NUMERICAL METHODS FOR PARTIAL DIFFERENTIAL EQUATIONS

Claudio Tessa

December 18, 2025

Contents

1 Introduction

Partial differential equations (PDEs) are differential equantions containing derivatives of the unknown function with respect to several variables.

Example

The differential model can be indicated as follows:

$$\mathcal{P}(u; g)$$

where

- $\mathcal{P}$ generically indicates the model
- u is the exact solution, a function of one or more independent variables
- g indicates the data

The following are typical differential models involving PDEs. Specifically, we will consider linear models with a scalar unknown solution.

Boundary value problem in 1D

It is a stationary differential model with a single independent variable x, representing the space coordinate in an interval $\Omega = (a, b) \subset \mathbb{R}$. The problem involves second order derivatives of the unknown solution $u = u(x)$ with respect to x.

The value of u, or the value of its first derivative, is set at the two boundaries of the domain (interval) Ω, that is at $x = a$ and $x = b$ (the domain boundary is $\partial\Omega = \{a, b\}$).

For example, let us consider the following *Poisson* problem with (homogeneous) Dirichlet boundary conditions: find $u : \Omega \subset \mathbb{R} \to \mathbb{R}$ such that

$$\begin{cases} -\dfrac{d^2u}{dx^2}(x) = f(x) & x \in \Omega = (a, b) \\ u(a) = u(b) = 0 \end{cases}$$

This equation models a stationary phenomenon: the time variable does not appear. Note that the boundary value problem in 1D is a particular case of PDEs, even if it involves only derivatives with respect to a single independent variable x, as one condition is set at $x = a$ and another condition is set at $x = b$. The conditions in the boundary value problem determine the so-called *global nature of the model.*

Initial and boundary value problem in 1D

These problems concern equations that depend on space and time. The unknown solution $u = u(x, t)$ depends on both the space coordinate $x \in \Omega \subset \mathbb{R}$ in 1D, and the time variable $t \in I$.

In this case, there must be prescribed the initial conditions at $t = 0$, as well as the boundary conditions at the end of the 1D interval. For example, consider the *heat equation* (or *diffusion equation*) with Dirichlet boundary conditions: find $u : \Omega \times I \to \mathbb{R}$ such that

$$\begin{cases} \dfrac{\partial u}{\partial t}(x, t) - \mu \dfrac{\partial^2 u}{\partial x^2}(x, t) = f(x, t) & x \in \Omega = (a, b), t \in I \\ u(a, t) = u(b, t) = 0 & t \in I \\ u(x, t_0) = u_0(x) & x \in \Omega = (a, b) \end{cases}$$

In this example, the unknown function $u(x,t)$ describes the temperature in a point $x \in \Omega = (a,b)$ and time $t \in I$ of a metallic bar covering the space interval Ω, and the diffusion coefficient μ represents the thermal response of the material. The boundary conditions express the fact that the ends of the bar are kept at a reference temperature (0 degrees in this case), while at time $t = t_0$, the temperature is assigned in each point $x \in \Omega$.

Boundary value problem in multidimensional domains $\Omega \subset \mathbb{R}^d$, with $d = 2,3$

We can extend the 1D boundary value problem in multidimensional domains $\Omega = \mathbb{R}^d$, with $d = 2,3$. The solution is then $u(\mathbf{x})$, where $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$. This leads to the following Poisson problem with (homogeneous) Dirichlet boundary conditions: find $u : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ such that:

$$\begin{cases} -\Delta u = f & \text{in } \Omega \text{ (i.e. } \mathbf{x} \in \Omega) \\ u = 0 & \text{on } \partial\Omega \text{ (i.e. on } \mathbf{x} \in \partial\Omega) \end{cases}$$

where

$$\Delta u(\mathbf{x}) := \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}(\mathbf{x})$$

is the *Laplace operator*, the domain $\Omega \subset \mathbb{R}^d$ is endowed with boundary $\partial\Omega$, and $f = f(x)$ is the external forcing term.

Initial and boundary value problem in multidimensional domains $\Omega \subset \mathbb{R}^d$, with $d = 2,3$

The multidimensional counterpart of the head equation reads: find $u : \Omega \times I \to \mathbb{R}$ such that

$$\begin{cases} \dfrac{\partial u}{\partial t} - \mu\Delta u = f & \mathbf{x} \in \Omega, t \in I \\ u(\mathbf{x}, t) = 0 & \mathbf{x} \in \partial\Omega, t \in I \\ u(\mathbf{x}, t_0) = u_(\mathbf{x}) & \mathbf{x} \in \Omega \end{cases}$$

where u_0 is the initial datum.

Classification of PDEs

A PDE is a relationship among the partial derivatives of a function $u = u(\mathbf{x}, t)$, the PDE solution (which typically depends on the spatial coordinates $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ if the problem is defined in a spatial domain $\Omega \subset \mathbb{R}^d$), and the time variable t.

Example

In general, a PDE can be written as:

$$\mathcal{P}\left(u, \frac{\partial u}{\partial t}, \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_d}, \ldots, \frac{\partial^{p_1+\cdots+p_d+p_t} u}{\partial x_1^{p_1} \ldots \partial x_d^{p_d} \partial t^{p_t}}, \mathbf{x}, t; g\right) = 0$$

where $p_1, \ldots, p_d, p_t \in \mathbb{N}$ and g is the data.

The PDE *order* is the maximum order of derivation that appears in $\mathcal{P}$, that is $q = p_1 + \cdots + p_d + p_t$. The PDE is *linear* if $\mathcal{P}$ linearly depends on u and its derivatives.

PDEs can be classified in three groups: *elliptic*, *parabolic*, and *hyperbolic* PDEs.

1.1 Physical, mathematical, and numerical problem

In most cases, we cannot analytically solve a PDE. We have to use *numerical methods* that allow to construct an approximation u_h of the exact solution u, for which the corresponding error $u - u_h$ can be quantified or estimated.

Here h indicates a discretization parameter that characterizes the numerical approximation. Conventionally, the smaller is h, the better is the approximation of u made by u_h (that is, the error $u - u_h$ tends to zero as h gets smaller).

Let us consider a *physical problem* (PP) endowed with a physical solution u_{ph}, and dependent on data g. The *mathematical problem* (MP) is the mathematical formulation for the PP and has a mathematical solution u. We indicate the MP as

$$\mathcal{P}(u; g) = 0$$

where $u \in \mathcal{U}$ and $g \in \mathcal{G}$, with $\mathcal{U}$ and $\mathcal{G}$ two suitable sets or spaces. The error between the physical and mathematical solution is called *model error*, say $e_m := u_{ph} - u$.

Before solving a MP, it is required to ensure that it is *well-posed*.

> **Definition**
>
> The MP $\mathcal{P}(u; g) = 0$ is *well-posed* if and only if there exists a unique solution $u \in \mathcal{U}$ that continuously depends on the data $g \in \mathcal{G}$.

$\mathcal{G}$ is the set of admissible data, i.e. those for which the MP admits a unique solution.

The *numerical problem* (NP) is an approximation of the MP. We indicate its numerical solution as u_h, where h is a suitable *discretization parameter*. We state the NP as

$$\mathcal{P}_h(u_h; g_h) = 0$$

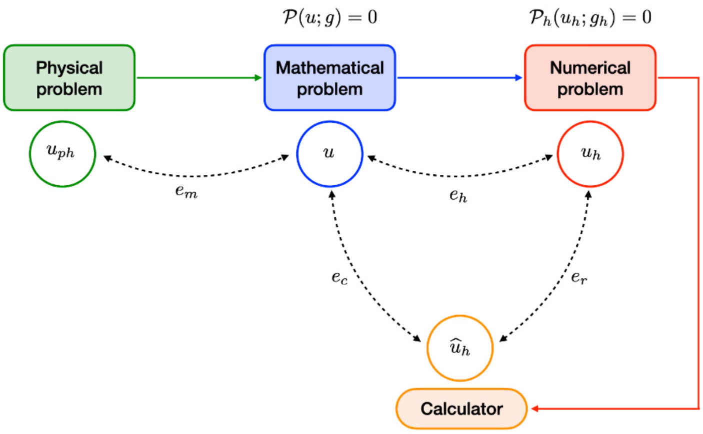
where $u_h \in \mathcal{U}_h$ and $g_h \in \mathcal{G}_h$, with $\mathcal{U}_h$ and $\mathcal{G}_h$ two suitable sets or spaces g_h is the representation of the data in the NP. The error between the mathematical and numerical solution is called *truncation error* $e_h := u - u_h$. This is the error that stems from the discretization of the MP.

If the numerical solution is computed by execuuting the algorithm on a *calculator*, then the final solution is $\hat{u}_h$ and is affected by the *round-off error* $e_r := u_h - \hat{u}_h$. Both the truncation and round-off errors concur to determine the *computational error*, say $e_c := u - \hat{u}_h = e_h + e_r$. For some NP, we can have $|e_r| << |e_h|$, for which $e_c \approx e_h$.

As for the MP, we have to ensure that the NP is well-posed and we have to assess the condition number of the NP.

> **Definition**
>
> The NP $\mathcal{P}_h(u_h; g_h) = 0$ is *well-posed* if and only if there exists a unique solution $u_h \in \mathcal{U}_h$ that continuously depends on the data $g_h \in \mathcal{G}_h$.

We are interested in NPs that allow to obtain computational errors that tend to zero as the discretization parameter h goes to zero.



Definition

The NP is *convergent* when the computational error tends to zero as h tends to zero, that is

$$\lim_{h \to 0} e_c = 0$$

A crucial aspect is to determine the *convergence order* of the the NP.

Definition

If $|e_c| \leq Ch^p$, with C positive constant independent of h and p, then the NP is convergent with *order* p.

Note that a well-posed NP is not necessarily convergent. To ensure convergence of the NP, this must satisfy the *consistency* property: the NP must be a "faithful copy" of the original MP.

Definition

The NP is *consistent* if and only if $\lim_{h\to 0} \mathcal{P}_h(u;g) = \mathcal{P}(u;g) = 0$, with $g \in \mathcal{G}_h$.

The NP is *strongly consistent* if and only if $\mathcal{P}_h(u;g) \equiv \mathcal{P}(u;g) = 0$ for all $h > 0$, with $g \in G_h$.

It is clear that the NP must be well-posed, consistent, and convergent. The concepts of well-posedness and convergence are indeed strongly connected and encoded in the following theoretical result.

Theorem: equivalence

If the NP $\mathcal{P}_h(u_h;g_h) = 0$ with $u_h \in \mathcal{U}_h$ and $g_h \in \mathcal{G}_h$ is consistent, then it is well-posed if and only if it is also convergent.

It follows that, if the NP is well-posed and consistent, then the NP is also convergent. Equivalently, if the NP is consistent and convergent, then the NP is also well-posed. The equivalence theorem is very useful as it allows us to verify only two of the properties of a NP to obtain the third.

2 Formulation of partial differential equations

2.1 From the strong formulation to the weak formulation

Here we will focus on the mathematical formulation of elliptic PDEs, in 1D or multidimensional domains (2D or 3D), with particular emphasis on the *weak formulation*.

Let us consider an elliptic PDE defined in a domain $\Omega \subset \mathbb{R}^d$, for $d = 1, 2, 3$. We can express such PDE as

$$\mathcal{L}u = f \qquad \text{in } \Omega$$

where $\mathcal{L}$ is a suitable differential operator, and appropriate boundary conditions are set on the boundary $\partial\Omega$ of the domain Ω. We say that this is the *strong formulation* of the PDE, and u is its *strong solution* if u satisfies the equation in every point $\mathbf{x} \in \Omega$ and every partial derivative appearing in $\mathcal{L}u$ is well-defined.

However, the strong formulation cannot represent many physically feasible solutions. We need to transition to a *weak formulation* by follwing two steps:

1. Multiply both sides by a suitable test function v:

$$(\mathcal{L}u)v = fv$$

2. Integrate in Ω applying integration by parts:

$$\int_\Omega (\mathcal{L}u)v \; d\Omega = \int_\Omega fv \; d\Omega$$

2.1.1 Example: Poisson problem in 1D

Let us consider again the Poisson problem with homogeneous Dirichlet boundary conditions: find $u : \Omega \subset \mathbb{R} \to \mathbb{R}$ such that

$$\begin{cases} -u''(x) = f(x) & x \in \Omega = (a, b) \\ u(a) = u(b) = 0 \end{cases}$$

let us now introduce a function $v \in C^1(0, 1)$ such that $v(0) = v(1) = 0$. We say that $v \in C^1_0(0, 1)$ and we call this function a *test function*. We multiply the differential problem by v and we integrate over $(0, 1)$, for which

$$\int_0^1 -u''(x)v(x)dx = \int_0^1 f(x)v(x)dx$$

must hold for every test function $v \in C^1_0(0, 1)$. By *partial integration* of the left hand side for a general domain $\Omega = (a, b) \subset \mathbb{R}$, we have

$$\begin{aligned} \int_a^b -u'' \; v \; dx &= \int_a^b -(u'v)' \; dx + \int_a^b u' \; v' \; dx = [-u' \; v]_{x=a}^b + \int_a^b u' \; v' \; dx \\ &= [u'(b)v(b) - u'(a)v(a)] + \int_a^b u' \; v' \; dx \end{aligned}$$

As $\Omega = (a, b) = (0, 1)$ and $v(0) = v(1) = 0$, we have in this case that

$$\int_0^1 -u'' \ v \ dx = \int_0^1 u' \ v' \ dx$$

This yields the *weak formulation* of the problem:

$$\text{find } u \quad : \quad \int_0^1 u' \ v' \ dx = \int_0^1 f \ v \ dx \qquad \text{for all } v \in C_0^1(0, 1)$$

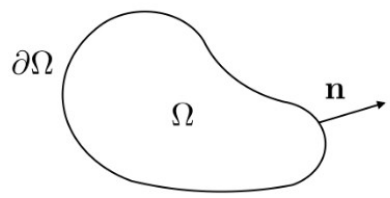
Note how in the weak formultation of the problem, second derivatives of u do not appear. Therefore, it is no longer required that $u \in C^2(0, 1)$. Instead, u' and v' do appear, but under the integral: it is therefore not necessary that u' and v' are continuous functions.

2.1.2 Example: Poisson problem in 2D

Let us consider again the Poisson problem with homogeneous Dirichlet boundary conditions: find $u : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ such that

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

where $d = 2, 3$, the domain $\Omega \subset \mathbb{R}^d$ is endowed with boundary $\partial\Omega$. We indicate by $\mathbf{n}$ the outward directed, unit vector normal to the boundary $\partial\Omega$.



We write the weak formulation of this problem. First, we introcuce the test function $v \in C_0^1(\Omega)$, i.e. such that ∇v is a continuous vector-valued function in Ω and $v = 0$ on $\partial\Omega$. Then, we multiply the PDE by v and we integrate, thus obraining:

$$\int_\Omega -\Delta u(\mathbf{x}) v(\mathbf{x}) d\mathbf{x} = \int_\Omega f(\mathbf{x}) v(\mathbf{x}) d\mathbf{x}$$

Partial integration of the left hand side yields the following formula:

$$\begin{aligned} \int_\Omega -\Delta u \ v \ d\mathbf{x} &= \int_\Omega -\nabla \cdot (v \nabla u) d\mathbf{x} + \int_\Omega \nabla u \cdot \nabla v \ d\mathbf{x} \\ &= \oint_{\partial\Omega} -v \nabla u \cdot \mathbf{n} \ d\mathbf{x} + \int_\Omega \nabla u \cdot \nabla v \ d\mathbf{x} \end{aligned}$$

The term $\nabla u \cdot \mathbf{n}$ on the boundary $\partial\Omega$ is often indicated as $\dfrac{\partial u}{\partial n}$.

By recalling that $v = 0$ on $\partial\Omega$, we obtain the weak formulation of the Poisson problem:

$$\text{find } u \quad : \quad \int_\Omega \nabla u \cdot \nabla v \ d\mathbf{x} = \int_\Omega f \ v \ d\mathbf{x} \qquad \text{for all } v \in C_0^1(\Omega)$$

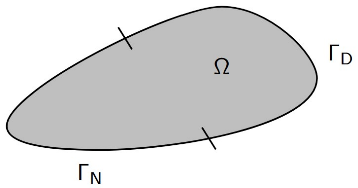
However, we still have to identify the function space to which the weak solution u belongs.

2.1.3 Example

$$\begin{cases} \mathcal{L}u = -\text{div}\,(\mu\nabla u) + \mathbf{b}\cdot\nabla u + \sigma u = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_D \\ \mu\nabla u\cdot\mathbf{n} = g & \text{on } \Gamma_N \end{cases}$$

where:

- $f \in L^2(\Omega)$ is given, $\mathcal{L}$ is the 2nd order differential operator, and u is the unknown
- $\partial\Omega = \Gamma_D \cup \Gamma_N,\ \Gamma_D \cap \Gamma_N \neq \emptyset$
- $\mu \in L^\infty(\Omega),\ \mu(\mathbf{x}) \geq \mu_0 > 0$
- $\mathbf{b} \in (L^\infty(\Omega))^d$
- $\sigma \in L^2(\Omega)$
- $g \in L^2(\Gamma_N)$



div$(\cdot)$ is the divergence operator, defined below.

Definition: divergence operator

$$\text{div}(\mathbf{v}) = \nabla\cdot\mathbf{v} = \sum_{i=1}^{d}\frac{\partial v_i}{\partial x_i}, \qquad \mathbf{v} = (v_1,\ldots,v_d)^T$$

To obtain the weak formulation, we multiply by the test function v and integrate in Ω:

$$\int_\Omega -\text{div}(\mu\nabla u)v\ d\Omega + \int_\Omega(\mathbf{b}\cdot\nabla u)v + \int_\Omega \sigma uv = \int_\Omega fv \tag{1}$$

we can now apply integration by part to the first integral using Green's formula.

Green's formula

$$\int_\Omega \mathbf{q}\cdot\nabla\varphi = -\int_\Omega \text{div}(\mathbf{q})\cdot\varphi + \int_{\partial\Omega}(\mathbf{q}\cdot\mathbf{n})\varphi$$

where $\mathbf{n}$ is the *normal* to the boundary surface $\partial\Omega$ at every specific point.

With this formula we obtain that

$$\int_\Omega -\underbrace{\text{div}\,(\mu\nabla u)}_{\nabla\varphi}\cdot\underbrace{v}_{\mathbf{q}} = \int_\Omega \mu\nabla u\cdot\nabla v - \int_{\partial\Omega}(\mu\nabla u\cdot\mathbf{n})\,v$$

We can split the boundary integral into two parts Γ_D and Γ_N:

$$\int_{\partial\Omega} (\mu\nabla u\cdot\mathbf{n})\, v = \underbrace{\cancelto{0}{\int_{\Gamma_D} (\mu\nabla u\cdot\mathbf{n})\, v}}_{0 \text{ if } v|_{\Gamma_D}=0} + \int_{\Gamma_N} \underbrace{(\mu\nabla u\cdot\mathbf{n})}_{g}\, v$$

Substituting these into the starting equation (1) and rearanging terms we get:

$$\underbrace{\int_\Omega \mu\nabla u\cdot\nabla v + \int_\Omega \mathbf{b}\cdot\nabla u v + \int_\Omega \sigma u v}_{a(u,v)} = \underbrace{\int_\Omega f v + \int_{\Gamma_N} g v}_{F(v)}$$

Therefore the problem becomes

$$\boxed{\begin{array}{c}\text{Find } u\in V : a(u,v) = F(v),\ \forall v\in V \\ \text{where } V = \{v\in H^1(\Omega),\ v|_{\Gamma_D}=0\}\end{array}}$$

2.2 Lax-Milgram lemma

The **Lax-Milgram lemma** is the fundamental theorem that guarantees the solution you are looking for in the weak formulation actually exists and is unique.

Lax-Milgram lemma

Assume that
- (i) V Hilbert space with norm $\|\cdot\|$ and inner product $(\cdot,\cdot)$
- (ii) $F\in V' : |F(v)| \le \|F\|_{V'}\|V\|,\ \forall v\in V$
- (iii) a continuous: $\exists M>0 : |a(u,v)| \le M\|u\|\|v\|,\ \forall u,v\in V$
- (iv) a coercive: $\exists \alpha>0 : a(v,v) \ge \alpha\|v\|^2,\ \forall v\in V$

Then, there exists a unique solution u

Remark that V' is the *dual space* of V. It is the set of all linear and continuous functionals that map elements from V into a real number $\mathbb{R}$, with norm $\|F\|_{V'} = \sup\limits_{v\in V\setminus\{0\}} \dfrac{F(v)}{\|v\|_V}$

Moreover, $\|u\| \le \dfrac{\|F\|_{V'}}{\alpha}$. α is called *coercivity constant.*

2.2.1 Proving the assumptions of the Lax-Milgram lemma

Consider the following elliptic problem in 1D

$$\begin{cases}\mathcal{L}u = -\left(\mu u'\right)' + bu' + \sigma u = f & \bar{a} < x < \bar{b}\\ u(\bar{a}) = 0 \\ \mu u'(\bar{b}) = \gamma v(\bar{b})\end{cases}$$

To prove if the assumptions of the Lax-Milgram lemma hold, let us first consider the weak formulation of the problem:

$$
\begin{aligned}
V &= \{v \in H^1(\bar{a}, \bar{b}) : v(\bar{a}) = 0\} \\
a(u, v) &= \int_{\bar{a}}^{\bar{b}} (\mu u'v' + bu'v + \sigma uv) \ dx \\
F(v) &= \int_{\bar{a}}^{\bar{b}} fv \ dx + \gamma v(\bar{b})
\end{aligned}
$$

(i) V is a Hilbert space

Proof: Since $H^1(\bar{a}, \bar{b})$ is a Hilbert space, and V is a subspace of $H^1(\bar{a}, \bar{b})$, V is a Hilbert space (there are other conditions required but we take this for granted). Furthermore,

$$
\begin{aligned}
\|v\| &= \sqrt{\|v\|^2_{L^2(\bar{a},\bar{b})} + \|v'\|^2_{L^2(\bar{a},\bar{b})}} \\
|||v||| &= \sqrt{\|v'\|^2_{L^2(\bar{a},\bar{b}a)}} = \|v'\|_{L^2(\bar{a},\bar{b})} = |v|
\end{aligned}
$$

(ii)