_{i=1}^{n} u_i e_{1,i}$.

In our case, a being linear in its first argument, the system can be written as:

$$\forall j \in [\![1, m]\!],\ \sum_{i=1}^{n} u_i a(e_{1,i}\ ,\ e_{2,j}) = l(e_{2,j})$$

Matrix form of the Galerkin equation (3)

Finally, the system of equations can be written as a matrix product:

The matrix form of a Galerkin equation

Find $U \in \mathscr{M}_{n,1}(\mathbb{C})$ such that

$$A^{\mathsf{T}} U = L$$

Where:

- $\forall (i,j) \in [\![1,n]\!] \times [\![1,m]\!],\ [A]_{i,j} = a(e_{1,i}\ ,\ e_{2,j})$
- $\forall j \in [\![1,m]\!],\ [L]_j = l(e_{2,j})$

Mesh (1)

The first step to finite element methods is the choice of the mesh. A mesh is composed of nodes that circumscribe elements.

Some properties are often required for a mesh composed of closed elements $(\Omega_i)_i$ over a domain Ω:

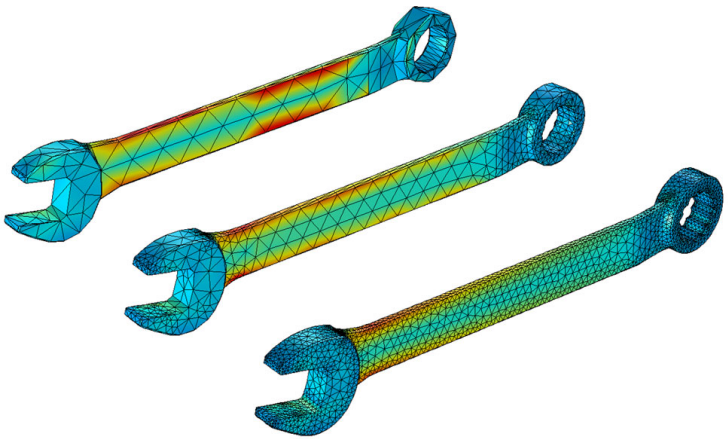
$$\bigcup_i \Omega_i = \overline{\Omega}$$

$$\forall i \neq j,\ \Omega_i \cap \Omega_j = \delta\Omega_i \cap \delta\Omega_j$$

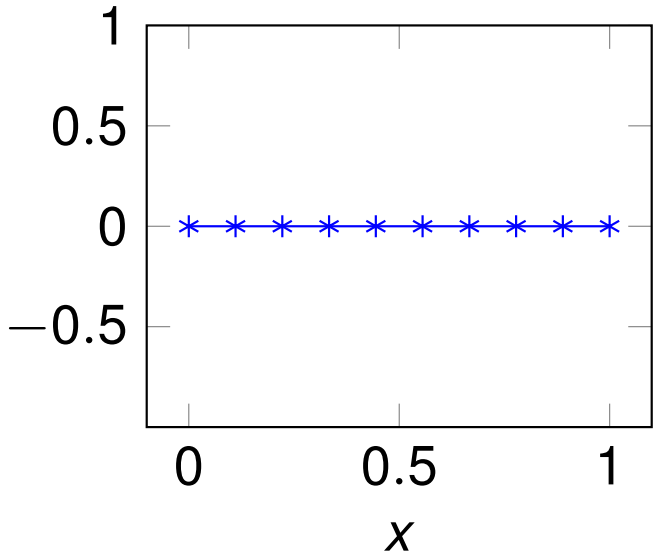
This means that nothing more but the whole domain is covered by the elements, and that elements only enventually overlap on their boundary (a hyperplane of inferior dimension).

Mesh (2)

Mesh design is a whole discipline in itself. However, for the sake of simplicity, we will use a uniform mesh of element size h over our unit interval.



(a) Iterations of a 3D mesh



(b) 1D uniform mesh (an asterisk is a node)

Linear interpolation

For our

Table of Contents

Bibliography I

Table of Contents

Appendix A (1)

We will now prove that, in order to prove that any $u^h \in V_1^h$ is a solution of the problem for every $v^h \in V_2^h$, it is sufficient to show it is a solution for every base function of V_2^h.

Writing v^h in its basis, this can be written as:

$$\forall j \in [\![1, m]\!],\ a\left(u^h,\ e_{2,j}\right) = l\left(e_{2,j}\right)$$

$$\Leftrightarrow$$

$$\forall (v_j)_j \in \mathbb{C}^m,\ a\left(u^h,\ \sum_{j=1}^{m} v_j e_{2,j}\right) = l\left(\sum_{j=1}^{m} v_j e_{2,j}\right)$$

Appendix A (2)

$\boxed{\Leftarrow}$ Setting $(v_j)_j$ to $(1,0,\ldots,0)$, $(0,1,0,\ldots,0)$, ..., $(0,0,\ldots,0,1)$ seals the deal.

$\boxed{\Rightarrow}$ Let $(v_j)_j \in \mathbb{C}^m$.

$$a\left(u^h,\ \sum_{j=1}^{m} v_j e_{2,j}\right) = \sum_{j=1}^{m} \overline{v_j}\, a\left(u^h,\ e_{2,j}\right) = \sum_{j=1}^{m} \overline{v_j}\, l(e_{2,j}) = l\left(\sum_{j=1}^{m} v_j e_{2,j}\right)$$

Respectively because a is right-antilinear, the hypothesis, and l is antilinear.