Solving Partial Differential Equations

From Finite Elements to Physics Informed Neural Networks

Eloi Martinet

Contents

PDEs and neural networks: the theoretical minimum

1. Variational formulation of a PDE

1.1 Introduction

Partial Differential Equations (PDEs) is the language used to describe the natural world. We can use this language to talk about nearly every phenomenon, from the intense gravitational pull of a black hole to the blurry motion of an electron, the flow of air around an airplane wing, the vibration of the Eiffel Tower under the effect of an earthquake, the evolution of financial markets or the spread of diseases. In this course, we will be interested in one of the simplest kind of PDEs, namely: elliptic PDEs.

Poisson equation

Let us review a few common elliptic PDEs. The simplest of all is the Poisson equation, and read as follows: for $\Omega \subset \mathbb{R}^n$ and $f : \Omega \to \mathbb{R}$, we say that $u : \Omega \to \mathbb{R}$ is the solution to the Poisson equation with homogeneous Dirichlet boundary conditions if

$$\begin{cases} -\Delta u = f \text{ in } \Omega \\ \quad u = 0 \text{ on } \partial\Omega \end{cases} \tag{1.1}$$

where $\Delta u := \partial^2_{11} u + \cdots + \partial^2_{nn} u$ is the *Laplacian* of u. This equation models, for instance, the vertical displacement of a membrane fixed on $\partial\Omega$ on which we exert a pressure f, or the temperature distribution inside Ω when imposing a temperature of 0 at the boundary and heating the domain with a heat source f. Other boundary conditions can be considered, like non-homogeneous Dirichlet boundary conditions (i.e. imposing $u = g$ of $\partial\Omega$ for a prescribed $g : \partial\Omega \to R$), or Neumann boundary conditions, that reads

$$\begin{cases} -\Delta u = f \text{ in } \Omega \\ \partial_n u = g \text{ on } \partial\Omega \end{cases} \tag{1.2}$$

where $\partial_n u := \nabla u \cdot \mathbf{n}$ where $\mathbf{n}$ is the outward normal vector to the boundary $\partial\Omega$. This conditions imposes a constraint on some *flux* across $\partial\Omega$; in the context of the heat interpretation and with homogeneous Neumann boundary conditions (i.e. $g = 0$), it physically means that the heat flux across the boundary is 0, in other words: the domain is perfectly insulated.

Linearized elasticity

Another important elliptic PDE is the one describing elastic structures in the regime of small deformations, which is called *linearized elasticity*. For a domain $\Omega \subset \mathbb{R}^n$ modeling an elastic material and a function $u : \Omega \to \mathbb{R}^n$ being the displacement field of this material, we define the *symmetrized gradient*

$$e(u) := \frac{1}{2}(\nabla u + (\nabla u)^T)$$

where

$$\nabla u := \begin{pmatrix} \partial_1 u_1 & \dots & \partial_1 u_n \\ \vdots & \ddots & \vdots \\ \partial_n u_1 & \dots & \partial_n u_n \end{pmatrix}$$

For a tensor field $\sigma : \Omega \to \mathbb{R}^{n\times n}$, we will denote

$$\mathrm{div}\sigma := \left(\sum_{j=1}^{n} \partial_j \sigma_{ij} \right)_{1\le i\le n},$$

which is the vector of the divergence of each row of σ. We suppose that the elastic body is fixed on a certain portion $\partial\Omega_D$ and subject to some surface force (like a pressure) $g : \partial\Omega_N \to \mathbb{R}^n$ on $\partial\Omega_N := \partial\Omega \setminus \partial\Omega_D$. Moreover, we suppose that there are body forces $f : \Omega \to \mathbb{R}^n$ (think about gravity). Then u is a solution of the linearized elasticity problem if it solves

$$\begin{cases} -\mathrm{div}(Ae(u)) = f \text{ in } \Omega \\ u = 0 \text{ on } \partial\Omega_D \\ Ae(u)\mathbf{n} = g \text{ on } \partial\Omega_N \end{cases} \tag{1.3}$$

where λ and μ are constants called the *Lamé coefficients* and

$$Ae(u) := 2\mu e(u) + \lambda Tr(e(u))I.$$

Stokes problem

Another vector-valued problem is the incompressible Stokes problem. This problem is the linearization of the incompressible Naviers-Stokes equation and describes the movement of a fluid with high viscosity (or equivalently a fluid at low velocity)(it can be useful if you try to design a submarine that moves in honey). Let Ω be a body of fluid, $u : \Omega \to \mathbb{R}^n$ be its velocity field and $p : \Omega \to \mathbb{R}$ its pressure which are the two unknowns. Given a force field $f : \Omega \to \mathbb{R}^n$, we want u and p to solve

$$\begin{cases} \nabla p - \Delta u = f \text{ in } \Omega \\ \mathrm{div} u = 0 \text{ in } \Omega \\ u = 0 \text{ on } \partial\Omega \end{cases}, \tag{1.4}$$

where $\Delta u = (\Delta u^1, \dots, \Delta u^n)^T$ is the component-wise Laplace operator.

Time-dependent PDEs

While we will not deal with time-dependent equations, some of them are closely related to the Poisson equation and their analysis can easily follow from the analysis of the Poisson equation. Let us mention:

- The heat equation $\partial_t u = \Delta u$;
- The wave equation $\partial_{tt}^2 u = \Delta u$;
- The Schrödinger equation $i\partial_t u = -\Delta u + Vu$ where $V : \Omega \to \mathbb{R}$ is called the *potential*.

Eigenvalue problems

Finally, one can be interested in *eigenvalue problems*, which are of interest in a wide variety of ways. In the case of the Laplace operator with homogeneous Dirichlet boundary conditions, we are interested in finding all functions $u : \Omega \to \mathbb{R}$ and scalar $\lambda \in \mathbb{R}$ such that

$$\begin{cases} -\Delta u = \lambda u \text{ in } \Omega \\ \quad u = 0 \text{ on } \partial\Omega. \end{cases} \tag{1.5}$$

We call u and eigenfunction and λ its associated eigenvalue. One physical interpretation for this equation is that it allows to find all the natural frequencies of a membrane of shape Ω, and there frequencies are given by $\sqrt{\lambda}$. In the one dimensional case, it models a vibrating string.

Knowing all couple of eigenfunctions and eigenvalues of this problem (called the *spectral decomposition* of the Laplacian) also allow to solve the time-dependent heat and wave equations, using the method of *separation of the variables*.

But the most striking application is maybe the use of eigenvalue problems in quantum mechanics: indeed, the third principle of quantum mechanics stipulates that the result of a measurement must be the eigenvalue of some operator. For instance, by solving

$$\begin{cases} -\Delta u + Vu = \lambda u \text{ in } \Omega \\ \quad u = 0 \text{ on } \partial\Omega \end{cases} \tag{1.6}$$

one can determine the energy levels λ that a particle is allowed to take in a potential V, along with the probability $|u|^2(x)$ of finding the particle at a certain point $x \in \Omega$.

R In general, finding an analytic solution to a PDE is a difficult task; moreover, such solution might not even exist in a classical sense. All this chapter is devoted to crafting some of the tools necessary to show that in a certain sense, (1.1) has a solution.

1.2 Divergence formula

The Divergence formula is the multidimensional counterpart of the 1-dimensional integration by parts. It is the name of a general result of differential geometry that implies the Green-Gauss, Green-Ostrogradski, and Green-Riemann formulas. For a regular domain $\Omega \subset \mathbb{R}^n$ and a vector field $X \in C^1(\bar{\Omega}, \mathbb{R}^n)$, it takes the following form:

$$\int_{\partial\Omega} X \cdot \mathbf{n} = \int_{\Omega} \operatorname{div} X$$

where $\operatorname{div} X := \partial_1 X_1 + \cdots + \partial_n X_n$ and $\mathbf{n}$ is the normal vector to the boundary $\partial\Omega$. The first purpose of this section is to make sense of all the quantities involved. We follow the line of [2].

1.2.1 Integral over the boundary

We first define the integral of a function over the boundary of a graph-like domain. The general case follows from decomposing any domain Ω into graph-like domains.

Let $\omega = (a_1, b_1) \times \ldots \times (a_{n-1}, b_{n-1}) \subset \mathbb{R}^{n-1}$, (a,b) be an interval in $\mathbb{R}$ and $\phi \in C^1(\omega, (a,b))$. For a point $x \in \mathbb{R}^n$, we will denote $x = (x', x_n)$ where $x' \in \mathbb{R}^{n-1}$, $x_n \in \mathbb{R}$. We define

$$\begin{aligned} \Omega &:= \{x \in \omega \times (a,b) : x_n > \phi(x')\} \\ \tilde{\Omega} &:= \{x \in \omega \times (a,b) : x_n \geq \phi(x')\} \\ \delta\Omega &:= \{x \in \omega \times (a,b) : x_n = \phi(x')\}. \end{aligned}$$

We say that Ω is a *graph-like domain*. A depiction of such domain is given in Figure 1.2.1.

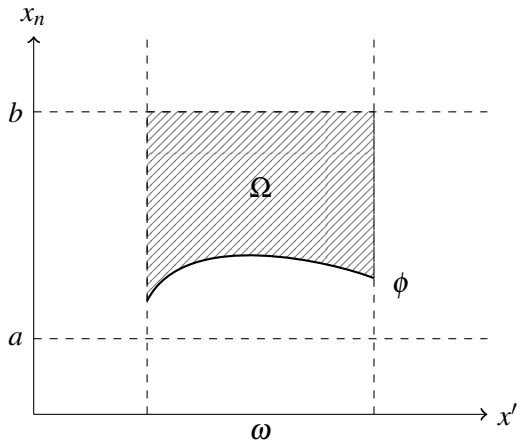

Figure 1.1: Example of 2D graph-like domain (generated by ChatGPT from a hand drawing!).

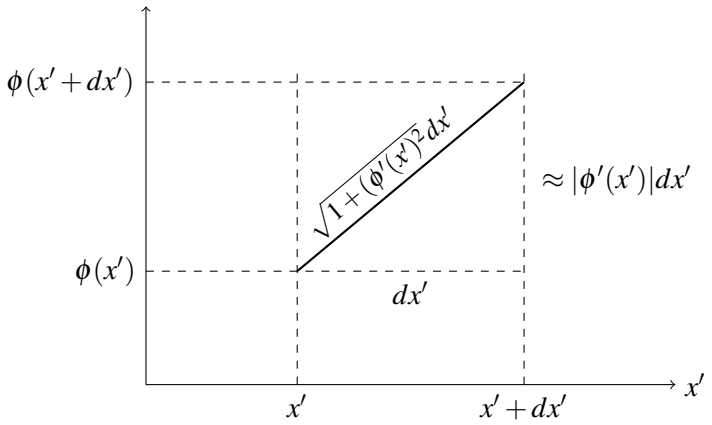

Figure 1.2: Derivation of the length of the boundary of a 2D domain, "physicist style".

Definition 1.2.1 — Surface integral on a graph. Let $f \in C_c(\delta\Omega, \mathbb{R})$. We define

$$\int_{\partial\Omega} f := \int_{\omega} f(x', \phi(x')) \sqrt{1+|\nabla\phi(x')|^2} dx'. \tag{1.7}$$

Intuitively, the term $\sqrt{1+|\nabla\phi(x')|^2}$ represents the area of a small patch of the surface $\delta\Omega$ around x', as illustrated in Figure 1.2. We also define the *normal vector* to the boundary $\delta\Omega$, which appears in the Divergence formula:

Definition 1.2.2 — Normal vector. The outward normal vector at $x \in \delta\Omega$ is given by

$$\mathbf{n}(x) := \frac{1}{\sqrt{1+|\nabla\phi(x')|^2}} \begin{pmatrix} \nabla\phi(x') \\ -1 \end{pmatrix} \tag{1.8}$$

Definition 1.2.3 — Regular domain. An open set $\Omega \subset \mathbb{R}^n$ is said to be of class C^1 if for all $x \in \partial\Omega$, there exists a system of coordinates $(y_1, ..., y_n)$, a set $\omega = (a_1, b_1) \times \ldots \times (a_{n-1}, b_{n-1})$, an interval (a, b) and a function $\phi \in C^1(\omega, (a, b))$ such that

$$\Omega \cap Q = \{y \in Q : y_n > \phi(y')\}$$

where $Q := \omega \times (a,b)$

Hence, a regular domain is the domain for which the boundary is locally the graph of a regular function. Such a definition also imposes that the domain lies on only one side of its boundary (fractured domains do not meet this definition). Also, it allows us to define the outward normal at each point of $\partial\Omega$.

Definition 1.2.4 — Normal vector. Let Ω be a regular domain and $x \in \partial\Omega$. With the same notations used in Definition 1.2.3, the outward normal is given by (1.8).

R We should prove that the definition of the normal vector does not depend on the system of coordinates $(y_1,...,y_n)$ or on the function ϕ. It is indeed the case but the proof is of limited interest. The interested reader is advised to look at [2], Annexe A.

We now want to give a meaning to the integral over the boundary $\partial\Omega$. For this purpose, we need to recall the concept of partition of unity, which allows to "glue" together the different parts of the boundary $\partial\Omega$.

Theorem 1.2.1 — Partition of unity. Let $K \subset \mathbb{R}^n$ be compact and $(\omega_i)_{0\leq i\leq N}$ be a family of open sets such that $K \subset \cup_{i=0}^N \omega_i$. There exists a *partition of unity* subordinated to $(\omega_i)_{0\leq i\leq N}$, i.e. a set of functions $(\theta_i)_{0\leq i\leq N}$ in $C_c^\infty(\mathbb{R}^n,[0,1])$ such that $\text{supp}\theta_i \subset \omega_i$ and

$$\sum_{i=0}^{N} \theta_i(x) = 1$$

for all $x \in K$.

Now consider a bounded regular domain Ω. Its boundary $\partial\Omega$ being compact, it can be covered by a finite family $(Q_i)_{1\leq i\leq N}$ of open sets as defined in Definition 1.2.3. There exists a partition of unity $(\theta_i)_{1\leq i\leq N}$ associated to $(Q_i)_{1\leq i\leq N}$. Let us denote $\Omega_i = \Omega \cap Q_i$ for all i. On each graph-like domain Ω_i, the quantity

$$\int_{\delta\Omega_i} f\theta_i$$

is well defined in the sense of Definition 1.2.1.

Definition 1.2.5 — Integral over the boundary. With the previous notations, we define

$$\int_{\delta\Omega} f := \sum_{i=0}^{N} \int_{\delta\Omega_i} f\theta_i.$$

R Once again, we should prove that this definition is independent of all the quantities we introduced: the sets Q_i, the partition of unity, etc.

We will often talk about the "surface measure" of $\partial\Omega$. This is not an abuse of language, thanks to the following classical (and wonderful) theorem of measure theory:

Theorem 1.2.2 — Riesz-Markov. Let X be a locally compact Hausdorff space and L a positive linear functional on $C_c(X)$ (i.e. $f \geq 0 \implies L(f) \geq 0$). Then there exists a unique positive Borel

measure μ on X such that for all $f \in C_c(X)$,

$$L(f) = \int_X f d\mu.$$

Proof. See [3]. ■

According to the previous definition, we can show that for all $f \in C(\partial\Omega)$,

$$L(f) := \int_{\partial\Omega} f$$

is a linear functional and is positive (since each θ_i is positive). Hence the Riesz-Markov theorem states that there exists a measure σ such that

$$\int_{\partial\Omega} f = \int_{\partial\Omega} f d\sigma,$$

and σ is called the "surface measure" of $\partial\Omega$. When it is not necessary, we will continue not to write this measure in the integrals.

1.2.2 Divergence formula

We can now move on to the proof of the Divergence formula. In the case where the vector field vanishes at the boundary, it can easily be proved:

Proposition 1.2.3 Let Ω be an open bounded set and $X \in C_c^1(\Omega, \mathbb{R}^n)$. Then

$$\int_\Omega \mathrm{div} X = 0.$$

Proof. Take l such that $\Omega \subset (-l,l)^n$ and extend X by 0 on $(-l,l)^n$. Then

$$\int_\Omega \mathrm{div} X = \int_{(-l,l)^n} \mathrm{div} X = \sum_{i=1}^n \int_{-l}^{l} \cdots \int_{-l}^{l} \partial_i X_i dx_1 \dots dx_n.$$

But $\int_{-l}^{l} \partial_i X_i dx_i = 0$ for all i hence the result. ■

The main lemma for the full proof reduces to the case of a graph-like domain. It makes use of the change of variables formula in $\mathbb{R}^n$:

Theorem 1.2.4 — Change of variables. Let $\Omega \subset \mathbb{R}^n$ be an open set and $\Phi : \Omega \to \mathbb{R}^n$ be a C^1-diffeomorphism from Ω to $\Phi(\Omega)$. Let $f : \Phi(\Omega) \to \mathbb{R}$ be a measurable function. Then

$$\int_{\Phi(\Omega)} f = \int_\Omega (f \circ \Phi) |\det D\Phi|.$$

Lemma 1.1 — Divergence formula for a graph. Let $\Omega = \{x \in \omega \times (a,b) : x_n > \phi(x')\}$ be a graph-like domain and $X \in C_c^1(\omega \times (a,b), \mathbb{R}^n)$. Then

$$\int_{\delta\Omega} X \cdot \mathbf{n} = \int_\Omega \mathrm{div} X.$$

Proof. First, we show that

$$\begin{aligned}
\int_\Omega \partial_n X_n dx &= \int_\omega \int_{\phi(x')}^{b} \partial_n X_n(x', x_n) dx_n dx' \\
&= \int_\omega \underbrace{X_n(x', b)}_{=0} - X_n(x', \phi(x')) dx' \\
&= -\int_\omega X_n(x', \phi(x')) dx'.
\end{aligned}$$

Next, let $i \in \{1,...,n-1\}$ and $h(x',t) := X_i(x',t+\phi(x'))$. Hence $X_i(x',x_n) = h(x',x_n-\phi(x'))$ and

$$\partial_i X_i(x',x_n) = \partial_i h(x',x_n-\phi(x')) - \partial_n h(x',x_n-\phi(x'))\partial_i\phi(x').$$

Let

$$\Phi : \Omega \to \omega \times (0,\infty), (x',x_n) \mapsto (x',x_n-\phi(x')).$$

Then $\det D\Phi(x) = 1$ for all $x \in \Omega$ hence by the formula of change of variables we get

$$\begin{aligned}\int_\Omega \partial_i X_i dx &= \int_\Omega \partial_i h(x',x_n-\phi(x')) - \partial_n h(x',x_n-\phi(x'))\partial_i\phi(x')dx \\ &= \int_{\omega\times(0,\infty)} \partial_i h(x',t)dtdx' - \int_{\omega\times(0,\infty)} \partial_n h(x',t)\partial_i\phi(x')dtdx'\end{aligned}$$

The first term of the right-hand side is equal to zero. Indeed, integrating in x_i first (using Fubini) leads to

$$\int_{\omega\times(0,\infty)} \partial_i h(x',t)dtdx' = \int_{a_1}^{b_1}\dots\int_{a_{n-1}}^{b_{n-1}} \underbrace{\left(\int_{a_i}^{b_i} \partial_i h(x_1,\dots,x_{n-1},t)dx_i\right)}_{=0} dx_1\dots dx_{n-1}dt = 0.$$

By integrating first in t, the second term can be re-expressed as

$$\int_{\omega\times[0,\infty)} \partial_n h(x',t)\partial_i\phi(x')dtdx' = -\int_\omega h(x',0)\partial_i\phi(x')dtdx' = -\int_\omega X_i(x',\phi(x'))\partial_i\phi(x')dx'$$

hence

$$\int_\Omega \partial_i X_i dx = \int_\omega X_i(x',\phi(x'))\partial_i\phi(x')dx'.$$

Now, by summing for $i \in \{1,\dots,n\}$, we get

$$\begin{aligned}\int_\Omega \mathrm{div}X &= \int_\omega X_1(x',\phi(x'))\partial_1\phi(x') + \dots + X_{n-1}(x',\phi(x'))\partial_{n-1}\phi(x') - X_n(x',\phi(x'))dx' \\ &= \int_\omega X \cdot \begin{pmatrix}\nabla\phi(x') \\ -1\end{pmatrix} \\ &= \int_{\partial\Omega} X\cdot\mathbf{n}.\end{aligned}$$

■

We now want to prove the Divergence formula for a class of regular domains in $\mathbb{R}^n$. We first need to define what it means for a function to be smooth up to the boundary:

Definition 1.2.6 Let $\Omega \subset \mathbb{R}^n$. A function u is of class $C^k(\overline{\Omega})$ if there exists $\tilde{u} \in C^k(\mathbb{R}^n)$ such that $u = \tilde{u}|_\Omega$.

R For Ω smooth enough, it is the equivalent of saying that u is in $C^k(\Omega)$ and all its derivatives up to order k continuously extend to $\overline{\Omega}$. This is however not trivial and beyond the scope of these notes.

Theorem 1.2.5 — Divergence formula. Let Ω be a C^1 open bounded set and $X \in C^1(\overline{\Omega},\mathbb{R}^n)$. We have

$$\int_\Omega \mathrm{div}X = \int_{\partial\Omega} X\cdot n.$$

Proof. Let $(Q_i)_{1\leq i\leq N}$ be an open cover of $\partial\Omega$ where each Q_i verifies Definition 1.2.3. Let $Q_0=\Omega$. Then $(Q_i)_{0\leq i\leq N}$ is an open cover of $\overline{\Omega}$. Consider a partition of unity $(\theta_i)_{0\leq i\leq N}$ on $\overline{\Omega}$ associated to $(Q_i)_{0\leq i\leq N}$. Since $\theta_0=0$ on $\partial\Omega$, $(\theta_i)_{1\leq i\leq N}$ is also a partition of unity on $\partial\Omega$ associated to $(Q_i)_{1\leq i\leq N}$. Knowing this, we can decompose :

$$\begin{aligned}
\int_{\partial\Omega} X\cdot n &:= \sum_{i=1}^{N}\int_{\partial\Omega_i}(X\cdot n)\theta_i \\
&= \sum_{i=1}^{N}\int_{\Omega_i}\operatorname{div}(X\theta_i) && \text{using Stokes for graphs} \\
&= \sum_{i=0}^{N}\int_{\Omega_i}\operatorname{div}(X\theta_i) && \text{since } X\theta_0\in C_c^1(\Omega,\mathbb{R}^n) \\
&= \sum_{i=0}^{N}\int_{\Omega}\operatorname{div}(X\theta_i) \\
&= \int_{\Omega}\operatorname{div}(X\underbrace{\sum_{i=0}^{N}\theta_i}_{=1}) = \int_{\Omega}\operatorname{div}(X)
\end{aligned}$$

■

There are many variants of the Divergence formula, that we will often call the "Green's Formulas". In the following exercise, we prove some of them.

Exercise 1.1 — Green's formulas. Show that for $u,v\in C^1(\overline{\Omega})$, we have

$$\int_{\Omega}(\partial_i u)v=\int_{\partial\Omega}uv\mathbf{n}_i-\int_{\Omega}u\partial_i v.$$

Deduce that for $u\in C^2(\overline{\Omega})$,

$$\int_{\Omega}(\Delta u)v=\int_{\partial\Omega}(\partial_{\mathbf{n}}u)v-\int_{\Omega}\nabla u\cdot\nabla v.$$

with $\Delta u=\partial_1^2u+\cdots+\partial_n^2u$ and $\partial_n u=\nabla u\cdot n$. ■

The Divergence formula also allows to deduce the pretty useful formula of polar integration:

Exercise 1.2 Let $f\in C_c^1(\mathbb{R}^n)$. We denote by B_r the ball centered at 0 of radius r.

1. Using the change of variable formula

$$\int_{\psi(\Omega)}f(x)dx=\int_{\Omega}f(\psi(x))|\det D\psi(x)|dx$$

and the Divergence formula, show that

$$\frac{d}{dt}\left(\int_{B_{r+t}}f\right)_{t=0}=\int_{\partial B_r}f.$$

2. Deduce that we can integrate "in polar coordinates", i.e.:

$$\int_{\mathbb{R}^n} f = \int_{r=0}^{\infty} \left(\int_{\partial B_r} f \right) dr.$$

■

The Divergence formula has a lot of other applications: establishment of the Maxwell equations, proof of the Brouwer fixed point theorem and is essential to the definition of the so-called *variational formulation* of PDEs.

1.3 Variational formulation

Here we mostly follow [1].

1.3.1 Transforming the Poisson equation

Let's get back to Equation (1.1). Our goal is to transform it into an integral equation. To this purpose, let

$$X := \left\{ u \in C^1(\bar{\Omega}) \text{ s.t. } u = 0 \text{ on } \partial\Omega \right\}.$$

Proposition 1.3.1 Let $u \in C^2(\bar{\Omega})$ and $f \in C(\overline{\Omega})$. Then u is solution of (1.1) if and only if

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v \quad \text{for all } v \in X. \tag{1.9}$$

The equation (1.9) is called the *variational formulation* or *weak formulation* of (1.1).

The proof of this proposition requires the following lemma:

Proposition 1.3.2 Let $g \in C(\Omega)$ Suppose that for all $\phi \in C_c^\infty(\Omega)$, we have

$$\int_\Omega g\phi = 0.$$

Then $g = 0$.

Proof. For the sake of contradiction, suppose that there exists x s.t. $g(x) \neq 0$ (for instance $g(x) > 0$). Since g is continuous, there exists a ball B_x on which $g > 0$. Now take a test function $\phi \in C_c^\infty(B_x)$ such that $\phi \geq 0, \phi \neq 0$. Then

$$\int_\Omega g\phi > 0.$$

■

Proof of Proposition 1.3.1. The direct implication uses the Divergence formula. The reverse implication uses the previous proposition. See [1], Proposition 3.2.7. ■

R We will see in the future that it is common to take the weak formulation as the definition of a PDE. Being used to the strong formulation, it may seem strange at first; however, there are good reasons for that. First, the variational formulation requires less regularity to be defined than the strong one and coincides with it when it is possible to show that the solution of the weak one is regular enough. Hence the weak formulation is strictly **more general** than the strong one. Another motivation is that the weak formulation may in some sense be *more physical* than the strong one (namely, when we can show that the solution u minimizes an energy functional).

Exercise 1.3 For Ω in $\mathbb{R}^n$ bounded, open, and $f \in C(\overline{\Omega})$, consider the Poisson problem with homogeneous *Neumann* boundary conditions :

$$\begin{cases} -\Delta u = f \text{ in } \Omega \\ \partial_n u = 0 \text{ on } \partial\Omega \end{cases} \tag{1.10}$$

Show that a function $u \in C^2(\overline{\Omega})$ is a solution of (3.13) if and only if for all $v \in C^1(\overline{\Omega})$,

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv.$$

Show that f must verify a certain integral condition for u to exist. ■

Exercise 1.4 For Ω in $\mathbb{R}^n$ bounded, open, and $f \in C(\overline{\Omega})$, consider the *plate problem*:

$$\begin{cases} \Delta^2 u = f \text{ in } \Omega \\ u = 0 \text{ on } \partial\Omega \\ \partial_n u = 0 \text{ on } \partial\Omega \end{cases} \tag{1.11}$$

where $\Delta^2 u = \Delta(\Delta u)$. Show that a function $u \in C^4(\overline{\Omega})$ is a solution of (3.13) if and only if for all $v \in Y$,

$$\int_\Omega \Delta u \Delta v = \int_\Omega fv.$$

for a certain space Y. ■

We can rewrite (1.9) as: find $u \in C^2$ such that,

$$a(u,v) = L(v) \quad \text{for all} \quad v \in X.$$

where

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v$$

$$L(v) := \int_\Omega fv.$$

An abstract result on Hilbert spaces called the *Lax-Milgram theorem* allows us to solve this kind of equation, under certain assumptions on a and L.

1.3.2 Lax-Milgram theory

Let V be an Hilbert space with scalar product $\langle , \rangle$ and norm $\|.\|$. We consider the following variational formulation:

$$\text{Find } u \in V \text{ such that for every } v \in V,\ a(u,v) = L(v). \tag{1.12}$$

where a and L verifies the following hypothesis:

1. L is a continuous linear form on V, i.e. $L : V \to \mathbb{R}$ is linear and there exists $C > 0$ such that for all $v \in V$,
$$|L(v)| \leq C\|v\|;$$
2. a is a continuous bilinear form on V, i.e. $a : V \times V \to \mathbb{R}$ is such that $a(.,w)$ is linear for all w, $a(v,.)$ is linear for all v and there exists $M > 0$ such that for all$v, w \in V$,
$$|a(v,w)| \leq M\|v\|\|w\|;$$

3. a is coercive (also called elliptic), i.e. there exists $\nu > 0$ such that for all $v \in V$,

$$a(v,v) \geq \nu \|v\|^2.$$

Theorem 1.3.3 — Lax Milgram. Let V be a real Hilbert space, L a continuous linear form on V and a a continuous, coercive bilinear form on V. Then the problem (1.12) has a unique solution depending continuously on L.

Proof. For all $w \in V$, $v \mapsto a(w,v)$ is a linear continuous form. By Riesz theorem, there exist $A(w) \in V$ such that $\langle A(w), v\rangle = a(w,v)$ for all v. By bilinearity of a and uniqueness in the Riesz theorem, $w \mapsto A(w)$ defines a linear map. Moreover,

$$\|A(w)\|^2 \leq a(w, A(w)) \leq M\|w\|\|A(w)\|$$

hence A is continuous. By the Riesz theorem again, there exists $f \in V$ such that for all $v \in V$,

$$\langle f, v\rangle = L(v).$$

The problem (1.12) then reduce to

$$\text{Find u} \in V \text{ such that } A(u) = f.$$

Let $\mu = \frac{\nu}{M^2}$ and define

$$\begin{aligned} T : V &\to V \\ w &\mapsto w - \mu\,(A(w) - f) \end{aligned}$$

We show that T is a contraction :

$$\begin{aligned} \|T(v) - T(w)\| &= \|v - w - \mu A(v-w)\|^2 \\ &= \|v-w\|^2 - 2\mu\langle A(v-w), v-w\rangle + \mu^2\|A(v-w)\|^2 \\ &= \|v-w\|^2 - 2\mu \underbrace{a(v-w, v-w)}_{\geq \nu\|v-w\|^2} + \mu^2\|A(v-w)\|^2 \\ &\leq (1 - 2\mu\nu\mu^2 M^2)\|v-w\|^2 \\ &\leq \underbrace{(1 - \frac{\nu^2}{M^2})}_{\leq 1}\|v-w\|^2 \end{aligned}$$

By the Banach fixed point theorem, there exists a unique $u \in V$ such that $T(u) = u$ hence $A(u) = f$. The continuity w.r.t. L comes from the fact that $\|f\| = \|L\|$ (dual norm of L) and the continuity of A^{-1}; indeed, for all $w \in V$,

$$\nu\|w\|^2 \leq a(w,w) \leq \langle A(w), w\rangle \leq \|A(w)\|\|w\|$$

hence by taking $w = u = A^{-1}(f)$,

$$\|u\| = \|A^{-1}(f)\| \leq \nu^{-1}\|f\|.$$

■

The following exercise links the Lax-Milgram theorem to a *variational principle*, namely the minimization of a quantity (often interpreted as an energy).

Exercise 1.5 In addition to the hypotheses of the Lax-Milgram theorem, we suppose that a is symmetric (i.e. $a(u,v) = a(v,u)$ for all $u,v \in V$). Let us define for all $v \in V$ the functional

$$J(v) := \frac{1}{2}a(v,v) - L(v).$$

Show that we have

$$J(u) = \min_{v \in V} J(v)$$

where u is the unique solution of (1.12). Reciprocally, show that if $u \in V$ is a minimum of J then it is the unique solution of (1.12). As a hint, you can consider the function $t \mapsto J(u+tv)$ and use first-order optimality conditions. ■

R Consequently, when a is symmetric, we can show that there exists a minimizer for J instead of using the Lax-Milgram theorem to get a solution of (1.12). This is usually done using the convexity and lower-semicontinuity of J with respect to the *weak convergence*. This kind of method is the prototype of the field of **Calculus of Variations**, which is a way to study mechanical structures, vibrating membranes and...soap bubbles (and honeycombs, and fractures, and a lot of other things that all have in common of optimizing an energy).

1.3.3 Back to the Poisson equation

We want to use the Lax-Milgram theorem to prove that the Poisson equation (1.9) has a solution. To this purpose, we have to define an inner product on the space X. The one that naturally comes to mind is

$$\langle u, v \rangle := \int_\Omega \nabla u \cdot \nabla v.$$

The norm associated with this product is

$$\|v\| = \left(\int_\Omega |\nabla v|^2 \right)^{1/2}.$$

It is indeed an inner product: first, it is obviously bilinear, symmetric and positive. The definiteness comes from the fact that

$$\|v\| = 0 \implies |\nabla v| = 0 \implies v \text{ is constant on every connected component of } \Omega$$

and v being zero at the boundary, it is zero everywhere.

R To show the last $\implies$, we can first show (by the mean value theorem) that if $\nabla v = 0$ on a ball then it is constant on this ball. Then it is enough to show that a function that is locally constant on a connected set is constant, by considering the set $\{x \in \Omega : f(x) = f(x_0)\}$ is open and closed (for a certain $x_0 \in \Omega$).

We now prove that a and L verify the conditions of the Lax-Milgram theorem.

Proposition 1.3.4 1. The function

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v$$

is a coercive, continuous bilinear form on X.

2. The function
$$L(v) := \int_\Omega fv.$$
is a continuous linear form on X.

Proof. 1. All the properties of a come from the fact that it is also the scalar product defining our Hilbert space.

2. L is a linear form. We want to show that it is continuous. Using the Cauchy-Schwarz Inequality, we have for all $v \in X$:
$$|L(v)| = \left|\int_\Omega fv\right| \leq \|f\|_{L^2}\|v\|_{L^2}.$$
The domain Ω being bounded and f being continuous on $\overline{\Omega}$, $\|f\|_{L^2} < \infty$. To conclude, we would like to show that there exists $C > 0$ such that
$$\|v\|_{L^2} \leq C\|v\|;$$
for this we need the following *Poincaré Inequality*.

■

Theorem 1.3.5 — Poincaré Inequality for C^1 functions. Let Ω be an open bounded set of $\mathbb{R}^n$. There exists $C > 0$ such that for all $v \in X$,
$$\int_\Omega |v|^2 \leq C\int_\Omega |\nabla v|^2.$$

Proof. Since Ω is bounded then we can suppose $\Omega \subset (a,b)^n$. By extending v by 0, we can then consider $v \in C_0^1([a,b]^n)$. Then for $x = (x_1,\dots,x_n) \in \Omega$,
$$v(x) = \int_a^{x_1} \partial_1 v(t,x_2,\dots,x_n)dt.$$
Then by Chauchy-Schwarz,
$$|v(x)|^2 \leq |b-a|\int_a^b |\partial_1 v(t,x_2,\dots,x_n)|^2 dt \leq \int_a^b |\nabla v(t,x_2,\dots,x_n)|^2 dt$$
so by integrating on Ω,
$$\begin{aligned}
\int_\Omega |v(x)|^2 &\leq |b-a|\int_\Omega\int_a^b |\nabla v(t,x_2,\dots,x_n)|^2 dtdx \\
&\leq |b-a|\int_{[a,b]^n}\int_a^b |\nabla v(t,x_2,\dots,x_n)|^2 dx_1dx_2\dots dx_n dt \\
&\leq |b-a|\int_a^b\int_{[a,b]^n} |\nabla v(t,x_2,\dots,x_n)|^2 dtdx_2\dots dx_n dx_1 \\
&\leq |b-a|\int_a^b\int_\Omega |\nabla v(x)|^2 dxdx_1 \\
&\leq |b-a|^2\int_\Omega |\nabla v(x)|^2
\end{aligned}$$

■

R Actually, we only need Ω to be bounded in one direction, i.e. being contained between two parallel hyperplanes.

Now we should be able to apply the Lax-Milgram theorem! Wait...there is a hypothesis that we did not verify: we need X to be complete! Unfortunately, this is not the case (as shown in the following exercise). Hence, we will need to introduce a new family of spaces that are naturally complete with respect to the inner product we have defined: the *Sobolev spaces*.

Exercise 1.6 Let $\Omega = B_1$, the open unit ball of $\mathbb{R}^n$

1. For $n = 1$, put

$$u_n(x) = \begin{cases} -x-1 & \text{if } -1 < x < -1/n \\ (n/2)x^2 - 1 + 1/(2n) & \text{if } -1/n \leq x \leq 1/n \\ x-1 & \text{if } 1/n < x < 1 \end{cases}.$$

Draw the function u_n. Show that u_n is a Cauchy sequence in $(X, \langle , \rangle)$, i.e.

$$\int_0^1 |u_n' - u_m'|^2 \xrightarrow[n,m\to\infty]{} 0.$$

Show that u_n can not converge to a continuous function. Conclude.

2. For $n \geq 3$ and $0 < \beta < (n-2)/n$, put

$$u_n(x) = (|x|^2 + 1/n)^{-\beta/2} - (1 + 1/n)^{-\beta/2}.$$

Using the integration in polar coordinates and the Dominated Convergence Theorem, show that u_n is Cauchy. Show that it can not converge to a continuous function.

■

R During the final project, you will be asked to solve some elliptic PDE using Fenics, a PDE solver available in Python based on the Finite Element Method (FEM). While you do not know yet how the FEM works, you can already have a look at the Introduction and Fundamentals parts of the Fenics tutorial (link), and play around with the examples. It will save you some time for the project, for which you will also need to learn the neural network library Pytorch.

2. Generalities on Neural Networks

A neural network is a tool coming from the field of Machine Learning, which goal is to use data to produce an output. It can consist for instance in using the data to make *predictions*, or finding *structure* in the data. There are numerous subfields of machine learning, however the two main ones are *supervised* and *unsupervised* machine learning:

- **Supervised:** relies on *labeled* data.
  - **Regression:** given points $(x_i, y_i)_i \subset \mathbb{R}^n \times \mathbb{R}^m$, find $f : \mathbb{R}^n \to \mathbb{R}^m$ such that $f(x_i) \approx y_i$ for all i (e.g. linear regression, neural networks...)
  - **Classification:** similar to regression but y_i lies in a discrete set of *classes* (for instance, classification of pictures of cats vs dogs)(e.g. Support Vector Machines, neural networks...).
- **Unsupervised:** makes use of *unlabeled data*.
  - **Clustering:** given points $(x_i)_i$ and a number $K \in \mathbb{N}$, find a meaningful partition of $(x_i)_i$ into K sets (e.g. k-means, gaussian mixtures, spectral clustering...);
  - **Dimensionality reduction:** given points $(x_i)_i \subset \mathbb{R}^n$, find some m-dimensional manifold $\mathscr{M}$ such that $m << n$ and $(x_i)_i \subset \mathscr{M}$ (e.g. PCA, diffusion maps...)

In this chapter, we will examine the elementary theory of neural networks.

2.1 Neural Networks

2.1.1 Definition and examples

Consider the previous regression task. The main question is: how to represent the function f ? In the case of linear regression, we impose $f(x) = Mx$, $M \in \mathbb{R}^{m \times n}$, but what if we want to model more complex relationships between the input x_i and the output y_i ?

Definition 2.1.1 — Neural Network. A neural network is a function $f_\theta : \mathbb{R}^n \to \mathbb{R}^m$ of the form $f(x) = g^L \circ \cdots \circ g^1(x)$ where for all $l \in \{1, \dots, L\}$, $g^l : \mathbb{R}^{n_l} \to \mathbb{R}^{n_{l+1}}$ (with $n_1 = n$ and $n_L = m$) is of the form

$$g^l(z) = \sigma^l(W^l z + b^l)$$

where $W^l \in \mathbb{R}^{n_{l+1} \times n_l}$ are the *weights*, $b^l \in \mathbb{R}^{n_{l+1}}$ are the *biases* and $\sigma^l : \mathbb{R} \to \mathbb{R}$ (applied element-wise) is the *activation function* of the *l-layer* of the network. Moreover, $\theta := \{(W^l, b^l)\}_l$ represents the *parameters* of the network.

R Note that this definition is more precisely the definition of a *Multi-layer Perceptron* also called *deep fully-connected neural network*. This is the simplest non-trivial architecture of neural networks. Depending on the problem, more complex architectures are used:

- Convolutional neural networks are the go-to type of neural networks when it comes to imaging;
- Residual Neural networks have layers of the form

$$g^l(z) = \sigma^l(W^l z + b^l) + z$$

  and have been proven effective in building very-deep neural networks;
- Transformers are famous for its effectiveness in natural language processing with the Generative Pre-trained Transformers
- ... and millions of others.

In what follows, we will make the assumption (often used in practice) that $\sigma^1 = \sigma^{L-1} =: \sigma$ and $\sigma^L = \mathrm{Id}$. Our goal is the following: given $(x_i, y_i)_i \subset \mathbb{R}^n \times \mathbb{R}^m$, find θ such that $f_\theta(x_i) \approx y_i$ for all i.

2.2 How neural networks learn

2.2.1 Loss functions

In the context of neural network, the learning process amounts at solving an optimization problem. Namely, we aim at minimizing the disparity between each label y_i and the corresponding prediction of the network $f_\theta(x_i)$. One way (among plenty of others) to quantify this disparity is via the *mean squared error*, which reads

$$\mathscr{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} |y_i - f_\theta(x_i)|^2.$$

Training the neural network amounts at solving the optimization problem

$$\min_\theta \mathscr{L}(\theta).$$

R In the neural network lingo, the quantity we aim at minimizing in order to improve the performance of the network is called the *loss function*.

The question that naturally arises is how to numerically solve the previous problem ? In the vast majority of cases, it is simply impossible due to the fact that the function $\theta \mapsto \mathscr{L}(\theta)$ is highly non-convex and high dimensional. However, we can still look for a local minimum with the help of gradient descent.

2.2.2 Gradient descent

Let $F : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. For any $\theta \in \mathbb{R}^n$, the gradient $-\nabla F(\theta)$ is the direction of steepest descent of F. Indeed, for a small step $\tau > 0$, at the first order, we have that

$$F(\theta - \tau \nabla F(\theta)) \approx F(\theta) - \tau |\nabla F(\theta)|^2 \leq F(\theta)$$

which means that for a small enough τ, taking a step in the direction of $-\nabla F(\theta)$ will make F decrease. This is the idea behind the gradient descent algorithm, which starts from $\theta_0 \in \mathbb{R}^n$ and iterates according to:

$$\theta_{k+1} = \theta_k - \tau \nabla F(\theta_k).$$

This is the prototypical algorithm for training a neural network, by taking F to be the loss function $\mathscr{L}$.

R There exists a vast bestiary of gradient based optimization algorithm. To name a few that are important in the context of neural network, we can find stochastic gradient descent (SGD), batch gradient descent, adaptive moment estimation (ADAM)...

2.2.3 Backpropagation

In order to perform gradient descent over $\mathscr{L}$, we need to be able to compute $\nabla\mathscr{L}$, i.e. every partial derivatives $\frac{\partial\mathscr{L}}{\partial w_{ij}^l}$ and $\frac{\partial\mathscr{L}}{\partial b_i^l}$ for i, j and l. By linearity of the loss, we can restrict ourselves to the case where $\mathscr{L}(\theta) = |y - f_\theta(x)|^2$ for some $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. To make the computations easier, let us introduce for $l \in \{1, \dots, L\}$ the quantities $z^0 := x$,

$$a^l := W^l z^{l-1} + b^l \qquad z^l := \sigma(a^l) = g^l(z^{l-1})$$

and $z^L := a^L = g^L \circ \cdots \circ g^1(x) = f_\theta(x)$. Coordinate-wise, the quantities expands as:

$$a_i^l := \sum_{j=1}^{n_l} w_{ij}^l z_j^{l-1} + b_i^l \quad \text{for } i \in \{1, \dots, n_{l+1}\},$$

and

$$z_i^l := \sigma(a_i^l) \quad \text{for } i \in \{1, \dots, n_{l+1}\}.$$

We will focus on the computation of $\frac{\partial\mathscr{L}}{\partial w_{ij}^l}$. By the chain rule, we have that

$$\frac{\partial\mathscr{L}}{\partial w_{ij}^l} = \sum_k \frac{\partial\mathscr{L}}{\partial a_k^l} \underbrace{\frac{\partial a_k^l}{\partial w_{ij}^l}}_{=0 \text{ if } k \neq i} = \frac{\partial\mathscr{L}}{\partial a_i^l}\frac{\partial a_i^l}{\partial w_{ij}^l} = \underbrace{\frac{\partial\mathscr{L}}{\partial a_i^l}}_{=:\delta_i^l} z_j^{l-1} = \delta_i^l z_j^{l-1}.$$

It is possible to get all the z_j^{l-1} simply by evaluating the neural network (this is what machine learners call a *forward pass*). We will show now that the δ_i^l can be recursively computed starting from the last layer δ_i^L all the way back to the input layer (this is the *backward pass*, hence the name *backpropagation*). Indeed, for the last layer:

$$\delta_i^L = \frac{\partial\mathscr{L}}{\partial a_i^l} = \frac{\partial}{\partial a_i^l}\left(|y - z^L|^2\right) = \frac{\partial}{\partial a_i^l}\left(|y - a^L|^2\right) = 2(y_i - a_i^L).$$

Let $1 \leq l \leq L-1$. In this case,

$$\delta_i^l = \frac{\partial\mathscr{L}}{\partial a_i^l} = \sum_j \frac{\partial\mathscr{L}}{\partial a_j^{l+1}}\frac{\partial a_j^{l+1}}{\partial a_i^l} = \sum_j \delta_j^{l+1}\frac{\partial a_j^{l+1}}{\partial a_i^l}$$

but

$$\frac{\partial a_j^{l+1}}{\partial a_i^l} = \frac{\partial}{\partial a_i^l}\left(\sum_k w_{jk}^{l+1} z_k^l + b_j^{l+1}\right) = \sigma'(a_i^l) w_{ji}^{l+1}$$

which finally leads to

$$\delta_i^l = \sum_j \delta_j^{l+1}\sigma'(a_i^l) w_{ji}^{l+1}.$$

R Do not worry, you will not have to implement this by yourself (although it is a very good exercise), as every neural network library does all these computations (and way more) automagically.

Exercise 2.1 Following the same logic as before, compute $\frac{\partial \mathscr{L}}{\partial b_i^l}$. ■

2.2.4 Notebook

You can use this notebook to experiment with you first neural network: (click here).

2.3 The Universal Approximation Theorem

We now know how to minimize the MSE loss given some data and a neural network f_θ. However, we do not know if f_θ is *expressive* enough to actually fit the data correctly (think of a linear f_θ trying to fit non-linear data). In this part, we will prove the so-called *Universal Approximation Theorem*, stating that any continuous function can be approximated 1-layer neural network.

2.3.1 Reminders of functional analysis

We will show the original proof of the UAT, which was derived by George Cybenko in 1989 [7]. His famous proof makes use of the Hahn-Banach theorem, a fundamental result of functional analysis:

Theorem 2.3.1 — Hahn-Banach. Let V be a normed vector space and $A, B \subset V$ two closed convex disjoint sets. There exists $f \in V'$ (the vector space of continuous linear forms over V) and $\alpha \in \mathbb{R}$ such that

$$f(x) < \alpha \text{ for } x \in A \quad \text{and} \quad f(x) > \alpha \text{ for } x \in B. \tag{2.1}$$

Proof. See [6], Chapter 5. ■

Corollary 2.3.2 Let V be a normed vector space and $U \subset V$ be a subspace such that $\overline{U} \neq V$. Then there exists $f \in V'$ such that $f \not\equiv 0$ and $f = 0$ on U.

Proof. In theorem 2.3.1, take $A = \overline{U}$ and $B = \{z\}$ for $z \notin \overline{U}$. There exists $f \in V'$ and $\alpha > 0$ such that $f(z) > \alpha > f(x)$ for all $x \in \overline{U}$. By linearity, or all $x \in \overline{U}$,

$$f(\lambda x) = \lambda f(x) < \alpha$$

implying that $f(x) = 0$. ■

The proof of Cybenko uses the Riesz-Markov representation theorem for signed measure, which is a generalisation of the one we saw in the previous chapter for positive measures:

Theorem 2.3.3 — Riesz-Markov. Let X be a locally compact Hausdorff space and L a continuous linear functional on $C_0(X)$ (the space of continuous functions vanishing at infinity). Then there exists a unique signed, regular Borel measure μ on X such that for all $f \in C_0(X)$,

$$L(f) = \int_X f d\mu.$$

Proof. See [6], Chapter 6. ■

The last ingredient of Cybenko's proof is the Fourier transform of signed measures, which generalizes the usual Fourier transform on the set of Borel measures:

Definition 2.3.1 — Fourier transform. Let μ be a signed, regular Borel measure on $\mathbb{R}^n$. The Fourier transform of μ is the function

$$\begin{aligned}\hat{\mu}: \quad & \mathbb{R}^n \to \mathbb{C} \\ & x \mapsto \int_{\mathbb{R}^n} e^{-ix\cdot y} d\mu(y)\end{aligned}$$

R

- $\hat{\mu}$ is well defined since μ is finite and $y \mapsto e^{-ix\cdot y}$ is bounded;
- Due to the Lebesgue Dominated Convergence theorem, we have that $\hat{\mu} \in C(\mathbb{R}^n, \mathbb{C})$.

Definition 2.3.2 — Schwartz space. The Schwartz space of *rapidly decreasing functions* is defined as

$$\mathscr{S}(\mathbb{R}^n) := \left\{ f \in \mathbb{C}^\infty(\mathbb{R}^n, \mathbb{C}) : \|x^\alpha \partial^\beta f\|_\infty < \infty \text{ for all } \alpha, \beta \in \mathbb{N}^n \right\}$$

R $\alpha, \beta \in \mathbb{N}^n$ are called *multi-indices* and the expression $x^\alpha \partial^\beta f(x)$ must be understood in the following way:

$$x^\alpha \partial^\beta f(x) := x_1^{\alpha_1} \dots x_n^{\alpha_n} \partial_1^{\alpha_1} \dots \partial_n^{\alpha_n} f(x).$$

The Fourier transform has the following very useful property over the Schwartz space:

Theorem 2.3.4 Let $f \in \mathscr{S}(\mathbb{R}^n)$ and define

$$\mathscr{F}(f)(y) := \int_{\mathbb{R}^n} f(x) e^{-ix\cdot y} dx$$

(in term of Fourier transform of measures, we have that $\mathscr{F}(f)$ is the Fourier transform of the measure $f(x)dx$). Then $\mathscr{F} : \mathscr{S}(\mathbb{R}^n) \to \mathscr{S}(\mathbb{R}^n)$ is bijective.

Theorem 2.3.5 The Fourier transform of signed Borel measures is injective, meaning that $\hat{\mu} \equiv 0 \implies \mu = 0$.

Proof. First, let $\phi \in \mathscr{S}(\mathbb{R}^n)$ and let $\hat{\phi}(y) = \int_{\mathbb{R}^n} \phi(x) e^{-ix\cdot y} dx$. Then by Fubini

$$\begin{aligned}\int_{\mathbb{R}^n} \hat{\phi}(y) d\mu(y) &= \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \phi(x) e^{-ix\cdot y} dx d\mu(y) \\ &= \int_{\mathbb{R}^n} \phi(x) \int_{\mathbb{R}^n} e^{-ix\cdot y} d\mu(y) dx \\ &= \int_{\mathbb{R}^n} \phi(x) \hat{\mu}(x) dx\end{aligned}$$

Now, if $\hat{\mu} \equiv 0$ then $\int_{\mathbb{R}^n} \hat{\phi}(y) d\mu(y) = 0$ for all $\phi \in \mathscr{S}(\mathbb{R}^n)$. Since the Fourier transform is bijectinve on $\mathscr{S}(\mathbb{R}^n)$, we have that $\int_{\mathbb{R}^n} \phi(y) d\mu(y) = 0$ for all $\phi \in \mathscr{S}(\mathbb{R}^n)$. Finally, $\mathscr{S}(\mathbb{R}^n)$ is dense in $C_0(\mathbb{R}^n)$ (since it contains $C_c^\infty(\mathbb{R}^n)$) hence

$$\int_{\mathbb{R}^n} \phi(y) d\mu(y) = 0 \text{ for all } \phi \in C_0(\mathbb{R}^n).$$

By uniqueness in the Riesz-Markov theorem, we get that $\mu = 0$. ∎

2.3.2 Back to the UAT

We will show that a single-layer neural network with a certain class of activation function can approximate any function in $C([0,1]^n)$. Let $\sigma \in C(\mathbb{R})$ and denote by

$$\Sigma(\sigma) := \text{span}\{x \mapsto \sigma(w \cdot x + b) : w \in \mathbb{R}^n, b \in \mathbb{R}\}$$

the space of 1-layer neural networks with activation function σ. We have that $\Sigma(\sigma) \subset C([0,1]^n)$. We aim at showing that $\Sigma(\sigma)$ is dense in $C([0,1]^n)$.

Definition 2.3.3 — Discriminatory functions. We say that $\sigma : \mathbb{R} \to \mathbb{R}$ is discriminatory if for every signed Borel measure μ,

$$\int_{[0,1]^n} \sigma(w \cdot x + b) d\mu(x) = 0 \text{ for all } w \in \mathbb{R}^n, b \in \mathbb{R} \implies \mu = 0.$$

Theorem 2.3.6 — Universal Approximation theorem. If σ is discriminatory then $\Sigma(\sigma)$ is dense in $C([0,1])$ (for the $\|.\|_\infty$ norm). In other words, for all $\varepsilon > 0$ and $g \in C([0,1])$, there exists $N \in \mathbb{N}, a_1, \dots, a_N, b_1, \dots, b_N \in \mathbb{R}$ and $w_1, \dots, w_N \in \mathbb{R}^n$ such that $f_\theta(x) = \sum_{i=1}^N a_i \sigma(w_i \cdot x + b_i)$ verifies

$$\|f_\theta - g\|_\infty < \varepsilon.$$

Proof. We will prove the contraposition. Suppose that $\Sigma(\sigma)$ is not dense in $C([0,1])$, i.e. $\overline{\Sigma(\sigma)} \neq C([0,1])$. By the Corollary 2.3.2, there exists $F \in C([0,1]^n)'$ such that $F \not\equiv 0$ and $F = 0$ on $\overline{\Sigma(\sigma)}$. By the Riesz-Markov theorem, there exists a measure μ such that for all $f \in C([0,1]^n)$ (remark that $[0,1]^n$ is compact hence $C([0,1]^n) = C_0([0,1]^n)$), we have:

$$F(f) = \int_{\mathbb{R}^n} f d\mu.$$

However, for all $w \in \mathbb{R}^n$, $b \in \mathbb{R}$, the function $x \mapsto \sigma(w \cdot x + b) \in C([0,1]^n)$ which implies that

$$\int_{[0,1]^n} \sigma(w \cdot x + b) d\mu(x) = 0.$$

Hence, σ is not discriminatory, otherwise we would have $\mu = 0$ and consequently $F \equiv 0$. ■

We now that the UAT holds for the class of discriminatory functions. However, this class is rather obscure in the sense that it might be difficult to determine if a given function σ is discriminatory, restricting the applicability of the theorem. Nevertheless, we can show that this discriminatory functions contains a subclass of functions that are easier to characterize:

Definition 2.3.4 — Sigmoidal function. We say that $\sigma \in C(\mathbb{R})$ is sigmoidal when

$$\sigma(x) \xrightarrow[x \to +\infty]{} 1 \quad \text{and} \quad \sigma(x) \xrightarrow[x \to -\infty]{} 0.$$

Theorem 2.3.7 Every sigmoidal function is discriminatory.

Proof. Let σ be sigmoidal and μ be a signed Borel measure such that

$$\int_{[0,1]^n} \sigma(w \cdot x + b) d\mu(x) = 0.$$

We aim at showing that $\mu = 0$.

First, for $\lambda, \phi \in \mathbb{R}$, let

$$\gamma(x) := \lim_{\lambda \to \infty} \sigma(\lambda(w \cdot x + b) + \phi) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{if } w \cdot x + b < 0 \\ \sigma(\phi) & \text{if } w \cdot x + b = 0 \end{cases}$$

By defining

$$H^+_{w,b} := \{x \in [0,1]^n : w \cdot x + b > 0\}$$

and

$$\Pi_{w,b} := \{x \in [0,1]^n : w \cdot x + b = 0\},$$

we can re-write

$$\gamma(x) = \mathbf{1}_{H^+_{w,b}}(x) + \sigma(\phi)\mathbf{1}_{\Pi_{w,b}}(x).$$

Using the Dominated Convergence Theorem,

$$\begin{aligned} 0 = \int_{[0,1]^n} \sigma(\lambda(w \cdot x + b) + \phi) d\mu(x) &\xrightarrow[\lambda \to \infty]{} \int_{[0,1]^n} \gamma(x) d\mu(x) \\ &= \int_{[0,1]^n} \mathbf{1}_{H^+_{w,b}}(x) + \sigma(\phi)\mathbf{1}_{\Pi_{w,b}}(x) d\mu(x) \\ &= \mu(H^+_{w,b}) + \sigma(\phi)\mu(\Pi_{w,b}). \end{aligned}$$

When ϕ goes to $-\infty$, we have $\mu(H^+_{w,b}) = 0$. On the other hand, $\phi \to -\infty$ leads to $\mu(H^+_{w,b}) + \mu(\Pi_{w,b}) = \mu(\overline{H^+_{w,b}}) = 0$.

For $w \in \mathbb{R}^n$, let

$$\begin{aligned} F : \quad L^\infty(\mathbb{R}) &\to \mathbb{R} \\ h &\mapsto \int_{[0,1]^n} h(w \cdot x) d\mu(x). \end{aligned}$$

Using that $\mathbf{1}_{[-b,\infty)(w \cdot x)} = \mathbf{1}_{H^+_{w,b}}(x)$, the previous analysis shows that $F\left(\mathbf{1}_{[-b,\infty)}\right) = 0$. Similarly, $F\left(\mathbf{1}_{(-b,\infty)}\right) = 0$. Hence, for any characteristic function of an interval h, $F(h) = 0$. By regularity, we can approximate any measurable set in $\mathbb{R}$ by a union of intervals, hence $F(h) = 0$ for the characteristic function f of any measurable set, and hence for any simple function. Finally, simple functions being dense in $L^\infty(\mathbb{R})$, we have $F \equiv 0$. Since $\sin, \cos \in L^\infty(\mathbb{R})$ we have that

$$0 = F(cos) + iF(sin) = \int_{[0,1]^n} \cos(w \cdot x) + i \sin(w \cdot x) d\mu(x) = \int_{[0,1]^n} e^{iw \cdot x} d\mu(x) = \hat{\mu}(-w)$$

which implies that $\hat{\mu} \equiv 0 \implies \mu = 0$ by the injectivity of the Fourier transform. Hence, σ is discriminatory. ■

The sigmoid activation function $S(x) = \frac{1}{1+e^{-x}}$ is obviously sigmoidal. The common tanh activation function is not, but a simple scaling shows that it is discriminatory. The ReLU function, which is the most widely used activation function, is not sigmoidal either. However, we can show that it is still discriminatory:

Exercise 2.2 Show that $\text{ReLU}(x) = \max(0, x)$ is discriminatory. For this, build a simple sigmoidal function using two ReLU. ■

One can wonder what kind of activation functions makes the UAT fail. This is the purpose of this exercise:

Exercise 2.3 Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a polynomial. Show that a 1-layer neural network with activation function σ does not have the universal approximation property. ■

Exercise 2.4 Show that there exists a single-neuron neural network of the form $f_b(x) = \sigma(x+b)$ (where σ is not necessarily continuous), $b \in \mathbb{R}$ such that for all $g \in C([0,1])$, and $\varepsilon > 0$ there exists $b \in \mathbb{R}$ such that

$$\|g - f_b\|_\infty \leq \varepsilon.$$

■

In order to conclude this chapter, let us mention that the UAT is not restricted to the space of continuous functions and to sigmoidal activation functions. For instance, in 1990, Hornik and co-authors showed that the Universal Approximation property also holds in Sobolev norms by using some fine property of the Fourier transform [10], allowing the approximation of both a function and its derivatives. In 1993 Leshno and co-authors [13] showed that a single layer neural network with non-polynomial activation function can approximate any continuous function, thus generalizing the result of Cybenko. More recent results also study the UAT for deep neural networks with bounded width (see for instance [15]). The moral of the story is that if you need some UAT, there exists one that fits your purpose almost surely.

3. A short introduction to Sobolev spaces

As we saw previously the space of C^1 functions was not the right space to work with variational formulations of PDEs, as it is not complete for the natural inner product. Here we introduce the Sobolev space H^1, which is a space of *weakly differentiable* functions.

3.1 Reminders of functional analysis

3.1.1 The Lebesgue space L^2

Definition 3.1.1 Let Ω be an open set of$\mathbb{R}^n$. The space $L^2(\Omega)$ is the space of measurable functions that are square integrable in Ω (modulo the "almost everywhere" equivalence).

Theorem 3.1.1 $L^2(\Omega)$ is a Hilbert space for the scalar product

$$\langle f, g \rangle := \int_\Omega fg.$$

The associated norm is denoted $\|.\|_{L^2}$.

Theorem 3.1.2 — Density of C_c^∞ in L^2. For all $f \in L^2(\Omega)$, there exists a sequence $f_n \in C_c^\infty(\Omega)$ such that

$$\|f_n - f\|_{L^2} \xrightarrow{n\to\infty} 0.$$

Theorem 3.1.3 — Fundamental lemma of the calculus of variations. Let $f \in L^2(\Omega)$ such that for all $\phi \in C_c^\infty(\Omega)$,

$$\int f\phi = 0.$$

Then $f = 0$ almost everywhere.

Proof. Take a sequence $f_n \in C_c^\infty(\Omega)$ such that $f_n \xrightarrow{n\to\infty} f$ in L^2. Then

$$0 = \int_\Omega f_n f \to \int_\Omega |f|^2 = \|f\|_{L^2}.$$

■

3.1.2 The notion of weak derivative

Definition 3.1.2 — Weak derivative. Let $v \in L^2(\Omega)$. We say that v is weakly differentiable if there exists $w_1, \dots, w_n \in L^2(\Omega)$ such that for all $\phi \in C_c^\infty(\Omega)$,

$$\int_\Omega v \partial_i \phi = -\int_\Omega w_i \phi.$$

The function w_i is the i-th weak derivative of v and is denoted by $\partial_i v$.

R This is a special case of the theory of distributions of Laurent Schwartz, but historically was introduced first by Sergueï Lvovitch Sobolev.

Exercise 3.1 Show that :

1. In the previous definition, the weak derivatives are unique.
2. For a bounded open set Ω and $v \in C^1(\Omega)$, the strong derivatives and the weak derivatives coincide almost everywhere (from this, we can say that the weak differentiability generalizes the strong one).
3. The Heaviside function $1_{(0,1)}$ is not weakly differentiable on $(-1,1)$.
4. A function which is in $C^0(\mathbb{R})$ and piecewise C^1 is weakly differentiable but not necessarily differentiable.

■

Weak differentiability shares some properties of common differentiability, for instance:

Proposition 3.1.4 Let $v \in L^2(\Omega)$ be weakly differentiable and such that $\partial_i v = 0$ for all i. Then v is constant on every connected component of Ω.

Proof. Let $Q = (-l,l)^n \subset \Omega$ and let $\theta \in C_c^\infty(-l,l)$ verify

$$\int_{-l}^{l} \theta(t)dt = 1.$$

For every $\phi \in C_c^\infty(Q)$, define

$$\psi(x', x_i) = \int_{-l}^{x_i} \left(\theta(t) \int_{-l}^{l} \phi(x', s)ds - \phi(x', t) \right) dt,$$

with the notation $x = (x', x_i)$ for $x = (x_1, \dots, x_i, \dots, x_n)$. Then $\psi \in C_c^\infty(Q)$ and

$$\partial_i \psi(x', x_i) = \theta(x_i) \int_{-l}^{l} \phi(x', s)ds - \phi(x', t).$$

Since $\partial_i v = 0$ then

$$\int_\Omega v \partial_i \psi = 0$$

which implies by Fubini

$$\int_Q v\phi = \int_Q v(x)\theta(x_i) \left(\int_{-l}^{l} \phi(x', s)ds \right) dx'dx_i = \int_Q \phi(x', s) \left(\int_{-l}^{l} v(x', x_i)\theta(x_i)dx_i \right) dx'ds.$$

As ϕ is arbitrary then by the Fundamental Lemma of the Calculus of Variations,

$$v(x) = \int_{-l}^{l} v(x',s)\theta(s)ds$$

almost everywhere on Q, which is constant w.r.t. x_i. By repeating the same argument on every x_j then v is constant on Q. Every pair of points in a connected component of Ω can be linked by a chain of such Q hence v is constant on every connected component of Ω. ■

Exercise 3.2 Let $v \in L^2(\Omega)$ be weakly differentiable and suppose that each $\partial_i v$ is also weakly differentiable. We denote $\partial^2_{ij} v \in L^2(\Omega)$ those second-order derivatives. Show that if for all i, j, $\partial^2_{ij} v = 0$ then v is a polynomial of degree at most 1. ■

3.2 The space H^1

Definition 3.2.1 Let Ω be an open set of $\mathbb{R}^n$. The Sobolev space $H^1(\Omega)$ is defined by

$$H^1(\Omega) := \left\{ v \in L^2(\Omega) : \forall i, \partial_i v \in L^2(\Omega) \right\},$$

i.e. each weak partial derivative exists and is in $L^2(\Omega)$.

Theorem 3.2.1 $H^1(\Omega)$ is a Hilbert space for the scalar product

$$\langle u, v \rangle := \int_\Omega \nabla u \cdot \nabla v + uv$$

where $\nabla u = (\partial_1 u, \dots, \partial_n u)^T$ is the *weak gradient* of u.

Proof. ■

Exercise 3.3 Let $n = 2$. Considering the function $u(x) = |\log(|x|)|^\alpha, 0 < \alpha < 1/2$ on B_1, show that H^1 functions are not necessarily continuous. Using this function, construct a function in $H^1(\Omega)$ on an open set Ω which is unbounded on all open subsets of Ω. ■

R We can actually prove that for $n = 1$, the functions in H^1 are indeed continuous, but not for any $n > 1$. This is a consequence of the Morrey inequalities (see [4], Chapter 5, Theorem 4), which among other things implies the Rademacher theorem stating that every Lipschitz function is differentiable a.e.

Exercise 3.4 Let $v \in H^1(0,1)$. Show that

$$w(x) := \int_0^x v'$$

is in H^1. Deduce that, almost everywhere,

$$v(y) = v(x) + \int_x^y v'$$

and that v is continuous (precisely: equal to a continuous function a.e.). ■

3.2.1 Density of smooth functions

This section aims to prove a result asserting that smooth functions are dense in H^1. We begin with a simple case.

Theorem 3.2.2 — Density in $\mathbb{R}^n$. $C_c^\infty(\mathbb{R}^n)$ is dense in $H^1(\mathbb{R}^n)$.

Proof. The proof is by regularization and truncation. Let $v \in H^1(\mathbb{R}^n)$. Consider $\rho \in C_c^\infty(B)$ such that $\rho \geq 0$ and $\int_B \rho = 1$, $\rho_k(x) := k^n \rho(kx)$ and define

$$v_k(x) := v * \rho_k(x) = \int_{\mathbb{R}^n} \rho_n(x-y)v(y)dy.$$

Using differentiation under the integral sign, we can prove that v_k is $C^\infty(\mathbb{R}^n)$ and that $\partial_i v_k = \rho_k * \partial_i v$ for all $i \in \{1,\dots,n\}$. Indeed, for all $i \in \{1,\dots,n\}$,

$$\begin{aligned}
\partial_i v_k(x) &= \partial_i \int_{\mathbb{R}^n} \rho_n(x-y)v(y)dy \\
&= \int_{\mathbb{R}^n} \partial_i \rho_n(x-y)v(y)dy \\
&= -\int_{\mathbb{R}^n} \rho_n(x-y)\partial_i v(y)dy \quad \text{by definition of the weak derivative} \\
&= \int_{\mathbb{R}^n} \rho_n(y)\partial_i v(x-y)dy \quad \text{by the change of variable } y \leftarrow x-y \\
&= \rho_k * \partial_i v.
\end{aligned}$$

Moreover, it is easy to verify that v_k and $\partial_i v_k$ are in L^2 for all i and it is a standard result that for $u \in L^2(\mathbb{R}^n)$,

$$u * \rho_k \xrightarrow[k\to\infty]{L^2} u.$$

Hence we deduce that

$$v_k \xrightarrow[k\to\infty]{H^1} v$$

is a sequence of functions in $C^\infty(\mathbb{R}^n)$ that goes to v.

To make the functions of the sequence compactly supported, consider a function $\chi \in C_c^\infty(\mathbb{R}^n)$ such that $0 \leq \chi \leq 1$ and $\chi(x) = 1$ for $|x| < 1$. The sequence $\tilde{v}_k(x) := v_k(x)\chi(x/n)$ verifies the theorem. ■

Exercise 3.5 — Alternative definition via the Fourier transform. Let $u \in L^2(\mathbb{R}^n)$. Using the previous theorem, show that $u \in H^1(\mathbb{R}^n)$ if and only if

$$x \mapsto (1+|x|^2)\hat{u}(x) \in L^2(\mathbb{R}^n),$$

where $\hat{u}$ is the Fourier transform of u. (You can show the result for $n = 1$, the idea is the same). ■

The following theorem is of great interest by itself.

Theorem 3.2.3 — Extension of H^1 functions. Let Ω be a C^1 bounded open set or $\Omega = \mathbb{R}^n_+$ where

$$\mathbb{R}^n_+ := \{(x_1,\dots,x_n) \in \mathbb{R}^n \text{ such that } x_n > 0\}.$$

There exists a continuous linear operator $P : H^1(\Omega) \to H^1(\mathbb{R}^n)$ such that for all $v \in H^1(\Omega)$,

$$Pv|_\Omega = v.$$

Proof. Sketch of proof First, let $\Omega = \mathbb{R}^n_+$. Denote $x = (x', x_n) \in \mathbb{R}^n$ and for $v \in H^1(\mathbb{R}^n_+)$, define

$$Pv(x) := \begin{cases} v(x', x_n) & \text{if } x_n > 0 \\ v(x', -x_n) & \text{if } x_n < 0 \end{cases}.$$

This defines a linear operator such that $Pv|_{\mathbb{R}^n_+} = v$. Let us show that it is bounded.

First, it is obvious that

$$\|Pv\|_{L^2(\mathbb{R}^n)} \leq \sqrt{2}\|v\|_{L^2(\mathbb{R}^n_+)}.$$

Now, for $1 \leq i < n$ we have

$$\partial_i(Pv)(x) := \begin{cases} \partial_i v(x', x_n) & \text{if } x_n > 0 \\ \partial_i v(x', -x_n) & \text{if } x_n < 0 \end{cases}$$

and

$$\partial_n(Pv)(x) := \begin{cases} \partial_n v(x', x_n) & \text{if } x_n > 0 \\ -\partial_n v(x', -x_n) & \text{if } x_n < 0 \end{cases}.$$

This implies that

$$\|\nabla Pv\|_{L^2(\mathbb{R}^n)} \leq \sqrt{2}\|\nabla v\|_{L^2(\mathbb{R}^n_+)}$$

which leads to

$$\|Pv\|_{H^1(\mathbb{R}^n)} \leq C\|v\|_{H^1(\mathbb{R}^n_+)}$$

and P is continuous.

For the case of Ω being a C^1 bounded open set, we use a partition of unity to go back to the previous case.

Since $\partial\Omega$ is compact, it can be covered by a finite number of sets $Q_1, \ldots, Q_N$ as defined in Definition 1.2.3, each Q_i being associated to a function ϕ_i. Not put $Q_0 = \Omega$ and let $\theta_0, \ldots, \theta_N$ be a partition of unity associated with $Q_0, \ldots, Q_N$.

Let $1 \leq i \leq N$ and for $x \in Q_i$, define $\psi_i(x', x_n) = (x', x_n - \psi_i(x'))$. Thanks to the inverse function theorem, we can make the hypothesis that ψ_i is a C^1-diffeomorphism from Q_i to $\psi_i(Q_i)$ (by taking Q_i smaller if needed, because $\det \psi_i = 1$). Moreover, for all $\psi_i(Q_i \cap \Omega) = \psi_i(Q_i) \cap \mathbb{R}^n_+$. Hence for $v \in H^1(\Omega)$, the function w_i defined by

$$w_i(x) = \begin{cases} (\theta_i v)\left(\psi_i^{-1}(x)\right) & \text{if } x \in \psi(Q_i) \cap \mathbb{R}^n_+ \\ 0 & \text{elsewhere in } \mathbb{R}^n_+. \end{cases}$$

is in $H^1(\mathbb{R}^n_+)$. Hence, we can use the prolongation operator defined on $\mathbb{R}^n_+$ to get a function $Pw_i \in H^1(\mathbb{R}^n)$.

Now set $P_0(\theta_0 v)$ to be the extension by 0 on $\mathbb{R}^n$ of $\theta_0 v$ and for all $1 \leq i \leq N$,

$$P_i(\theta_i v)(x) = \begin{cases} (Pw_i)(\psi_i(x)) & \text{for } x \in Q_i \\ 0 & \text{elsewhere on } \mathbb{R}^n. \end{cases}$$

One consequence of this definition is that for $x \in Q_i \cap \Omega$, $\psi_i(x) \in \mathbb{R}^n_+$ so

$$P_i(\theta_i v)(x) = (Pw_i)(\psi_i(x)) = w_i(\psi_i(x)) = \theta_i v(x).$$

Moreover, we can show thanks to the change of variable formula that there exists $C_i > 0$ such that

$$\|P_i(\theta_i v)\|_{H^1(\mathbb{R}^n)} \leq C_i \|v\|_{H^1(\Omega)}$$

Finally, we can show that the extension operator defined by

$$\tilde{P}v := \sum_{i=0}^{N} P_i(\theta_i v).$$

works. ■

Theorem 3.2.4 — Density. Let Ω be an open bounded set of class C^1, or $\Omega = \mathbb{R}^n$ or $\Omega = \mathbb{R}^n_+$. Then $C^\infty(\overline{\Omega})$ is dense in $H^1(\Omega)$.

Proof. Let $v \in H^1(\Omega)$. Thanks to the two previous theorems, there exists a sequence $v_n \in C_c^\infty(\mathbb{R}^n)$ such that

$$v_n \xrightarrow[n\to\infty]{H^1(\mathbb{R}^n)} Pv.$$

This implies

$$v_n|_\Omega \xrightarrow[n\to\infty]{H^1(\Omega)} v.$$

■

R This theorem is also true if Ω is a domain with Lipschitz boundary. This will be useful when we treat the case of polygonal domains in numerical experiments.

3.2.2 The space H_0^1

We are now interested in the the $H_0^1(\Omega)$. Roughly speaking, this set has to be understood as the set of H^1 functions that vanish at the boundary. Such functions being defined only up to a set of measure 0, we will rigorously define the trace of such function at the boundary.

Definition 3.2.2 The space $H_0^1(\Omega)$ is the closure of $C_c^\infty(\Omega)$ in $H^1(\Omega)$.

R Thanks to the previous density result, we have that $H_0^1(\mathbb{R}^n) = H^1(\mathbb{R}^n)$.

Proposition 3.2.5 $H_0^1(\Omega)$ is a Hilbert space for the same scalar product as $H^1(\Omega)$.

Proof. It is a closed subspace of a Hilbert space. ■

Theorem 3.2.6 — Poincaré Inequality. Let $\Omega \subset \mathbb{R}^n$ be an open set that is bounded in at least one direction. There exists $C > 0$ such that for all $v \in H_0^1(\Omega)$,

$$\int_\Omega |v|^2 \leq C \int_\Omega |\nabla v|^2.$$

Proof. We will use the Poincaré inequality for C^1 functions proved in the previous chapter, and conclude using a density argument.

Let $v \in H^1_0(\Omega)$. There exists a sequence $v_n \in C^\infty_c(\Omega)$ such that $\|v_n - v\|_{H^1} \xrightarrow[n\to\infty]{} 0$ and in particular this implies $\|v_n - v\|_{L^1} \xrightarrow[n\to\infty]{} 0$ and $\|\nabla v_n - \nabla v\|_{L^1} \xrightarrow[n\to\infty]{} 0$. By the Poincaré inequality in C^1, there exists a constant C such that

$$\int_\Omega |v_n|^2 \leq C \int_\Omega |\nabla v_n|^2.$$

By passing to the limit, we have the result. ■

Exercise 3.6 Using the Poincaré inequality, show that

$$v \mapsto \left(\int_\Omega |\nabla v|^2 \right)^{1/2}$$

is a norm over $H^1_0(\Omega)$ but not $H^1(\Omega)$. ■

3.2.3 Notion of trace

Since H^1 functions are not necessarily continuous, it seems that one can not impose boundary conditions as in (1.1). We will see that it is possible to define the value of an H^1 function at the boundary of a domain.

Theorem 3.2.7 — Trace. Let Ω be an open bounded C^1 set of $\mathbb{R}^n$ or $\Omega = \mathbb{R}^n_+$. Let

$$\begin{aligned} \gamma : H^1(\Omega) \cap C(\overline{\Omega}) &\to L^2(\partial\Omega) \cap C(\partial\Omega) \\ v &\mapsto v|_{\partial\Omega} \end{aligned}.$$

We can extend γ by continuity to a continuous linear map

$$\gamma : H^1(\Omega) \to L^2(\partial\Omega).$$

To prove this theorem in the case of a C^1 open bounded set, we will need this little lemma :

Proposition 3.2.8 Let Ω be an open bounded C^1 set of $\mathbb{R}^n$. There exists a vector field $V \in C^\infty_c(\mathbb{R}^n, \mathbb{R}^n)$ such that for all $x \in \partial\Omega$,

$$V(x) \cdot \mathbf{n}(x) \geq 1.$$

In other words, V is a smooth vector field pointing outward on $\partial\Omega$.

Proof. Let $x_0 \in \partial\Omega$. According to Definition 1.2.3, there exists a set $Q = \omega \times (a,b)$, a function $\phi \in C^1(\omega, (a,b))$ and a set of coordinates $(x_1, ..., x_n)$ such that $\partial\Omega$ is locally the graph of ϕ. Up to taking a smaller ω we can suppose that $\phi \in C^1(\bar{\omega}, (a,b))$. Hence $|\nabla\phi|$ is bounded on ω and we denote

$$L_Q := \sup_{x' \in \omega} |\nabla\phi(x')|.$$

Take the constant vector field defined on Q as $\tilde{V}(x) = (0, \dots, 0, -1)^T$. Then from the definiton of the normal, for $x = (x', \phi(x')) \in Q \cap \partial\Omega$:

$$\tilde{V}(x) \cdot \mathbf{n}(x) = \left(1 + |\nabla\phi(x')|^2\right)^{-1/2} \geq \left(1 + L_Q^2\right)^{-1/2}$$

so the vector field $V := \sqrt{1+L_Q^2}\tilde{V}$ is such that $V(x)\cdot n(x) \geq 1$ on $Q\cap\partial\Omega$.

Since $\partial\Omega$ is bounded, $\partial\Omega$ is compact. Hence $\partial\Omega$ can be covered by a finite amount of open sets Q_i as defined in Definition 1.2.3. On each Q_i, we can define a vector field V_i as previously such that for all i,

$$V_i \cdot \mathbf{n} \geq 1$$

on $Q_i \cap \partial\Omega$.

Take a partition of unity θ_i associated to Q_i. Hence for all i, $\theta_i V_i \in C_c^\infty(\mathbb{R}^n, \mathbb{R}^n)$. Defining

$$V := \sum_i \theta_i V_i,$$

we compute for $x \in \partial\Omega$:

$$V(x)\cdot\mathbf{n}(x) = \sum_i \theta_i(x) \underbrace{V_i(x)\cdot\mathbf{n}(x)}_{\geq 1} \geq \sum_i \theta_i = 1.$$

■

We can now proceed to give the proof of the theorem.

Proof of the Trace Theorem. Let $V \in C_c^\infty(\mathbb{R}^n, \mathbb{R}^n)$ as defined previously. We start by showing the estimate

$$\|\gamma v\|_{L^2(\partial\Omega)} \leq C\|v\|_{H^1(\Omega)}.$$

for $v \in C_c^\infty(\overline{\Omega})$. The general result follows by the density of $C_c^\infty(\overline{\Omega})$ in $H^1(\Omega)$.

For such v, we have using Stokes

$$\|\gamma u\|^2_{L^2(\partial\Omega)} = \int_{\partial\Omega} u^2 \leq \int_{\partial\Omega} u^2 (V\cdot\mathbf{n}) \leq \int_\Omega \operatorname{div}(u^2 V). \tag{3.1}$$

Denoting

$$\|V\|_{C^1} := \sup_{x\in\mathbb{R}^n} |V(x)| + \sum_{i=1}^n \sup_{x\in\mathbb{R}^n} |\partial_i V(x)|,$$

we show that

$$\begin{aligned}
\int_\Omega \operatorname{div}(u^2 V) &= \int_\Omega \nabla(u^2)\cdot V + (\operatorname{div} V)u^2 && (3.2)\\
&\leq \int_\Omega 2|u(\nabla u)\cdot V| + \int_\Omega |(\operatorname{div} V)|u^2 && (3.3)\\
&\leq \|V\|_{C^1}\int_\Omega \underbrace{2|u||\nabla u|}_{\leq |u|^2+|\nabla u|^2} + \|V\|_{C^1}\int_\Omega u^2 && (3.4)\\
&\leq (\|V\|_{C^1}+1)\|u\|_{L^2(\Omega)} + \|V\|_{C^1}\|\nabla u\|_{L^2(\Omega)} && (3.5)\\
&\leq C\|v\|_{H^1(\Omega)}. && (3.6)
\end{aligned}$$

■

R For obvious reasons, we will denote

$$\int_{\partial\Omega} \gamma v ds = \int_{\partial\Omega} v ds.$$

Definition 3.2.3 We will denote $H^{1/2}(\partial\Omega)$ the image of the trace operator, i.e.

$$H^{1/2}(\partial\Omega) := \gamma(H^1(\Omega))$$

Theorem 3.2.9 — Green's formula. Let Ω be a C^1 open bounded set. For all $u, v \in H^1(\Omega)$,

$$\int_\Omega u\partial_i v = \int_{\partial\Omega} uv\mathbf{n}_i - \int_\Omega v\partial_i u. \tag{3.7}$$

Proof. This formula has already been proven for functions in $C^1(\overline{\Omega})$. Using the density of those functions and the continuity of the trace operator, we have the desired result. ■

The following theorem states that the space H_0^1 is indeed the spaces of H^1 functions that "vanish at the boundary".

Theorem 3.2.10 — *. Let Ω be a C^1 open bounded set. Then

$$H_0^1(\Omega) = \left\{u \in H^1(\Omega) \text{ s.t. } \gamma u = 0\right\}$$

Proof. Let

$$V := \left\{u \in H^1(\Omega) \text{ s.t. } \gamma u = 0\right\}.$$

Since every function of $H_0^1(\Omega)$ is a limit of functions in $C_c^\infty(\Omega)$ and since the trace is continuous, we have $H_0^1(\Omega) \subset V$.

Now let $v \in V$. We would like to prove that v is the limit of a sequence of functions of $C_c^\infty(\Omega)$. By taking a partition of unity, we can reduce the proof to the case where $\Omega = \mathbb{R}^n_+$. The final argument is based on truncation and regularization. All the details can be found in [4], Chapter 5, Section 5, Theorem 2. ■

3.3 Application to elliptic problems

"Are we there yet?" - You

We have developed all the necessary materials to tackle elliptic PDEs. Indeed, H^1 is a Hilbert space, so applying the Lax-Milgram theorem should now be possible. Moreover, by having defined the trace of a function on $\partial\Omega$, the boundary value problems are well-posed. Let's go back to our toy problem (1.1).

3.3.1 Going back to the Poisson equation

Let rewrite the problem we want to solve :

$$\begin{cases} -\Delta u = f \text{ in } \Omega \\ \quad u = 0 \text{ on } \partial\Omega \end{cases}. \tag{3.8}$$

Proceeding as in Chapter 1 (this time considering the space $H_0^1(\Omega)$ instead of the space of C^1 functions vanishing at the boundary), we can show that

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv$$

for all $v \in H_0^1(\Omega)$.

We have the following theorem of existence and uniqueness :

Theorem 3.3.1 Let Ω be an open bounded set and $f \in L^2(\Omega)$. There exists a unique $u \in H_0^1(\Omega)$ such that for all $v \in H_0^1(\Omega)$,

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv. \tag{3.9}$$

Proof. The proof is the same as in Chapter 1: for $u, v \in H_0^1(\Omega)$, take

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \quad \text{and} \quad L(v) = \int_\Omega fv.$$

Both a and L satisfy the conditions of the Lax-Milgram theorem, hence the result. ■

R In the theorem, there is no need for Ω to be regular! It will be necessary however to show that if the solution is regular enough, it is a strong solution.

Proposition 3.3.2 If Ω is C^1 and $u \in C^2(\overline{\Omega})$ is a solution to (3.9) then

$$\begin{cases} -\Delta u = f & \text{a.e. in } \Omega \\ u = 0 & \sigma - \text{a.e. on } \partial\Omega \end{cases}. \tag{3.10}$$

Proof. For all $v \in C_c^\infty(\Omega)$, we have that

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv.$$

Integrating by parts, it leads to

$$\int_\Omega (\Delta u + f)v = 0$$

for all $v \in C_c^\infty(\Omega)$ hence according to the Fundamental Lemma of Calculus of Variations,

$$\Delta u + f = 0$$

a.e. in Ω.

Moreover, u is continuous hence its trace $\gamma u = u|_{\partial\Omega} = 0$ σ-a.e. ■

We can actually prove a stronger result, namely that u belongs to a certain space $H^2(\Omega)$ which allows to define Δu. This requires so-called *regularity results*, that are extremely useful but outside the scope of this course.

Exercise 3.7 Let Ω be an open bounded set. Let $V \in C^1(\Omega, \mathbb{R}^n)$ be a vector field such that $\operatorname{div} V = 0$ in Ω. Put the problem

$$\begin{cases} -\Delta u + V \cdot \nabla u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

in a variational form. Show that it has a solution in the weak sense. ■

We now treat the case of non-homogeneous boundary conditions.

Theorem 3.3.3 Let Ω be a C^1 open bounded set, $f \in L^2(\Omega)$ and $u_0 \in H^{1/2}(\partial\Omega)$. Show that the

problem

$$\begin{cases} -\Delta u = f \text{ in } \Omega \\ u = u_0 \text{ on } \partial\Omega \end{cases} \tag{3.11}$$

has a unique solution $u \in H^1(\Omega)$ in the weak sense.

Proof. Suppose that u_0 can be extended in the interior of Ω by a smooth C^2 function. By putting $\tilde{u} = u - u_0$, we have $(\tilde{u} - u_0)|_{\partial\Omega} = 0$ and formally, if u is solution :

$$-\Delta\tilde{u} = -\Delta u + \Delta u_0 = f + \Delta u_0.$$

Once again multiplying by $v \in C_c^\infty(\Omega)$ and integrating by parts, we have

$$\int_\Omega \nabla\tilde{u} \cdot \nabla v = \int_\Omega fv - \nabla u_0 \cdot \nabla v.$$

Now consider $u_0 \in H^{1/2}(\partial\Omega) = \gamma(H^1(\Omega))$: it is the trace of a function in $H^1(\Omega)$ (which we will still denote u_0). Moreover, for $f \in L^2(\Omega)$ and $\tilde{u}, v \in H_0^1(\Omega)$, the previous expression still make sense. Hence, we take it as the variational formulation of (3.11).

Let us show that this VF has a solution using Lax-Milgram. By putting

$$a(\tilde{u}, v) := \int_\Omega \nabla\tilde{u} \cdot \nabla v,$$

we know that a is a continuous, coercive bilinear form on $H_0^1(\Omega)$. Let

$$L(v) = \int_\Omega fv - \nabla u_0 \cdot \nabla v.$$

Using Cauchy-Schwarz, we show that

$$|L(v)| \leq \max(\|f\|_{L^2}, \|u_0\|_{H^1})\|v\|_{H^1}$$

hence it is a continuous linear form on $H_0^1(\Omega)$. Using Lax-Milgram, we show that the VF has a solution. ■

3.3.2 Poisson equation with Neumann boundary conditions

The Dirichlet boundary conditions are not the only ones we can consider. Another very useful one is the so-called Neumann BC. It physically emerges as a condition on a certain "flux" through the surface $\partial\Omega$. For instance, the problem

$$\begin{cases} -\Delta u + u = f & \text{in } \Omega \\ \partial_n u = 0 & \text{on } \partial\Omega \end{cases} \tag{3.12}$$

can model the heat distribution u on a domain Ω with a source f, such that the domain Ω is perfectly insulated: the condition $\partial_n u$ on $\partial\Omega$ states that the heat flux is zero at the boundary of Ω.

In greater generality, we can consider the problem

$$\begin{cases} -\Delta u + u = f & \text{a.e. in } \Omega \\ \partial_n u = g & \text{on } \partial\Omega \end{cases} \tag{3.13}$$

for a certain function g. Now, let us find the variational formulation of (3.13).

In the sequel, we suppose that all the functions are sufficiently regular for the computations to make sense. Multiplying (3.13) by $v \in C^\infty(\overline{\Omega})$, we have

$$\begin{aligned}\int_\Omega fv &= \int_\Omega (-\Delta u + u)v \\ &= \int_\Omega \nabla u \cdot \nabla v + uv - \int_{\partial\Omega} (\partial_n u)v \\ &= \int_\Omega \nabla u \cdot \nabla v + uv - \int_{\partial\Omega} gv\end{aligned}$$

We see that the integrals make sens if Ω is C^1, $u, v \in H^1(\Omega)$, $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$. Hence, we can prove the following:

Theorem 3.3.4 Let Ω be a C^1 open bounded set, $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$. There exists a unique $u \in H^1(\Omega)$ such that ,

$$\int_\Omega \nabla u \cdot \nabla v + uv = \int_\Omega fv + \int_{\partial\Omega} gv \quad \text{for all } v \in H^1(\Omega). \tag{3.14}$$

Proof. Let

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v + uv \quad \text{and} \quad L(v) = \int_\Omega fv + \int_{\partial\Omega} gv.$$

Since a is the standard inner product of $H^1(\Omega)$, it is a bilinear continuous and coercive form on $H^1(\Omega)$.

L is a linear form on $H^1(\Omega)$ Let us show that it is bounded. For $v \in H^1(\Omega)$,

$$\begin{aligned}|L(v)| &\le \int_\Omega |fv| + \int_{\partial\Omega} |gv| \\ &\le \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}\|v\|_{L^2(\partial\Omega)} && \text{using Cauchy-Schwarz} \\ &\le \|f\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)} + C\|g\|_{L^2(\partial\Omega)}\|v\|_{H^1(\Omega)} && \text{thanks to the trace theorem} \\ &\le C'\|v\|_{H^1(\Omega)}.\end{aligned}$$

Hence by the Lax-Milgram theorem, there exists a unique solution. ■

R Even the homogeneous Neumann boundary conditions are taken into account in the VF, where Dirichlet BC are forced into the Hilbert space. It is because the space

$$V = \{v \in H^1 : \partial_n v = 0 \text{ on } \partial\Omega\}$$

can not be defined (more precisely: $\partial_n v$ can not). We could be tempted to take

$$\{v \in H^2 : \partial_n v = 0 \text{ on } \partial\Omega\}$$

but then a would not be coercive.

As for the Dirichlet problem, it is possible to prove that almost everywhere, u is a solution of (3.13).

Theorem 3.3.5 Let Ω be an open bounded set, $f \in L^2(\Omega)$, and $g \in H^{1/2}(\Omega)$. Then the solution

u of (3.14) is in fact in $H^2(\Omega)$ and verifies

$$\begin{cases} -\Delta u + u = f & \text{a.e. in } \Omega \\ \partial_n u = g & \sigma - \text{a.e. on } \partial\Omega \end{cases} \tag{3.15}$$

Proof. We won't prove that here. We need regularity results to show that u actually belongs in H^2. Once we have that, the first line of (3.15) follows from the Green formula. For the second line, we need to show that $H^{1/2}(\partial\Omega)$ is dense in $L^2(\partial\Omega)$. ■

3.3.3 Other common elliptic problems

Variable coefficients

There are a lot of other elliptic problems other than the one of Poisson. A first generalization could be

$$\begin{cases} -\text{div}(A\nabla u) = \rho f \text{ in } \Omega \\ u = 0 \text{ on } \partial\Omega \end{cases}. \tag{3.16}$$

where $A : \Omega \to \mathbb{R}^{n\times n}$ and $\rho : \Omega \to \mathbb{R}$. This can for instance model the heat in an inhomogeneous, anisotropic media.

Exercise 3.8 Let Ω be a bounded open set. Suppose that A is measurable and there exists $\alpha, \beta > 0$ such that for all $x \in \Omega$ and all $\zeta \in \mathbb{R}^n$,

$$A(x)\zeta \cdot \zeta \geq \alpha|\zeta|^2 \quad \text{and} \quad |A(x)\zeta| \leq \beta|\zeta|.$$

Moreover, let $\rho \in L^\infty(\Omega)$. Show that there exists a unique $u \in H_0^1(\Omega)$ such that

$$\int_\Omega A\nabla u \cdot \nabla v = \int_\Omega \rho f v \quad \text{for all } v \in H_0^1(\Omega).$$

■

Stokes problem

Other problems require to consider vector solutions. The Stokes problem, briefly mentionned in the introduction, is one of them. Let Ω be a body of fluid and $u : \Omega \to \mathbb{R}^n$ be the velocity field of this fluid and $p : \Omega \to \mathbb{R}$ which are the two unknowns. Given a force field $f : \Omega \to \mathbb{R}^n$, we want u and p to solve

$$\begin{cases} \nabla p - \Delta u = f \text{ in } \Omega \\ \text{div} u = 0 \text{ in } \Omega \\ u = 0 \text{ on } \partial\Omega \end{cases}, \tag{3.17}$$

The natural Hilbert space for this problem is

$$V := \left\{ v \in H_0^1(\Omega)^n \text{ s.t. } \text{div} v = 0 a.e. \right\}.$$

By taking the dot product of (3.17) with a function $v \in V$ and integrating, we get that the corresponding VF is: find $u \in V$ such that

$$\int_\Omega \nabla u : \nabla v = \int_\Omega f \cdot v \quad \text{for all } v \in V,$$

where $A : B = Tr(AB^T)$ is the usual inner product between matrices. We can show using the Lax-Milgram theorem that, once again, this problem has a solution. But there is something strange: **the pressure term has disappeared**!

In fact, it is "hidden" in the divergence-free condition $\text{div} v = 0$. There is a theorem, called the DeRham theorem, that can "bring it back" (see [1], Section 5.3.2). It can also be seen as the Lagrange multiplier associated with this divergence-free condition (see this link).

Linearized elasticity

Finally, we can derive the weak fomulation of the linearized elasticity problem. We suppose that the elastic body is fixed on a certain portion $\partial\Omega_D$ and subject to some surface force (like a pressure) $g : \partial\Omega_N \to \mathbb{R}^n$ on $\partial\Omega_N := \partial\Omega \setminus \partial\Omega_D$. Moreover, we suppose that there are body forces $f : \Omega \to \mathbb{R}^n$ (think about gravity). Then u is a solution of the linearized elasticity problem if it solves For a domain Ω modeling an elastic domain we look for a function $u : \Omega \to \mathbb{R}^n$ which solves

$$\begin{cases} -\text{div}(Ae(u)) = f \text{ in } \Omega \\ u = 0 \text{ on } \partial\Omega_D \\ Ae(u)\mathbf{n} = g \text{ on } \partial\Omega_N \end{cases} \tag{3.18}$$

where $g : \partial\Omega_N \to \mathbb{R}^n$ on $\partial\Omega_N := \partial\Omega \setminus \partial\Omega_D$, $f : \Omega \to \mathbb{R}^n$ and

$$Ae(u) := 2\mu e(u) + \lambda Tr(e(u))I.$$

with λ and μ the Lamé coefficients of the material.

Using the Hilbert space

$$V := \left\{ v \in H^1(\Omega)^n \text{ such that } v = 0 \text{ on } \partial\Omega_D \right\}$$

we can show that the VF corresponding to (3.18) is: find $u \in V$ such that

$$\int_\Omega Ae(u) : e(v) = \int_\Omega f \cdot v + \int_{\partial\Omega_N} g \cdot v \quad \text{for all} v \in V.$$

To prove that this variational formulation has a solution, we need to use the *Korn's inequality* to prove the H^1 continuity of the bilinear form. Roughly speaking, Korns Inequality allows to control the gradient of a vector field using only the symmetrized gradient (which is not trivial). See [1], Section 5.3.1 for more information.

3.4 Sobolev spaces H^m

We can build more regular Hilbert spaces in the same fashion as H^1. First, let us call a *multi-index* a vector $\alpha = (\alpha_1, ..., \alpha_n)$ such that $\alpha_i \in \mathbb{N}$ for all i. We denote $|\alpha| = \alpha_1 + \cdots + \alpha_n$. We denote

$$\partial^\alpha = \partial_1^{\alpha_1} \dots \partial_n^{\alpha_n}.$$

Definition 3.4.1 For an open set Ω in $\mathbb{R}^n$, we denote

$$H^m(\Omega) := \left\{ v \in L^2(\Omega) \text{ s.t. } \partial^\alpha v \in L^2(\Omega) \text{ for all } |\alpha| \leq m \right\}.$$

This space is a Hilbert space for the inner product

$$\langle u, v \rangle_{H^m} := \sum_{|\alpha| \leq m} \int_\Omega \partial^\alpha u \partial^\alpha v.$$

R For $m < m'$, we have $H^{m'}(\Omega) \subset H^m(\Omega)$.

Theorem 3.4.1 — Density. Let Ω be a bounded open set of class C^1. Then $C^\infty(\overline{\Omega})$ is dense in $H^m(\Omega)$

Proof. For $\Omega = \mathbb{R}^n$, the proof is similar to H^1. For an open bounded set of class C^1, see [4], Chapter 5, Section 5.3.3, Theorem 3 (maybe give an idea of the proof for $\mathbb{R}^n_+$ on a drawing). ■

Theorem 3.4.2 — Trace. Let Ω be a C^1 open bounded set. Let

$$\gamma_1 : H^2(\Omega) \cap C^1(\Omega) \to L^2(\partial\Omega) \cap C(\partial\Omega)$$
$$v \mapsto \partial_n v$$

. This mapping can be extended by continuity as a mapping

$$\gamma_1 : H^2(\Omega) \to L^2(\partial\Omega).$$

Idea of proof. It is a consequence of the previous trace theorem applied to ∇v. ■

Theorem 3.4.3 — Green's Formula. Let Ω be a C^2 open bounded set. If $u \in H^2(\Omega)$ and $v \in H^1(\Omega)$ then

$$\int_\Omega (\Delta u)v = \int_{\partial\Omega} (\partial_n u)v - \int_\Omega \nabla u \cdot \nabla v. \tag{3.19}$$

R In the previous formula, $\int_{\partial\Omega}(\partial_n u)v$ is an abuse of notation for $\int_{\partial\Omega} \gamma_1(u)\gamma(v)$.

Idea of proof. We know that the formula is true for smooth functions; we finish by density and continuity of γ_1 ■

Theorem 3.4.4 — Regularity *. Let Ω be a Lipschitz bounded open set and $m > n/2$. Then $H^m(\Omega) \subset C(\overline{\Omega})$ and the canonical injection is continuous, meaning there exists $C > 0$ such that for all $u \in H^m(\Omega)$,

$$\|u\|_{C(\overline{\Omega})} \leq C\|u\|_{H^m(\Omega)}.$$

Proof. See [4]. This is a consequence of the more general "Sobolev inequalities". These equalities also leads to the famouse "Rademarcher theorem".

MIGHT BE INTERESTING TO GIVE A PROOF FOR $\Omega = \mathbb{R}^n$. Then can prove the case of a Lipschitz domain by extension theorem. ■

3.5 Other useful results

So far, we have seen pretty much all the results that will be useful to present the Finite Element Method for elliptic PDEs. However, there are some results of first importance when it comes to the theoretical study of PDEs. The most important one is maybe the Rellich Theorem, a wonderfully powerful compactness property.

Theorem 3.5.1 — Rellich *. Let Ω be a Lipschitz open bounded set. Then for every bounded sequence in $H^{m+1}(\Omega)$ there exists a subsequence which converges in $H^m(\Omega)$. We say that $H^{m+1}(\Omega)$ is compactly embedded in H^m.

Proof. The proof is based on the Arzela-Ascoli theorem. All the details can be found in [4], Chapter 5, Section 7, Theorem 1. ■

Exercise 3.9 — Poincaré-Wirtinger inequality. Let Ω be a C^1 bounded open connected set. Let us denote $\bar{u} := \frac{1}{|\Omega|}\int_\Omega u$. Using the Rellich theorem, show that there exist $C > 0$ such that for all $u \in H^1(\Omega)$,

$$\|u - \bar{u}\|_{L^2} \leq C\|\nabla u\|_{L^2}.$$

■

Theorem 3.5.2 — Maximum principle. Let Ω be an open bounded set and let u be the weak solution of (3.8). Suppose $f \geq 0$ a.e. in Ω. Then $u \geq 0$ a.e. in Ω.

Proof. See [1], Section 5.5.4, Theorem 5.2.22. Uses the not trivial fact that if $v \in H_0^1$ then $v^+ := max(0, v) \in H_0^1$ and a.e. in Ω,

$$\nabla v^+ = 1_{\{v>0\}}\nabla v.$$

■

This theorem has a very intuitive interpretation: if we apply a positive pressure beneath a membrane Ω, then the displacement is everywhere toward the top. It can for instance be used to prove the uniqueness of equations like (3.8).

Exercise 3.10 Suppose that there exists a weak solution to (3.8). Using the Maximum Principle, show that this solution is unique. ■

Theorem 3.5.3 — Spectral theorem*. Let Ω be a C^1 open bounded set. There exists a sequence

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \to \infty$$

and a basis $(u_k)_{k\geq 1}$ of $H_0^1(\Omega)$ such that in the weak sense,

$$\begin{cases} -\Delta u_k = \lambda_k u_k & \text{in } \Omega \\ u_k = 0 & \text{on } \partial\Omega \end{cases} \tag{3.20}$$

. The numbers λ_k are called the eigenvalues of the eigenfunctions u_k.

Proof. See [1], Chapter 7, Theorem 7.3.5. Basically, it is a consequence of the Rellich theorem and the Spectral theorem for compact self-adjoint operators. ■

R The spectral theorem is insanely powerful. It is at the basis of Quantum Mechanics. It is the foundation of Spectral Geometry, which studies the links between the geometry of Ω and the eigenvalues. Practically, the knowledge of all eigen elements of an operator allows to fully solve time-dependent problems. For instance, suppose that we search $u : [0,T] \times \Omega \to \mathbb{R}, (t,x) \mapsto u(t,x)$, the solution of the time-dependent heat equation :

$$\begin{cases} \partial_t u = \Delta u & \text{in } [0,T] \times \Omega \\ u = 0 & \text{on } [0,T] \times \partial\Omega \\ u(0,.) = v_0 & \text{in } \Omega \end{cases} . \tag{3.21}$$

If $v_0 = u_k$ then

$$u(t,x) := e^{-\lambda_k t} u_k(x)$$

is solution. Hence, let us take a general $v_0 \in H_0^1(\Omega)$. Since (u_k) is a Hilbert basis of H_0^1, we can write $v_0 = \sum_k v^k u_k$. By linearity of the heat equation, the function

$$u(t,x) := \sum_k v^k e^{-\lambda_k t} u_k(x)$$

is solution.

II Numerical methods for PDEs: old and new

4. Solving PDEs with neural networks

4.1 Physics Informed Neural Networks

4.1.1 General principles

Let say that we want to approximate some function $u : \Omega \subset \mathbb{R}^n \to \mathbb{R}$. The usual approach using neural networks is to generate data $(x_i, u(x_i))_i$ and regress a neural network u_θ on the data by minimizing

$$\mathscr{L}(\theta) = \frac{1}{N}\sum_i |u_\theta(x_i) - u(x_i)|^2.$$

Now, suppose that we know that the function u is the solution to some PDE, for instance the Poisson equation

$$\begin{cases} -\Delta u = f \text{ in } \Omega \\ u = 0 \text{ on } \partial\Omega \end{cases} \tag{4.1}$$

In this case, we can incorporate this knowledge into the loss in the following way :

$$\mathscr{L}(\theta) = \underbrace{\frac{1}{N}\sum_i |u_\theta(x_i) - u(x_i)|^2}_{\text{data-fitting loss}} + \underbrace{\|\Delta u_\theta + f\|^2_{L^2(\Omega)}}_{\text{PDE loss}} + \underbrace{\|u_\theta\|^2_{L^2(\partial\Omega)}}_{\text{boundary loss}}.$$

The idea of incorporating the PDE into the loss dates back to 1997, with the seminal work of Lagaris and co-authors [12]. While relatively unnoticed for a while (see Figure 4.1.1), the method became widely popular after the publication of two papers in August and November 2017. The first one, by Sirignano and Spiliopoulos [17] presented this method under the name of Deep Galerking Method. The second work [16] by Raissi and co-authors [16] coined the term Physics Informed Neural Networks (PINNs), which is more commonly used. The idea is indeed pretty natural: given that u_θ has a C^2 activation function, then it is C^2 itself and the laplacian Δu_θ is well defined. Moreover, the derivatives of u_θ can be automatically computed in all usual neural network library.

However, if knowing that u_θ verifies the Poisson equation makes the data information $(x_i, u(x_i))_i$ unnecessary; hence, we can try to solve the PDE by only using the loss

$$\|\Delta u_\theta + f\|^2_{L^2(\Omega)} + \|u_\theta\|^2_{L^2(\partial\Omega)}.$$

Figure 4.1: Amount of citations per year of [12]

In this paradigm, u_θ becomes a parametrization of the solution and the backpropagation can be seen as a numerical PDE solver !

In greater generality, one can put a PDE in abstract form

$$\begin{cases} \mathscr{N}(u) = 0 & \text{in } \Omega \\ \mathscr{B}(u) = 0 & \text{on } \partial\Omega \end{cases} \tag{4.2}$$

and set the loss to be

$$\|\mathscr{N}(u_\theta)|^2_{L^2(\Omega)} + \|\mathscr{B}(u_\theta)\|^2_{L^2(\partial\Omega)}.$$

In general, this loss is not computable as is. However, it can be approximated by Monte-Carlo type methods. The simplest way consists in drawing uniform iid samples $(x_i)_{1\leq i\leq N_\Omega} \subset \Omega$ and $(\tilde{x}_i)_{1\leq i\leq N_{\partial\Omega}} \subset \partial\Omega$ and approximate the previous loss (up to multiplicative factors) by

$$\mathscr{L}(\theta) = \frac{1}{N_\Omega}\sum_{i=1}^{N_\Omega} |\mathscr{N}(u_\theta)(x_i)|^2 + \frac{1}{N_{\partial\Omega}}\sum_{i=1}^{N_{\partial\Omega}} |\mathscr{B}(u_\theta)(\tilde{x}_i)|^2.$$

While this approach may seem sound at first, it comes with some serious problems:

1. In practice, $\mathscr{L}(\theta) << 1$ does **not** imply that u_θ is close to the solution of , even for simple PDEs.
2. Moreover, there is no guarantee that the optimizer finds the global minimizer (and, in general, does not).

In other words, we have to this day no proof of converge of the method, even if we let N_Ω, $N_{\partial\Omega}$ and the size of the network go to infinity. [COUNTER EXAMPLE FOR THE FIRST ITEM IN CONTINUOUS L2 norm]

4.1.2 Imposition of boundary conditions

Consider the following Poisson PDE with non-homogeneous Dirichlet boundary condition:

$$\begin{cases} -\Delta u = f \text{ in } \Omega \\ u = g \text{ on } \partial\Omega \end{cases} \tag{4.3}$$

In the context of PINNs, there exists two main way of imposing the Dirichlet boundray conditon.

Penalty method

The penalty method is the one we already used the define the loss $\mathscr{L}$. In general, we will tune the penalization of the boundary condition thanks to a parameter $\beta > 0$ and put

$$\mathscr{L}(\theta) = \frac{1}{N_\Omega}\sum_{i=1}^{N_\Omega} |\Delta u_\theta(x_i) - f(x_i)|^2 + \beta\frac{1}{N_{\partial\Omega}}\sum_{i=1}^{N_{\partial\Omega}} |u_\theta(\tilde{x}_i) - g(\tilde{x}_i)|^2.$$

where larger β will enforce the boundary condition more strongly.

Exercise 4.1 Convergence of Robin to Dirichlet ? Maybe in the part of deep ritz ■

Exact imposition

Suppose that you already know some functions $G, \phi : \mathbb{R}^n \to \mathbb{R}$ smooth enough such that $G|_{\partial\Omega} = g$ and $\phi(x) = 0 \iff x \in \partial\Omega$. We can put

$$u_\theta := G + \phi v_\theta$$

where v_θ is a neural network. In this case, u_θ automatically satisfies the boundary condition, and the loss can be taken as

$$\mathscr{L}(\theta) = \frac{1}{N_\Omega} \sum_{i=1}^{N_\Omega} |\Delta u_\theta(x_i) - f(x_i)|^2.$$

■ **Example 4.1** There is different way of choosing ϕ:

- If Ω is the unit ball of $\mathbb{R}^n$, we can set $\phi(x) = |x|^2 - 1$;
- If $\Omega = [0,1]^n$, we can take $\phi(x) = \Pi_{i=1}^n \sin(2\pi x_i)$;
- If Ω is a more complex domain and we can sample from the boundary and the domain, we can pre-train a neural network ϕ_θ to fulfill the condition $\phi_\theta(x) = 0 \iff x \in \partial\Omega$ (for instance, by fitting ϕ_θ to the *signed distance function* of Ω).

In general, the important property is that ϕ must be as smooth as required by the PDE (same for G).

■

4.1.3 A note on the choice of the architecture

In the context of PINNs, one must take some special care about the choice of the architecture (and especially the activation), due to the differential operator acting on u_θ. Indeed, consider the simple one-dimensional problem

$$\begin{cases} -u'' = 1 \text{ in } (0,1) \\ u(0) = u(1) = 0 \end{cases}$$

and suppose that we want to solve it using the loss

$$\mathscr{L}(\theta) = \frac{1}{N_\Omega} \sum_{i=1}^{N_\Omega} |u_\theta''(x_i) - 1|^2 + \beta \left(|u_\theta(0)|^2 - |u_\theta(1)|^2 \right)$$

with a 1-layer ReLU network. In this case, $u_\theta'' = 0$ almost everywhere hence the loss reduces to

$$\mathscr{L}(\theta) = 1 + \beta \left(|u_\theta(0)|^2 + |u_\theta(1)|^2 \right)$$

which can never converge to 0.

To avoid this kind of issue, the PINN community often uses tanh, sigmoid or ReLU^2 (or higher power) activation functions.

4.1.4 Notebook

Have a look on the following notebook (click here). Try to understand how it works, and follow the directions given in the end of the file.

4.2 Alternative approaches

4.2.1 Energy methods

As we have seen, a lot of PDEs admits a formulation in terms of the minimization of an energy. For this type of PDE, a natural idea is to try to solve them by directly using the energy functional as a loss function. In the literature, this approach takes different names, like Deep Ritz Method (DRM) [9] or Deep Energy Method [14].

■ **Example 4.2 — Poisson problem.** As we know, $u \in H_0^1(\Omega)$ is a solution to the homogeneous Poisson equation 4.1 if and only if it is minimizes

$$J(v) := \int_\Omega \frac{1}{2}|\nabla v|^2 - fv$$

for $v \in H_0^1(\Omega)$. Suppose that we chose a neural network u_θ such that the boundary conditions are exactly imposed, i.e. $u_\theta = 0$ on $\partial\Omega$, we can use the loss

$$\mathscr{L}(\theta) = \frac{1}{N_\Omega}\sum_{i=1}^{N_\Omega} \frac{1}{2}|\nabla u_\theta(x_i)|^2 - f(x_i)u_\theta(x_i)$$

■

R As previously discussed, there exist no proof of convergence for PINN-like techniques. However, one can upper bound the error made by the approximation error in a way that "decouples" the different sources of errors. Namely, for the Poisson equation solved by the DRM, if u denotes the solution to the PDE and $\theta^\star := \arg\min_\theta \mathscr{L}(\theta)$, one can show [8] that

$$\|u_\theta - u\|^2_{H^1(\Omega)} \leq \underbrace{C \inf_{\tilde{\theta}} \|u_{\tilde{\theta}} - u\|^2_{H^1(\Omega)}}_{\mathscr{E}_{app}} + \underbrace{2 \sup_{\tilde{\theta}} |\mathscr{L}(\tilde{\theta}) - J(u_{\tilde{\theta}})|^2}_{\mathscr{E}_{stat}} + \underbrace{|\mathscr{L}(\theta) - \mathscr{L}(\theta^\star)|^2}_{\mathscr{E}_{opt}}. \quad (4.4)$$

Then:

- $\mathscr{E}_{app}$ measures the *approximation error*, which depends on the expressivity of u_θ. Thanks to the UAT, this error can be driven to 0;
- $\mathscr{E}_{stat}$ measures the *statistical error*, i.e. the error that arises from the Monte Carlo approximation of the integrals. This error can also be estimated using statistical tools like the *Rademacher complexity*;
- $\mathscr{E}_{opt}$ is the *optimization error*. It measures the error made by the optimizer, and can not in general be put to 0. This is the critical source of error and, in the current state of the research, can not be avoided.

Exercise 4.2 Prove the inequality 4.4. ■

■ **Example 4.3 — Eigenvalue problem.** Suppose that we want to find the smallest eigenvalue $\lambda \in \mathbb{R}$ of the eigenvalue problem

$$\begin{cases} -\Delta u = \lambda u \text{ in } \Omega \\ u = 0 \quad \text{on } \partial\Omega \end{cases}$$

along with its associated eigenfunction $u \in H_0^1(\Omega)$. The Courant-Fisher theorem gives the following variational formula for the computation of the eigenvalue:

$$\lambda = \min_{v \in H_0^1(\Omega)\setminus\{0\}} \frac{\int_\Omega |\nabla v|^2}{\int_\Omega v^2}$$

and stipulates that any minimizer is an eigenfunction. Similarly to the Poisson problem, we can take the loss

$$\mathscr{L}(\theta) = \frac{\sum_{i=1}^{N_\Omega} |\nabla u_\theta(x_i)|^2}{\sum_{i=1}^{N_\Omega} u_\theta(x_i)^2} + \left(\frac{1}{N_\Omega}\sum_{i=1}^{N_\Omega} u_\theta(x_i)^2 - 1\right)^2$$

(the rightmost term is to ensure that u_θ does not go to 0 during the optimization process). ■

One can remark that this approach requires less regularity than the PINN one, which might be a possible advantage.

4.2.2 Variational PINNs

Introduced in [11], the principle of vPINNs is to use the weak formulation of the PDE instead of the strong form. Once again in the context of 4.1, the weak formulation is :

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v \quad \text{for all } v \in H_0^1(\Omega). \tag{4.5}$$

Let $(\phi_n)_{n\in\mathbb{N}}$ be a Hilbert basis of $H_0^1(\Omega)$. Then 4.5 is equivalent to

$$\int_\Omega \nabla u \cdot \nabla \phi_n = \int_\Omega f \phi_n \quad \text{for all } n \in \mathbb{N}. \tag{4.6}$$

The idea of vPINNs is, similarly to PINNs, to minimize the residual of the (weak form) of the PDE

$$\mathscr{L}(\theta) := \sum_{n=1}^{\infty} \left| \int_\Omega \nabla u_\theta \cdot \nabla \phi_n - f\phi_n \right|.$$

In order to compute it, one must approximate the integrals using a Monte Carlo method and truncate the sum to some order N. A usable loss is

$$\mathscr{L}(\theta) := \sum_{n=1}^{N} \left| \frac{1}{N_\Omega} \nabla u_\theta(x_i) \cdot \nabla \phi_n(x_i) - f(x_i)\phi_n(x_i) \right|.$$

As the DRM, the vPINN approach requires less regularity that the vanilla PINN approach. However, it also comes with some difficulties. For instance, the N must be taken big enough for accuracy purposes, but it quickly makes the computations very expensive. Another, perhaps more important aspect is that we need a Hilbert basis of $H_0^1(\Omega)$, which in the case of a general Ω is unknown.

4.2.3 Weak Adversarial Networks

For a fixed $u \in H_0^1(\Omega)$, consider the operator

$$\begin{aligned} \mathscr{A}[u] : & H_0^1(\Omega) \to \mathbb{R} \\ & v \mapsto \langle \mathscr{A}[u], v \rangle := \int_\Omega \nabla u \cdot \nabla v - fv \end{aligned}$$

and define its operator norm

$$\|\mathscr{A}[u]\|_{op} := \max_{v \in H_0^1(\Omega)\setminus\{0\}} \frac{|\langle \mathscr{A}[u], v\rangle|}{\|v\|_{H^1}}$$

Then u is a solution to the Poisson problem $\iff \mathscr{A}[u] = 0 \iff \|\mathscr{A}[u]\|_{op} = 0 \iff u$ minimizes $\|\mathscr{A}[u]\|_{op}$. Hence, u is a solution of the PDE if and only if it is a solution to

$$\min_{u \in H_0^1(\Omega)} \max_{v \in H_0^1(\Omega)\setminus\{0\}} \frac{|\langle \mathscr{A}[u], v\rangle|}{\|v\|_{H^1}}.$$

The WAN approach, introduced in [18], aims at solving this saddle point problem by training two *adversarial* neural networks u_θ and v_η to find a saddle point of the loss

$$\mathscr{L}(\theta, \eta) := \frac{|\langle \mathscr{A}[u_\theta], v_\eta\rangle|}{\|v_\eta\|_{H^1}}$$

or, equivalently,

$$\mathscr{L}(\theta, \eta) := \log |\langle \mathscr{A}[u_\theta], v_\eta\rangle|^2 - \log \|v_\eta\|_{H^1}^2.$$

4.3 Operator learning

5. The Finite Element Method

We finally got to the core of this course. Everything that has been introduced before will allow us to build numerical approximations to elliptic PDEs.

This method is now an old method: the premisses were proposed by Richard Courant in the 1940s, and then developed by mechanical engineers in the 1950s-1960s. Then the applied mathematicians brought the theoretical justifications of the method, which is still one of the numerical methods the most used today for mechanical purposes.

The idea is to approximate the Hilbert space V of the solution of a PDE by a finite-dimensional space V_h, having in some sense that $V_h \xrightarrow[h\to 0]{} V$. The PDE, when expressed on V_h, is reduced to a linear system that can be solved using classical algorithms.

Once again, we follow [1], Chapter 6.

5.1 Variational approximation

Let V be a Hilbert space, a a continuous, coercive bilinear form on V and L a continuous linear form on V. We know that there exists $u \in V$ such that

$$a(u,v) = L(v) \quad \text{for all } v \in V. \tag{5.1}$$

Let V_h be a finite-dimensional subspace of V. The idea of the internal approximation is to replace (5.1) by: find $u_h \in V_h$ such that

$$a(u_h,v_h) = L(v_h) \quad \text{for all } v_h \in V_h. \tag{5.2}$$

Once again, the Lax-Milgram theorem shows the existence and uniqueness of such u_h.

Exercise 5.1 Show that (5.2) has a unique solution without using the Lax-Milgram theorem. ■

Solution. Let $N_h = \dim(V_h)$. Take a basis $(\phi_i)_{1\le i\le N_h}$ of V_h and set

$$K_h = (a(\phi_i,\phi_j))_{1\le i,j\le N_h},$$

which is called the *stiffness matrix*. Using the properties of a, we show that K_h is invertible. For details, see [1], Lemma 6.1.1. ■

We now want to estimate the error $\|u-u_h\|$ where u is the solution to (5.1) and u_h the one of (5.2).

Proposition 5.1.1 — Céa's Lemma. Under the previous conditions, we have

$$\|u-u_h\| \leq \frac{M}{\nu} \inf_{v_h \in V_h} \|u-v_h\|. \tag{5.3}$$

Proof. ■

R Fun fact: Céa was my great-great Ph.D. advisor :)

Theorem 5.1.2 Let $\mathscr{V}$ be a dense subspace of V. Suppose that for all $h>0$, there exist a function $r_h : \mathscr{V} \to V_h$ (called an *interpolation operator*) such that

$$r_h v \xrightarrow[h\to 0]{V} v \quad \text{for all } v \in \mathscr{V}. \tag{5.4}$$

Then

$$u_h \xrightarrow[h\to 0]{V} u.$$

Proof. See [1], Lemma 6.1.3. ■

R The parameter h has no particular meaning. Actually, it needs not to be continuous either: in the sequel, we will often take $h=1/k$ with $k \to \infty$. The previous proof is still valid in this case.

The previous theorem gives the path to follow to build a good approximation of a solution u. First, we have to provide such spaces V_h, then find a $\mathscr{V}$ and a projection operator r_h. Easier said than done.

■ **Example 5.1 — The Galerkin Method.** A simple example is the following. Let V be a separable Hilbert space. In this case, there exists a Hilbert basis $(e_i)_{i\geq 1}$ of V, i.e. $V = \overline{\text{Span}\{e_1,\dots,e_k,\dots\}}$. Hence, we can choose

$$\mathscr{V} := \text{Span}\{e_1,\dots,e_k,\dots\},$$

the set of all finite combinations of basis vectors, which is dense in V. Putting $h=1/k$, we can define

$$V_h := \text{Span}\{e_1,\dots,e_k\}$$

and

$$\begin{aligned} r_h : V &\to V_h \\ v &\mapsto \sum_{i=1}^{k} \langle v, e_i\rangle e_i \end{aligned},$$

the orthogonal projection on V_h. Hence (5.4) is verified (thanks to Parseval theorem) and $u_h \xrightarrow[h\to 0]{} u$. ■

In practice, the dimensions of the stiffness matrix K_h will be enormous: several hundred, or thousand dimensions. Hence, it is of interest if the matrix K_h is *sparse*, meaning it has only a few non-zero coefficients: for such a sparse matrix, we have more efficient storage and computation algorithms. However, the matrix K_h given by the Galerkin method will often be *full*. Hence, the previous method is often unusable in practice. This is why we need to develop better approximations of V, leading us toward the Finite Element Method.

5.2 FEM for $n = 1$

The idea of the Finite Element Method is to use a certain discretization of the domain Ω (a *mesh*) to build the spaces V_h that will approximate $H^1(\Omega)$. In this section, we will focus on the one-dimensional FEM. While this is pretty useless for real word applications, it will allow us to get a good grasp on how the FEM works without having to tackle the problems linked to the geometry of the domain for $n \geq 2$.

5.2.1 Lagrange Finite Elements

In the sequel, we will put $\Omega = (0,1)$. A "mesh" is a sequence

$$x_0 = 0 < x_1 < \cdots < x_k < x_{k+1} = 1.$$

It is said to be uniform when $x_j = jh = \frac{j}{k+1}$, $0 \leq j \leq k+1$. This is what we will suppose in the sequel.

We want to solve the problem

$$\begin{cases} -u'' = f & \text{in } (0,1) \\ u(0) = u(1) = 0 \end{cases} \tag{5.5}$$

Definition 5.2.1 We denote by $\mathbb{P}_N$ the set of real-valued polynomials of degree less or equal to N.

The $\mathbb{P}_1$ element method consist in approximating $V = H^1(0,1)$ by

$$V_h := \left\{ v \in C([0,1]) \text{ s.t. } v|_{[x_i, x_{i+1}]} \in \mathbb{P}_1 \text{ for all } 0 \leq j \leq k \right\}, \tag{5.6}$$

the set of continuous, piecewise affine functions.

Let

$$\phi(x) := \begin{cases} 1 - |x| & \text{if } |x| < 1 \\ 0 & \text{if } |x| \geq 1. \end{cases}$$

This is called the "hat" function for obvious reasons. Suppose the mesh is uniform and put

$$\phi_j(x) = \phi\left(\frac{x - x_j}{h}\right).$$

Using $\phi_i(x_j) = \delta_{ij}$, we show that functions in $\mathrm{Span}(\phi_j)_{0 \leq j \leq k+1}$ are uniquely determined by their values at the mesh points (x_j). Hence $(\phi_j)_{0 \leq j \leq k+1}$ is a basis of V_h and for all $v_h \in V_h$,

$$v_h(x) = \sum_{j=0}^{n+1} v_h(x_j)\phi_j(x).$$

Since we know that C^0, piecewise C^1 functions are in H^1, we know that $V_h \subset H^1(0,1)$. Hence V_h is a subspace of dimension $k+2$ of $H^1(0,1)$.

In the same fashion, we can show that

$$V_{h0} := \{v \in V_h \text{ s.t. } v(0) = v(1) = 0\}$$

is a subspace of $H_0^1(0,1)$ of dimension k.

We can see that for h smaller and smaller, the space V_h will "approximate" $H^1(0,1)$ more and more. We will use it for internal approximation.

5.2.2 Practical resolution of the Poisson PDE with Dirichlet BC

Suppose that we know that V_h is indeed an internal approximation of $H^1(0,1)$. The VF of (5.13) is the following :

$$\text{Find } u \in H_0^1(0,1) \text{ such that for all } v \in H_0^1(0,1), \quad \int_0^1 u'v' = \int_0^1 fv. \tag{5.7}$$

The formulation on V_{h0} is naturally

$$\text{Find } u_h \in V_{h0} \text{ such that for all } v_h \in V_{h0}, \quad \int_0^1 u_h'v_h' = \int_0^1 fv_h. \tag{5.8}$$

Taking $u_h = \sum_{j=1}^k u_j\phi_j$ and $v_h = \phi_i$, (5.8) becomes

$$\sum_{j=1}^k u_j \int_0^1 \phi_i'\phi_j' = \int_0^1 \phi_i.$$

Taking $U_h = (u_j)$,

$$K_h = \left(\int_0^1 \phi_i'\phi_j'\right)_{i,j} \text{ and } b_h = \left(\int_0^1 \phi_i\right)_i,$$

we see that solving (**??**) is equivalent to solving the linear system

$$K_h U_h = b_h,$$

where

$$K_h = h^{-1}\begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix}$$

Exercise 5.2 Show it. ∎

R Quadrature is needed for the RHS

5.2.3 Same with Neumann BC

See [1] p 157.

5.2.4 Convergence

Definition 5.2.2 The $\mathbb{P}_1$ interpolation operator $r_h : H^1(0,1) \to V_h$ is defined as

$$(r_h v) := \sum_{j=0}^{k+1} v(x_j)\phi_j.$$

Make a drawing.

Proposition 5.2.1 — Uniform continuity. There exists $C > 0$ such that for all h and all $v \in H^1(0,1)$,

$$\|r_h v\|_{H^1} \leq C\|v\|_{H^1}.$$

Proof. Using Exercise 3.4, we know that $v \in H^1(0,1)$ can be written

$$v(y) = v(x) + \int_x^y v'(t)dt.$$

Hence

$$|v(x)| \leq |v(y)| + \sqrt{|x-y|}\left(\int_x^y |v'|^2\right)^{1/2} \leq |v(y)| + \|v'\|_{L^2(0,1)}.$$

Integrating in y, we get

$$|v(x)| \leq \int_0^1 v(y)dy + \|v'\|_{L^2(0,1)} \leq \|v\|_{L^2(0,1)} + \|v'\|_{L^2(0,1)} \leq \sqrt{2}\|v\|_{H^1(0,1)}$$

using $\sqrt{a}+\sqrt{b} \leq \sqrt{2}\sqrt{a+b}$. Hence

$$\max_{x\in[0,1]} |v| \leq C\|v\|_{H^1}$$

from which we deduce

$$\|r_h v\|_{L^2} \leq \max_{x\in[0,1]} |r_h v| \leq \max_{x\in[0,1]} |v| \leq C\|v\|_{H^1}.$$

Now,

$$\int_{x_i}^{x_{i+1}} |(r_h v)'|^2 = \frac{(v(x_{i+1}) - v(x_i))^2}{h} = \frac{1}{h}\left(\int_{x_i}^{x_{i+1}} |v'|\right)^2 \leq \int_{x_i}^{x_{i+1}} |v'|^2$$

hence by summing over i

$$\|(r_h v)'\|_{L^2} \leq C\|v\|_{H^1}$$

which, combined with the previous inequality, gives the result. ∎

Theorem 5.2.2 — Interpolation. For all $v \in H^1(0,1)$,

$$r_h v \xrightarrow[h\to 0]{H^1} v.$$

Moreover, if $v \in H^2(0,1)$, there exist $C > 0$ such that

$$\|v - r_h v\|_{H^1} \leq Ch\|v''\|_{L^2}. \tag{5.9}$$

Corollary 5.2.3 — Convergence. Let $u \in H_0^1(0,1)$ be the solution of (**??**) and $u_h \in V_{h0}$ be the solution of (5.8). Then the $\mathbb{P}_1$ FEM converges:

$$u_h \xrightarrow[h\to 0]{H^1} u.$$

Moreover, if $u \in H^2(0,1)$, there exist $C > 0$ such that

$$\|u - u_h\|_{H^1} \leq Ch\|u''\|_{L^2}.$$

(Need to introduce H^2 before)

Proof of the Theorem. Here Cauchy and Schwarz will be your best friends.

First Step: We start by showing (5.9), since it will be necessary for the general case.

We first prove that for a smooth function $v \in H^2$,

$$\|v - r_h v\|_{L^2} \leq Ch\|v''\|_{L^2}.$$

Let first $v \in C^\infty([0,1])$. For $x \in [x_i, x_{i+1}]$, we have:

$$\begin{aligned} r_h v(x) - v(x) &= v(x_i) - v(x) + \frac{x - x_i}{x_{i+1} - x_i}(v(x_{i+1} - v(x_i))) \\ &= (x_i - x)v'(\xi_1) + (x - x_i)v'(\xi_2) \quad \text{for certain } \xi_1, \xi_2 \text{ using the mean value theorem} \\ &= (x - x_i)(v'(\xi_2) - v'(\xi_1)) \\ &= (x - x_i)\int_{\xi_1}^{xi_2} v''(t)dt. \end{aligned}$$

We deduce by CS that

$$(r_h v(x) - v(x))^2 \leq (x - x_i)^2 \left(\int_{x_i}^{x_{i+1}} |v''(t)|dt\right)^2 \leq h^2 \left(\int_{x_i}^{x_{i+1}} 1^2\right) \|v''\|^2_{L^2(x_i, x_{i+1})} \leq h^3 \|v''\|^2_{L^2(x_i, x_{i+1})}.$$

Integrating between x_i and x_{i+1} and summing on i, we get:

$$\|v - r_h v\|_{L^2} \leq Ch\|v''\|_{L^2}.$$

Since $C^\infty([0,1])$ is dense in $H^2(0,1)$ and everything is continuous w.r.t. the H^2 norm, we have the result for $v \in H^2$.

Now we need to prove

$$\|v' - (r_h v)'\|_{L^2} \leq Ch\|v''\|_{L^2},$$

which with the previous result would imply (5.9). Once again, take $v \in C^\infty([0,1])$. Write:

$$(r_h v)'(x) - v'(x) = \frac{v(x_{i+1}) - v(x_i)}{x_{i+1} - x_i} - v'(x) = \frac{1}{h}\int_{x_i}^{x_{i+1}} v'(t) - v'(x)dt = \frac{1}{h}\int_{x_i}^{x_{i+1}} \int_x^t v''(y)dy$$

hence by CS+ Fubini (multiple times)

$$\begin{aligned} |(r_h v)'(x) - v'(x)|^2 &\leq \frac{1}{h^2}\left(\int_{x_i}^{x_{i+1}} \int_x^t v''(y)dy\right)^2 \\ &\leq \frac{1}{h^2}\left(\int_{x_i}^{x_{i+1}} 1^2\right)\left(\int_{x_i}^{x_{i+1}} \left|\int_x^t v''(y)dy\right|^2\right) \\ &\leq \frac{1}{h}\int_{x_i}^{x_{i+1}} \left(\int_x^t |v''(y)|dy\right)^2 \\ &\leq \frac{1}{h}\int_{x_i}^{x_{i+1}} \left(\int_{x_i}^{x_{i+1}} |v''(y)|dy\right)^2 \\ &\leq \left(\int_{x_i}^{x_{i+1}} 1^2\right)\left(\int_{x_i}^{x_{i+1}} |v''(y)|^2 dy\right) \leq h\|v''\|^2_{L_2(x_i, x_{i+1})}. \end{aligned}$$

By integrating and summing on i we get

$$\|v' - (r_h v)'\|_{L^2} \leq Ch\|v''\|_{L^2}$$

and by the same density argument as before, this inequality is valid in H^2. Therefore,

$$\|v - r_h v\|_{H^1} \leq Ch\|v''\|_{L^2}.$$

Second Step: Now let's show that for all $v \in H^1(0,1)$,

$$r_h v \xrightarrow[h\to 0]{H^1} v.$$

Let $\varepsilon > 0$. By density, chose $\phi \in C^\infty([0,1])$ such that $\|v - \phi\|_{H^1} \leq \varepsilon$. Then

$$\begin{aligned}\|r_h v - v\|_{H^1} &\leq \|r_h v - r_h\phi\|_{H^1} + \|r_h\phi - \phi\|_{H^1} + \|\phi - v\|_{H^1} \\ &\leq C\|\phi - v\|_{H^1} + Ch\|\phi''\|_{H^1} + \|\phi - v\|_{H^1} \\ &\leq C'\varepsilon + Ch\|\phi''\|_{H^1}.\end{aligned}$$

Now, for all $h < \frac{\varepsilon}{C\|\phi''\|_{H^1}}$, we get

$$\|r_h v - v\|_{H^1} \leq (C' + 1)\varepsilon$$

hence the result. ■

Exercise 5.3 Derive the stiffness matrix of the problem

$$\begin{cases} -u'' + u = f & \text{in } (0,1) \\ u(0) = u(1) = 0 \end{cases} \tag{5.10}$$

■

5.2.5 A word on $\mathbb{P}_2$ Finite Elements

Another common FE space is the space of $\mathbb{P}_2$ functions. More regular between the nodes (but NOT at the nodes: we don't have C^1)

Definition 5.2.3

$$V_h := \left\{ u \in C^0([0,1]) \text{ s.t. } u|_{[x_j, x_{j+1}]} \in \mathbb{P}_2 \text{ for all } j \right\} \tag{5.11}$$

$$V_{h0} := \{ u \in V_h \text{ s.t. } u(0) = u(1) = 0 \} \tag{5.12}$$

To build a basis, we will need to add some intermediate points between the node x_j; namely, $x_{j+1/2} := x_j + \frac{h}{2}$. The set of $\{x_j, x_{j+1/2}\}_j$ are called the *degrees of freedom* of V_h. So as we have seens dofs does not necessarily corresponds to nodes of the mesh.

We can define two functions

$$\phi(x) := \begin{cases} (1+x)(1+2x) & \text{if } -1 \leq x \leq 0 \\ (1-x)(1-2x) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\psi(x) := \begin{cases} 1 - 4x^2 & \text{if } -1/2 \leq x \leq 1/2 \\ 0 & \text{otherwise} \end{cases}.$$

Then the functions

$$\phi_j(x) = \phi(\frac{x-x_j}{h}) \quad \text{and} \quad \phi_{j+1/2}(x) = \psi(\frac{x-x_{j+1/2}}{h})$$

forms a basis of V_h, in which a function $v_h \in V_h$ can be written

$$v_h(x) := \sum_{j=0}^{k+1} v_h(x_j)\phi_j(x) + \sum_{j=0}^{k+1} v_h(x_{j+1/2})\psi_{j+1/2}(x).$$

We can show that in this basis, the stiffness matrix has the form

$$K_h = h^{-1}\begin{pmatrix} 16/3 & -8/3 & 0 & & & & 0 \\ -8/3 & 14/3 & -8/3 & 1/3 & & & \\ 0 & -8/3 & 16/3 & -8/3 & 0 & & \\ & 1/3 & -8/3 & 14/3 & -8/3 & 1/3 & \\ 0 & & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

This is a pentadiagonal matrix, contrary to the $\mathbb{P}_1$ case where the matrix was tridiagonal. Also, for a given h, this matrix has order $2k$ entries, contrary to k for the $\mathbb{P}_1$ matrix. The number of operations needed for the resolution of a m-diagonal system of size k is of order mk so in the case of $\mathbb{P}_1$ it is of order $3k$ and of order $5 \times 2k = 10k$ in the case of $\mathbb{P}_2$. So using the $\mathbb{P}_2$ FEM is not interesting, unless we can show that the solution $u \in H^3$, since in this case we have quadratic convergence toward the optimum :

$$\|u - u_h\|_{H^1} \leq Ch^2\|u'''\|_{L^2}.$$

Which means that to be 100 times more precise with $\mathbb{P}_2$ we would only need 10 times more nodes (and hence matrix entries) whereas we would need 100 more for $\mathbb{P}_1$.

5.2.6 Hermite finite elements

Suppose that we want to solve the fourth-order *plate problem*

$$\begin{cases} u'''' = f \quad \text{in } (0,1) \\ u(0) = u(1) = 0 \\ u'(0) = u'(1) = 0 \end{cases} \tag{5.13}$$

We can show with LM that this problem has a unique weak solution in

$$H_0^2(0,1) := \left\{u \in H^2(0,1) \text{ s.t. } u(0) = u(1) = u'(0) = u'(1) = 0\right\}.$$

We directly see that the space V_h previously build does not work since V_h is not included in H^2, even for $\mathbb{P}_2$ elements. Moreover, we need a space allowing to take into account the vanishing of the derivatives at the boundary. We define the *Hermite finite elements* as

$$V_h := \left\{v \in C^1([0,1]) \text{ s.t. } v|_{[x_j,x_{j+1}]} \in \mathbb{P}_3 \text{ for all } j\right\}.$$

which is included in H^2. Using the reference functions

$$\phi(x) := \begin{cases} (1+x^2)(1-2x) & \text{if } -1 \leq x \leq 0 \\ (1-x^2)(1+2x) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\psi(x) := \begin{cases} x(1+x)^2 & \text{if } -1 \le x \le 0 \\ x(1-x)^2 & \text{if } 0 \le x \le -1 \\ 0 & \text{otherwise} \end{cases}.$$

Then the functions

$$\phi_j(x) = \phi(\frac{x - x_j}{h}) \quad \text{and} \quad \psi_j(x) = \psi(\frac{x - x_{j+1/2}}{h})$$

fors a basis of V_h and all $v_h \in V_h$ can be written as

$$v_h(x) := \sum_{j=0}^{k+1} v_h(x_j)\phi_j(x) + \sum_{j=0}^{k+1} v_h'(x_j)\psi_j(x).$$

R In this case there is 2 dofs by node (the value of the function and the value of it's derivative).

By defining V_{h0} accordingly, we show that $(\phi_1, ..., \phi_k, \psi_1, ..., \psi_k)$ is a basis. We can solve the plate problem by computing

$$K_h = \begin{pmatrix} K_{\phi\phi} & K_{\phi\psi} \\ K_{\psi\phi} & K_{\psi\psi} \end{pmatrix}$$

where

$$K_{\phi\phi} := \left(\int_0^1 \phi_i'' \phi_j''\right)_{1 \le i,j \le k} \quad , \quad K_{\phi\psi} := \left(\int_0^1 \phi_i'' \psi_j''\right)_{1 \le i,j \le k}$$

etc.

R There is a HUGE diversity of finite elements. You can find them at `https://defelement.com/`. This site allows you to see the defintion of each element, see the position of their dofs w.r.t. the mesh and see their basis functions.

5.3 FEM for $n = 2$

In this section we will study finite elements in dimension $n = 2$. This restriction simplifies the notations while keeping the main ideas intact.

In the sequel, Ω will be a polygonal domain. This assuption is necessary since only sch domains can be perfectly meshed by triangles. A triangle in $\mathbb{R}^2$ is defined as the convex hull of three points. The triangle is said to be degenerate if the three points are aligned.

5.3.1 Definitions and elementary properties

Definition 5.3.1 — Triangular mesh. Les Ω be an open connected polyhedron of $\mathbb{R}^2$. A *triangular mesh* of $\overline{\Omega}$ is a set $\mathcal{T}_h$ of nondegenerate triangles $(K_i)_{1 \le i \le k}$ satisfying for all $i \ne j$:

1. $K_i \subset \overline{\Omega}$ and $\overline{\Omega} = \cup_{i=1}^k K_i$
2. $K_i \cap K_j$ is either empty, or reduced to a common vertex, or reduced to a common edge.

The *vertices* or *nodes* of the mesh are the vertices of the triangles that composes it.

R In the sequel, the parameter h denotes the maximal diameter of elements of $\mathcal{T}_h$, i.e.

$$h = \max_i \operatorname{diam}(K_i).$$

R Meshes that satisfies these conditions are said to be *conforming*. [Draw meshes that are not conforming.]

Definition 5.3.2 — Barycentric coordinates. Let $a_1, a_2, a_3 \in \mathbb{R}^2$ be the edges of a non-degenerate triangle. The baricentric coordinates $\lambda_1, \lambda_2, \lambda_3$ of a point $x \in \mathbb{R}^2$ are defined by

$$\lambda_1 + \lambda_2 + \lambda_3 = 1 \quad \text{and} \quad \lambda_1 a_1 + ... + \lambda_3 a_3 = x.$$

R The numbers $\lambda_1, ..., \lambda_3$ are well defined. Indeed, the non-degeneracy of the triangle is equivalent to the fact that the matrix

$$\begin{pmatrix} a_1^1 & a_2^1 & a_3^1 \\ a_1^2 & a_2^2 & a_3^2 \\ 1 & 1 & 1 \end{pmatrix}$$

is invertible. Indeed,

$$\begin{vmatrix} a_1^1 & a_2^1 & a_3^1 \\ a_1^2 & a_2^2 & a_3^2 \\ 1 & 1 & 1 \end{vmatrix} = \begin{vmatrix} a_1^1 - a_3^1 & a_2^1 - a_3^1 & a_3^1 \\ a_1^2 - a_3^2 & a_2^2 - a_3^2 & a_3^2 \\ 0 & 0 & 1 \end{vmatrix} = \begin{vmatrix} a_1^1 - a_3^1 & a_2^1 - a_3^1 \\ a_1^2 - a_3^2 & a_2^2 - a_3^2 \end{vmatrix} = 2\text{Area}(\widehat{a_1 a_2 a_3})$$

In particular, we can see the λ_i as functions of x : $\lambda_i : x \mapsto \lambda_i(x)$.

Definition 5.3.3 — Lattice of order k. We for a triangle K, define its *lattice of order k*

$$\Sigma_k := \left\{ x \in K \text{ s.t. } \lambda_j(x) \in \left\{ 0, \frac{1}{k}, \dots, \frac{k-1}{k}, 1 \right\} \text{ for all } 1 \leq j \leq 3 \right\}.$$

[Make a drawing]

Definition 5.3.4

$$\mathbb{P}_p := \left\{ p(x) = \sum_{|\alpha| \leq k} p_\alpha x_1^{\alpha_1} x_2^{\alpha_2}, x = (x_1, x_2) \right\}$$

is the set of polynomials of degree less or equal to k in $\mathbb{R}^2$.

Proposition 5.3.1 Let K be a triangle and $k \geq 1$. Then every polynomial in $\mathbb{P}_k$ is uniquely determined by its values at Σ_k.

Proof for $\mathbb{P}_1$. For $p \in \mathbb{P}_1$, we have

$$p(x) = p_1 x_1 + p_2 x_2 + p_3$$

hence $\dim(\mathbb{P}_1) = 3$. On the other hand, $\Sigma_1 = \{a_1, a_2, a_3\}$, the vertices of the triangle K. We want to show that

$$\begin{aligned} \phi : \mathbb{P}_1 &\to \mathbb{R}^3 \\ p &\mapsto (p(a_1), p(a_2), p(a_3)) \end{aligned}$$

is bijective. Indeed, we see that

$$\phi(p) = \begin{pmatrix} p_1 & p_2 & p_3 \end{pmatrix} \begin{pmatrix} a_1^1 & a_2^1 & a_3^1 \\ a_1^2 & a_2^2 & a_3^2 \\ 1 & 1 & 1 \end{pmatrix}$$

and we know that the matrix is invertible, hence the result. ■

Exercise 5.4 Every polynomial in $\mathbb{P}_1$ has the form

$$p(x) = p(a_1)\lambda_1(x) + p(a_2)\lambda_2(x) + p(a_3)\lambda_3(x)$$

■

Proof. Correction Let $\lambda_i(x)$ the numbers defined by

$$\begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix} = \begin{pmatrix} a_1^1 & a_2^1 & a_3^1 \\ a_1^2 & a_2^2 & a_3^2 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1(x) \\ \lambda_2(x) \\ \lambda_3(x) \end{pmatrix}.$$

If $p(x) = p_1x_1 + p_2x_2 + p_3$ then

$$\begin{aligned} p(x) &= \begin{pmatrix} p_1 & p_2 & p_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix} = \begin{pmatrix} p_1 & p_2 & p_3 \end{pmatrix} \begin{pmatrix} a_1^1 & a_2^1 & a_3^1 \\ a_1^2 & a_2^2 & a_3^2 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1(x) \\ \lambda_2(x) \\ \lambda_3(x) \end{pmatrix} \\ &= \begin{pmatrix} p(a_1) & p(a_2) & p(a_3) \end{pmatrix} \begin{pmatrix} \lambda_1(x) \\ \lambda_2(x) \\ \lambda_3(x) \end{pmatrix} = p(a_1)\lambda_1(x) + \cdots + p(a_3)\lambda_3(x). \end{aligned}$$

■

Definition 5.3.5 — Lagrange FE. Let $\mathscr{T}_h$ be a mesh of an open connected polygonal set Ω. The $\mathbb{P}_k$ finite element space (or Lagrange FE space) associated to $\mathscr{T}_h$ is

$$V_h := \left\{ v \in C(\overline{\Omega}) \text{ s.t. } v|_K \in \mathbb{P}_k \text{ for all } K \in \mathscr{T}_h \right\}.$$

The *nodes* or *degrees of freedom (dofs)* of V_h is the set of points of the lattices of each $K \in \mathscr{T}_h$ (repeated points being counted once).

We also define

$$V_{h0} := \{ v \in V_h \text{ s.t. } v|_{\partial\Omega} = 0 \}$$

R For $\mathbb{P}_1$, the nodes coincides with the vertices of the mesh.

Proposition 5.3.2 V_h is a finite-dimensional subspace of $H^1(\Omega)$ with dimension equal to the number of dofs $(\hat{a}_i)_{1 \le i \le n_d}$. Moreover, there exists $(\phi_i)_{1 \le i \le n_d}$, a basis of V_h defined by

$$\phi_i(\hat{a}_j) = \delta_{ij}$$

such that

$$v(x) = \sum_{i=1}^{n_d} v(\hat{a}_i)\phi_i(x)$$

for all $v \in V_h$.

Proof. Boooooring ■

5.3.2 Practical implementation

Suppose that we want to practically implement the FEM for $\mathbb{P}_1$ for the problem

$$\begin{cases} -\Delta u + u = f & \text{in } (0,1) \\ u(0) = u(1) = 0 \end{cases}. \tag{5.14}$$

Given a mesh $\mathscr{T}_h$, there is a basis $(\phi_i)_{1\leq i\leq n_d}$ of the space V_{h0} of continuous, piecewise affine functions which vanishes at $\partial\Omega$. We first focus on the stiffness matrix, which is of the form

$$K_h = \left(\int_\Omega \nabla\phi_i \cdot \nabla\phi_j + \int_\Omega \phi_i\phi_j\right)_{1\leq i,j\leq n_d}.$$

For i, j, we can decompose the contribution of ϕ_i, ϕ_j on each element of $\mathscr{T}_h$:

$$\int_\Omega \nabla\phi_i \cdot \nabla\phi_j = \sum_{K\in\mathscr{T}_h} \int_K \nabla\phi_i \cdot \nabla\phi_j \quad \text{and} \quad \int_\Omega \phi_i\phi_j = \sum_{K\in\mathscr{T}_h} \int_K \phi_i\phi_j$$

Hence, if we know how to compute the contribution on each K, we know how to assemble the matrix. To do so, we send each K to a "reference element" $\hat{K}$, which is the triangle of edges $\hat{a}_1 = (0,0), \hat{a}_2 = (0,1), \hat{a}_3 = (1,0)$. On this triangle, the basis functions are given by

$$\hat{\phi}_1(x) = 1 - x_1 - x_2 \quad \hat{\phi}_2(x) = x_1 \quad \hat{\phi}_3(x) = x_2.$$

Exercise 5.5 For all $1 \leq i, j \leq 3$, compute

$$\int_{\hat{K}} \hat{\phi}_i\hat{\phi}_j \quad \text{and} \quad \int_{\hat{K}} \nabla\hat{\phi}_i \cdot \nabla\hat{\phi}_j$$

■

Correction. ■

Now let $K \in \mathscr{T}_h$ be a triangle of edges a_1, a_2, a_3. Let

$$\begin{aligned} \Phi : \hat{K} &\to K \\ x &\mapsto a_1 + Ax \end{aligned}$$

where

$$A = \begin{pmatrix} a_2 - a_1 & a_3 - a_1 \end{pmatrix}$$

By a change of variables, we can express every function h on K as a function on $\hat{K}$:

$$\int_K h = \int_{\Phi(\hat{K})} h = \int_{\hat{K}} h\circ\Phi |\det D\Phi| = 2|K| \int_{\hat{K}} h\circ\Phi$$

since $\det D\Phi = \det A = \det\begin{pmatrix} a_2 - a_1 & a_3 - a_1 \end{pmatrix} = 2|K|$. Suppose that the basis functions are numbered on this triangle such that $\phi_i(a_j) = \delta_{ij}$, $1 \leq i, j \leq 3$ [MAKE A DRAWING]. Then we have $\hat{\phi}_i = \phi_i \circ \Phi$ (since $\phi_i \circ \Phi$ is affine and $\phi_i \circ \Phi(\hat{a}_j) = \delta_{ij}$). This means that

$$\int_K \phi_i\phi_j = 2|K| \int_{\hat{K}} (\phi_i\circ\Phi)(\phi_j\circ\Phi) = 2|K| \int_{\hat{K}} \hat{\phi}_i\hat{\phi}_j.$$

Since we already computed the last quantity, the assembly of this ter is almost "free". The term with the gradients is a bit more complex. In the same setting, using that

$$\nabla\hat{\phi}_i = \nabla(\phi_i\circ\Phi) = (D\Phi)^T(\nabla\phi_i\circ\Phi) = A^T\nabla\phi_i\circ\Phi$$

we can show that

$$\int_K \nabla\phi_i \cdot \nabla\phi_j = 2|K| \int_{\hat{K}} (\nabla\phi_i \circ \Phi) \cdot (\nabla\phi_j \circ \Phi) = 2|K| \int_{\hat{K}} A^{-1}A^{-T}\nabla\hat{\phi}_i \cdot \nabla\hat{\phi}_j = |K|A^{-T}A^{-1}\nabla\hat{\phi}_i \cdot \nabla\hat{\phi}_j$$

since $\nabla\hat{\phi}_i$ is constant on $\hat{K}$ and $|\hat{K}| = 1/2$.

R The matrix K_h is **sparse** since $\int_\Omega \nabla\phi_i \cdot \nabla\phi_j + \int_\Omega \phi_i\phi_j = 0$ as soon as the dofs i and j does not lie on the same triangle.

Exercise 5.6 Stiffness matrix of $-\Delta u = f$ + Neumann b.c. on a simple mesh (exo 6.3.9 p179 in [1]). ■

Knowing how to implement the stiffness matrix is not enough to numerically solve the system. We also need to compute the RHS

$$b_h = \left(\int_\Omega f\phi_i \right)_{1 \le i \le n_d}.$$

Unless f has a really particular form (for instance $f \in V_h$), these integrals can not be computed explicitely. Hence we need to rely on *quadrature formulas* to approximate them on each element $K \in \mathcal{T}_h$, for instance

$$\int_K h \approx \frac{|K|}{3}(h(a_1) + \cdots + h(a_3)).$$

We can then replace the RHS b_h by

$$\tilde{b}_h = \left(\sum_{K \in \mathcal{T}_h} \frac{|K|}{3} (f(a_1)\phi_i(a_1) + \cdots + f(a_3)\phi_i(a_3)) \right)_{1 \le i \le n_d}$$

R We can show that modifying the RHS in this way still allows to prove convergence if f is smooth enough (see the section hereafter).

Exercise 5.7 Show that te previous formula is exact for a $\mathbb{P}_1$ function $h \in V_h$. ■

5.3.3 Convergence with exact RHS

Now we will adress the proof of convergence of the $\mathbb{P}_1$ FEM in 2D for the now usual problem

$$\begin{cases} -\Delta u = f \text{ in } \Omega \\ u = 0 \text{ on } \partial\Omega \end{cases} \tag{5.15}$$

where Ω is a polygonal mesh and $f \in L^2(\Omega)$. We still follow Allaire.

Definition 5.3.6 — Diameter, inner ball. Let K be a triangle. Then we define

$$\text{diam}(K) := \max_{x,y \in K} |x - y| \quad \text{and} \quad \rho(K) := \max_{B_r \subset K} (2r).$$

Definition 5.3.7 — Regular meshes. Let $(\mathcal{T}_h)_h$ be a sequence of meshes of Ω. We say that it is a sequence of regular meshes if

$$h := \max_{K \in \mathcal{T}_h} \text{diam}(K) \xrightarrow[h \to 0]{} 0$$

and there exists C such that for all h and $K \in \mathcal{T}_h$,

$$\frac{\text{diam}(K)}{\rho(K)} \leq C.$$

R Condition 1 makes the meshes be thinner and thinner. Condition 2 prevents the triangles from flattening too much.

As usual, we consider

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \quad \text{and } L(v) = \int_\Omega fv.$$

Theorem 5.3.3 Let $(\mathcal{T}_h)_h$ be a sequence of regular meshes of Ω and u be the solution of (5.15). Let V_{h0} be the $\mathbb{P}_k$ FE space on $\mathcal{T}_h$ and u_h be the solution of

$$\text{Find } u_h \in V_{h0} \text{ s.t. } a(u_h, v_h) = L(v_h) \quad \text{for all } v_h \in V_{h0}.$$

Then the $\mathbb{P}_k$ FEM converges, i.e.

$$u_h \xrightarrow[h\to 0]{H^1} u.$$

Moreover, if $u \in H^{k+1}(\Omega)$ with $k \geq 1$ then

$$\|u - u_h\|_{H^1} \leq Ch^k \|u\|_{H^{k+1}}$$

where C depends neither on h nor on u.

To prove this theorem in the case of $\mathbb{P}_1$ (the case of a general $\mathbb{P}_k$ is actually exactly the same), we will need a few results. First, let $n_h = \dim(V_{h0})$, $(\hat{a}_i)_{1\leq i\leq n_h}$ be the dofs of V_{h0} and $(\phi_i)_i$ the associated basis functions. For $v \in C(\overline{\Omega})$ we define the interpolation operator

$$r_h v := \sum_{i=1}^{n_h} v(\hat{a}_i)\phi_i.$$

Theorem 3.4.4 tells us that this operator r_h is a continous linear operator from H^2 to V_{h0}. First, it is actually useful to restrict our analysis of the interpolation operator on each element of the mesh and then "glue" them together. Hence for $K \in \mathcal{T}_h$, by denoting its edges $\hat{a}_1, \hat{a}_2, \hat{a}_3$ and the associated $\mathbb{P}_1$ basis functions ϕ_1, ϕ_2, ϕ_3, we can define

$$r_K v = \sum_{i=1}^{3} v(\hat{a}_i)\phi_i$$

for all $v \in C(\overline{\Omega})$. It is simply the restriction of r_h on r_K.

Theorem 5.3.4 — Bramble-Hilbert. The operator

$$r_K : H^2(\Omega) \to H^2(\Omega)$$

is linear and continous. Moreover, there exists a constant $C > 0$ (depending on K) such that

$$\|v - r_K v\|_{H^2} \leq C|v|_{H^2} \tag{5.16}$$

where

$$|v|_{H^2}^2 = \sum_{1\leq i,j\leq 2} \int_K |\partial_{ij}^2 v|^2$$

Proof. It is clear that r_K is linear. To prove that it is continuous, take $v \in H^2(\Omega)$. Then

$$\|r_K v\|_{H^2(K)} \leq \sum_i |v(\hat{a}_i)| \|\phi_i\|_{H^2} \leq \|r_K v\|_{C(\bar{K})} \sum_i \|\phi_i\|_{H^2} \leq C \left(\sum_i \|\phi_i\|_{H^2} \right) \|v\|_{H^2(K)}$$

where the last inequality follows from the regularity theorem 3.4.4 stating that the inclusion $H^2(K) \subset C(\bar{K})$ is continuous. Now, let us prove that there exists $C > 0$ such that

$$\|v\|_{H^2} \leq C(|v|_{H^2} + \|r_K v\|_{H^2}).$$

For the sake of contradiction, suppose that it is not the case; hence, for all $k \in \mathbb{N}$, there exist $v_k \in H^2$ s.t.

$$1 = \|v_k\|_{H^2} > n(|v_k|_{H^2} + \|r_K v_k\|_{H^2}). \tag{5.17}$$

The sequence v_k being bounded in H^2, the Rellich theorem 3.5.1 asserts that there exists $v \in H^1$ such that (up to relabeling the indices)

$$v_k \xrightarrow[k\to\infty]{H^1} v.$$

But (5.17) implies that $|v_k|_{H^2} \to 0$ so the previous convergence is actually in H^2 so in particular $\|v\|_{H^2} = 1$. Passing to the limit in (5.17), we have that

$$|v|_{H^2} = 0 \quad \text{and} \quad \|r_K v\|_{H^2} = 0.$$

Using Proposition 3.1.4 (or the exercice just after), we know that $|v|_{H^2} = 0$ implies that $v \in \mathbb{P}_1$. But then $v = r_K v = 0$, which contradicts $\|v\|_{H^2} = 1$.

To finish, apply (5.17) to $v - r_K v$ and note that $r_K(v - r_K v) = 0$ and $|v - r_K v|_{H^2} = |v|_{H^2}$ (indeed $r_K v \in \mathbb{P}_1$ so all the derivatives of order 2 vanishes). ■

Corollary 5.3.5 Suppose that $\operatorname{diam}(K) \leq 1$. There exists C independent on K such that for all $v \in H^2(K)$,

$$\|v - r_K v\|_{H^1} \leq C \frac{\operatorname{diam}(K)^2}{\rho(K)} |v|_{H^2}.$$

Proof. The idea is to go back to the reference element $\hat{K}$. Remembering the function

$$\begin{aligned} \Phi : \hat{K} &\to K \\ x &\mapsto a_1 + Ax \end{aligned}$$

we can put $\hat{v} = v \circ \Phi$ and use the same computations as in Subsection 5.3.2 to get that

$$\|\hat{v}\|_{L^2(\hat{K})}^2 = \int_{\hat{K}} \hat{v}^2 = |\det A|^{-1} \int_{\hat{K}} v^2 \circ \Phi |\underbrace{\det D\Phi}_{=\det A}| = |\det A|^{-1} \int_K v^2 = |\det A|^{-1} \|v\|_{L^2(K)}^2.$$

Similarly, we compute

$$\begin{aligned}
|\hat{v}|^2_{H^1(\hat{K})} &= \int_{\hat{K}} |\nabla(v\circ\Phi)|^2 = \int_{\hat{K}} (AA^T\nabla v\cdot\nabla v)\circ\Phi \\
&= |\det A|^{-1}\int_{\hat{K}} (AA^T\nabla v\cdot\nabla v)\circ\Phi|\det D\Phi| \\
&= |\det A|^{-1}\int_{K} (AA^T\nabla v\cdot\nabla v) \\
&\leq |\det A|^{-1}\int_{K} \|A\|\|A^T\||\nabla v||\nabla v| \quad \text{using the matrix norm induced by the euclidean norm} \\
&\leq |\det A|^{-1}\|A\|^2|v|^2_{H^1(K)}.
\end{aligned}$$

In the same fashion, we can finally get the same kind of estimate for the H^2 seminorm. Considering Φ^{-1} we also obtain reversed inequalities. To sum up, we have for $l\in\{0,1,2\}$

$$\begin{aligned}
|\hat{v}|_{H^l(\hat{K})} &\leq C\|A\|^l|\det A|^{-1/2}|v|_{H^l(K)} \\
|v|_{H^l(K)} &\leq C\|A^{-1}\|^l|\det A|^{1/2}|\hat{v}|_{H^l(\hat{K})}.
\end{aligned}$$

with $C>0$ independent of K. Using equation (5.16), we find that for (another) C idependant of K,

$$\|v-r_Kv\|_{L^2(K)} \leq |\det A|^{1/2}\|\hat{v}-r_{\hat{K}}\hat{v}\|_{L^2(\hat{K})} \leq C|\det A|^{1/2}|\hat{v}|_{H^2(\hat{K})} \leq C\|A\|^2|v|_{H^2(K)}.$$

Similarly,

$$|v-r_Kv|_{H^1(K)} \leq C\|A\|^2\|A^{-1}\||v|_{H^2(K)}.$$

Finally, we can show that

$$\|A\| \leq \frac{\operatorname{diam}(K)}{\rho(\hat{K})} \quad \text{and} \quad \|A^{-1}\| \leq \frac{\operatorname{diam}(\hat{K})}{\rho(K)}.$$

This leads to

$$\|v-r_Kv\|_{L^2(K)} \leq C\left(\frac{\operatorname{diam}(K)}{\rho(\hat{K})}\right)^2|v|_{H^2(K)} \leq C\frac{\operatorname{diam}(K)^2}{\rho(K)}|v|_{H^2(K)}$$

using that $\rho(K)<1$ and putting $\rho(\hat{K})^2$ in C, and

$$|v-r_Kv|_{H^1(K)} \leq C\frac{\operatorname{diam}(K)^2}{\rho(K)}|v|_{H^2(K)}$$

leading to the result. ■

We can now proceed to prove the theorem

Proof of Theorem 5.3.3. Let $v\in H^2(\Omega)$. We have

$$\begin{aligned}
\|v-r_hv\|^2_{H^1(\Omega)} &= \sum_{K\in\mathscr{T}_h}\|v-r_Kv\|^2_{H^1(K)} \\
&\leq C\sum_{K\in\mathscr{T}_h}\left(\frac{\operatorname{diam}(K)^2}{\rho(K)}\right)^2|v|^2_{H^2(K)} && \text{using the previous Corollary} \\
&\leq C\sum_{K\in\mathscr{T}_h}\operatorname{diam}(K)^2|v|^2_{H^2(K)} && \text{since } \frac{\operatorname{diam}(K)}{\rho(K)}\leq C \text{ (regular meshes)} \\
&\leq Ch^2\|v\|^2_{H^2(\Omega)}
\end{aligned}$$

hence

$$\|v - r_h v\|_{H^1(\Omega)} \leq Ch\|v\|_{H^2(\Omega)}.$$

We will now apply the approximation theorem 5.1.2. Let $\mathscr{V} = C_c^\infty(\Omega)$ which is dense in $V = H_0^1(\Omega)$. In particular, $C_c^\infty(\Omega) \subset H^2(\Omega)$ so the previous estimation tells us that

$$r_h v \xrightarrow[h\to 0]{H^1} v$$

for all $v \in C_c^\infty(\Omega)$ hence by Theorem 5.1.2, we have

$$u_h \xrightarrow[h\to 0]{H^1} 0.$$

Now suppose that $u \in H^2(\Omega)$. Using Céa's Lemma (5.3) we know that

$$\|u - u_h\|_{H^1} \leq \inf_{v_h \in V_{h0}} \|v - v_h\|_{H^1} \leq C\|u - r_h u\|_{H^1} \leq Ch\|u\|_{H^2(\Omega)}$$

■

5.3.4 Convergence with quadrature

Follow [5].

In the previous convergence result, we assume that we were able to explicitely computate the RHS $\int_\Omega f v_h$. This is however rarely the case, because of the function f which may not have a nice analytical form. In this section we describe how, under some hypothesis on the regularity of u, we can show that the convergence still holds when approximating the RHS by a quadrature formula. For $r \in \mathbb{N}$ and $\psi \in C^r(\overline{\Omega})$, let us define

$$|\psi|_{C^r(\overline{\Omega})} = \sum_{x \in \overline{\Omega}} \max_{|\alpha| = r} |\partial^\alpha \psi(x)|$$

and

$$\|\psi\|_{C^r(\overline{\Omega})} = \sum_{i=0}^{r} |\psi|_{C^i(\overline{\Omega})}.$$

For an element $K \in \mathscr{T}_h$ of vertices $a_1, ..., a_3$, we will approximate $\phi \in C^r(\overline{\Omega})$ by

$$\int_K \phi \approx \frac{|K|}{3}(\phi(a_1) + \phi(a_2) + \phi(a_3)).$$

In this perspective, for $v_h \in V_{h0}$, we can can replace the RHS $L(v_h)$ by

$$\tilde{L}(v_h) = \sum_{K \in \mathscr{T}_h} \frac{|K|}{3}(f(a_1)v_h(a_1) + \cdots + f(a_3)v_h(a_3))$$

and denote $\tilde{u}_h$ the solution of the problem

$$\text{Find } \tilde{u}_h \in V_{h0} \text{ s.t. } a(\tilde{u}_h, v_h) = \tilde{L}(v_h) \quad \text{for all } v_h \in V_{h0}.$$

We prove the following theorem :

Theorem 5.3.6 Suppose that the solution u of (5.15) is in $C^3(\overline{\Omega})$. Then

$$\|u - \tilde{u}_h\|_{H^1} \leq Ch|u|_{C^3}.$$

This means that even if we approximate $\int fv$ by quadrature formulas, we still converge to the solution. To prove it, we will need some intermediary results.

Proposition 5.3.7 Let $K \in \mathscr{T}_h$ with vertices (a_1, a_2, a_3) and $\psi \in C^1(K)$. There exists $C > 0$ (independent of everything) such that

$$\left| \int_K \psi dx - \frac{|K|}{3} (\psi(a_1) + \psi(a_2) + \psi(a_3)) \right| \leq C|K|\text{diam}(K)|\psi|_{C^1}.$$

Proof.

$$\begin{aligned}
&\left| \int_K \psi dx - \frac{|K|}{3} (\psi(a_1) + \psi(a_2) + \psi(a_3)) \right| \\
&\leq \left| \int_K \psi - \psi(a_1) dx - \frac{|K|}{3} (\psi(a_2) - \psi(a_1) + \psi(a_3) - \psi(a_1)) \right| \\
&\leq |K| \sup_{x \in K} |\psi(x) - \psi(a_1)| + \frac{|K|}{3} |\psi(a_2) - \psi(a_1)| + \frac{|K|}{3} |\psi(a_3) - \psi(a_1)|
\end{aligned}$$

Using the Mean Value Theorem, we show that there exist $\xi_x \in [a_1, x]$ such that for all $x \in K$,

$$\psi(x) = \psi(a_1) + (x - a_1) \cdot \nabla \psi(\xi_x).$$

This leads to

$$|\psi(x) - \psi(a_1)| \leq |x - a_1| |\nabla \psi(\xi_x)| \leq \text{diam}(K) \sup_{x \in K} |\nabla \psi(x)| \leq 2\text{diam}(K) |\nabla \psi(x)|_{C^1(K)}$$

hence the result. ■

R With more regularity on ψ and suitable quadrature formulas, we could get a higher order term in $\text{diam}(K)$ by using Taylor expansion.

Proposition 5.3.8 Let $f \in C^1(\overline{\Omega})$ and $K \in \mathscr{T}_h$ and $p \in \mathbb{P}_1$ on K. Then

$$\left| \int_K fp dx - \frac{|K|}{3} (f(a_1)p(a_1) + \cdots + f(a_3)p(a_3)) \right| \leq C|K|^{1/2} \text{diam}(K) \|f\|_{C^1(\overline{\Omega})} \|p\|_{H^1(K)}.$$

Proof. The previous propopsition asserts that

$$\begin{aligned}
\left| \int_K fp dx - \frac{|K|}{3} (f(a_1)p(a_1) + \cdots + f(a_3)p(a_3)) \right| &\leq C|K|\text{diam}(K)|fp|_{C^1(\bar{K})} \\
&\leq C|K|\text{diam}(K)(|f|_{C^1(\bar{K})}|p|_{C^0(\bar{K})} + |f|_{C^0(\bar{K})}|p|_{C^1(\bar{K})})
\end{aligned}$$

Let Φ be the affine bijection from $\hat{K}$ to K and $\hat{p} = p \circ \Phi^{-1}$. Then

$$\begin{aligned}
|p|_{C^0(\bar{K})} = |\hat{p}|_{C^0(\bar{\hat{K}})} &\leq C\|\hat{p}\|_{L^2(\hat{K})} && \text{equivalent norms on } \mathbb{P}_1(\hat{K}) \text{ of finite dim} \\
&\leq C|K|^{-1/2}\|p\|_{L^2(K)} && \text{by change of variables}
\end{aligned}$$

Moreover, $\partial_i p$ is constant on K so

$$|p|_{C^1}^2 = \max\{|\partial_1 p|^2, |\partial_2 p|^2\} \leq \frac{1}{|K|} \int_K |\partial_1 p|^2 + |\partial_2 p|^2$$

or put another way,

$$|p|_{C^1} \leq C|K|^{-1/2} |p|_{H^1}(K).$$

Putting everything togheter, we get the result. ■

Corollary 5.3.9 Let $f \in C^1(\overline{\Omega})$ and $v_h \in V_{h0}$. Then

$$|L(v_h) - \tilde{L}(v_h)| \leq Ch\|f\|_{C^1(\overline{\Omega})}\|v_h\|_{H^1(\Omega)}.$$

Proof.

$$\begin{aligned}
|L(v_h) - \tilde{L}(v_h)| &\leq \sum_{K\in\mathcal{T}_h} \left| \int_K f p dx - \frac{|K|}{3}\left(f(a_1)p(a_1) + \cdots + f(a_3)p(a_3)\right)\right| \\
&\leq Ch\|f\|_{C^1(\overline{\Omega})} \sum_{K\in\mathcal{T}_h} |K|^{1/2}\|v_h\|_{H^1(K)} \\
&\leq Ch\|f\|_{C^1(\overline{\Omega})} \sqrt{\sum_{K\in\mathcal{T}_h} |K|}\sqrt{\sum_{K\in\mathcal{T}_h} \|v_h\|_{H^1(K)}} \qquad \text{by Cauchy-Schwarz} \\
&\leq Ch\|f\|_{C^1(\overline{\Omega})}\sqrt{|\Omega|}\|v_h\|_{H^1(\Omega)}
\end{aligned}$$

■

Proof of the Theorem. Using the Fundamental Trick of Analysis,

$$\|u - \tilde{u}_h\|_{H^1} \leq \|u - u_h\|_{H^1} + \|u_h - \tilde{u}_h\|_{H^1}.$$

Using previous Theorem, we have

$$\|u - u_h\|_{H^1} \leq Ch\|u\|_{H^2} \leq Ch\|u\|_{C^3}.$$

Moreover,

$$\alpha\|u_h - \tilde{u}_h\|_{H^1} \leq \frac{|a(u_h - \tilde{u}_h, u_h - \tilde{u}_h)|}{\|u_h - \tilde{u}_h\|_{H^1}} \leq \sup_{v_h\in V_{h0}} \frac{|a(u_h - \tilde{u}_h, v_h)|}{\|v_h\|_{H^1}} \leq \sup_{v_h\in V_{h0}} \frac{|L(v_h) - \tilde{L}(v_h)|}{\|v_h\|_{H^1}}.$$

Using the previous corollary (since $f = \Delta u \in C^1$), we get that

$$\|u_h - \tilde{u}_h\|_{H^1} \leq \frac{1}{\alpha}Ch\|f\|_{C^1} \leq Ch\|u\|_{C^3}$$

hence the result. ■

Bibliography

Books

[1] Grégoire Allaire et al. *Numerical Analysis and Optimization*. Oxford, England, UK: Oxford University Press, May 2007. ISBN: 978-0-19920522-6. URL: https://global.oup.com/academic/product/numerical-analysis-and-optimization-9780199205226 (cited on pages 17, 44, 46, 57, 58, 60, 69).

[2] Jean-Michel Bony. *Cours d'analyse: théorie des distributions et analyse de Fourier*. Editions Ecole Polytechnique, 2001 (cited on pages 11, 13).

[3] Lawrence C. Evans. *Measure Theory and Fine Properties of Functions, Revised Edition (Textbooks in Mathematics)*. Chapman and Hall/CRC, Apr. 2015. ISBN: 978-1-48224238-6. URL: https://www.amazon.de/Measure-Properties-Functions-Textbooks-Mathematics/dp/1482242389 (cited on page 14).

[4] Lawrence C. Evans. *Partial Differential Equations: Second Edition*. [Online; accessed 22. Mar. 2024]. Mar. 2024. URL: https://bookstore.ams.org/gsm-19-r (cited on pages 33, 39, 45, 46).

[5] Pierre-Arnaud Raviart and Jean-Marie Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles*. July 2004. ISBN: 978-2-10048645-8. URL: https://www.dunod.com/sciences-techniques/introduction-analyse-numerique-equations-aux-derivees-partielles-mathematiques (cited on page 73).

[6] Walter Rudin. *Real and Complex Analysis*. Maidenhead, England, UK: McGraw-Hill, 1987. ISBN: 978-0-07100276-9. URL: https://books.google.de/books?id=NmW7QgAACAAJ (cited on page 26).

Articles

[7] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Math. Control Signals Systems* 2.4 (Dec. 1989), pages 303–314. ISSN: 1435-568X. DOI: 10.1007/BF02551274 (cited on page 26).

[8] Chenguang Duan et al. “Convergence Rate Analysis for Deep Ritz Method”. In: *arXiv* (Mar. 2021). DOI: 10.4208/cicp.OA-2021-0195. eprint: 2103.13330 (cited on page 54).

[9] Weinan E and Bing Yu. “The Deep Ritz method: A deep learning-based numerical algorithm for solving variational problems”. In: *arXiv* (Sept. 2017). DOI: 10.48550/arXiv.1710.00211. eprint: 1710.00211 (cited on page 53).

[10] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. “Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks”. In: *Neural Networks* 3.5 (Jan. 1990), pages 551–560. ISSN: 0893-6080. DOI: 10.1016/0893-6080(90)90005-6 (cited on page 30).

[11] E. Kharazmi, Z. Zhang, and G. E. Karniadakis. “Variational Physics-Informed Neural Networks For Solving Partial Differential Equations”. In: *arXiv* (Nov. 2019). DOI: 10.48550/arXiv.1912.00873. eprint: 1912.00873 (cited on page 55).

[12] I. E. Lagaris, A. Likas, and D. I. Fotiadis. “Artificial Neural Networks for Solving Ordinary and Partial Differential Equations”. In: *arXiv* (May 1997). DOI: 10.1109/72.712178. eprint: physics/9705023 (cited on pages 51, 52).

[13] Moshe Leshno et al. “Multilayer feedforward networks with a nonpolynomial activation function can approximate any function”. In: *Neural Networks* 6.6 (Jan. 1993), pages 861–867. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(05)80131-5 (cited on page 30).

[14] Vien Minh Nguyen-Thanh, Xiaoying Zhuang, and Timon Rabczuk. “A deep energy method for finite deformation hyperelasticity”. In: *Eur. J. Mech. A Solids* 80 (Mar. 2020), page 103874. ISSN: 0997-7538. DOI: 10.1016/j.euromechsol.2019.103874 (cited on page 53).

[15] Sejun Park et al. “Minimum Width for Universal Approximation”. In: *arXiv* (June 2020). DOI: 10.48550/arXiv.2006.08859. eprint: 2006.08859 (cited on page 30).

[16] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. “Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations”. In: *arXiv* (Nov. 2017). DOI: 10.48550/arXiv.1711.10561. eprint: 1711.10561 (cited on page 51).

[17] Justin Sirignano and Konstantinos Spiliopoulos. “DGM: A deep learning algorithm for solving partial differential equations”. In: *arXiv* (Aug. 2017). DOI: 10.1016/j.jcp.2018.08.029. eprint: 1708.07469 (cited on page 51).

[18] Yaohua Zang et al. “Weak Adversarial Networks for High-dimensional Partial Differential Equations”. In: *arXiv* (July 2019). DOI: 10.1016/j.jcp.2020.109409. eprint: 1907.08272 (cited on page 55).