A brief introduction to weak formulations of PDEs and the finite element method

T. J. Sullivan[1,2]

June 29, 2020

1 Introduction

The aim of this note is to give a *very* brief introduction to the "modern" study of partial differential equations (PDEs), where by "modern" we mean the theory based in weak solutions, Galerkin approximation, and the closely-related finite element method. We will illustrate the main ideas by reference to an elliptic PDE of second order. The point is not to be totally rigorous about all details, but rather to give some motivation for an important connection between linear algebra and PDEs that has deep consequences both for the mathematical analysis of PDEs and their numerical solution on computers.

2 Prerequisite concepts and notation

Before proceeding any further, make sure that you are familiar with the following concepts and notation, and refresh your memory if necessary.

- The natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$.
- The real numbers $\mathbb{R}$ and the vector spaces $\mathbb{R}^2$, $\mathbb{R}^3$, and $\mathbb{R}^d$ for $d \in \mathbb{N}$.
- What it means for a function of a single variable to be differentiable.
- The partial derivatives of a function of d variables, $d \in \mathbb{N}$.
- Integration of functions of one or many variables.
- Integration by parts.
- Ideally, also the notion of a Hilbert space.

3 Strong and weak formulations of second-order elliptic PDEs

Fix $d \in \mathbb{N}$ and let $D \subseteq \mathbb{R}^d$ be a "nice" domain; ∂D denotes the boundary of D. A perfectly good example to have in mind is that the d-dimensional unit cube

$$D = (0,1)^d = \{x = (x_1, \dots, x_d) \in \mathbb{R}^d \mid 0 < x_i < 1 \text{ for each } i \in \{1, \dots, d\}\}.$$

In fact, almost all of the important ideas in this note are covered by the case $d = 1$, so that is a good place to start if your vector calculus is a bit rusty.

For a scalar-valued function $u \colon D \to \mathbb{R}$, $\nabla u \colon D \to \mathbb{R}^d$ denotes its *gradient*, i.e. the vector field whose components are the first-order partial derivatives of u:

$$\nabla u(x) = \left(\frac{\partial u}{\partial x_1}(x), \dots, \frac{\partial u}{\partial x_d}(x) \right) \in \mathbb{R}^d.$$

[1] Mathematics Institute and School of Engineering, University of Warwick, Coventry CV4 7AL, United Kingdom (t.j.sullivan@warwick.ac.uk)
[2] Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany (sullivan@zib.de)

Later on, we will discuss a so-called "weak" interpretation of this gradient. Similarly, for a vector field $v = (v_1, \ldots, v_d) \colon D \to \mathbb{R}^d$, we will write $\nabla \cdot v \colon D \to \mathbb{R}$ for its *divergence*, the scalar field given by

$$\nabla \cdot v(x) = \frac{\partial v_1}{\partial x_1}(x) + \cdots + \frac{\partial v_d}{\partial x_d}(x).$$

Let $a \colon D \to \mathbb{R}$ and $f \colon D \to \mathbb{R}$ be given, and consider the problem of finding $u \colon D \to \mathbb{R}$ such that

$$\begin{aligned} -\nabla \cdot (a(x) \nabla u(x)) &= f(x) && \text{for } x \in D, \\ u(x) &= 0 && \text{for } x \in \partial D. \end{aligned} \tag{3.1}$$

A standard example of such a problem is Poisson's equation, with $a(x) \equiv 1$:

$$\begin{aligned} -\Delta u(x) &= f(x) && \text{for } x \in D, \\ u(x) &= 0 && \text{for } x \in \partial D, \end{aligned} \tag{3.2}$$

where Δ is the Laplace operator

$$\Delta u(x) = \frac{\partial^2 u}{\partial x_1^2}(x) + \cdots + \frac{\partial^2 u}{\partial x_d^2}(x). \tag{3.3}$$

On the face of it, expressions like (3.1), (3.2), and (3.3) only make sense if u is twice differentiable and a is differentiable. This is the so-called *strong form* of the PDE. It is an unfortunate fact of mathematical life that the strong formulation is simply too strong in practice, in the sense that many physical problems of interest have no strong solutions.

Using integration by parts, however, we can relax the definition of a solution considerably. We will call $\varphi \colon D \to \mathbb{R}$ a *compactly supported test function* if φ is infinitely differentiable (i.e. it has all mixed partial derivatives of all orders) and there is a closed and bounded set $S \subset D$ such that φ and all its partial derivatives are zero on $D \setminus S$. Let

$$\mathcal{C}_{\mathrm{c}}^{\infty}(D) = \left\{ \varphi \colon D \to \mathbb{R} \,\middle|\, \begin{array}{c} \varphi \text{ is infinitely differentiable and, for some closed, bounded } S \subset D, \\ \varphi \text{ and all its partial derivatives are zero outside } S \end{array} \right\}. \tag{3.4}$$

denote the space of all such test functions.

Exercise 3.1. Show that $\mathcal{C}_{\mathrm{c}}^{\infty}(D)$ as defined above is a vector space, i.e. if $\alpha_1, \alpha_2 \in \mathbb{R}$ and $\varphi_1, \varphi_2 \in \mathcal{C}_{\mathrm{c}}^{\infty}(D)$, then $\alpha_1 \varphi_1 + \alpha_2 \varphi_2 \in \mathcal{C}_{\mathrm{c}}^{\infty}(D)$.

We will now derive the so-called *weak form* of the PDE (3.1). The motivation for this weak form is the following observation: any two finite-dimensional vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ are equal if and only if their inner products with an arbitrary $\boldsymbol{\varphi} \in \mathbb{R}^d$ are equal, i.e.

$$\boldsymbol{u} = \boldsymbol{v} \in \mathbb{R}^d \iff \forall \boldsymbol{\varphi} \in \mathbb{R}^d, \boldsymbol{u} \cdot \boldsymbol{\varphi} = \boldsymbol{v} \cdot \boldsymbol{\varphi} \in \mathbb{R}. \tag{3.5}$$

We call this "testing" the equation $\boldsymbol{u} = \boldsymbol{v}$ against the test direction $\boldsymbol{\varphi} \in \mathbb{R}^d$. In essence, the weak form of (3.1) consists of "testing" the equation (3.1) against an arbitrary test function φ. It turns out that we can do something similar for infinite-dimensional spaces of functions, the kinds of spaces in which a solution u to (3.1) might live; the only catch is that then the "$\iff$" in (3.5) is only an "$\implies$", and so the weak form is genuinely weaker than the "strong form" (3.1).

Exercise 3.2 (Deriving the weak form of (3.1))**.** Multiply (3.1) by an arbitrary test function $\varphi \in \mathcal{C}_{\mathrm{c}}^{\infty}(D)$ and integrate the resulting equation over all of D to get

$$\begin{aligned} -\int_D \nabla \cdot (a(x) \nabla u(x)) \varphi(x) \, \mathrm{d}x &= \int_D f(x) \varphi(x) \, \mathrm{d}x \\ u(x) \varphi(x) &= 0 \qquad \text{for } x \in \partial D. \end{aligned}$$

Then use integration by parts and the fact that both u and φ vanish on ∂D to obtain that, if u is a solution to (3.1), then

$$\int_D a(x) \nabla u(x) \cdot \nabla \varphi(x) \, \mathrm{d}x = \int_D f(x) \varphi(x) \, \mathrm{d}x \tag{3.6}$$

holds for all test functions $\varphi \in \mathcal{C}_{\mathrm{c}}^{\infty}(D)$.

If u is such that (3.6) holds for all test functions φ, then we call u a *weak solution* to the PDE (3.1). Note that, perhaps surprisingly, (3.6) makes no use of the second derivative of u, only first-order derivatives, and this is one reason that seeking a weak solution u that satisfies (3.6) is genuinely a weaker/easier problem than seeking a strong solution u to (3.1).

In fact, when people speak of "weak solutions", they usually mean that (3.6) holds for all φ in some Sobolev space, as defined later, but this definition will do for now. Three important facts here are that

(a) many PDEs have weak solutions even when they do not have strong solutions;

(b) weak solutions are much easier to describe in terms of linear algebra and its infinite-dimensional analogues than strong solutions are;

(c) even when a strong solution exists, it is often easier to first show that a weak solution exists and then show that it is in fact strong, than to find a strong solution directly.

To see why linear algebra is very relevant here, observe that (3.6) can be reformulated as the system of equations

$$B(u, \varphi) = \langle f, \varphi \rangle_{L^2} \tag{3.7}$$

where B is the bilinear function (meaning that it is a linear function of each of its two arguments)

$$B(u, v) = \int_D a(x) \nabla u(x) \cdot \nabla v(x) \, \mathrm{d}x$$

and

$$\langle u, v \rangle_{L^2} = \int_D u(x) v(x) \, \mathrm{d}x$$

is the L^2 inner product. Thus, solving the PDE in weak form looks very much like solving something like an infinite-dimensional matrix-vector equation "$B(u, \cdot) = f$".

Exercise 3.3 (A more general setting)**.** Let $\mathcal{L}$ be a second-order differential operator of the following so-called "divergence form":

$$\mathcal{L}u(x) = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i} \left(a_{ij}(x) \frac{\partial u}{\partial x_j}(x) \right) + \sum_{i=1}^{d} b_i(x) \frac{\partial u}{\partial x_i}(x)$$

for d^2 functions $a_{ij} \colon D \to \mathbb{R}$ and d functions $b_i \colon D \to \mathbb{R}$. Follow the model described above to derive an equation similar to (3.7) for a weak solution u of the equation $\mathcal{L}u = f$.

4 Sobolev spaces

The purpose of this section is to give various equivalent, though slightly informal, definitions of the Sobolev spaces $H^1(D)$ and $H^1_0(D)$.

Our first definition is one sentence long: the *Sobolev space* $H^1_0(D)$ consists of all functions $u \colon D \to \mathbb{R}$ that are square-integrable (i.e. $\int_D |u(x)|^2 \, \mathrm{d}x$ is finite) and have a square-integrable gradient, and that vanish on the boundary ∂D of D, and it is equipped with the inner product

$$\langle u_1, u_2 \rangle_{H^1} = \int_D u_1(x) u_2(x) \, \mathrm{d}x + \int_D \nabla u_1(x) \cdot \nabla u_2(x) \, \mathrm{d}x \tag{4.1}$$

$$= \int_D u_1(x) u_2(x) \, \mathrm{d}x + \sum_{i=1}^{d} \int_D \frac{\partial u_1}{\partial x_i}(x) \frac{\partial u_2}{\partial x_i}(x) \, \mathrm{d}x, \tag{4.2}$$

and norm

$$\|u\|^2_{H^1} = \int_D |u(x)|^2 \, \mathrm{d}x + \int_D |\nabla u(x)|^2 \, \mathrm{d}x \tag{4.3}$$

$$= \int_D |u(x)|^2 \, \mathrm{d}x + \sum_{i=1}^{d} \int_D \left| \frac{\partial u}{\partial x_i}(x) \right|^2 \mathrm{d}x. \tag{4.4}$$

This first definition leaves some important technical points unanswered. For example, does the gradient really have to exist everywhere in D?

A second, slightly better definition is given by first weakening the notion of a derivative, again using the trick of integrating by parts. We will say that $v \colon D \to \mathbb{R}$ is a *weak derivative* for $u \colon D \to \mathbb{R}$ in the i^{th} coordinate direction, $1 \leq i \leq d$, if

$$\int_D u(x)\varphi(x)\,\mathrm{d}x = -\int_D v(x)\frac{\partial \varphi}{\partial x_i}(x)\,\mathrm{d}x$$

for all compactly-supported test functions $\varphi \in \mathcal{C}_{\mathrm{c}}^{\infty}(D)$ (see (3.4)).

Exercise 4.1. Show that, if u has a genuine partial derivative $\frac{\partial u}{\partial x_i}$, then this is a weak derivative for u. Construct an example of a function that has a weak derivative but is not differentiable. (Hint: Try a function with a "kink" at one point, such as the absolute value function.)

We define the *Sobolev space* $H^1(D)$ to consist of all functions u that are square-integrable and also have a weak gradient (i.e. the d-dimensional vector field of weak derivatives in the d coordinate directions) that is also square-integrable.

We then define the Sobolev space $H_0^1(D)$, which consists of all $u \in H^1(D)$ such that u is identically zero on ∂D. [Actually, we have to be very careful with this statement, because we ∂D has no d-dimensional volume, and changing a function on such a set is regarded as not changing the function at all, so what really are the values of u on ∂D?]

In practice, perhaps a better definition is that $H_0^1(D)$ is the closure of the space $\mathcal{C}_{\mathrm{c}}^{\infty}(D)$ of test functions with respect to the norm (4.3), i.e. all possible limit function — with respect to the norm (4.3) — of convergent sequences of test functions. This is similar to the way that we could say that the space $L^2(D)$ of all square-integrable functions on D is the closure of $\mathcal{C}_{\mathrm{c}}^{\infty}(D)$ of test functions with respect to the norm

$$\|u\|_{L^2}^2 = \int_D |u(x)|^2\,\mathrm{d}x.$$

5 Galerkin's method and finite elements

In computational practice, we are usually tasked with approximating the solution to some desired level of accuracy and we do so using a finite-dimensional subspace of $H^1(D)$ or $H_0^1(D)$ spanned by a collection of finitely many basis functions, often associated to a mesh or grid of some kind.

Exercise 5.1 (Piecewise linear finite elements)**.** Consider the case $d = 1$ and let $D = (0, 1)$, the unit interval. Fix $N \in \mathbb{N}$ and consider the $N + 1$ points ("nodes")

$$x_n = \frac{n}{N} \text{ for } n = 0, \ldots, N.$$

Let ψ_n be the piecewise linear function that takes the value 1 at x_n and 0 at all x_m for $m \neq n$, and is linear in between adjacent nodes — we sometimes call this a "tent function".

Calculate the weak derivative $\psi_n' \equiv \frac{\partial \psi_n}{\partial x}$ of ψ_n. Show that each $\psi_n \in H^1(D)$ and that $\psi_1, \ldots, \psi_{N-1} \in H_0^1(D)$.

Calculate all the inner products $\langle \psi_n, \psi_m \rangle_{L^2}$ and $\langle \psi_n', \psi_m' \rangle_{L^2}$ for $m, n \in \{0, \ldots, N\}$.

The piecewise linear finite elements are a good motivating example, but we will now take a more abstract view and simply let $\psi_1, \ldots, \psi_N$ be some linearly independent functions in $H_0^1(D)$; they span an N-dimensional subspace $\mathcal{V}_N$ of $H_0^1(D)$.

In principle, we are charged with finding $u \in H_0^1(D)$ such that

$$\int_D a(x)\nabla u(x) \cdot \nabla v(x)\,\mathrm{d}x = \int_D f(x)v(x)\,\mathrm{d}x$$

for all $v \in H_0^1(D)$. Now we shall seek an approximate solution in $\mathcal{V}_N$ of the form $u = \sum_{n=1}^N u_n \psi_n$. Instead of testing u against arbitrary $v \in H_0^1(D)$ we shall only test against $v \in \mathcal{V}_N$ of the form $v = \sum_{n'=1}^N v_{n'} \psi_{n'}$. This is called *Galerkin's method* for solving the PDE.

By the linearity of integration, we seek $u = \sum_{n=1}^N u_n \psi_n \in \mathcal{V}_N$ such that

$$\sum_{n,n'=1}^N \int_D a(x) \nabla u_n \psi_n(x) \cdot v_{n'} \nabla \psi_{n'}(x) \, \mathrm{d}x = \sum_{n'=1}^N \int_D f(x) v_{n'} \psi_{n'}(x) \, \mathrm{d}x$$

for all $v = \sum_{n'=1}^N v_{n'} \psi_{n'} \in \mathcal{V}_N$. Equivalently,

$$\sum_{n,n'=1}^N u_n v_{n'} \int_D a(x) \nabla \psi_n(x) \cdot \nabla \psi_{n'}(x) \, \mathrm{d}x = \sum_{n'=1}^N v_{n'} \int_D f(x) \psi_{n'}(x) \, \mathrm{d}x.$$

In particular, it is enough to consider $v = \psi_m$ for $m = 1, \dots, N$, i.e. $v_m = \delta_{n'm}$, yielding

$$\sum_{n=1}^N u_n \int_D a(x) \nabla \psi_n(x) \cdot \nabla \psi_m(x) \, \mathrm{d}x = \int_D f(x) \psi_m(x) \, \mathrm{d}x.$$

In other words, writing

$$\begin{aligned} \boldsymbol{u} &= (u_n)_{n=1}^N \in \mathbb{R}^N, \\ \boldsymbol{b} &= \left(\int_D f(x) \psi_m(x) \, \mathrm{d}x \right)_{m=1}^N \in \mathbb{R}^N, \\ \boldsymbol{A} &= \left(\int_D a(x) \nabla \psi_n(x) \cdot \nabla \psi_m(x) \, \mathrm{d}x \right)_{n,m=1}^N \in \mathbb{R}^{N \times N}, \end{aligned}$$

the coefficients $u_1, \dots, u_N$ for u solve the following matrix-vector equation:

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{b} \tag{5.1}$$

The matrix $\boldsymbol{A}$ is called the *stiffness matrix* of the basis $\{\psi_1, \dots, \psi_N\}$ and the coefficient field a.

If furthermore $\boldsymbol{f} = (f_m)_{m=1}^N$, where $f = \sum_{m=1}^N f_m \psi_m$, and $\boldsymbol{M} = (\int_D \psi_n \psi_m \, \mathrm{d}x)_{n,m=1}^N$, then (5.1) can be equivalently expressed as

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

The matrix $\boldsymbol{M}$ is called the *mass matrix* of the basis $\{\psi_1, \dots, \psi_N\}$ and is independent of the coefficient field a. Note that the matrices $\boldsymbol{A}$ and $\boldsymbol{M}$ can be assembled ahead of time, before knowledge of the right-hand side f, and hence can be re-used for many different fs to generate the corresponding approximate solution coefficient vectors $\boldsymbol{u}$.

Exercise 5.2. Show that the matrices $\boldsymbol{A}$ and $\boldsymbol{M}$ are always symmetric matrices. Show that $\boldsymbol{M}$ is always positive-definite, and that $\boldsymbol{A}$ is positive-definite if there exists some $a_0 > 0$ such that $a(x) \geq a_0$ for all $x \in D$. Hence show that, if a is bounded below in this way, the matrix-vector equation $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{b}$ in (5.1) has a unique solution $\boldsymbol{u} \in \mathbb{R}^N$.

In view of the previous exercise, we should ask ourselves two questions:

(a) Does the matrix-vector equation (5.1) have an infinite-dimensional analogue in Sobolev spaces instead of $\mathbb{R}^N$, the solutions of which are weak solutions to the PDE?

(b) Do solutions to the the matrix-vector equation (5.1) converge to weak solutions of the PDE as $N \to \infty$?

The answer to the first question is given by the following result, which you should think of as the general version of the previous exercise:

Theorem 5.3 (Lax–Milgram lemma). *Let $\mathcal{V}$ be a Hilbert space (i.e. a complete inner product space). Suppose that $B \colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is a bilinear function that is*

(a) bounded, *in the sense that there is a constant $C > 0$ such that*

$$|B(u, v)| \leq C \|u\| \|v\| \text{ for all } u, v \in \mathcal{V};$$

(b) and coercive*, in the sense that there is a constant* $c > 0$ *such that*

$$B(u, u) \geq c\|u\|^2 \text{ for all } u \in \mathcal{V};$$

Then, for all $f \in \mathcal{V}$*, there is a unique* $u \in \mathcal{V}$ *such that*

$$B(u, v) = \langle f, v \rangle \text{ for all } v \in \mathcal{V}.$$

In order to determine the constants c and C for application of the Lax–Milgram lemma to our elliptic PDEs, with $\mathcal{V} = H_0^1(D)$, one generally needs to do some careful mathematical analysis, beyond the scope of these notes. However, an accessible special case is the following:

Exercise 5.4. Determine the constants c and C for the bilinear form B associated to Poisson's equation on the d-dimensional unit cube D. Hence show that, for every $f \in L^2(D)$, Poisson's equation $-\Delta u = f$ has a unique weak solution $u \in H_0^1(D)$.

The answer to the second question, convergence, is answered by a result called *Céa's lemma*, which states that the exact solution $u \in \mathcal{V}$ and the approximate solution $u_N \in \mathcal{V}_N$ satisfy

$$\|u - u_N\| \leq \frac{C}{c} \inf_{v \in \mathcal{V}_N} \|u - v\|$$

Thus, in some sense, the Galerkin / finite-element solution is a "quasi-optimal" approximation of u within the finite-dimensional subspace $\mathcal{V}_N$.

In fact, when the bilinear form B in the Lax–Milgram problem is symmetric (i.e. $B(u, v) = B(v, u)$, which is the case for our motivating example (3.1)), one can show a little bit more: the subspace solution u_N is the best approximation to the full-space solution u with respect to the *energy norm* $\|u\|_B \coloneqq \sqrt{B(u, u)}$ and is the orthogonal projection of u onto the subspace $\mathcal{V}_N$ with respect to the *energy inner product* $\langle u, v \rangle_B \coloneqq B(u, v)$.

Exercise 5.5. Write a computer program (e.g. in C/C++, Python, or Matlab) to approximately solve Poisson's equation $-\Delta u = f$ in dimension $d = 1$ in the piecewise linear finite element basis of Exercise 5.1. That is, build the stiffness and mass matrices and solve the Galerkin system (5.1).

References

R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003. ISBN 0-12-044143-8.

L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010. ISBN 978-0-8218-4974-3. doi:10.1090/gsm/019.

M. Renardy and R. C. Rogers. *An Introduction to Partial Differential Equations*, volume 13 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2004. ISBN 0-387-00444-0.