

Numerical methods for PDEs

1. | Finite element method

Variational formulation of elliptic PDEs

Definition 1. Consider a elliptic pde of the form:

$$-\sum_{i,j=1}^n \partial_j (a_{ij} \partial_i u) + \sum_{i=1}^n b_i \partial_i u + cu = f \quad (1)$$

in a bounded open subset $\Omega \subset \mathbb{R}^n$ with a_{ij} , b_i , c and f sufficiently regular functions. The variational formulation of this problem is:

$$\begin{aligned} \sum_{i,j=1}^n \langle a_{ij} \partial_i u, \partial_j v \rangle_{L^2(\Omega)} + \sum_{i=1}^n \langle b_i \partial_i u, v \rangle_{L^2(\Omega)} + \langle cu, v \rangle_{L^2(\Omega)} &= \\ &= \langle f, v \rangle_{L^2(\Omega)} + \sum_{i,j=1}^n \langle a_{ij} \partial_i u \nu_j, v \rangle_{L^2(\partial\Omega)} \end{aligned}$$

where v is a test function (sufficiently regular) and ν is the outward unit normal vector on $\partial\Omega \in \mathcal{C}^1$. The functions in the last inner product on $\partial\Omega$ are meant to be in the sense of traces. From now on, the left hand side of the variational formulation will be denoted by $a(u, v)$ and $\ell(v) := \langle f, v \rangle_{L^2(\Omega)}$.

Remark. This formulation makes sense for $a_{ij}, b_i, c \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$. We will look for a solution $u \in V$, where V is a suitable space of functions, such that the variational formulation holds for all $v \in V$.

Definition 2 (Dirichlet boundary conditions). Consider Eq. (1) with Dirichlet boundary conditions $u = g$ on $\partial\Omega$. If $g = 0$, the boundary term disappears and we can choose $V = H_0^1(\Omega)$. If $g \neq 0$, and both g and $\partial\Omega$ are smooth (i.e. $g \in H^1(\Omega)$), by ?? we can find $u_g \in H^1(\Omega)$ such that $\text{Tr}(u_g) = g$. Then, we set $\tilde{u} := u - u_g \in H_0^1(\Omega)$ and satisfies:

$$a(\tilde{u}, v) = \ell(v) - a(u_g, v) \quad \forall v \in H_0^1(\Omega)$$

Definition 3 (Neumann boundary conditions). Consider Eq. (1) with Neumann boundary conditions $\sum_{i,j=1}^n a_{ij} \partial_i u \nu_j = g$ on $\partial\Omega$, for $g \in L^2(\partial\Omega)$. In this case, we take $V = H^1(\Omega)$ and look for a solution $u \in V$ such that:

$$a(u, v) = \ell(v) + \langle g, v \rangle_{L^2(\partial\Omega)} \quad \forall v \in V$$

Definition 4 (Robin boundary conditions). Consider Eq. (1) with Robin boundary conditions $\gamma u + \sum_{i,j=1}^n a_{ij} \partial_i u \nu_j = g$ on $\partial\Omega$, for $g \in L^2(\partial\Omega)$ and $\gamma \in L^\infty(\partial\Omega)$. In this case, we take $V = H^1(\Omega)$ and look for a solution $u \in V$ such that:

$$a(u, v) + \langle \gamma u, v \rangle_{L^2(\partial\Omega)} = \ell(v) + \langle g, v \rangle_{L^2(\partial\Omega)} \quad \forall v \in V$$

Remark. Recall that for these problems to have a unique solution, we need to impose the coercivity and continuity in ?? ??.

Proposition 5. Consider the homogeneous Dirichlet problem from Eq. (1) and set $\beta = \alpha^{-1} \sum_{i=1}^n \|b_i\|_{L^\infty(\Omega)}^2$, where α is the ellipticity constant of the PDE. Then, the homogeneous Dirichlet problem has a unique solution u in $H_0^1(\Omega)$ if $\forall x \in \Omega$ we have $c - \frac{\beta}{2} \geq 0$. In this case, $\exists C > 0$ such that:

$$\|u\|_{H^1(\Omega)} \leq C \|f\|_{L^2(\Omega)}$$

Consequently, the non-homogeneous Dirichlet problem for $g \in H^1(\partial\Omega)$ has a unique solution u in $H^1(\Omega)$ such that:

$$\|u\|_{H^1(\Omega)} \leq \tilde{C} (\|f\|_{L^2(\Omega)} + \|g\|_{H^1(\partial\Omega)})$$

Proposition 6. Consider the Neumann problem from Eq. (1) for $g \in L^2(\partial\Omega)$ and set $\beta = \alpha^{-1} \sum_{i=1}^n \|b_i\|_{L^\infty(\Omega)}^2$, where α is the ellipticity constant of the PDE. Then, the Neumann problem has a unique solution u in $H^1(\Omega)$ if $\forall x \in \Omega$ we have $c - \frac{\beta}{2} \geq \delta > 0$. In this case, $\exists C > 0$ such that:

$$\|u\|_{H^1(\Omega)} \leq C (\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)})$$

Proposition 7. Consider the Robin problem from Eq. (1) for $g \in L^2(\partial\Omega)$ and $\gamma \in L^\infty(\partial\Omega)$ and set $\beta = \alpha^{-1} \sum_{i=1}^n \|b_i\|_{L^\infty(\Omega)}^2$, where α is the ellipticity constant of the PDE. Then, the Robin problem has a unique solution u in $H^1(\Omega)$ if $\forall x \in \Omega$ we have $c - \frac{\beta}{2} \geq \delta \geq 0$ and $\gamma \geq \eta \geq 0$ with either $\delta > 0$ or $\eta > 0$. In this case, $\exists C > 0$ such that:

$$\|u\|_{H^1(\Omega)} \leq C (\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)})$$

Galerkin method

Definition 8. The *conforming Galerkin method* consists in looking for a solution $u_h \in V_h \subset V$ such that:

$$a_h(u_h, v_h) = \ell_h(v_h) \quad \forall v_h \in V_h$$

where V_h is a closed finite-dimensional subspace of V and a_h and ℓ_h are the approximations of a and ℓ in V_h .

Remark. Here h is meant to be a discretization parameter.

Theorem 9 (C ea's lemma). Let $u \in V$ be the solution of the variational formulation of Eq. (1) and $u_h \in V_h$ be the solution of the Galerkin method. Then, $\exists C > 0$ such that:

$$\|u - u_h\|_V \leq C \inf_{v_h \in V_h} \|u - v_h\|_V$$

Finite element spaces

Remark. Finite element methods are a special case of Galerkin methods in which the finite-dimensional subspace V_h consists of piecewise polynomial functions over a mesh.

Definition 10. A *finite element* is a triplet $(K, \mathcal{P}, \mathcal{N})$ where:

- $K \subseteq \mathbb{R}^n$ is a simply connected bounded open set with piecewise smooth boundary (the *geometric element*).

- $\mathcal{P}$ is a finite-dimensional space of functions defined on K , whose elements are called *basis functions*.
- $\mathcal{N} = \{N_1, \dots, N_d\}$ is a basis of $\mathcal{P}^*$.

Lemma 11. Let $\mathcal{P}$ be d -dimensional vector space and let $\{N_1, \dots, N_d\} \subset \mathcal{P}^*$. The following statements are equivalent:

1. $\{N_1, \dots, N_d\}$ is a basis of $\mathcal{P}^*$.
2. For all $v \in \mathcal{P}$ such that $N_i(v) = 0$ for all $i = 1, \dots, d$, then $v = 0$.

Remark. To construct a finite element, we usually proceed as follows:

1. We choose a geometric element K .
2. We choose a polynomial space $\mathcal{P}$ up to a given degree k .
3. We choose the degrees of freedom $\mathcal{N} = \{N_1, \dots, N_d\}$, where $d = \dim \mathcal{P}$, such that the corresponding interpolation problem is well-posed.
4. We compute the dual basis of $\mathcal{N}$, which gives a basis of $\mathcal{P}$.

Definition 12. Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element and $\{\psi_1, \dots, \psi_d\}$ be the corresponding basis of $\mathcal{P}$. For a given function v such that $N_i(v)$ is defined $\forall i \in \{1, \dots, d\}$, we define the *local interpolant* of v as:

$$I_K v := \sum_{i=1}^d N_i(v) \psi_i$$

Lemma 13. Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element and I_K be the local interpolant operator associated to it. Then, the following properties hold:

1. I_K is linear.
2. $N_i(I_K v) = N_i(v) \forall i \in \{1, \dots, d\}$.
3. $I_K v = v \forall v \in \mathcal{P}$, i.e. I_K is a projection.

Definition 14. A *subdivision* of a bounded open set $\Omega \subset \mathbb{R}^n$ is a collection $\mathcal{T}$ of open sets K_i such that:

1. $K_i \cap K_j = \emptyset \forall i \neq j$.
2. $\bar{\Omega} = \bigcup_{K \in \mathcal{T}} \bar{K}$.

Definition 15. Let $\mathcal{T}$ be a subdivision of Ω such that for each $K \in \mathcal{T}$ there exists a finite element $(K, \mathcal{P}, \mathcal{N})$ with local interpolant I_K . Let m be the order of the highest partial derivative appearing in any of the degrees of freedom of $\mathcal{N}$. We define the *global interpolant* $I_{\mathcal{T}} v$ of $\mathcal{T}$, for $v \in \mathcal{C}^m(\bar{\Omega})$, as:

$$I_{\mathcal{T}} v|_K := I_K v \quad \forall K \in \mathcal{T}$$

Definition 16. A *triangulation* of a bounded open set $\Omega \subset \mathbb{R}^2$ is a subdivision $\mathcal{T}$ of Ω such that:

1. Each $K \in \mathcal{T}$ is a triangle.

2. The intersection of two triangles is either empty or a common vertex or a common edge.

Definition 17. Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$, $(K, \mathcal{P}, \mathcal{N})$ be finite elements and $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be an affine transformation. We say that these finite elements are *affinely equivalent* by T if:

1. $K = T(\hat{K})$.
2. $\mathcal{P} = \{\hat{p} \circ T^{-1} : \hat{p} \in \hat{\mathcal{P}}\}$.
3. $\mathcal{N} = \{N_i\}$, where $N_i(p) = \hat{N}_i(p \circ T) \forall p \in \mathcal{P}$.

Lemma 18. Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$, $(K, \mathcal{P}, \mathcal{N})$ be two affine equivalent finite elements by the affine transformation $\mathbf{T}_K$. Then:

$$I_{\hat{K}}(v \circ \mathbf{T}_K) = I_K v \circ T$$

Polygonal interpolation in Sobolev spaces

Lemma 19 (Bramble-Hilbert lemma). Let $F : W^{k,p}(\Omega) \rightarrow \mathbb{R}$ be such that:

1. $|F(v)| \leq c_1 |v|_{W^{k,p}(\Omega)} \forall v \in W^{k,p}(\Omega)$, where

$$|v|_{W^{k,p}(\Omega)} := \begin{cases} \left(\sum_{|\alpha|=k} \|\partial^\alpha v\|_{L^p(\Omega)} \right)^{1/p} & \text{if } p < \infty \\ \max_{|\alpha|=k} \|\partial^\alpha v\|_{L^\infty(\Omega)} & \text{if } p = \infty \end{cases}$$

2. $|F(u+v)| \leq c_2(|F(u)| + |F(v)|) \forall u, v \in W^{k,p}(\Omega)$.
3. $|F(q)| = 0 \forall q \in \mathcal{P}_{k-1}(\Omega)$, where $\mathcal{P}_\ell(\Omega)$ is the space of polynomials of degree less than ℓ .

Then, $\exists C > 0$ such that $\forall v \in W^{k,p}(\Omega)$:

$$|F(v)| \leq C |v|_{W^{k,p}(\Omega)}$$

Theorem 20. Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element such that $\mathcal{P}_{k-1} \subseteq \mathcal{P}$ for some $k \in \mathbb{N}$ and all $N \in \mathcal{N}$ be bounded in $W^{k,p}(K)$ for some $p \in [1, \infty]$. Then, $\exists C > 0$ such that $\forall v \in W^{k,p}(K)$:

$$|v - I_K v|_{W^{\ell,p}(K)} \leq C |v|_{W^{k,p}(K)} \quad \forall \ell \in \{0, \dots, k\}$$

Remark. Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element and $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be the reference element. From now on, if they are affine equivalent by $\mathbf{T}_K : \hat{K} \rightarrow K$, we will assume that $\mathbf{T}_K \hat{x} = \mathbf{A}_K \hat{x} + \mathbf{b}_K$, with $\mathbf{A}_K$ invertible.

Lemma 21. Let $k \in \mathbb{N}$ and $p \in [1, \infty]$. Then, $\exists C > 0$ such that $\forall K \subset \Omega$ and $\forall v \in W^{k,p}(\hat{K})$:

$$|v|_{W^{k,p}(K)} \leq C \|\mathbf{A}_K\|^k |\det \mathbf{A}_K|^{-1/p} |v|_{W^{k,p}(\hat{K})}$$

$$|v|_{W^{k,p}(K)} \leq C \|\mathbf{A}_K^{-1}\|^k |\det \mathbf{A}_K|^{1/p} |v|_{W^{k,p}(\hat{K})}$$

Definition 22. Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element. We define the *diameter* of K as:

$$h_K := \max_{x,y \in K} \|x - y\|$$

We define the *insphere diameter* of K as:

$$\rho_K := 2 \max\{\rho > 0 : B(x, \rho) \subset K \text{ for some } x \in K\}$$

We define the *condition number* of K as $\sigma_K := \frac{h_K}{\rho_K}$.

Lemma 23. Let $(K, \mathcal{P}, \mathcal{N})$, $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be affine equivalent finite elements by $\mathbf{T}_K : \hat{K} \rightarrow K$. Then, $|\det \mathbf{A}_K| = \frac{\text{vol}(K)}{\text{vol}(\hat{K})}$, $\|\mathbf{A}_K\| \leq \frac{h_K}{\rho_{\hat{K}}}$ and $\|\mathbf{A}_K^{-1}\| \leq \frac{\hat{\rho}_K}{\rho_K}$.

Theorem 24 (Local interpolation error). Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a finite element with $\mathcal{P}_{k-1} \subseteq \mathcal{P}$ for some $k \in \mathbb{N}$ and all $N \in \mathcal{N}$ be bounded in $W^{k,p}(\hat{K})$ for some $p \in [1, \infty]$. Then, for all finite element $(K, \mathcal{P}, \mathcal{N})$ affine equivalent to $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ by $\mathbf{T}_K : \hat{K} \rightarrow K$, $\exists C > 0$ (independent of K) such that $\forall v \in W^{k,p}(K)$:

$$|v - I_K v|_{W^{\ell,p}(K)} \leq C h_K^k \sigma_K^\ell |v|_{W^{k,p}(K)} \quad \forall \ell \in \{0, \dots, k\}$$

Definition 25. A subdivision $\mathcal{T}$ of $\Omega \in \mathbb{R}^n$ is called *regular* if $\exists C > 0$ such that $\forall K \in \mathcal{T}$ we have $\sigma_K \leq C$.

Theorem 26 (Global interpolation error). Let $\mathcal{T}$ be a regular subdivision of $\Omega \in \mathbb{R}^n$ and $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a reference finite element with $\mathcal{P}_{k-1} \subseteq \mathcal{P}$ for some $k \in \mathbb{N}$ and all $N \in \mathcal{N}$ be bounded in $W^{k,p}(\hat{K})$ for some $p \in [1, \infty]$. Let $h := \max_{K \in \mathcal{T}} h_K$. Then, $\exists C > 0$ (independent of h) such that $\forall v \in W^{k,p}(\Omega)$:

$$|v - I_{\mathcal{T}} v|_{W^{\ell,p}(\Omega)} + \sum_{\ell=1}^k \left(h^\ell \sum_{K \in \mathcal{T}} |v - I_K v|_{W^{\ell,p}(K)}^p \right)^{1/p} \leq C h^k |v|_{W^{k,p}(\Omega)}$$

if $p < \infty$ and:

$$|v - I_{\mathcal{T}} v|_{W^{\ell,\infty}(\Omega)} + \sum_{\ell=1}^k h^\ell \max_{K \in \mathcal{T}} |v - I_K v|_{W^{\ell,\infty}(K)} \leq C h^k |v|_{W^{k,\infty}(\Omega)}$$

if $p = \infty$.

Error estimates for finite element approximation

Theorem 27. Let $\Omega \subset \mathbb{R}^n$ be open and bounded, $u \in H^1(\Omega)$ be the solution of the boundary value problem and $\mathcal{T}$ be a regular triangulation of Ω with reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ such that $\mathcal{P}_{k-1} \subseteq \mathcal{P}$ for some $k \in \mathbb{N}$. Let $u_h \in V_h$ be the solution of the Galerkin method. Then, if $u \in H^m$, with $\frac{n}{2} < m < k$, then $\exists C > 0$ (independent of h and u) such that:

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^{m-1} \|u\|_{H^m(\Omega)}$$

2. | Spectral methods

Definition 28. The idea of *spectral methods* is to approximate the solution u of a boundary value problem by an expression in terms of the so-called *trial functions*:

$$u(x) \approx \sum_{i=1}^N \tilde{u}_i \phi_i(x)$$

We will impose the following requirements on the trial functions:

1. The approximation should converge rapidly as $N \rightarrow \infty$.
2. The computation of the coefficients $\tilde{u}_i$ and the reconstruction of u should be efficient.
3. Given the coefficients of some function u , it should be easy to determine the coefficients of the derivative of u .

Remark. When using spectral methods, it is generally assumed that the solution of the problem of interest is very smooth, and thus, the trial functions are *globally smooth*, i.e. algebraic or trigonometric polynomials.

Definition 29. The choice of the test functions distinguishes between three types of spectral methods:

1. *Galerkin methods*: the test functions are the same as the trial functions. These test functions usually satisfy some or all the boundary conditions. The PDE is enforced by requiring that the residual is orthogonal to the test functions.
2. *Collocation methods*: the test functions are Dirac delta distributions centered at the so-called *collocation points*. This approach requires the PDE to be satisfied exactly at the collocation points. A supplementary set of equations may be imposed to satisfy the boundary conditions.
3. *τ methods*: these are similar to the Galerkin methods in the way the PDE is enforced, but the test functions don't need to satisfy the boundary conditions.

Remark. From what follows, we will focus collocation methods.

Periodic problem

When considering a problem with periodic boundary conditions, we can use trigonometric polynomials as trial functions.

Remark. It can be seen that the Fourier basis functions $\{e^{ikx} : k \in \mathbb{Z}\}$ and its coefficients satisfy the requirements of **Theorem 28**, using the FFT for the computation of the coefficients.

Definition 30. Given a complex-valued periodic function u defined in $[0, 2\pi]$ and $N \in \mathbb{N}$, we define the *interpolant* of u as:

$$I_N u(x) := \sum_{k=-N+1}^N \tilde{u}_k e^{ikx}$$

where $\tilde{u}_k = \sum_{j=0}^{2N-1} u(x_j) e^{-ikx_j}$ are the discrete Fourier coefficients and $x_j = \frac{2\pi j}{2N}$.

Proposition 31. Let $P_N u = \frac{1}{2\pi} \sum_{k=-N+1}^N \hat{u}_k e^{ikx}$ be the truncated Fourier series of u . Then, we have:

$$\|u - I_N u\|^2 = \|u - P_N u\|^2 + \|R_N u\|^2$$

with:

$$R_N u(x) = \frac{1}{2\pi} \sum_{k=-N+1}^N \sum_{m \in \mathbb{Z}^*} \hat{u}_{k+2mN} e^{ikx}$$

because

$$\tilde{u}_k = \frac{2N}{2\pi} \sum_{m \in \mathbb{Z}} \hat{u}_{k+2mN}$$

Remark. Note that $(P_N u)' = P_N u'$, but $(I_N u)' \neq I_N u'$. What we do in general is to pass to the Fourier space, differentiate and then come back to the physical space.

Non-periodic problems

Remark. Recall that using trigonometric polynomials as trial functions for problems with non-periodic boundary conditions can lead to the Gibbs phenomenon. To prevent that from happening, we will use algebraic polynomials as trial functions. But in that case, we need to choose the collocation points carefully, to prevent the so-called Runge phenomenon (see ??).

In this section we will only consider one case of polynomial trial functions, the so-called Chebyshev polynomials (see ??).

Definition 32. Given a real-valued function u defined in $[-1, 1]$ and $N \in \mathbb{N}$, we define the *interpolant* of u with orthogonal polynomials $\{p_k\}_{k \in \mathbb{N} \cup \{0\}}$ and weight function $\omega(x)$ as:

$$I_N u(x) = \sum_{k=0}^N \tilde{u}_k p_k(x)$$

where $\tilde{u}_k = \frac{1}{\gamma_k} \sum_{j=0}^N u(x_j) p_k(x_j) \omega_j$, $\gamma_k = \sum_{j=0}^N p_k(x_j)^2 \omega_j$, x_j

are the chosen nodes and ω_j are the weights corresponding to the Gauß-Lobatto formula:

$$\sum_{j=0}^N x_j^k \omega_j = \int_{-1}^1 x^k \omega(x) dx$$

Remark. Recall that from Gauß quadrature, we have

$$\sum_{j=0}^N u(x_j) v(x_j) \omega_j = \int_{-1}^1 u(x) v(x) \omega(x) dx$$

for all $uv \in \mathcal{P}_{2N-1}$ (space of polynomials of degree less than $2N-1$).

Remark. Recall that the Chebyshev polynomials are those defined by being the family of orthogonal polynomials with respect to the weight function $\omega(x) = \frac{1}{\sqrt{1-x^2}}$ in $[-1, 1]$.

Lemma 33. Chebyshev polynomials satisfy the following properties:

1. $T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x) \forall k \in \mathbb{N}$, and $T_0(x) = 1$ and $T_1(x) = x$.
2. $\int_{-1}^1 (T_k(x))^2 \frac{1}{\sqrt{1-x^2}} dx = \frac{\pi}{2} c_k$, where $c_k = 2 - \mathbf{1}_{k>0}$.
3. $2T_k(x) = \frac{1}{k+1} (T_{k+1})'(x) - \frac{1}{k-1} (T_{k-1})'(x) \forall k \in \mathbb{N}$, and $(T_0)'(x) = 0$ and $(T_1)'(x) = 1$.

Proposition 34. For Chebyshev polynomials, we have $\omega_j = \frac{\pi}{N \bar{c}_j}$ with $\bar{c}_j = 2 - \mathbf{1}_{0 < j < N}$ and $x_j = \cos \frac{j\pi}{N}$. Moreover, the Chebyshev transform is given by:

$$\tilde{u}_k = \frac{2}{\pi \bar{c}_k} \sum_{j=0}^N \frac{u(x_j)}{\bar{c}_j} \cos \frac{jk\pi}{N}$$

Remark. From this last expression, we can see that the Chebyshev transform is equivalent to the discrete cosine transform (DCT), and so it can be computed efficiently using the FFT.

Remark. To differentiate a function $u = \sum_{k \in \mathbb{N}} \hat{u}_k T_k$ using the Chebyshev transform, we first compute the Chebyshev transform of u , then we differentiate the coefficients using the formula

$$c_k \hat{u}'_k = \hat{u}'_{k+1} + 2(k+1) \hat{u}_{k+1}$$

in a backward sweep (since $\hat{u}'_k = 0$ for $k \geq N$) and finally we compute the inverse Chebyshev transform of the result.