Konputazio Ingeniaritza eta Sistema Adimentsuak Unibertsitate Masterra
Máster Universitario en Ingeniería Computacional y Sistemas Inteligentes

University of the Basque Country
Euskal Herriko Unibertsitatea

MASTER'S DEGREE IN COMPUTATIONAL ENGINEERING AND INTELLIGENT SYSTEMS

THESIS

Massive Finite Element Computations for Deep Learning Inversion

---

Author:
Carlos Uriarte Baranda

Supervisors:
Dr. David Pardo Zubiaur
Dr. Elisabete Alberdi Celaya

A thesis submitted in fulfillment of the requirements
for the Master's degree in Computational Engineering and Intelligent Systems

of the

Master and Doctoral School - UPV/EHU

taught in the

Faculty of Informatics - UPV/EHU

by the

Department of Computer Science and Artificial Intelligence

SEPTEMBER, 2019

«*Mathematics is the language in which God has written the universe.*»

Galileo Galilei

Acknowledgments

First, I would like to thank infinitely my supervisor David Pardo for his time and dedication this year and for everything he has taught me. I also want to strongly thank my supervisor Elisabete Alberdi for advising me and helping me whenever I have needed. Furthermore, I wish to express my gratitude to *BCAM - Basque Center for Applied Mathematics* for the grant and to all its members for their hospitality. In particular, I especially appreciate the close and professional treatment of my colleagues in the *Simulation of Wave Propagation* group to which I belonged at *BCAM* during the development of this thesis. Finally, I wish to thank my parents and brother for their support.

Abstract

We focus on the inversion of borehole resistivity measurements in real time. To perform this task, we propose the use of Deep Learning methods. One critical task on this endeavor is to produce a large database that can be used to train Deep Neural Networks. In this work, we explore different venues to obtain such database conforming the ground truth data via massive finite element computer simulations of the so-called forward problem. This consists of solving multiple Boundary Value Problems governed by a Partial Differential Equation with different material coefficients. After describing the Finite Element Method, we investigate a venue to achieve high performance for performing a large amount of simulations using a Fourier approximation based Finite Element Method. The idea is to exploit the orthogonality of Fourier basis functions under reasonable assumptions often satisfied in our geophysics applications to reduce the computational cost of building the corresponding systems of linear equations. Solving such systems requires the use of advanced iterative solvers, which will be analyzed during the Ph.D. studies of Carlos Uriarte.

Contents

Acknowledgments **iii**

Abstract **v**

1 Introduction **1**
1.1 Motivation and background 1
1.2 Main contributions and structure of the thesis 2

2 From Partial Differential Equations to Neural Networks **5**
2.1 Partial Differential Equations and mathematical modeling 5
2.2 Forward and inverse problems 6
2.3 Deep Learning as a method for solving inverse Problems 8
2.3.1 Basic architecture of Neural Networks 8
2.3.2 Construction of Neural Networks 9
2.4 Solving a large number of forward problems 11
2.5 A simple model problem: heat propagation 11

3 A mathematical review of the Finite Element Method **13**
3.1 Basic concepts 13
3.1.1 Variational formulation of Boundary Value Problems 13
3.1.2 Ritz-Galerkin approximation and error estimates 15
3.1.3 Picewise polynomial spaces for the FEM 17
3.2 Sobolev spaces 18
3.2.1 Lebesgue integration theory 18
3.2.2 Weak derivatives 19
3.2.3 Sobolev norms, Sobolev spaces and some properties 20
3.2.4 Review of section 3.1 22
3.3 Variational Formulation of elliptic Boundary Value Problems 22
3.3.1 Inner-product and Hilbert spaces 23
3.3.2 Projections onto subspaces and Riesz's Representation Theorem . . . 24
3.3.3 Formulation of symmetric variational problems 25
3.3.4 Formulation of non-symmetric variational problems 27
3.3.5 Error estimates for the general Finite Element Method 28
3.4 Variational formulation of Poisson's equation BVP 28

4 Fourier summation approximation for Finite Element computations **31**
4.1 Variational Formulation in an arbitrary Coordinate System 31
4.2 Fourier summation approximation 33
4.3 Poisson's equation with Fourier summation approximation 34
4.3.1 General development 34
4.3.2 Example in one dimension 37
4.3.3 Some comments for higher dimension problems 40
4.3.4 Construction of a two dimensional tensor 41
4.3.5 Example in general dimension. Methodology 42
4.3.6 Rapid generation of stiffness matrices 43

5 Conclusions and Future Work **45**
5.1 Review and conclusions 45
5.2 Future work 46

A Construction of a Finite Element space **49**
A.1 Finite Element 49
A.2 Examples of triangular FEs in two dimensions 49
A.3 The interpolant 51

Bibliography **53**

Chapter 1

Introduction

> «There is nothing more difficult to take in hand, more perilous to conduct, or more certain in its success, than to take lead in the introduction of a new order of things.»
>
> *Niccolo Machiavelli*

This chapter aims to motivate the work and present its structure. The first section raises a case of application and exposes the state of the art of the research work. The second section presents the contribution of the developed work and describes the thesis' structure.

1.1 Motivation and background

The exploration of the surface and subsoil is fundamental in today's society. Its knowledge allows us to take measures in: (i) earthquake predictions or seismic hazards [6,31], (ii) mining [38], (iii) geothermal energy production [15], and (iv) massive construction projects [13, 14], among other applications. Surface measurements are routinely acquired through different measurement systems or techniques (e.g., Logging-While-Drilling –LWD–, Ground-Penetrating Radar –GPR–, or Electrical Resistivity Tomography –ERT–). These measurements are subsequently interpreted (inverted) using different numerical methods or machine learning techniques to generate maps of the subsoil composition.

The above application is given by a *Boundary Value Problem* (abbrev., BVP) governed by *Partial Differential Equations* (abbrev., PDEs). Measurements are thus associated with samples of solutions of the PDE, and their corresponding inversions, with parameters involved in the equation. The problem of predicting results in the form of measurements is called *forward problem*, and an *inverse problem* consists in using given measurements to infer the values of the parameters that characterize the system [28,53]. For example, in the aforementioned geoterrestrial application framework, the phenomena correspond to wave propagations, and parameters to characteristics of the materials through which the waves are propagated (e.g., resistivity, electromagnetic permeability, Lame coefficients, etc.) [33].

Thanks to the current computing capabilities, one way to address these inverse problems is through the use of *Deep Learning* architectures [52,53]. The current interest in these computational models is big, growing and with a great scientific prospect. In particular, in the MATHMODE research group of the UPV/EHU there are some of the first proposals of these inversion models for subsoil applications [48]. In addition, this type of models are relatively new and there is still a large room for improvement. The topic that concerns this master's thesis is the creation of synthetic databases so that later they can be used for the creation of inverse problem models based on Deep Learning architectures.

A simple way to generate these databases is to perform simulations (i.e., solve forward problems) with computers when the parameters of the PDEs have been fixed. That is: first we select some parameters, then we perform a simulation for that parameter selection, and as a result, we obtain measurements. In this way, each parameters-measurements pair corresponds to a sample of the database. Repeating this process iteratively a sufficiently large number of times (e.g., performing one million simulations) modifying the choice of parameters, we obtain the desired database. A possible numerical method to carry out these simulations is the Finite Element Method (abbrev., FEM) [26,32]. Performing a single simulation with this method does not generally require excessive execution times in the considered applications (e.g., a few minutes). But if this procedure needs to be repeated a considerable number of times (e.g., one million repetitions), the execution time becomes unaffordable (in the previous example, several years). Therefore, in a context in which problems must be solved in real times of execution, the way to deal with the problem must be different from simply repeating simulations iteratively with the FEM.

Examples of current proposals that favor execution times in Finite Element simulations are: (i) Dimensionally Adaptive Methods (e.g., [2]), (ii) hp-Adaptivity (e.g., [42]), (iii) Proper Generalized Decomposition (abbrev., PGD) (e.g., [61]), and (iv) Alternating Direction Implicit (abbrev., ADI) method (e.g., [10]), among others. Generally, these proposals are usually presented in the scientific literature in a sense of "individual" optimization times rather than in a "global" sense of solving large families of forward problems. This is the starting point and research motivation of the proposed work.

1.2 Main contributions and structure of the thesis

The work has been developed in a preliminary sense with the aim to describe the mathematical and computational frameworks, and to present and analyze the very first proposals of solution methods for the concerned problem. No computational experiments have been carried out in this project. This means that the approach along the thesis is theoretical, although by its nature it is expected to obtain the first numerical results in a short time.

The current work has two main contributions:

(i) **Study and review of the concerned mathematical and computational frameworks, and formalization of the involved mathematical and computational ingredients.** It is divided in chapters 2 and 3. Chapter 2 introduces the mathematical and computational general frameworks more exhaustively. We give formal definitions of essential ingredients such as forward and inverse problems and Deep Learning architectures. At the end of the chapter we motivate and deduce a model problem from a physical phenomenon: Poisson's equation Boundary Value Problem. This model problem is used in the remaining chapters as an illustrative example of the developed theory. In chapter 3, we present, study and describe the FEM from a high mathematical and abstract perspective. In particular, we review the theory of Lebesgue integration, weak derivatives, and Hilbert and Sobolev spaces. Then, we introduce the variational formulation concept and we apply it to the aforementioned model problem. Thereafter, we also establish existence and uniqueness of solutions to the model problem.

(ii) **Proposal and analysis of a Fourier summation based Finite Element Method for massive computations.** This part is contained in chapter 4. In it, we consider the previously worked model problem in Fourier summation terms. Then, we make the necessary calculations so as to express the model problem in a finite element variational formulation. Then, we particularize the calculations for a 1D case and comment some important observations of the obtained results. Later, we generalize the used methodology to an arbitrary dimension Poisson's equation problem. At the end of the chapter, we emphasize the main characteristics of the obtained results and their relation with massive computations.

In the last chapter, we summarize the general and particular conclusions of the developed work, and we present a future research project.

This thesis has been mainly developed at *BCAM - Basque Center for Applied Mathematics* within the *Simulation of Wave Propagation* research group. The current work is intended to be continued by a Ph.D. program in the next academic year at the same center and under the supervision of the same directors of this master's thesis, Dr. David Pardo and Dr. Elisabete Alberdi.

Chapter 2

From Partial Differential Equations to Neural Networks

> «The mathematician plays a game in which he himself invents the rules while the physicist plays a game in which the rules are provided by nature, but as time goes on it becomes increasingly evident that the rules which the mathematician finds interesting are the same as those which nature has chosen.»
>
> *Paul A.M. Dirac*

This chapter provides formal definitions of some critical aspects of the research such as forward and inverse problems, and Neural Networks architectures. It is divided in five sections.

The first and second sections introduce the mathematical framework of the work, and the third section does so with the computational one. The fourth section presents the general problem we aim to solve and the fifth one deduces from a physical phenomenon a particular model problem in which the whole project will later be focused on.

2.1 Partial Differential Equations and mathematical modeling

From the very basic linear equations learned at school to the still unsolved millenium prize Navier-Stokes' equation [12], equations are, without doubt, the quintessential ingredient of mathematics [51].

In simple terms, equations are equalities where some unknowns are involved. Determining which values do solve the considered equality (in case it is solvable), finding general patterns for solving similar model equations, and discussing their solutions' properties, are still some of the main concerns in mathematics.

In this project, we will deal with Partial Differential Equations [11]. This kind of equations specify a relation between functions of more than one variable and their (partial)

derivatives. In mathematical terms: if $u = u(x_1, x_2, \ldots, x_d)$ is a function with d variables, a PDE for u is an equation of the form,

$$\mathcal{D}(x_1, x_2, \ldots, x_d; u; u_{x_1}, u_{x_2}, \ldots, u_{x_d}; u_{x_1x_1}, u_{x_1x_2}, \ldots, u_{x_dx_{d-1}}, u_{x_dx_d}; \ldots) = C, \tag{2.1}$$

for a certain function $\mathcal{D}$, a constant C, and where $u_{x_{i_1}x_{i_2}\cdots x_{i_k}}$ is the abbreviate form of denoting the k-th order derivative of u with respect to the $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ variables, i.e.,

$$u_{x_{i_1}x_{i_2}\cdots x_{i_k}} = \frac{\partial^k u}{\partial x_{i_1}\partial x_{i_2}\cdots\partial x_{i_k}}. \tag{2.2}$$

The *order* of a Partial Differential Equation is the order of the highest derivative involved. A *particular solution* is a function that verifies the PDE. A solution is called *general* if it contains all the particular solutions of the concerned PDE. Moreover, if $\mathcal{D}$ is a linear mapping with respect to u, we say that the corresponding PDE is *linear*.

Many ideas in mathematics arose because of the interest generated in certain physical phenomena. For example, calculus has its origins in efforts to accurately describe the motion of bodies [4]. In consequence, mathematical equations have provided a context to formulate concepts in physics. In recent years, the interest in promoting and using mathematics as a tool to explain phenomena in natural sciences has increased notably because of the desire to understand, and as consequence also be able to take advantage of, the world in which we live in. This effort to understand and interpret the environment that surrounds us from a mathematical point of view has crystallized in a growing field called *mathematical modeling*.

A mathematical model is an equation or a set of equations whose solutions describe the physical behavior of a related physical system [56]; and finding the appropriate model involves physical observation, selection of the relevant physical variables, formulation and analysis of the proposed equations, simulation, and finally, its validation.

In this context, and as we will later see, PDEs are useful for modeling a large class of physical phenomena [58]; and therefore, being able to solve PDEs is a powerful tool to understand the behavior of important physical systems.

2.2 Forward and inverse problems

Given a complete description of a physical system, we can predict the outcome of some measurement functions. This problem of predicting results in the form of measurements is called *forward problem*. An *inverse problem* consists in using the actual measurements to infer the values of some parameters that characterize the system [28, 53].

If we assume that the considered system can be modeled by a partial differential equation, we can describe the forward and inverse problems in the following way: let u be a

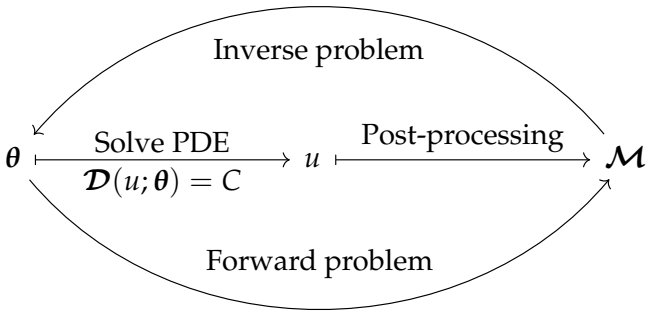

FIGURE 2.1: Forward and inverse problems sketch

function representing a certain physical phenomenon, let $\mathcal{D}$ be a partial differential operator in the sense explained at (2.1), and let $\boldsymbol{\theta}$ be a set of parameters taking part in the considered model. We briefly denote by

$$\mathcal{D}(u;\boldsymbol{\theta}) = C \tag{2.3}$$

to the considered PDE. Then, two kind of problems arise from model (2.3):

- *Forward problem.* Given a set of parameters $\boldsymbol{\theta}$, find a (numerical) solution u of (2.3). Later, the solution u is post-processed in order to extract certain useful information from it. We call *measurements* to the post-processed values of u and denote them by $\boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{M}}(u)$. In some cases, and depending on the context in which the problem is presented, solving the forward problem refers directly to obtain the measurements from the model where the involved parameters are known.

- *Inverse problem.* Given a set of measurements $\boldsymbol{\mathcal{M}}$, find the set of parameters $\boldsymbol{\theta}$ such that when solving equation (2.3) with it, the obtained u solution satisfies $\boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{M}}(u)$.

There is a fundamental difference between forward and inverse problems. The inverse problem is often *ill-posed* in the Hadamard's sense, while the forward problem is *well-posed*. In his lectures [20], Hadamard claims that the mathematical model of a physical problem is well-posed when it satisfies the following three conditions: (i) a solution exists, (ii) it is unique, and (iii) the solution's behavior changes continuously with the initial conditions (given data).

In this sense, the inverse problem is not well-defined since, for a given measurements set, there may not exist a corresponding set of parameters or, as it is more usual, there could exist multiple sets of parameters. These undesirable properties of inverse problems (e.g., see [53, 57]) make them very difficult to treat, since there is not an "effectively clear" procedure to follow in the aim of finding good solutions to them.

In a physical phenomenon problem modeled with PDEs, parameters usually describe properties about the media or material. If in particular we consider the heat equation, the

involved set of parameters could be the thermal conductivity of the media. Being able to characterize those parameters from the existing measurements is the goal of inverse problems; and having these characterizations could later be useful for other fields of knowledge.

2.3 Deep Learning as a method for solving inverse Problems

Let us assume we have a PDE model from which we are able to perform several simulations (solve forward problems) for distinct set of parameters and obtain simulated measurements. Let us denote by $\{\boldsymbol{\theta}_i, \boldsymbol{\mathcal{M}}_i\}_{i=1}^N$ to the obtained database, where N is the number of simulations performed and $\boldsymbol{\mathcal{M}_i}$ denotes the set of measurements obtained in the i-th simulation employing the $\boldsymbol{\theta}_i$ set of parameters.

Then, if N is large enough, it would be possible to use the corresponding database to generate a model that, based on the measurements, provides the associated parameters. Thanks to the current computation capabilities, these kind of problems could be efficiently approximated with an appropriate Deep Learning architecture.

2.3.1 Basic architecture of Neural Networks

In simple terms, a *Neural Network* (abbrev., NN) is a mapping whose architecture has been designed as a composition of multiple mappings, i.e.,

$$\boldsymbol{\mathcal{N}} = \boldsymbol{\mathcal{L}}^{(r)} \circ \cdots \circ \boldsymbol{\mathcal{L}}^{(i)} \circ \cdots \circ \boldsymbol{\mathcal{L}}^{(2)} \circ \boldsymbol{\mathcal{L}}^{(1)}, \tag{2.4}$$

where r is the number of mappings involved in the composition; and if $r \geq 2$, we often refer to it as a *Deep Neural Networks* (abbrev., DNN) instead of a NN.

Each $\boldsymbol{\mathcal{L}}^{(i)}$ is commonly called a *layer* of the NN. In the most fundamental case, when the NN is a *Multilayer Perceptron* (abbrev., MLP) [16, 45], each $\boldsymbol{\mathcal{L}}^{(i)}$ is of the form:

$$\boldsymbol{\mathcal{L}}^{(i)}\left(\boldsymbol{x}^{(i-1)}\right) = \boldsymbol{a}^{(i)}\left(\boldsymbol{W}^{(i)}\boldsymbol{x}^{(i-1)} + \boldsymbol{b}^{(i)}\right), \tag{2.5}$$

where $\boldsymbol{W}^{(i)}$ and $\boldsymbol{b}^{(i)}$ are, respectively, a matrix and a vector, suitable for the input vector $\boldsymbol{x}^{(i-1)}$. Thus, $\boldsymbol{W}^{(i)}\boldsymbol{x}^{(i-1)} + \boldsymbol{b}^{(i)}$ is an affine transformation where the output vector's size, commonly known as *layer's size* or *dimension*, is chosen according to a certain convenient architectural criterion. The $\boldsymbol{a}^{(i)}$ mapping is non-linear and componentwise, commonly called an *activation function*, which typically is a so-known *rectified linear unit* (abbrev., ReLU), *sigmoid* (σ), *hyperbolic tangent* (tanh), etc.

In this way, given an input vector for the MLP, the result is another vector which has been produced according to a series of paired linear and non-linear transformations [17, 48].

Figure 2.2 shows an example of a MLP with 6 layers ($r = 6$), with a 4-dimensional input, 5-dimensional intermediate (hidden) layers, and a 1-dimensional output.

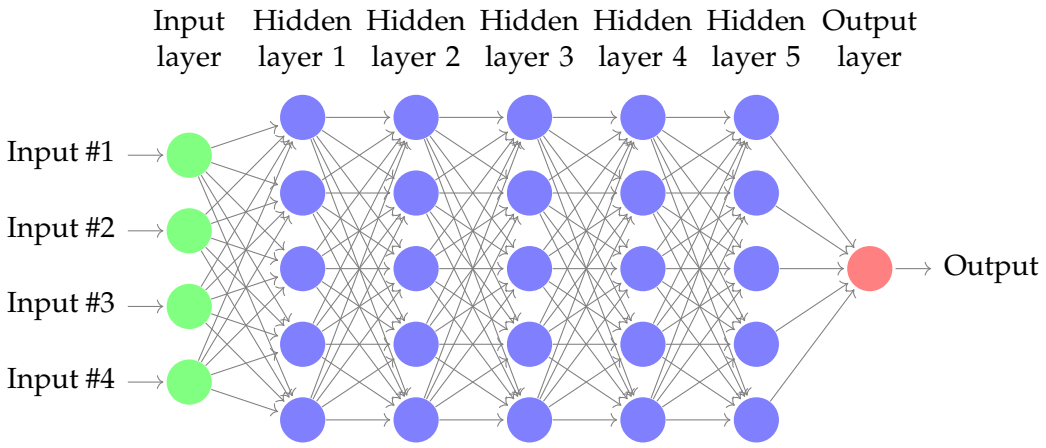

FIGURE 2.2: Illustration of a Multilayer Perceptron

The amount of needed parameters to set (weights —$\boldsymbol{W}^{(i)}$— and biases —$\boldsymbol{b}^{(i)}$—) in figure 2.2 is: $20+5$ for the first layer; $25+5$ for the second, third, fourth and fifth layers; and $5+1$ for the sixth. That is, a total of 151 parameters.

2.3.2 Construction of Neural Networks

Let us consider the database introduced at the beginning of section 2.3, $\{\boldsymbol{\theta}_i, \boldsymbol{\mathcal{M}}_i\}_{i=1}^{N}$. The idea of a NN in our concerned kind of problems is to approximate a hypothetical mapping $\boldsymbol{\mathcal{I}}$ such that $\boldsymbol{\mathcal{I}}(\boldsymbol{\mathcal{M}}_i) = \boldsymbol{\theta}_i$ for all $i = 1, 2, \ldots, N$ [25]. Nevertheless, as we mentioned in section 2.2, the problem we want to address is ill-posed. That is, such mapping is not ensured to be well-defined. However, because of the great adaptation capability NNs have for fitting a dataset, these models have shown in practice to be efficient enough in these tasks so as to obtain very acceptable solutions to problems that are not necesarily well-defined.

The key point for obtaining a good NN model is to set an appropriate *cost* or *loss function* [30]. This function aims at measuring the discrepancy between the solutions obtained by the NN and the real ones. Depending on the magnitude of the discrepancy obtained in each case, the parameters of the NN are re-adjusted in order to minimize such discrepancy.

Gradient descent algorithm

Let ℓ be a real-valued loss function of a given NN. Assume that both ℓ and all non-linear mappings taking part in the NN's architecture are differentiable and easy to derivate. Then, the chain rule allows us to calculate the partial derivatives of ℓ with respect to the adjustable parameters (entries of $\boldsymbol{W}^{(i)}$ and $\boldsymbol{b}^{(i)}$) [55], i.e.,

$$\frac{\partial \ell}{\partial \boldsymbol{W}^{(i)}} \quad \text{and} \quad \frac{\partial \ell}{\partial \boldsymbol{b}^{(i)}}, \qquad \text{for } i = 1, 2, \ldots, r. \tag{2.6}$$

In this way, after initialized the parameters of the NN, forward-passed the data and computed the corresponding derivatives, we re-adjust the parameters toward the negative gradient sense:

$$\boldsymbol{W}^{(i)}_{\text{new}} := \boldsymbol{W}^{(i)}_{\text{current}} - \lambda \frac{\partial \ell}{\partial \boldsymbol{W}^{(i)}}\bigg|_{\text{data}} \quad \text{and} \quad \boldsymbol{b}^{(i)}_{\text{new}} := \boldsymbol{b}^{(i)}_{\text{current}} - \lambda \frac{\partial \ell}{\partial \boldsymbol{b}^{(i)}}\bigg|_{\text{data}}, \tag{2.7}$$

where $\frac{\partial \ell}{\partial \boldsymbol{W}^{(i)}}|_{\text{data}}$ and $\frac{\partial \ell}{\partial \boldsymbol{b}^{(i)}}|_{\text{data}}$ denote the evaluated derivatives of ℓ with respect to the corresponding parameters, and λ is a sufficiently small and positive hyperparameter of the NN, commonly called *learning-rate* [29]. Each time the parameters are re-adjusted with the backwards criterion, it is said that the NN is *backpropagated* [3, 23].

The main idea in this method is that, since all the employed functions are differentiable (except perhaps at some points of the activation function whose derivative is approximated), we are able to minimize the loss function by following small steps against the gradient. Following this path allows us to approximate to a zero-valued point of the derivative (in case it exists), where if the cost function has been conveniently posed, it will coincide with a local minimum of the loss function.

Training and validating Neural Networks

The most common way to train and validate the NN requires a data partition into three sets: *training*, *validating* and *testing sets*.

The first set is used to iteratively re-adjust the parameters. If the training data is partitioned into smaller (and possibly equal sized) subsets, known as *batches*, and the backpropagation is performed after the forward-pass of each batch, we say that the gradient-based algorithm is a *stochastic gradient descent* [3, 60].

The validation data is used to perform some high-level NN design decision, e.g., to modify the network architecture (layers' dimensions) [27] or numerical optimization criteria. For example, since NNs have a great capability to adapt to the training data (memorization), in order to avoid a disproportionate adjustment in detriment of the generalization capacity (*overfitting*) [22, 54], the validation data is employed to control it and stop if necessary, e.g., when the NN does not improve significantly the accuracy of the validation data when the trainig data does so (*early-stopping*) [43].

The testing data is employed only when the final model is constructed (no more parameters are re-adjusted). We say that the NN generalizes properly when the accuracy of forward-passing the testing set is similar to the accuracy obtained when doing it with the training and validating sets.

2.4 Solving a large number of forward problems

In section 2.3 we briefly explained one of the current methods for solving inverse problems, provided a paired database of parameters and measurements is given.

The main concern is then that, for applying the previous method, a huge amount of forward problems need to be solved, which in essence it reduces to solving the associated PDE problem modifying in each case the corresponding parameters set $\boldsymbol{\theta}_i$.

Traditional methods of application in engineering for solving PDE problems are the Finite Difference Method (abbrev., FDM) [49] and the Finite Element Method (abbrev., FEM) [19, 59]. The latter is the most used in academia and we focus on it on this work. The FEM partitions a possibly irregular domain in a finite number of more simple subdomains, commonly called *elements*. Next, a set of approximating functions are constructed systematically for each of these subdomains in order to approximate the solution of the given problem with a desired degree of precision.

2.5 A simple model problem: heat propagation

Let $\Omega \subset \mathbf{R}^d$ be our domain of interest in a d spatial dimension and let $u : \Omega \longrightarrow \mathbf{R}$ be the temperature at each point $x = (x_1, x_2, \ldots, x_d) \in \Omega$. Let $\mathbf{q} : \Omega \longrightarrow \mathbf{R}^d$ be the heat flux in the domain and let $f : \Omega \longrightarrow \mathbf{R}$ be the heat source. If $\omega \subset \Omega$ is a small test volume, conservation of energy principle gives:

$$\frac{dE}{dt} = \int_{\partial\omega} \mathbf{q} \cdot \mathbf{n} \, ds - \int_{\omega} f \, dx = 0, \tag{2.8}$$

where $\mathbf{n} \in \mathbf{R}^d$ denotes the outer normal vector in Ω and $\cdot : \mathbf{R}^d \times \mathbf{R}^d \longrightarrow \mathbf{R}$ denotes the Euclidean scalar product. In other words, equation (2.8) represents that the outflow of energy over the boundary $\partial\omega$, $\mathbf{q} \cdot \mathbf{n}$, equals the energy emitted by the heat source function, f. Fourier's law relates this heat flux to the temperature in the following way:

$$\mathbf{q} = -\sigma \nabla u, \tag{2.9}$$

where $\nabla := (\partial_{x_1}, \partial_{x_2}, \cdots, \partial_{x_d})^T$ is the gradient operator and $\sigma = \sigma(x_1, x_2, \ldots, x_d) \in \mathbf{R}^{d \times d}$ is a coefficients matrix according to the domain heat propagation properties. Therefore,

$$\int_{\partial\omega} -\sigma \nabla u \cdot \mathbf{n} \, ds = \int_{\omega} f \, dx. \tag{2.10}$$

Because of *Gauss' Divergence Theorem*, we can rewrite this expression as follows:

$$\int_{\partial\omega} -\sigma \nabla u \cdot \mathbf{n} \, ds = \int_{\omega} \nabla \cdot (-\sigma \nabla u) \, dx \implies \int_{\omega} (-\nabla \cdot (\sigma \nabla u) - f) \, dx = 0. \tag{2.11}$$

Since equality (2.11) holds for all test volumes $w \subset \Omega$, if u, σ and f are regular enough, we recover the so-called *Poisson's Equation* for the heat propagation:

$$-\nabla \cdot (\sigma \nabla u) = f \quad \text{in } \Omega. \tag{2.12}$$

Now assume that $\partial\Omega = \partial\Omega_N \dot{\cup}\, \partial\Omega_D$ and that some information about the temperature (and its derivative) is known on these two subboundaries (i.e., sections of the boundary):

$$u = u_D \quad \text{in } \partial\Omega_D, \tag{2.13}$$

$$-\sigma \nabla u \cdot \mathbf{n} = g \quad \text{in } \partial\Omega_N, \tag{2.14}$$

with $u_D : \partial\Omega_D \longrightarrow \mathbf{R}$ and $g : \partial\Omega_N \longrightarrow \mathbf{R}$. This way, the problem of knowing the temperature at each point of the body is reduced to solving the following Boundary Value Problem:

$$-\nabla \cdot (\sigma \nabla u) = f \quad \text{in } \Omega, \tag{2.15}$$

$$u = u_D \quad \text{in } \partial\Omega_D, \tag{2.16}$$

$$-\sigma \nabla u \cdot \mathbf{n} = g \quad \text{in } \partial\Omega_N. \tag{2.17}$$

Equations (2.16) and (2.17) are known as *Dirichlet* and *Neumann boundary conditions*, respectively. Similarly, $\partial\Omega_D$ and $\partial\Omega_N$ are the corresponding Dirichlet and Neumann boundaries of the BVP.

This model problem will be exhaustively used in the subsequent chapters.

Chapter 3

A mathematical review of the Finite Element Method

«In theory, there is no difference between theory and practice. But in practice, there is.»

Albert Einstein

The Finite Element Method employs a formalism for generating discrete (finite) algorithms for approximating the solutions of differential equations. It is a method that given a Partial Differential Equation, it delivers an approximation of the corresponding solution. Such a task could conceivably be done automatically by a computer, but it requires an amount of mathematical skill that today still requires human involvement. The purpose of this chapter is to help on the comprehension of this method, presenting (and sometimes justifying) its mathematical bases.

3.1 Basic concepts

In this section we present and develop a one-dimensional BVP. We leave many loose ends (indicated in footnotes along the section), which most of them will be tied up in the subsequent sections, in order to motivate and briefly present the necessary ingredients so as to later properly introduce the FEM.

3.1.1 Variational formulation of Boundary Value Problems

Consider the following two-point BVP (case $d = 1$ of Poisson's Equation):

$$\begin{cases} -(\sigma\, u')' = f \quad \text{in } [0,1] \\ u(0) = 0, \quad (\sigma u')(1) = 0, \end{cases} \tag{3.1}$$

where $u : [0,1] \longrightarrow \mathbf{R}$ is the solution and $\sigma, f : [0,1] \longrightarrow \mathbf{R}$ are given. Then, if v is any (sufficiently regular) function such that $v(0) = 0$, integration by parts yields:

$$\int_0^1 -(\sigma\, u')'\, v\, dx = \int_0^1 f\, v\, dx, \tag{3.2}$$

$$\int_0^1 \sigma\, u'\, v'\, dx = \int_0^1 f\, v\, dx. \tag{3.3}$$

We denote by $\mathfrak{L}(u,v)$ and by $\mathfrak{R}(f,v)$ to the left and right hand side of (3.3) respectively. Let us define (formally for the moment since the notion of derivative has not been made precise)*

$$V := \{v \in L^2([0,1])^\dagger : \mathfrak{L}(v,v) < \infty \text{ and } v(0) = 0\}. \tag{3.4}$$

Then, we can characterize the solution u of (3.1) —in a necessary condition sense— by

$$u \in V \quad \text{such that} \quad \mathfrak{L}(u,v) = \mathfrak{R}(f,v) \quad \text{for all } v \in V, \tag{3.5}$$

which is called the *weak* or *variatiational formulation* of (3.1). The relationship (3.5) is called "variational" because the function v is allowed to vary arbitrarily. It seems somewhat unusual at first, but later we will see that it has a natural interpretation in the setting of Hilbert spaces.

The following theorem states that, under some strong regularity assumptions, finding u verifying (3.5) provides a solution for (3.1).

Theorem 3.1. *Let $f \in C^0([0,1]), \sigma \in C^1([0,1]), \sigma > 0$ and $u \in C^2([0,1])$ verifying* (3.5)*. Then, u solves* (3.1)*.*

Proof. Let $v \in V \cap C^1([0,1])$. Then, since $\mathfrak{L}(u,v) = \mathfrak{R}(f,v)$, integration by parts gives:

$$\mathfrak{R}(f,v) = \int_0^1 -(\sigma\, u')'\, v + \sigma(1)\, u'(1)\, v(1). \tag{3.6}$$

Thus, if we write $w = f + (\sigma\, u')'$, we get $\mathfrak{R}(w,v) = 0$ for all $v \in V \cap C^1([0,1])$ such that $v(1) = 0$. If $w \not\equiv 0$, then there exists an interval $[x_0, x_1] \subset [0,1]$ where $w(x)$ maintains the sign (because $w \in C^0([0,1])$). Take $v(x) = (x - x_0)^2(x - x_1)^2$ in $[x_0, x_1]$ and $v \equiv 0$ otherwise. Since $\mathfrak{R}(w,v) \neq 0$, we arrive to a contradiction. Therefore $w \equiv 0$, or equivalently, $-(\sigma u')' = f$. Now, we apply (3.6) with $v(x) = x$ to obtain $(\sigma u')(1) = 0$ and see that $u(0) = 0$ because $u \in V$. □

Assumptions $f \in C^0([0,1]), \sigma \in C^1([0,1])$ and $u \in C^2([0,1])$ in the theorem allow us to interpret (3.1) in the *classic* sense (*strong formulation*). Nevertheless, these requirements for the solution and the given functions are not always possible to assure. Therefore, the variational formulation (3.5) will provide an alternative interpretation to the solution of (3.1) that, on the one hand, it requires less restrictive assumptions on the concerned functions, and on the other hand, it generalizes the approach to the same problem from a weaker perspective (i.e., it maintains the strong formulation interpretation when the involved functions meet the continuity/derivability conditions).

*Later we will see that $\emptyset \neq V \subsetneq \{v \in L^2([0,1]) : \mathfrak{L}(v,v) < \infty \text{ and } v(0) = 0\}$.

†In section 3.2 we will formally define the L^p spaces for $p \geq 1$.

3.1.2 Ritz-Galerkin approximation and error estimates

Let $S \subset V$ be a finite dimensional space and let us consider (3.5) replacing V by S:

$$u_S \in S \quad \text{such that} \quad \mathfrak{L}(u_S, v) = \mathfrak{R}(f, v) \quad \text{for all } v \in S. \tag{3.7}$$

We will now see that (3.7) does actually define an object. In fact, we will show that (3.7) can be represented as a system of equations and prove that u_S is the best approximation to u. The existence and uniqueness of u_S is given in the following theorem:

Theorem 3.2. *Let $f \in L^2([0,1])$ and $\sigma > 0$. Then,* (3.7) *has a unique solution.*

Proof. Let $\{\phi_i\}_{i=1}^n$ be a basis for S and let us write $u_S = \sum_{j=1}^n U_j \phi_j$. Let $K_{ij} = \mathfrak{L}(\phi_j, \phi_i)$ and let $F_i = \mathfrak{R}(f, \phi_i)$ for $i, j \in \{1, 2, \dots, n\}$. Set $\mathbf{U} = [U_j], \mathbf{K} = [K_{ij}]$ and $\mathbf{F} = [F_i]$. Then, solving (3.7) is equivalent to solving $\mathbf{KU} = \mathbf{F}$. Since it represents a finite dimensional square linear system, uniqueness and existence are equivalent. If the solution was not unique, it would imply that there exists a non-zero $\mathbf{V} = [V_j]$ such that $\mathbf{KV} = \mathbf{0}$. We write $v = \sum_{j=1}^n V_j \phi_j$ and check that the previous homogeneous system implies $\mathfrak{L}(v, \psi_j) = 0$ for all $j = 1, 2, \dots, n$ because of the linearity of $\mathfrak{L}(\cdot, \cdot)$ in the first component. Noticing that $\mathfrak{L}(\cdot, \cdot)$ is also linear on the second component, multiplying each term by V_j and summing over j yields $0 = \mathfrak{L}(v, v) = \int_0^1 \sigma\, (v')^2 dx$. Then $v' \equiv 0$ and since $v(0) = 0$, we conclude $v \equiv 0$.[‡] That is, $\mathbf{V} = \mathbf{0}$, which means that the solution to $\mathbf{KU} = \mathbf{F}$ is unique (and hence it must exist). Equivalently, the solution u_S to (3.7) exists and it is unique. □

Observation 3.1. The matrix $\mathbf{K}$ is often referred to as the *stiffness* matrix, a name coming from the context of structural problems. It is clearly symmetric since so is $\mathfrak{L}(\cdot, \cdot)$.

Now let us check the Galerkin orthogonality relation between u and u_S. To do so, we check that

$$\mathfrak{L}(u - u_S, w) = \mathfrak{L}(u, w) - \mathfrak{L}(u_S, w) = \mathfrak{R}(f, w) - \mathfrak{R}(f, w) = 0, \quad \text{for all } w \in S. \tag{3.8}$$

We define by $\|v\|_E := \sqrt{\mathfrak{L}(v, v)}$ for $v \in V$ to the *energy norm*. An interesting relationship between the energy norm and $\mathfrak{L}$ is thanks to Schwarz's inequality[§]: $|\mathfrak{L}(v, w)| \leq \|v\|_E \, \|w\|_E$ for all $v, w \in V$. Then, for any $v \in S$, we have:

$$\|u - u_S\|_E^2 = \mathfrak{L}(u - u_S, u - u_S) = \mathfrak{L}(u - u_S, u - v) + \mathfrak{L}(u - u_S, v - u_S) = \tag{3.9}$$

$$= \mathfrak{L}(u - u_S, u - v) \leq \|u - u_S\|_E \, \|u - v\|_E \,. \tag{3.10}$$

[‡]Why $v' \equiv 0$ implies v being constant? For those familiar with the Cantor function, whose derivative is zero almost everywhere but it is certainly not constant, the previous implication is not sufficiently clear. However, V is a subspace of a Sobolev space and we will see that such situations do not occur in these spaces.

[§]Schwarz's inequality is stated and proved in section 3.3 under the hypothesis of the corresponding linear functional being an inner-product. Is $\mathfrak{L}(\cdot, \cdot)$ an inner product in V? No. However, as it will later be clarified, Schwarz's inequality also holds in a lesser restrictive kind of functionals than inner-products.

If $\|u-u_S\|_E^2 \neq 0$, we divide inequality (3.10) by it to obtain $\|u-u_S\|_E \leq \|u-v\|_E$ for any $v \in S$. If $\|u-u_S\|_E^2 = 0$, this inequality is trivial. Taking the infimum over $v \in S$ yields

$$\|u-u_S\|_E \leq \inf\{\|u-v\|_E : v \in S\}. \tag{3.11}$$

The converse inequality is trivial because $u_S \in S$. Therefore,

$$\|u-u_S\|_E = \inf\{\|u-v\|_E : v \in S\}. \tag{3.12}$$

Since there is an element for which the infimum is attained, u_S, we have proved the following result:

Theorem 3.3. *The solution u_S of* (3.7) *verifies* $\|u-u_S\|_E = \min\{\|u-v\|_E : v \in S\}$.

This is the basic error estimate for the Ritz-Galerkin method, which states that the error is optimal in the energy norm. Now we consider another norm and study its error estimate: let $\|v\|_{L^2} := \sqrt{\mathfrak{R}(v,v)} = \sqrt{\int_0^1 v^2 dx}$ be the $L^2([0,1])$ norm. In an intuitive approach, one could think that the L^2-norm is weaker than the energy norm, as the latter is a kind of L^2-norm of the derivative.

To estimate $\|u-u_S\|_{L^2([0,1])}$, we proceed using what it is known as a "duality" argument: let w be the solution of $-(\sigma w')' = u-u_S$ on $[0,1]$ with $w(0) = (\sigma w')'(1) = 0$. Then, integration by parts lets us find:

$$\|u-u_S\|_{L^2([0,1])}^2 = \mathfrak{R}(u-u_S, -(\sigma w')') = \mathfrak{L}(u-u_S, w) = \mathfrak{L}(u-u_S, w-v), \tag{3.13}$$

for all $v \in S$ because (3.8). Thus, assuming $\|u-u_S\|_{L^2([0,1])} \neq 0$ (the case equal to zero is trivial), Schwarz's inequality implies

$$\|u-u_S\|_{L^2([0,1])} \leq \frac{\|u-u_S\|_E \|w-v\|_E}{\|u-u_S\|_{L^2([0,1])}} \tag{3.14}$$

$$\leq \frac{\|u-u_S\|_E \|w-v\|_E}{\|(\sigma w')'\|_{L^2([0,1])}}. \tag{3.15}$$

Taking the infimum over $v \in S$, we obtain

$$\|u-u_S\|_{L^2([0,1])} \leq \frac{\|u-u_S\|_E}{\|(\sigma w')'\|_{L^2([0,1])}} \inf\{\|w-v\|_E : v \in S\}. \tag{3.16}$$

Now assume that we can take $v \in S$ close to $w \in V$ in the following sense:

$$\inf\{\|w-v\|_E : v \in S\} \leq \epsilon \left\|(\sigma w')'\right\|_{L^2([0,1])} \quad \text{for some } \epsilon > 0. \tag{3.17}$$

Then, we conclude:

$$\|u-u_S\|_{L^2([0,1])} \leq \epsilon \|u-u_S\|_E. \tag{3.18}$$

Retaking (3.17), replacing w by u, considering theorem 3.3 and recalling (3.1), we obtain:

$$\|u - u_S\|_E \leq \epsilon \left\|(\sigma u')'\right\|_{L^2([0,1])} = \epsilon \|f\|_{L^2([0,1])} . \tag{3.19}$$

The following theorem summarizes the error estimate relationships developed so far.

Theorem 3.4. *Assumption* (3.16) *implies*

$$\|u - u_S\|_{L^2} \leq \epsilon \|u - u_S\|_E \leq \epsilon^2 \|f\|_{L^2} . \tag{3.20}$$

The point is that $\|u - u_S\|_E$ is of order ϵ whereas $\|u - u_S\|_{L^2}$ is of order ϵ^2. In the following subsection we will introduce a family of spaces S for which ϵ may be arbitrarily small.

3.1.3 Picewise polynomial spaces for the FEM

Let $0 = x_0 < x_1 < \cdots < x_n = 1$ and let S be the set of picewise linear space functions over $[0,1]$ such that: (i) $v \in C^0([0,1])$, (ii) $v|_{[x_{i-1},x_i]}$ is a linear polynomial for $i = 1, 2, \ldots, n$, and (iii) $v(0) = 0$.[¶]

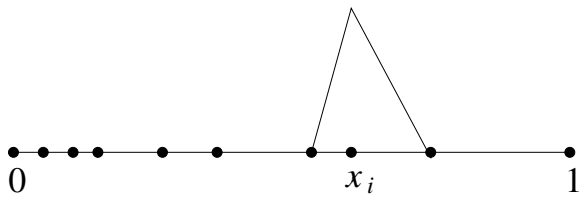

FIGURE 3.1: Piecewise linear basis function ϕ_i

Define $\phi_i \in S$ for each $i = 1, 2, \ldots, n$ by the requirement $\phi_i(x_j) = \delta_{ij}$, i.e., $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise (δ_{ij} is known as *Kronecker's delta*). Then:

Theorem 3.5. $\{\phi_i\}_{i=1}^n$ *is a basis for* S.

Proof. We need to see that $\{\phi_i : i = 1, 2, \ldots, n\}$ is both a linearly independent and a spanning set of S. It is linearly independent because $\sum_{i=1}^n c_i \phi_i(x_j) = 0$ implies $c_j = 0$. To see it spans S, we consider an arbitrary $v \in S$ and define $v_I = \sum_{i=1}^n v(x_i)\phi_i \in S$. Since $v - v_I \in S$ is linear on each $[x_{i-1}, x_i]$ and equal to zero at the endpoints, it must be identically zero, i.e., $v \equiv v_I$. □

In our case, each segment $[x_{i-1}, x_i]$ is called an *element* and the set of all the elements, $\{[x_{i-1}, x_i] : i = 1, 2, \ldots, n\}$, is known as a *grid* for $[0,1]$. The x_i endpoints are called *nodes*, the set $\{\phi_i\}_{i=1}^n$ is a *nodal basis* for S, and given any $v \in S$, the values $v(x_i)$ are called the *nodal values* of v. In general, when v does not belong to S, we call the *interpolant* of v by S to the function defined by $v_I := \sum_{i=1}^n v(x_i)\phi_i \in S$.

The following theorem connects the relation between the error estimate of (3.20) and the grid size.

[¶]We will later show that $S \subset V$.

Theorem 3.6. *Let* $h = \max_{1\leq i\leq n}\{x_i - x_{i-1}\}$. *Then,*

$$\|u - u_S\|_E \leq \frac{h}{\sqrt{2}} \left\|(-\sigma u')'\right\|_{L^2([0,1])} = \frac{h}{\sqrt{2}} \|f\|_{L^2([0,1])}, \qquad \text{for all } u \in V. \tag{3.21}$$

Proof. See [7, p. 7-9] □

Noticing that the above inequality's upper bound depends on h, we realize that we can obtain the desired accuracy on $\|u - u_S\|_E$ just by adjusting the grid size, i.e., considering a properly distributed (e.g., uniformly) large number of nodes.

3.2 Sobolev spaces

This section introduces proper function spaces that are used in the variational formulations of differential equations. We begin with a review of Lebesgue integration theory, upon which the notion of "variational" or "weak" derivatives rests. Functions with such "generalized derivatives" conform the so-called Sobolev spaces.

3.2.1 Lebesgue integration theory

We start reviewing the basic concepts of Lebesgue integration, [21, 44, 46]. In what follows, by "domain" we refer to a Lebesgue-measurable subset of $\mathbf{R}^d$ with non-empty interior and by "function" to a real valued and Lebesgue-measurable function. Then, for a function f with these characteristics we denote by

$$\int_\Omega f(x)dx \tag{3.22}$$

to its *Lebesgue integral* over the domain Ω (dx denotes the Lebesgue measure). We denote by

$$\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f(x)|^p dx\right)^{1/p}, \qquad \text{if } 1 \leq p < \infty, \tag{3.23}$$

$$\|f\|_{L^\infty(\Omega)} := \sup\{|f(x)| : x \in \Omega\} \tag{3.24}$$

to the L^p*-functionals* of f over Ω and by

$$L^p(\Omega) := \{f : \Omega \longrightarrow \mathbf{R} : \|f\|_{L^p(\Omega} < \infty\} \tag{3.25}$$

to its associated *Lebesgue spaces*.

To avoid "almost everywhere" differences, we identify couples of functions f and g as "the same" (with respect to the concerned L^p-functional) when they satisfy $\|f - g\|_{L^p(\Omega)} = 0$. With a small ambiguity of notation, we think of $L^p(\Omega)$ as a set of equivalence classes of functions with respect to this identification.

We cite below some famous and useful inequalities that hold for the above defined functionals:

$$\|f+g\|_{L^p(\Omega)} \leq \|f\|_{L^p(\Omega)} + \|g\|_{L^p(\Omega)}\,, \tag{3.26}$$

$$\|f\,g\|_{L^1(\Omega)} \leq \|f\|_{L^p(\Omega)}\,\|g\|_{L^q(\Omega)}\,. \tag{3.27}$$

In both cases, the hypotheses are that the functions taking part in the right hand side of the inequalities belong to the corresponding Lebesgue space (i.e., have a finite functional value), $1 \leq p \leq \infty$ and $1/p + 1/q = 1$. Inequality (3.26) is known as *Minkoski's inequality* and (3.27) is called *Hölder's inequality*. The particular case $p = q = 2$ of (3.27) is more commonly known as *Schwartz's inequality*, and later its equivalent form will be proved in the framework of inner-product spaces.

One interesting consequence of (3.26) is that it shows that Lebesgue spaces are closed under linear combinations, and thereby, it proves that they are vector spaces. Moreover, the L^p-functionals have the sufficient properties to endow the L^p-spaces of being *normed spaces*. In addition, we obtain that the L^p-spaces are *complete* for $1 \leq p \leq \infty$ (i.e., every Cauchy sequence is convergent). Complete normed spaces are called *Banach spaces*, and thereby, so are L^p-spaces endowed with L^p-norms. This is the key reason that "explains" why the Lebesgue integral is preferred over the Riemann integral: limits of integrable functions are integrable in the Lebesgue's sense, but this is not always the case in the Riemann's sense.

3.2.2 Weak derivatives

The so-typical definition of derivative for a univariate function $u(x)$ is

$$u'(x) = \lim_{h\to 0} \frac{u(x+h) - u(x)}{h}. \tag{3.28}$$

As we may see, this definition is "local" in the sense that it deals with each x point and its proximities. The variational formulation developed in section 3.1 takes a more global point of view: pointwise values of derivatives are not needed at all points. In the previous section, we have seen that pointwise values of functions in Lebesgue spaces are irrelevant since "almost everywhere" differences are omitted. Hence, it is sensible to develop a global notion of derivative, more suitable to Lebesgue spaces. To do so, we introduce the following three ingredients:

- A *multi-index* is a d-tuple $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ with $\alpha_i \in \mathbf{N} \cup \{0\}$ for each $i = 1, 2, \ldots, d$. Its *order* is given by $|\alpha| := \sum_{i=1}^{d} \alpha_i$. Then, if $f = f(x)$ with $x = (x_1, x_2, \ldots, x_d)$ is a sufficiently differentiable function (for the moment, for the classical sense), we denote its *α-th derivative* by

$$\partial^\alpha f := \frac{\partial^\alpha f}{\partial x^\alpha} := \left(\frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \circ \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \circ \cdots \circ \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} \right)(f). \tag{3.29}$$

  Note that the order of this derivative is given by the order of α, $|\alpha|$.

- Given a function f defined over a domain Ω, we call *support* of f to the set

$$\operatorname{supp} f := \overline{\{x \in \Omega : f(x) \neq 0\}}, \tag{3.30}$$

  where the top bar denotes the closure. If this set is bounded and it is a subset of the interior of Ω, we say that the support is *compact*. When Ω is bounded, it is equivalent to say that f vanishes in a neighborhood of $\partial\Omega$. We denote by $C_0^\infty(\Omega)$ to the set of functions in $C^\infty(\Omega)$ with compact support.

- We restrict to the following set of *locally integrable* functions (for a more general definition, see [47]):

$$L^1_{\text{loc}}(\Omega) := \left\{ f : \Omega \longrightarrow \mathbf{R} : f_{|_K} \in L^1(K) \text{ for all } K \subset \operatorname{int}(\Omega) \text{ with } K \text{ compact} \right\}. \tag{3.31}$$

Under these conditions, we define weak derivatives as follows: we say $f \in L^1_{\text{loc}}(\Omega)$ is *weakly derivable* when there exists a function $g \in L^1_{\text{loc}}(\Omega)$ such that

$$\int_\Omega g(x)\phi(x)\, dx = (-1)^{|\alpha|} \int_\Omega f(x)\partial^\alpha \phi \, dx, \qquad \text{for all } \phi \in C_0^\infty(\Omega). \tag{3.32}$$

If this is the case, we define by $\partial^\alpha_{\text{w}} f := g$ to the *weak derivative* of f.

The following fact shows that this definition is a generalization of the classical notion of derivative.

Theorem 3.7. *Let α be a multi-index according to the function $f \in C^{|\alpha|}(\Omega)$. Then, f is weakly derivable and $\partial^\alpha_w f$ is given by $\partial^\alpha f$.*

As a consequence of this theorem, we ignore henceforth the differences in notation between $\partial^\alpha_{\text{w}} f$ and $\partial^\alpha f$, denoting always the derivative by the latter. That is, the derivative symbol will refer to weak derivatives in general, but with the possibility of interpreting them with the classical definition if the function is derivable in that sense.

3.2.3 Sobolev norms, Sobolev spaces and some properties

Using this new notion of (general/weak) derivative, we generalize the Lebesgue norms and spaces in order to include the derivatives.

Let $k \geq 0$ and let $f \in L^1_{\text{loc}}(\Omega)$. Suppose that the (weak) derivatives $\partial^\alpha f$ exist for all α such that $|\alpha| \leq k$. Then, we define by

$$\|f\|_{W^k_p(\Omega)} := \left(\sum_{|\alpha| \leq k} \|\partial^\alpha f\|^p_{L^p(\Omega)} \right)^{1/p}, \quad 1 \leq p < \infty, \tag{3.33}$$

$$\|f\|_{W^k_\infty(\Omega)} := \max \left\{ \|\partial^\alpha f\|_{L^\infty(\Omega)} : |\alpha| \leq k \right\}, \tag{3.34}$$

to the *Sobolev W^k_p-functional*. In each case, we define the associated *Sobolev space* via

$$W^k_p(\Omega) := \left\{ f \in L^1_{\text{loc}}(\Omega) : \|f\|_{W^k_p(\Omega)} < \infty \right\}. \tag{3.35}$$

As it happens with Lebesgue spaces, it is easy to see that the W_p^k-functionals are *norms*. Therefore, Sobolev spaces are *normed vector spaces* and it can also be checked that they are *Banach spaces*.

There is an alternative potential definition of Sobolev spaces for $1 \leq p < \infty$: the corresponding closures of $C^k(\Omega)$ with respect to the W_p^k-functional. In the case $p = \infty$, that closure coincides with $C^k(\Omega)$, which is not the same as $W_\infty^k(\Omega)$. We state below some inclusion relations among Sobolev spaces:

Theorem 3.8. *Let Ω be a domain, let $1 \leq p \leq q \leq \infty$ be real numbers and let $0 \leq k \leq m$ be integers. Then,*

$$W_p^m(\Omega) \subset W_p^k(\Omega), \tag{3.36}$$

$$W_q^k(\Omega) \subset W_p^k(\Omega). \tag{3.37}$$

The following result was proved in [37] and shows that infinitely-derivable continuous functions are dense in Sobolev spaces:

Theorem 3.9. *Let Ω be an open set. Then $C^\infty(\Omega) \cap W_p^k(\Omega)$ is dense in $W_p^k(\Omega)$ for $1 \leq p < \infty$, i.e., for each $f \in W_p^k(\Omega)$ and each $\varepsilon > 0$, there exists some $g \in C^\infty(\Omega) \cap W_p^k(\Omega)$ (depending on f and ϵ) such that $\|f - g\|_{W_p^k(\Omega)} < \epsilon$.*

Unfortunately, this result does not hold in general when Ω is not open, e.g., with $C^\infty(\overline{\Omega})$ and $\emptyset \subsetneq \Omega \subsetneq \mathbf{R}^d$. The density does not happen whenever part of the boundary belongs to the domain, as it occurs with a slit domain that is frequently used to model crack propagation problems. In order for this stronger density result to hold, some sort of regularity conditions must also hold. For example, it is known to be valid when Ω satisfies the *segment condition*, i.e., if for each $x \in \partial\Omega$ there exists an open ball B_x centered at x and a non-zero vector n_x such that if $z \in \overline{\Omega} \cap B_x$, then $z + tn_x \in \Omega$ for $t \in (0, 1)$ [1].

The following kind of domains satisfy the segment condition, and thereby, the density of the infinitely derivable functions is ensured in them. A bounded domain $\Omega \subset \mathbf{R}^d$ is said to have a *Lipschitz boundary* $\partial\Omega$ if there exist constants $\alpha, \beta > 0$, a finite number of local coordinate system $\{(x_1^r, x_2^r, \cdots, x_d^r)\}_{r=1}^R$ and local Lipschitz continuous mappings

$$a_r : \{\hat{x}^r = (x_2^r, \cdots, x_d^r) \in \mathbf{R}^{d-1} : |x_i^r| \leq \alpha \text{ for } 2 \leq i \leq d\} \longrightarrow \mathbf{R}, \qquad 1 \leq r \leq R, \tag{3.38}$$

such that

$$\partial\Omega = \bigcup_{r=1}^{R} \{(x_1^r, \hat{x}^r) : x_1^r = a_r(\hat{x}^r) \text{ with } |\hat{x}^r| < \alpha\}, \tag{3.39}$$

$$\{(x_1^r, \hat{x}^r) : a_r(\hat{x}^r) < x_1^r < a_r(\hat{x}^r) + \beta \text{ with } |\hat{x}^r| < \alpha\} \subset \Omega, \quad 1 \leq r \leq R, \tag{3.40}$$

$$\{(x_1^r, \hat{x}^r) : a_r(\hat{x}^r) - \beta < x_1^r < a_r(\hat{x}^r) \text{ with } |\hat{x}^r| < \alpha\} \subset \mathbf{R}^d \setminus \overline{\Omega}, \quad 1 \leq r \leq R. \tag{3.41}$$

Roughly speaking, a domain $\Omega \subset \mathbf{R}^d$ is called a *Lipschitz domain* (with a Lipschitz boundary) if its boundary can be locally represented by a Lipschitz continuous function, i.e., if for any $x \in \partial\Omega$, there exists an open neighborhood of x, $U \subset \mathbf{R}^d$, such that $U \cap \partial\Omega$ is the graph

of a Lipschitz continuous function under a proper local coordinate system. Of course, all smooth domains are Lipschitz. Significant non-smooth and Lipschitz domain examples are polygonals in $\mathbf{R}^2$ or polyhedrons in $\mathbf{R}^3$. A more interesting example are convex domains in $\mathbf{R}^d$. A simple example of non-Lipschitz domains is two polygons touching at one vertex only.

The following theorem, that holds for Lipschitz domains, provides a rule for "viewing" sufficiently weak derivable functions as continuous and bounded. For a proof of this result, as well as more details concerning other material in this sections, see [50]:

Theorem 3.10 (Sobolev's Inequality)**.** *Let $\Omega \subset \mathbf{R}^d$ be a Lipschitz domain, let $k \in \mathbf{N}$ and let $1 \leq p < \infty$ such that $k \geq d$ if $p = 1$ or $k > d/p$ if $p > 1$. Then there exists a constant C such that for every $u \in W_p^k(\Omega)$ it satifies*

$$\|u\|_{L^\infty(\Omega)} \leq C \ \|u\|_{W_p^k(\Omega)}. \tag{3.42}$$

Furthermore, there is a continuous function in the $L^\infty(\Omega)$ equivalence class of u.

3.2.4 Review of section 3.1

At this point, we can tie up many loose ends of the previous section. First, we see that the V space introduced there can now rigorously be set as (reduced to)

$$V := \{v \in W_2^1(\Omega) : v(0) = 0\}, \tag{3.43}$$

with $\Omega = [0,1]$. We check that this makes sense since Sobolev's inequality ($d = 1, k = 1, p = 2$) guarantees that pointwise values are well defined for functions in $W_2^1(\Omega)$.

The derivation of the variational formulation of (3.3) is now also rigorous in the setting of weak derivatives, i.e., $u' = \partial_w^1 u$ and $v' = \partial_w^1 v$. Moreover, it is easy to check that piecewise linear functions have piecewise constant and bounded weak derivatives. Thus, we can now ensure that the S spaces constructed in the previous section satisfy $S \subset W_\infty^1(\Omega)$. Because of inclusion relation (3.37), we get $S \subset V$, a fact that was not proved until now.

Because Sobolev's inequality, we also deduce that w in the duality argument leading to theorem 3.4 is well defined. In fact, in the error estimate of $u - u_S$ we used the $L^2(\Omega)$ norm of the second derivative of functions (of w). Now, we can re-state the approximation assumption (3.17) by

$$\exists \epsilon > 0 \quad \text{such that} \quad \inf\{\|w - v\|_E : v \in S\} \leq \epsilon \left\|(\sigma w')'\right\|_{L^2(\Omega)} \quad \text{for all } w \in W_2^2(\Omega). \tag{3.44}$$

3.3 Variational Formulation of elliptic Boundary Value Problems

This third section of the chapter is devoted to the functional analysis tools required for developing the variational formulation of differential equations. Its objective is to provide

a framework in which existence and uniqueness of solutions to variational problems is established.

3.3.1 Inner-product and Hilbert spaces

Let V be a vector space. We say that $b : V \times V \longrightarrow \mathbf{R}$ is a *bilinear functional* if each univariate mapping $u \mapsto b(u, v)$ and $v \mapsto b(u, v)$ is linear. Moreover, if the bilinear functional verifies $b(v, u) = b(u, v)$ for all $v, u \in V$, we say that it is *symmetric*. An *inner-product* is a symmetric bilinear functional over V that verifies: (i) $b(v, v) \geq 0$ for all $v \in V$ and (ii) $b(v, v) = 0$ if and only if $v \equiv 0 \in V$. A vector space endowed with an inner-product is called an *inner-product space*.

Examples of inner product spaces are $L^2(\Omega)$ and $W_2^k(\Omega)$ with $\Omega \subset \mathbf{R}^d$ endowed by $\langle u, v\rangle_{L^2(\Omega)} := \int_\Omega u(x)v(x)dx$ and $\langle u, v\rangle_{W_2^k(\Omega)} := \sum_{|\alpha|\leq k}\langle \partial^\alpha u, \partial^\alpha v\rangle_{L^2(\Omega)}$, respectively. More commonly, the latter inner-product space is denoted by $H^k(\Omega)$ instead of by $W_2^k(\Omega)$.

Theorem 3.11 (Schwarz's inequality)**.** *Let $(V, \langle\cdot,\cdot\rangle)$ be an inner product space. Then,*

$$|\langle u, v\rangle| \leq \sqrt{\langle u, u\rangle}\,\sqrt{\langle v, v\rangle} \quad \textit{for all } u, v \in V, \tag{3.45}$$

and the equality holds if and only if u and v are linearly dependent.

Proof. Let $t \in \mathbf{R}$. Then, $0 \leq \langle u - tv, u - tv\rangle = \langle u, u\rangle - 2t\langle u, v\rangle + t^2\langle v, v\rangle$. If $\langle v, v\rangle = 0$, we have $0 \leq \langle u, u\rangle - 2t\langle u, v\rangle$. Since t is arbitrary, for that inequality to hold, necessarily $\langle u, v\rangle = 0$ because $\langle u, u\rangle \geq 0$. Then the inequality is trivial in this case (and in fact it is an equality). Now assume $\langle v, v\rangle > 0$ and substitute $t = \langle u, v\rangle / \langle v, v\rangle$ into the previous inequality. We obtain $0 \leq \langle u, u\rangle - |\langle u, v\rangle|^2 / \langle v, v\rangle$, which is equivalent to (3.45).

Let us now check the equality result. If $u = \lambda v$ for some $\lambda \in \mathbf{R}$, doing the replacement in (3.45) one gets the equality. To check the converse, assume the equality and first suppose $v = 0$. Then $v = 0 \cdot u$ is a linear combination of u. Otherwise, take $\lambda = \langle u, v\rangle / \langle v, v\rangle$ and consider $\langle u - \lambda v, u - \lambda v\rangle$ under the equality assumption of (3.45). Then, $\langle u - \lambda v, u - \lambda v\rangle = 0$. This implies $u - \lambda v = 0$ because of property (ii) of inner-products. □

Remark. In the proof of Schwarz's inequality (not in the equality part) it does not require property (ii) of inner products, and in consequence, other less restrictive bilinear functionals than inner-products can verify this inequality. For example, $b(u, v) = \int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx$ on $H^1(\Omega)$, which in case $\Omega = [0, 1]$, it coincides with the bilinear functional $\mathfrak{L}(\cdot,\cdot)$ of section 3.1. Thus,

$$b(u, v) \leq \sqrt{b(u, u)}\,\sqrt{b(v, v)} \quad \text{for all } u, v \in H^1(\Omega), \tag{3.46}$$

even though $b(\cdot,\cdot)$ is not an inner-product.

An interesting fact of inner-product spaces $(V, \langle\cdot,\cdot\rangle)$ is that they are also normed spaces. This is because the induced functional defined by $\|v\| := \sqrt{\langle v, v\rangle}$ for all $v \in V$ is a norm. In case the induced normed space $(V, \|\cdot\|)$ derived from the previous law is a Banach space,

we say that the corresponding inner-product space $(V, \langle\cdot,\cdot\rangle)$ is a *Hilbert space*. The converse, i.e., that normed spaces are inner-product spaces, is false in general. However, there is a sufficent condition that assures normed spaces being inner-product spaces: the *Parallelogram Law*, i.e., $\|v+u\|^2 + \|v-u\|^2 = 2(\|v\|^2 + \|u\|^2)$ for all $v, u \in V$ with $(V, \|\cdot\|)$ a normed space.

If $(H, \langle\cdot,\cdot\rangle)$ is a Hilbert space and $S \subset H$ is linearly closed, i.e., $\alpha u + \beta v \in S$ for all $u, v \in S$ and $\alpha, \beta \in \mathbf{R}$, we say that S is a subspace of H. Furthermore, because of the mentioned closedness, it verifies that S is a Hilbert space. Examples of subspaces of H are (i) $\{0\}$ and H (trivial subspaces), (ii) kernels of continuous linear mappings between H and another vector space, (iii) the set $x^\perp := \{v \in H : \langle x, v\rangle = 0\}$ for each $x \in H$, and (iv) the set $M^\perp := \{v \in H : \langle x, v\rangle = 0 \text{ for all } x \in M\} = \bigcap_{x\in M} x^\perp$ where $M \subset H$.

Example (iv), which is a generalization of example (iii) when $M = \{x\}$, is a remarkable case which conforms the starting point of the next subsection. Elements in $x^\perp$ are called *orthogonal* to x and $M^\perp$ is said to be the *orthogonal set* of M. The following theorem about orthogonality relations concludes this subsection.

Theorem 3.12. *Let H be a Hilbert space. Then,*

(i) *for $M, N \subset H$ such that $M \subset N$, it verifies $N^\perp \subset M^\perp$;*

(ii) *for $M \subset H$ such that $0 \in M$, it verifies $M \cap M^\perp = \{0\}$;*

(iii) *$H^\perp = \{0\}$ and $\{0\}^\perp = H$.*

3.3.2 Projections onto subspaces and Riesz's Representation Theorem

We begin stating the following theorem that connects the notion of closest elements in Hilbert subspaces with the notion of orthogonality introduced above.

Theorem 3.13. *Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space and let $M \subset H$ be a subspace. Let $v \in H \setminus M$ and set $\delta := \inf\{\|v - w\| : w \in M\}$ with $\|\cdot\|$ being the induced norm of the inner-product. Then, there exists $w_0 \in M$ such that* (i) *$\|v - w_0\| = \delta$ and* (ii) *$v - w_0 \in M^\perp$.*

The previous theorem states that whenever we consider a subspace of a Hilbert space, there exists a closest element to another element non-belonging to the concerned subspace, and whose difference is orthogonal to it. Equivalently, any $v \in H$ can be written in the form $v = w_0 + w_1$ with $w_0 \in M$ and $w_1 \in M^\perp$. Furthermore, the decomposition is *unique*. To show it, assume there exist two possible decomposition of v, $w_0 + w_1 = v = z_0 + z_1$ with $w_0, z_0 \in M$ and $w_1, z_1 \in M^\perp$, and check that $M \ni w_0 - z_0 = -(w_1 - z_1) \in M^\perp$. Because $M \cap M^\perp = \{0\}$, we deduce the uniqueness. The following theorem is then a corollary of the previous one.

Theorem 3.14. *Let H be a Hilbert space and $M \subset H$ a subspace. Then, $H = M \oplus M^\perp$.*

At this point, we introduce $P_M : H \longrightarrow M$ and $P_{M^\perp} : H \longrightarrow M^\perp$ defined by

$$P_M(v) := \begin{cases} v, & \text{if } v \in M, \\ w_0, & \text{if } v \in H \setminus M, \end{cases} \qquad P_{M^\perp}(v) := \begin{cases} 0, & \text{if } v \in M, \\ v - w_0, & \text{if } v \in H \setminus M, \end{cases} \tag{3.47}$$

where w_0 is the closest element to v in M according to theorem (3.13)'s sense. These two operators are called *orthogonal projections* of H onto M and $M^\perp$, respectively.

The *dual space* V^* of a vector space V is the set of all linear functionals on V, i.e., $V^* = \{L : V \longrightarrow \mathbf{R} : L \text{ is linear}\}$. If we now consider more particularly a Banach space $(B, \|\cdot\|)$, we distinguish between the set of all linear functionals on B, B^*, and the subspace $B' \subset B^*$ of all continuous functionals on B. One interesting characterization of these linear functionals on Banach spaces is the fact that they are continuous if and only if they are bounded, i.e., if there exists some $C \geq 0$ such that $|L(v)| \leq C\|v\|$ for all $v \in B$. Then, $B' = \{L : B \longrightarrow \mathbf{R} : L \text{ is bounded}\}$. We define the norm of B' by $\|L\|_{B'} := \inf\{C \geq 0 : |L(v)| \leq C\|v\|_B \text{ for all } v \in B\}$ for each $L \in B'$.

If $(H, \langle\cdot,\cdot\rangle)$ is a Hilbert space, it is clear that $L_u : H \longrightarrow \mathbf{R}$ defined by $L_u(v) := \langle u, v\rangle$ for some $u \in H$ is a continuous linear functional. The following results provides the converse fact: every continuous linear functional on a Hilbert space can be represented uniquely by a one-fixed-element inner-product.

Theorem 3.15 (Riesz's Representation Theorem)**.** *Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space and let L be a continuous linear functional on H. Then, there exists a unique $u \in H$ such that $L(v) = \langle u, v\rangle$ for all $v \in H$. Moreover, $\|L\|_{H'} = \|u\|_H$.*

Proof. See, for example, [9, p. 11-13] □

3.3.3 Formulation of symmetric variational problems

The remaining of section 3.3 applies all the abstract Hilbert theory developed so far to obtain the existence and uniqueness for variational formulations of BVPs.

From the example developed in section 3.1, we recall that $H^1([0,1]) = W_2^1([0,1])$ is a Hilbert space under the inner product

$$\langle v_1, v_2\rangle := \int_0^1 v_1(x)\, v_2(x)\, dx + \int_0^1 v_1'(x)\, v_2'(x)\, dx. \tag{3.48}$$

In subsection 3.2.4, we set $V := \{v \in H^1([0,1]) : v(0) = 0\}$. To see that it is indeed a subspace of $H^1([0,1])$, we consider $\delta_0 : H^1([0,1]) \longrightarrow \mathbf{R}$ defined by $\delta_0(v) = v(0)$. From Sobolev's inequality 3.10, we obtain that δ_0 is bounded and linear and thus continuous. Hence, $V = \delta_0^{-1}(\{0\})$ is closed in $H^1([0,1])$ and in particular a subspace. Nevertheless, the bilinear symmetric functional $\mathfrak{L}(u,v) = \int_0^1 \sigma\, u'\, v'\, dx$ is not an inner-product because $\mathfrak{L}(1,1) = 0$ (it does not fulfill property (ii) of inner-products). The following property will

solve the problem of symmetric bilinear functionals on subspaces of Hilbert spaces that are not inner-products.

A bilinear functional $b(\cdot,\cdot)$ on a normed space $(V, \|\cdot\|)$ is said to be *bounded* (or *continuous*) if there exists $C \geq 0$ such that $|b(u,v)| \leq C \|u\| \|v\|$ for all $u, v \in V$. If $U \subset V$, we say that $b(\cdot,\cdot)$ is *coercive* on U if there exists $\alpha > 0$ such that $b(v,v) \geq \alpha \|v\|^2$ for all $v \in U$.

Theorem 3.16. *Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space and suppose $b(\cdot,\cdot)$ is a symmetric continuous bilinear functional on H and coercive on a subspace V of H. Then $(V, b(\cdot,\cdot))$ is a Hilbert space.*

Proof. The coercivity of $b(\cdot,\cdot)$ implies that if $v \in V$ such that $b(v,v) = 0$, then $v \equiv 0$. Thus, $b(\cdot,\cdot)$ is an inner product over V.

Now define $\|v\|_V = b(v,v)$ and let $\{v_n\}_n$ be a Cauchy sequence in $(V, \|\cdot\|_V)$. By coercivity, $\{v_n\}_n$ is also Cauchy in $(H, \|\cdot\|)$. Because H is complete, there exists v limit element of $\{v_n\}_n$ in the $\|\cdot\|$ norm. The closedness of V in H implies $v \in V$, and the boundedness of $b(\cdot,\cdot)$ implies that exists some $C > 0$ such that $\|v - v_n\| \leq \sqrt{C} \|v - v_n\|_H$. Hence, v is also a limit point of $\{v_n\}$ in the norm $\|\cdot\|_V$, which means that $(V, \|\cdot\|_V)$ is complete. □

In general, a *symmetric variational problem* is posed in the following way: let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space, let $V \subset H$ be a subspace of H and let $b(\cdot,\cdot)$ be a continuous, symmetric and coercive bilinear functional on V. Then, given $F \in V'$, we want to find

$$u \in V \quad \text{such that} \quad b(u,v) = F(v) \text{ for all } v \in V. \tag{3.49}$$

In fact, because all the previous discussion, the solution to the problem *exists* and it is *unique* as a consequence of Riesz's Representation Theorem on $(V, b(\cdot,\cdot))$.

Similarly, the *Ritz-Galerkin approximation problem* to the symmetric variational problem is posed as follows: let $(V, b(\cdot,\cdot))$ be a Hilbert space, let $V^h \subset V$ be a finite-dimensional subspace and let $b(\cdot,\cdot)$ be a continuous, symmetric and coercive bilinear functional on V^h. Then, given $F \in V'$, we want to find

$$u^h \in V^h \quad \text{such that} \quad b(u^h, v^h) = F(v) \text{ for all } v^h \in V^h. \tag{3.50}$$

As before, the solution to it *exists* and it is *unique* because $F|_{V^h} \in (V^h)'$.

Error estimates for $u - u^h$ are a consequence of *Galerkin's Orthogonality*, i.e., $b(u - u_h, v) = 0$ for all $v \in V^h$ where u and u^h are the solutions to (3.49) and (3.50), respectively. In fact, $\|u - u_h\|_E = \min\{\|u - v\|_E : v \in V^h\}$ where $\|\cdot\|_E$ denotes the energy norm introduced at section 3.1. Moreover, it can be checked that u^h minimizes the quadratic functional $Q : V^h \longrightarrow \mathbf{R}$ defined by $Q(v) = b(v,v) - 2F(v)$, i.e., *Ritz's Method*.

3.3.4 Formulation of non-symmetric variational problems

The purpose of this subsection is to obtain the same result as before, existence and uniqueness of variational problems, but without the symmetry condition.

A *non-symmetric variational problem* is posed in the same way as in (3.49) but without the symmetry condition: let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space, let $V \subset H$ de a subspace and let $b(\cdot,\cdot)$ be a continuous and coercive bilinear functional on V. Then, given $F \in V'$, we want to find

$$u \in V \quad \text{such that} \quad b(u,v) = F(v) \text{ for all } v \in V. \tag{3.51}$$

The *Galerkin approximation* (note that we do not say Ritz-Galerkin approximation because of the non-symetry) is then the same as before but replacing the subspace V of H by a finite-dimensional subspace V^h of V.

An example of a non-symmetric formulation of a boundary value problem is the following: let us consider the BVP given by

$$-u'' + u' + u = f \quad \text{on } [0,1], \tag{3.52}$$

$$u'(0) = u'(1) = 0, \tag{3.53}$$

and consider the Hilbert space $H = H^1([0,1])$, the subspace $V = H^1([0,1])$, the bilinear form $b(u,v) = \int_0^1 (u'v' + u'v + uv)\,dx$, and the linear functional $F \in H'$ defined by $F(v) = \langle f, v\rangle_{H^1([0,1])}$. Note that $b(\cdot,\cdot)$ is not symmetric because of the $u'v$ term. However, it is continuous because

$$\begin{aligned} |b(u,v)| &\le \left|\langle u,v\rangle_{H^1([0,1])}\right| + \left|\int_0^1 u'v dx\right| \le \|u\|_{H^1([0,1])}\|v\|_{H^1([0,1])} + \|u'\|_{L^2([0,1])}\|v\|_{L^2([0,1])} \le \\ &\le 2\|u\|_{H^1([0,1])}\|v\|_{H^1([0,1])}, \end{aligned} \tag{3.54}$$

and coercive since we can write

$$\begin{aligned} b(v,v) &= \int_0^1 (v'^2 + v'v + v^2)\,dx = \frac{1}{2}\left(\int_0^1 (v'+v)^2\,dx + \int_0^1 (v'^2+v^2)\,dx\right) \ge \\ &\ge \frac{1}{2}\|v\|^2_{H^1([0,1])}. \end{aligned} \tag{3.55}$$

Note that if the above differential equation is changed to $-u'' + ku' + u = f$, then the corresponding bilinear functional $b(\cdot,\cdot)$ may not be coercive if $k \in \mathbf{R}$ is large enough.

Next theorem guarantees the existence and uniqueness of non-symmetric variational problems. It is interesting to mention that the proof of this result is based on the *Contraction Mapping Principle*. To see a detailed proof of both theorems, see [7, p. 60-63].

Theorem 3.17 (Lax-Milagram). *Let $(V, \langle\cdot,\cdot\rangle)$ be a Hilbert space, let $b(\cdot,\cdot)$ be a continuous and coercive bilinear functional on V, and let $F \in V'$. Then, there exists a unique $u \in V$ such that $b(u,v) = F(v)$ for all $v \in V$.*

3.3.5 Error estimates for the general Finite Element Method

Let u be the solution to the (symmetric or non-symmetric) variational problem on V and let u^h be the solution to its associated approximation problem (on V^h). Then, the following result provides an estimate for the error term $\|u - u^h\|_V$.

Theorem 3.18 (Céa). *In the previous set conditions, it verifies*

$$\left\|u - u^h\right\|_V \leq \frac{C}{\alpha} \min\left\{\left\|u - v^h\right\|_V : v^h \in V^h\right\}, \tag{3.56}$$

where C and α are the continuity and coercivity constants of $b(\cdot,\cdot)$, respectively.

Proof. Since $b(u,v) = F(v)$ for all $v \in V$ and $b(u^h, v^h) = F(v^h)$ for all $v^h \in V^h \subset V$, we have $b(u - u^h, v^h) = 0$ for all $v^h \in V^h$. Moreover, for each $v^h \in V^h$, we have

$$\alpha \left\|u - u^h\right\|_V^2 \leq b(u - u^h, u - u^h) = b(u - u^h, u - v^h) + b(u - u^h, v^h - u^h) = \tag{3.57}$$

$$= b(u - u^h, u - v^h) \leq C \left\|u - u^h\right\|_V \left\|u - v^h\right\|_V. \tag{3.58}$$

Hence, assuming $\|u - u^h\|_V \neq 0$, we obtain $\|u - u^h\|_V \leq \frac{C}{\alpha} \|u - v^h\|_V$ for all $v^h \in V^h$ (the case $\|u - u^h\|_V = 0$ is trivial). We take the infimum over $v^h \in V^h$ in the previous expression and, since V^h is closed, we get the desired inequality (the infimum is attained and becomes a minimum). □

Céa's theorem shows that u^h is quasi-optimal, i.e., the error $\|u - u^h\|_V$ is proportional to the best approximation error over the subspace V^h. In the symmetric and coercive case, we showed $\|u - u^h\|_E = \min\{\|u - v^h\|_E : v^h \in V^h\}$. Then,

$$\left\|u - u^h\right\|_V \leq \alpha^{-1/2} \left\|u - u^h\right\|_E = \alpha^{-1/2} \min\left\{\left\|u - v^h\right\|_E : v^h \in V^h\right\} \leq \tag{3.59}$$

$$\leq \sqrt{\frac{C}{\alpha}} \min\left\{\left\|u - v^h\right\|_V : v^h \in V^h\right\} \leq \frac{C}{\alpha} \min\left\{\left\|u - v^h\right\|_V : v^h \in V^h\right\}, \tag{3.60}$$

which is really the remark about the relationships between the two formulations, i.e., that one can be derived from the other.

3.4 Variational formulation of Poisson's equation BVP

We now illustrate the above theory to a particular BVP. Let us recall the heat propagation BVP introduced at subsection 2.5:

$$\begin{cases} -\nabla \cdot (\sigma \nabla u) = f & \text{in } \Omega \subset \mathbf{R}^d, \\ u = u_D & \text{in } \partial\Omega_D, \\ -\sigma \nabla u \cdot \mathbf{n} = g & \text{in } \partial\Omega_N. \end{cases} \tag{3.61}$$

We now follow a similar path to the previously followed one with (3.1). We multiply by a *test function* $v \in V := H_0^1(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ in } \partial\Omega_D\}$, which is a subspace of

the Sobolev space $H = H^1(\Omega)$, and integrate over Ω to obtain:

$$\int_\Omega -\nabla \cdot \sigma \nabla u \, v \, dx = \int_\Omega f \, v \, dx, \qquad \text{for all } v \in V. \tag{3.62}$$

We remark that $\nabla \cdot \sigma \nabla u \, v := (\nabla \cdot (\sigma \nabla u)) \, v \neq \nabla \cdot ((\sigma \nabla u) \, v) =: \nabla \cdot (\sigma \nabla u \, v)$. Then, since it verifies $\nabla \cdot (\sigma \nabla u \, v) = \sigma \nabla u \cdot \nabla v + \nabla \cdot (\sigma \nabla u \, v)$, we obtain:

$$\int_\Omega \sigma \nabla u \cdot \nabla v \, dx - \int_\Omega \nabla \cdot (\sigma \nabla u \, v) \, dx = \int_\Omega f \, v \, dx, \qquad \text{for all } v \in V. \tag{3.63}$$

The second integral of the left hand side can be replaced by its equivalent form applying Gauss' divergence theorem:

$$\int_\Omega \sigma \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} (\sigma \nabla u \cdot \mathbf{n}) \, v \, ds = \int_\Omega f \, v \, dx, \qquad \text{for all } v \in V. \tag{3.64}$$

Because the test functions are set equal to zero in $\partial\Omega_D$, the second integral of the left hand side becomes:

$$\int_{\partial\Omega} -(\sigma \nabla u \cdot \mathbf{n}) \, v \, dx = \int_{\partial\Omega_N} -(\sigma \nabla u \cdot \mathbf{n}) \, v \, ds = \int_{\partial\Omega_N} g \, v \, ds \qquad \text{for all } v \in V. \tag{3.65}$$

In consequence, we arrive to the following formulation:

Find $u \in U := \{u \in H^1(\Omega) : u = u_D \text{ in } \partial\Omega_D\}$ such that

$$\int_\Omega \sigma \nabla u \cdot \nabla v \, dx = \int_\Omega f \, v \, dx - \int_{\partial\Omega_N} g \, v \, ds \tag{3.66}$$

holds for all $v \in V := H^1_0(\Omega) := \{v \in H^1(\Omega) : v = 0 \text{ in } \partial\Omega_D\}$.

There are two main differences between the variational formulation of Poisson's BVP and the variational formulation developed in the previous subsections:

(i) The space U of possible solutions (*trial space*) is a translation of the space V of testing functions (*testing space*). Such translation is performed according to a *lift* of the Dirichlet condition function (i.e., $U := \tilde{u}_D + V = \{\tilde{u}_D + v : v \in V\}$ where $\tilde{u}_D|_{\partial\Omega_D} = u_D$).

(ii) There appears an extra integral term in the right hand side of equation (3.66).

To write equation (3.66) in terms of a bilinear functional on $V \times V$ in the left hand side and a linear functional in the right hand side, we proceed as follows. We decompose $u = \tilde{u}_D + w$ with $w \in V$ and consider the bilinear functional $b(w, v) := \int_\Omega \sigma \nabla w \cdot \nabla v \, dx$ on $V \times V$, and the linear functional $F \in V^*$ defined by $F(v) := \int_\Omega f \, v \, dx - \int_{\partial\Omega_N} g \, v \, ds - \int_\Omega \sigma \nabla \tilde{u}_D \cdot \nabla v \, dx$. Hence, the variational formulation to the concerned problem is rewritten in the following more usual terms:

Find $w \in V$ such that $b(w, v) = F(v)$ for all $v \in V$ where

$$b(w, v) := \int_\Omega \sigma \nabla w \cdot \nabla v \, dx, \tag{3.67}$$

$$F(v) := \int_\Omega f \, v \, dx - \int_{\partial\Omega_N} g \, v \, ds - \int_\Omega \sigma \nabla \tilde{u}_D \cdot \nabla v \, dx. \tag{3.68}$$

The solution to the variational formulation of Poisson's BVP is then $u := \tilde{u}_D + w$.

To guarantee existence and uniquness to the problem, it rests to check the continuity of $F(\cdot)$ and $b(\cdot, \cdot)$, and the coercivity of the latter. To see the continuity of $b(\cdot, \cdot)$, we assume that the entries of σ are bounded by a certain $C > 0$, and consider Hölder's inequality conveniently:

$$|b(w, v)| \leq C \left| \langle \nabla w, \nabla v \rangle_{L^2(\Omega)} \right| \leq C \, \|\nabla w\|_{L^2(\Omega)} \, \|\nabla v\|_{L^2(\Omega)} \leq C \, \|\nabla w\|_{H^1(\Omega)} \, \|\nabla v\|_{H^1(\Omega)} .$$

The continuity of $F(\cdot)$ may be showed following a similar argument as above under the assumption that f and g are bounded in Ω. Nevertheless, proving the coercivity of $b(\cdot, \cdot)$ requires technical details that go beyond the theory developed so far (sequences of functions in $H^1(\Omega)$, compactness arguments, Friedrich's inequality, etc.). In any case, the coercivity property is satisfied for $b(\cdot, \cdot)$ when the domain is sufficiently regular (find a proof of coercivity of Laplace's Equation with Robin boundary conditions at [5]).

In conclusion, the existence and uniqueness for the variational formulation of Poisson's equation BVP is assured. Furthermore, for any (finite-dimensional) subspace V^h of $H^1(\Omega)$, there exists a unique solution $u^h = \tilde{u}_D + w$ to the given problem with $w \in V^h$ and whose error estimate verifies $\left\| u - u^h \right\|_E = \min\{ \|u - v\|_E : v \in \tilde{u}_D + V^h \}$ because of the symmetry of $b(\cdot, \cdot)$.

Chapter 4

Fourier summation approximation for Finite Element computations

> «Mathematics compares the most diverse phenomena and discovers the secret analogies that unite them.»
>
> *Joseph Fourier*

In [41], D. Pardo adresses a 3D problem making a cylindrical change of coordinates and considers a Fourier summation expression in one of the three variables for the functions involved in the variational formulation. The orthogonality of the Fourier system and its exponential convergence becomes useful for a self-adaptive goal-oriented hp-FEM.

In this chapter we consider a Fourier summation approximation in order to take advantage of the orthogonality of the Fourier system. To do so, we organize it as follows. At first and second sections we make a brief review on change of coordinates systems and Fourier summation approximation theory, respectively. Then, we perform a Fourier approximation in all the variables of Poisson's equation BVP, and as a result, we find an explicit analytic expression for all the entries in the stiffness matrix. We describe the structure of the resulting stiffness matrix in one dimension and the corresponding generalization is performed for arbitrary dimension employing tensors. At the end of the chapter, we emphasize the benefits of the obtained structures for massive computations.

4.1 Variational Formulation in an arbitrary Coordinate System

Let $x = (x_1, x_2, \cdots, x_d)$ be the Cartesian coordinate system for $\mathbf{R}^d$ and let $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_d)$ be another (possibly non-orthogonal) coordinate system related to the first by $x = \psi(\alpha)$. We assume ψ to be an almost everywhere diffeomorphism with non-zero Jacobian determinant $\mathcal{J} = \left(\frac{\partial x_i}{\partial \alpha_j}\right)$, i.e., $\det \mathcal{J} \neq 0$ [36, p. 139-160].

If $w = w(x)$ is a real-valued function on the Cartesian coordinate system, we denote by $\hat{w} := w \circ \psi$ to the corresponding function in the new coordinate system. The hat notation will be used in what follows to express the coordinate system change in the subsequent functions.

Applying the chain rule, we calculate the gradient of w in terms of the new coordinate system:

$$\nabla w = \left(\sum_{j=1}^{d} \frac{\partial \hat{w}}{\partial \alpha_j} \frac{\partial \alpha_j}{\partial x_1}, \sum_{j=1}^{d} \frac{\partial \hat{w}}{\partial \alpha_j} \frac{\partial \alpha_j}{\partial x_2}, \cdots, \sum_{j=1}^{d} \frac{\partial \hat{w}}{\partial \alpha_j} \frac{\partial \alpha_j}{\partial x_d} \right)^T = (\boldsymbol{\mathcal{J}}^{-1})^T \nabla \hat{w}, \tag{4.1}$$

where $\nabla w = (\partial_{x_1} w, \partial_{x_2} w, \ldots, \partial_{x_d} w)^T$ and $\nabla \hat{w} = (\partial_{\alpha_1} \hat{w}, \partial_{\alpha_2} \hat{w}, \ldots, \partial_{\alpha_d} \hat{w})^T$.

Now, we retake Poisson's d-dimensional BVP in its variational formulation and write it in the new coordinate system. To do so, we first write the bilinear form (3.67) and the linear functional (3.68) making use of the *Change of Variable* theorem [36, p. 139-160]:

$$b(w, v) = \int_{\Omega} \sigma \nabla w \cdot \nabla v \, dx = \int_{\widehat{\Omega}} \hat{\sigma} (\boldsymbol{\mathcal{J}}^{-1})^T \nabla \hat{w} \cdot (\boldsymbol{\mathcal{J}}^{-1})^T \nabla \hat{v} \, |\det \boldsymbol{\mathcal{J}}| \, d\alpha = \tag{4.2}$$

$$= \int_{\widehat{\Omega}} \boldsymbol{\mathcal{J}}^{-1} \hat{\sigma} (\boldsymbol{\mathcal{J}}^{-1})^T \nabla \hat{w} \cdot \nabla \hat{v} \, |\det \boldsymbol{\mathcal{J}}| \, d\alpha = \int_{\widehat{\Omega}} \hat{s} \nabla \hat{w} \cdot \nabla \hat{v} \, d\alpha =: \hat{b}(\hat{w}, \hat{v}), \tag{4.3}$$

where $\hat{s} := \boldsymbol{\mathcal{J}}^{-1} \hat{\sigma} (\boldsymbol{\mathcal{J}}^{-1})^T |\det \boldsymbol{\mathcal{J}}|$ and

$$F(v) = \int_{\Omega} f \, v \, dx - \int_{\partial \Omega_N} g \, v \, ds - \int_{\Omega} \sigma \nabla \tilde{u}_D \cdot \nabla v \, dx = \tag{4.4}$$

$$= \int_{\widehat{\Omega}} \hat{f}_s \, \hat{v} \, d\alpha - \int_{\partial \widehat{\Omega}_N} \hat{g}_n \, \hat{v} \, d\alpha_n - \int_{\widehat{\Omega}} \hat{s} \nabla \hat{u}_D \cdot \nabla \hat{v} \, d\alpha =: \hat{F}(\hat{v}), \tag{4.5}$$

where $\hat{f}_s := \hat{f} \, |\det \boldsymbol{\mathcal{J}}|$, $\hat{g}_n := \hat{f} \, |\det \boldsymbol{\mathcal{J}}_n|$ with $\boldsymbol{\mathcal{J}}_n$ being the Jacobian associated to the change of variables of the $(d-1)$-dimensional $\partial \Omega_N$ surface, and $\hat{u}_D$ is a shortcut for $\tilde{u}_D \circ \psi$.

At this point, the testing space $V = H_0^1(\Omega)$ is actualized by $\widehat{V} = \{v \circ \psi : v \in H_0^1(\Omega)\}$ and the coordinate system change equivalent variational formulation for Poisson's BVP becomes as follows: find $\hat{w} \in \widehat{V}$ such that $\hat{b}(\hat{w}, \hat{v}) = \hat{F}(\hat{v})$ holds for all $\hat{v} \in \widehat{V}$.

Once such $\hat{w}$ is found, the solution to the problem in the original Cartesian coordinate system is $u = \hat{w} \circ \psi^{-1} + \tilde{u}_D$.

Summarizing, following the previously developed path, we are able to: (i) take the considered BVP's variational formulation in the Cartesian coordinate system, (ii) translate it to another (possibly) more convenient coordinate system, (iii) solve the equivalent problem taking advantage of the properties of the new system, and finally, (iv) recover the obtained solution but in the original terms.

In what follows, we will assume that the BVP's domain Ω verifies the *Cartesian product splitting property*, i.e., $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_d \subset \mathbf{R}^d$. In case the domain is not of this form, we could assume that there exists a certain coordinate change system whose corresponding domain satisfies so (i.e., there exists an almost everywhere diffeomorphism $\psi : D \longrightarrow \Omega$ such that $D = D_1 \times D_2 \times \cdots \times D_d$). Commonly used examples for transforming circular (in 2D), spherical or cylindrical (in 3D) domains into Cartesian product domains are the

polar, spherical and cylindrical change coordinate systems, respectively [39]. In the current project we will not deal with the problem of finding the appropriate change of variables system.

4.2 Fourier summation approximation

Fourier summation approximation theory allows us to approximate periodic functions as sums of trigonometric functions. In case the functions are defined in a bounded domain, they could be thought as periodic with a domain periodic extension (e.g., if f has a bounded domain $\Omega \subset \mathbf{R}$, we can periodic-extend it to $\mathbf{R}$ utilizing a function $\widetilde{f}$ such that $\widetilde{f}(x) = \widetilde{f}(x + \Omega)$ for all $x \in \mathbf{R}$ —in particular, $\widetilde{f}|_{\Omega} = f$—).

If we restrict to one-dimensional functions, we formally define the Fourier sum approximation as follows: let $s : I \longrightarrow \mathbf{R}$ with $I \subset \mathbf{R}$ being a bounded interval, assume that $s(x)$ belongs to $L^2(I)$, and denote by L to the length of I. Then, the *Fourier's N-th sum approximation* of $s(x)$ is given by the expression

$$s_N(x) := \frac{a_0}{2} + \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n x}{L}\right) + b_n \sin\left(\frac{2\pi n x}{L}\right)\right), \tag{4.6}$$

where a_n and b_n are the *Fourier coefficients* given by

$$a_n := \frac{2}{L}\int_I s(x)\cos\left(\frac{2\pi n x}{L}\right)dx, \qquad b_n := \frac{2}{L}\int_I s(x)\sin\left(\frac{2\pi n x}{L}\right)dx. \tag{4.7}$$

In theory, integer N may be infinite even so the series might not converge or exactly be equal to $s(x)$ at all values of x in I (e.g., at discontinuity single-points). Avoiding these almost everywhere exceptions (which we already commented at subsection 3.2.2 that they are not important in our established mathematical framework), the bigger N is, the better the approximation performs in general.

Thanks to Euler's identity, we are able to write (4.6) in a more abbreviated way:

$$s_N(x) = \sum_{n=-N}^{N} c_n\, e^{\frac{2\pi i n x}{L}} = \sum_{n=-N}^{N} c_n\, \exp\left(\frac{2\pi i\, n\, x}{L}\right), \tag{4.8}$$

where the coefficients are given by $c_n = \frac{1}{L}\int_I s(x)\ \exp(-2\pi i\, n\, x/L)dx$, which in the previous developed terms equals to

$$c_n = \begin{cases} a_0/2, & \text{if } n = 0, \\ \frac{1}{2}(a_n - i b_n), & \text{if } n > 0, \\ \overline{c_{|n|}} = \frac{1}{2}(a_n + i b_n), & \text{if } n < 0, \end{cases} \tag{4.9}$$

where the top bar denotes complex conjugation.

For a more detailed theory development about Fourier series and their convergence results, see [34].

4.3 Poisson's equation with Fourier summation approximation

We now consider solutions to Poisson's equation BVP in the Fourier approximation form and take advantage of the orthogonality of the employed system and the domain's Cartesian product structure. To do so, we consider certain simplifying assumptions with respect to the initial ones considered at subsection 2.5. The first one involves the σ parameter. We relax its structural complexity by reconsidering it as a scalar function $\sigma(x) \in \mathbf{R}$. This way, the $\sigma\nabla u$ product is understood as componentwise between the scalar σ and the vector ∇u. Furthermore, the source function $-f-$ and the boundary condition functions $-g, u_D$ and $h-$ are going to be displayed (assumed) in a finite Fourier sum form.

4.3.1 General development

Let $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_d \subset \mathbf{R}^d$ be a domain with corresponding Cartesian product domain lengths $L_1, L_2, \ldots, L_d$, and let us consider the variational formulation of Poisson's equation developed at chapter 3: find $w \in V$ such that $b(w, v) = F(v)$ for all $v \in V$ where $b(\cdot,\cdot)$ and $F(\cdot)$ are given by (3.67) and (3.68), respectively. In our case, V^h will denote the subspace of $H_0^1(\Omega)$ such that its elements are expressible in finite Fourier summation terms. Henceforth, we drop the superindex h notation in the approximation space denoting it simply by V.

Formally, $w \in V$ if and only if for a certain variable indexes reordering $\{j_1, j_2, \ldots, j_d\}$, there exist running indexes $\{n_{(j_1)}, n_{(j_1,j_2)}, \ldots, n_{(j_1,j_2,\cdots,j_d)}\}$ with corresponding running ranges $\{R_1, R_2, \ldots, R_d\}$ such that

$$\begin{aligned} w &= \sum_{n_{(j_1)}\in R_1} \sum_{n_{(j_1,j_2)}\in R_2} \cdots \sum_{n_{(j_1,j_2,\cdots,j_d)}\in R_d} w_{n_{(j_1,j_2,\cdots,j_d)}} \prod_{k=1}^{d} \exp\left(\frac{2\pi i\, n_{(j_1,j_2,\cdots,j_k)}\, x_{j_k}}{L_{j_k}}\right) \\ &= \sum_{n_{(j_1)}\in R_1} \sum_{n_{(j_1,j_2)}\in R_2} \cdots \sum_{n_{(j_1,j_2,\cdots,j_d)}\in R_d} w_{n_{(j_1,j_2,\cdots,j_d)}} \exp\left(2\pi i \sum_{k=1}^{d} \frac{n_{(j_1,j_2,\cdots,j_k)}\, x_{j_k}}{L_{j_k}}\right) \in H_0^1(\Omega), \end{aligned}$$

where each $w_{n_{(j_1,j_2,\cdots,j_d)}}$ denotes the Fourier coefficient given by the last term of the following recurrence relation:

$$w_{n_{(j_1,j_2,\cdots,j_k)}} = \frac{1}{L_{j_k}} \int_{\Omega_{j_k}} w_{n_{(j_1,j_2,\cdots,j_{k-1})}} \exp\left(\frac{-2\pi i\, n_{(j_1,j_2,\cdots,j_k)}\, x_{j_k}}{L_{j_k}}\right) dx_{j_k}, \quad \text{for } 1 \le k \le d. \tag{4.10}$$

Check that from the above definition it follows that a basis of V is given by

$$\left\{ \exp\left(2\pi i \sum_{k=1}^{d} \frac{n_{(j_1,j_2,\ldots,j_d)} x_{j_k}}{L_{j_k}}\right) : -N \le n_{(j_1,j_2,\ldots,j_k)} \le N \text{ for all } 1 \le k \le d \right\}. \tag{4.11}$$

Moreover, because all the elements $w \in V$ need to satisfy the condition $w = 0$ in the Dirichlet boundary, we must have into account that probably some exponential functions may not be considered in the above summation. For instance, in the so far developed 1D problem (3.1), the function $\exp(2\pi i\ 0\ x) \equiv 1$ must not be considered as a basis function in the Fourier summation because it does not belong to $H_0^1([0,1])$.

For the notation to be easy to follow in the subsequent development, we consider the following criteria: for distinct finite Fourier summation expressions (where distinct indexes summation ranges may be considered), we will assume all of them to have the same running range (assuming the biggest range for all the summations and completing, if necessary, with null Fourier coefficients). The Fourier summations ranges are always going to be considered symmetric with respect to zero (i.e., they will always run from $-N$ to N for a certain $N \in \mathbf{N}$). In addition, we are going to drop (queued) subindexes of indexes notation. This way, the j_k subindexes are simply replaced by k and $n_{(j_1,j_2,\cdots,j_k)}$ by n_k. Thereby, without loss of generality, the variables ordering selection is assumed to be the typical (natural) one. Furthermore, queues of summatories such as $\sum_{n_1}\sum_{n_2}\cdots\sum_{n_d}$ are going to be expressed more compactly by $\sum_{n_1,n_2,\ldots,n_d}$.

Under the previous considerations, a general element w of V is written by

$$w = \sum_{-N\leq n_1,n_2,\ldots,n_d\leq N} W_{n_1,n_2,\cdots,n_d}\ \exp\left(\sum_{k=1}^{d}\frac{2\pi i\ n_k\ x_k}{L_k}\right), \tag{4.12}$$

with possibly more restriction conditions in the indexes running ranges.

To simplify even more the notation, we substitute $n_1, n_2, \ldots, n_d$ by the multindex n so as to write W_n instead of $W_{n_1,n_2,\ldots,n_d}$ and consider e_n as a shortcut of $\exp(\sum_{k=1}^{d}\ 2\pi i\ n_k\ x_k/L_k)$. Hence, w is written in simple terms by $\sum_{-N\leq n\leq N} W_n\ e_n$ which, by abuse of notation, the $-N \leq n \leq N$ expression refers to $-N \leq n_1, n_2, \ldots, n_d \leq N$.

Let $w = \sum_{-N\leq m\leq N} W_m\ e_m$ and $v = \sum_{-N\leq n\leq N} V_n\ \overline{e_n}$ belong to V. Taking conjugate exponentials in the test function bases is done for convenience to obtain at the end a better final presentation of the solution. Recall that $\{e_n : -N \leq n \leq N\}$ and $\{\overline{e_n} : -N \leq n \leq N\}$ conform bases of the same space V since they are sets containing the same elements but ordered reversely, one with respect to the other (i.e., $\overline{e_n} = e_{-n}$ for all $0 \leq n \leq N$). Then, if we denote by

$$\mathbf{W} = \Big[W_m\Big], \qquad \mathbf{K} = \Big[b(e_m, \overline{e_n})\Big]^T \qquad \text{and} \qquad \mathbf{F} = \Big[F(\overline{e_n})\Big], \tag{4.13}$$

with $\mathbf{W}$ and $\mathbf{F}$ being column vectors, $\mathbf{K}$ being a matrix, and $-N \leq n, m \leq N$, we have that equality $b(w, v) = F(v)$ is equivalent to $\mathbf{KW} = \mathbf{F}$. In what rests of subsection, we assemble the $\mathbf{K}$ stiffness matrix and $\mathbf{F}$ vector.

To calculate the entries of **K**, we first check that

$$\nabla e_m = e_m \left[\frac{2\pi i\, m_1}{L_1}, \frac{2\pi i\, m_2}{L_2}, \ldots, \frac{2\pi i\, m_d}{L_d}\right]^T, \tag{4.14}$$

with the product being componentwise between the exponential and the vector. Then,

$$\nabla e_m \cdot \nabla \overline{e_n} = e_{m-n} \sum_{s=1}^{d} \frac{4\pi^2\, m_s\, n_s}{L_s^2}, \tag{4.15}$$

where e_{m-n} is a shortcut for $\exp(\sum_{k=1}^{d} 2\pi i\, (m_k - n_k)\, x_k / L_k)$.

We write now the parameter function in Fourier summation terms, $\sigma = \sum_{-N\leq l\leq N} S_l\, e_l$, and thus need to calculate the integral

$$b(e_m, \overline{e_n}) = \int_\Omega \sigma \nabla e_m \cdot \nabla \overline{e_n}\, dx = \left(\sum_{s=1}^{d} \frac{4\pi^2\, m_s\, n_s}{L_s^2}\right)\left(\sum_{-N\leq l\leq N} S_l \int_\Omega e_{m-n+l}\, dx\right), \tag{4.16}$$

where $l = (l_1, l_2, \ldots, l_d)$ is a multindex and s (and implicitly k in e_{m-n+l}) are ordinary indexes.

If we assume that $\sigma \nabla e_m \cdot \nabla \overline{e_n}$ is integrable over Ω (i.e., $\sigma \nabla e_m \cdot \nabla \overline{e_n}$ is measurable and its corresponding absolute value function has finite integral over Ω), then we can apply *Fubini's Theorem* in (4.16) and thereby take advantage of the Cartesian product domain assumption,

$$\begin{aligned} b(e_m, \overline{e_n}) &= \left(\sum_{s=1}^{d} \frac{4\pi^2\, m_s\, n_s}{L_s^2}\right)\left(\sum_{-N\leq l\leq N} S_l \int_{\Omega_1}\int_{\Omega_2}\cdots\int_{\Omega_d} e_{m-n+l}\, dx_d \cdots dx_2 dx_1\right) = \\ &= \left(\sum_{s=1}^{d} \frac{4\pi^2\, m_s\, n_s}{L_s^2}\right)\left(\sum_{-N\leq l\leq N} S_l \prod_{k=1}^{d}\int_{\Omega_k} \exp\left(\frac{2\pi i\, (m_k - n_k + l_k)\, x_k}{L_k}\right) dx_k\right). \end{aligned}$$

Because each $\int_{\Omega_k} \exp(2\pi i\, (m_k - n_k + l_k)\, x_k / L_k)\, dx_k$ equals 1 if $m_k - n_k + l_k = 0$ and equals 0 otherwise, we have that all the terms of the sum associated to the multindex $l = (l_1, l_2, \ldots, l_d)$ vanish except for those that satisfy $l_k = n_k - m_k$.

Hence,

$$b(e_m, \overline{e_n}) = 4\pi^2 \left(\sum_{s=1}^{d} \frac{n_s\, m_s}{L_s^2}\right) S_{n-m}, \tag{4.17}$$

where $n - m$ denotes the multindex $(n_1 - m_1, n_2 - m_2, \ldots, n_d - m_d)$.

This way, displaying the (4.17) values in a matrix, we assemble the desired stiffness matrix. To assemble the **F** vector, we need to calculate three integrals per entry of the vector:

$$F(\overline{e_n}) = \int_\Omega f\, \overline{e_n}\, dx - \int_{\partial\Omega_N} g\, \overline{e_n}\, ds - \int_\Omega \sigma \nabla \tilde{u}_D \cdot \nabla \overline{e_n}\, dx. \tag{4.18}$$

To do so, we proceed as above writing the concerned functions in their Fourier summation forms: $f = \sum_{-N\leq l\leq N} F_l\, e_l$, $g = \sum_{-N\leq l\leq N} G_l\, e_l$ and $\widetilde{u}_D = \sum_{-N\leq l\leq N} U_l\, e_l$. Then, the first and second integrals are immediate and the third integral equals the linear combination of bilinear forms $\sum_{-N\leq l\leq N}\ U_l\, b(e_l, \overline{e_n})$ that are already solved:

$$\int_\Omega f\, e_n\, dx = \sum_{-N\leq l\leq N} F_l \int_\Omega e_{l-n}\, dx = F_n, \tag{4.19}$$

$$\int_{\partial\Omega_N} g\, e_n\, ds = \sum_{-N\leq l\leq N} G_l \int_{\partial\Omega_N} e_{l-n}\, ds = G_n, \tag{4.20}$$

$$\int_\Omega \sigma\nabla\widetilde{u}_D\cdot\nabla e_n\, dx = 4\pi^2 \sum_{-N\leq l\leq N} U_l\, S_{l-n} \sum_{s=1}^{d} \frac{l_s\, n_s}{L_s^2}. \tag{4.21}$$

Hence,

$$F(\overline{e_n}) = F_n - G_n - 4\pi^2 \sum_{s=1}^{d} \left(\frac{l_s\, n_s}{L_s^2}\right) \sum_l U_l\, S_{l-n}. \tag{4.22}$$

Displaying these results in each entry of the right hand side vector of the matricial equation, we obtain the linear system to be solved.

4.3.2 Example in one dimension

If we go to problem (3.1) in chapter 3 (with $d = 1$), the calculations reduce to

$$b(e_m, \overline{e_n}) = 4\pi^2\, n\, m\, S_{n-m} \qquad \text{and} \qquad F(\overline{e_n}) = F_n, \qquad -N \leq n, m \leq N, \tag{4.23}$$

with $n \neq 0 \neq m$ because of the Dirichlet condition at zero.

The resulting $2N \times 2N$ sized stiffness matrix $\mathbf{K}$, with N being the number of modes considered in Fourier summation, satisfies the following structure: it has a four block structure (**TL**, top-left; **BL**, bottom-left; **TR**, top-right; and **BR**, bottom-right), i.e.,

$$\mathbf{K} = 4\pi^2 \begin{bmatrix} \mathbf{TL} & \mathbf{TR} \\ \mathbf{BL} & \mathbf{BR} \end{bmatrix}, \tag{4.24}$$

such that $\mathbf{BL} = \mathbf{TR}^*$ is triangular superior with null entries in the main diagonal, and if $\mathbf{TL} = [C_1, C_2, \ldots, C_N]$ with C_j being column vectors, then $\mathbf{BR} = [C_N, \ldots, C_2, C_1]^T$. Hence, for determining the whole matrix, it suffices to (for instance) simply calculate the entries of (4.23) when they belong to the lower triangular part of **TL** (including the main diagonal) and when they belong to the upper triangular part of **BL** (excluding the main diagonal), and take into account the previous considerations.

Below we include an example of the stiffness matrix for this problem when $N = 4$. We mark in bold the previously commented as "sufficient to calculate" entries.

$$\mathbf{K} = 4\pi^2 \begin{bmatrix} \mathbf{16S_0} & 12S_{-1} & 8S_{-2} & 4S_{-3} & 0 & 0 & 0 & 0 \\ \mathbf{12S_1} & \mathbf{9S_0} & 6S_{-1} & 3S_{-2} & -3S_{-4} & 0 & 0 & 0 \\ \mathbf{8S_2} & \mathbf{6S_1} & \mathbf{4S_0} & 2S_{-1} & -2S_{-3} & -4S_{-4} & 0 & 0 \\ \mathbf{4S_3} & \mathbf{3S_2} & \mathbf{2S_1} & \mathbf{S_0} & -S_{-2} & -2S_{-3} & -3S_{-4} & 0 \\ 0 & \mathbf{-3S_4} & \mathbf{-2S_3} & \mathbf{-S_2} & S_0 & 2S_{-1} & 3S_{-2} & 4S_{-3} \\ 0 & 0 & \mathbf{-4S_4} & \mathbf{-2S_3} & 2S_1 & 4S_0 & 6S_{-1} & 8S_{-2} \\ 0 & 0 & 0 & \mathbf{-3S_4} & 3S_2 & 6S_1 & 9S_0 & 12S_{-1} \\ 0 & 0 & 0 & 0 & 4S_3 & 8S_2 & 12S_1 & 16S_0 \end{bmatrix}$$

Rows correspond to index n and columns correspond to index m. The first row (respectively, column) is for $n = -4$ (respectively, $m = -4$), the second for $n = -3$ $(m = -3), \ldots,$ the fourth for $n = -1$ $(m = -1)$, the fifth for $n = 1$ $(m = 1), \ldots,$ and the eight for $n = 4$ $(m = 4)$.

Let us call by *t-diagonal*, for $1 \leq \pm t \leq 2N - 1$, to the set of entries of $\mathbf{K}$ given by the following rule: let $\mathcal{O} = (-N, -N+1, \ldots, -2, -1, 1, 2, \ldots, N-1, N)$ be an array/ordering. If $t \geq 0$, the t-diagonal consists on the elements

$$K_{(-N,-N+r)}, \ldots, K_{(n,m)}, K_{(n^*,m^*)}, \ldots, K_{(N-r,N)}, \tag{4.25}$$

where n^* and m^* denote the following elements of n and m in $\mathcal{O}$, respectively. If $t < 0$, the t-diagonal consists on the elements

$$K_{(-N-r,-N)}, \ldots, K_{(n,m)}, K_{(n^*,m^*)}, \ldots, K_{(N,N+r)}, \tag{4.26}$$

with n^* and m^* having the same interpretation as before.

If $t = 0$, the 0-diagonal is the main diagonal of $\mathbf{K}$. If $t = 1$, it is the diagonal immediately above the main one. If $t = -1$, it is the diagonal immediately below the main one. The rest of t-diagonals follow the same rule but being "more far away" (above or below) from the main diagonal. If $t \neq 0$, we also call *semidiagonals* to the t-diagonals.

According to this formalization, the stiffness matrix has the following "diagonalwise" structure: if $t = 0$, the main diagonal consists on multiples of S_0 Fourier coefficients. If $0 < \pm t < N$, the corresponding t-diagonals are formed by multiples of S_{-t} and $S_{-t(1+1/|t|)}$ coefficients. The amount of S_{-t} and $S_{-t(1+1/|t)}$ coefficients in the t-diagonal is of $N - 2|t|$ and $|t|$, respectively, being the latters distributed in the middle of it. When $\pm t \geq N$, the corresponding t-diagonals have null entries. Hence, the resulting matrix is said to have a $(2N-1)$-*diagonal* structure (i.e., with non-null $2N - 1$ middle diagonals).

In consequence, we can express the $\mathbf{K}$ matrix as a sum of semidiagonal matrices:

$$\mathbf{K} = 4\pi^2 \sum_{r=-N}^{N} \mathbf{K}_r, \tag{4.27}$$

where each term of the sum can be split as a componentwise product between a Fourier coefficient and a matrix with at most two non-null semidiagonals: $\mathbf{K}_r = S_r\ \mathbf{C}_r$ for $-N \leq r \leq N$. At the last section of the chapter we are going to explain the importance of the $\mathbf{C}_r$ matrices that we have obtained. In this context, we call these matrices as *elemental matrices*

Below we show the representations of the $\mathbf{C}_r$ matrices of the previous example. In representations $\mathbf{C}_r$ for $r \neq 0$ we have utilized a color criterion to depict both matrices in the same display. If $r > 0$ (colored in blue), the purple upper semidiagonal must be thought of with null entries; and when $r < 0$ (in purple), the blue lower semidiagonal must be thought of with null entries. The gray zeros indicate diagonal gaps that are filled in a posterior indexed matrix.

$$\mathbf{C}_0 = \begin{bmatrix} 16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 16 \end{bmatrix} \qquad \mathbf{C}_{-1}, \mathbf{C}_1 = \begin{bmatrix} 0 & 12 & 0 & 0 & 0 & 0 & 0 & 0 \\ 12 & 0 & 6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 6 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 & 12 \\ 0 & 0 & 0 & 0 & 0 & 0 & 12 & 0 \end{bmatrix} \tag{4.28}$$

$$\mathbf{C}_{-2}, \mathbf{C}_2 = \begin{bmatrix} 0 & 0 & 8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 8 & 0 & 0 \end{bmatrix} \tag{4.29}$$

$$\mathbf{C}_{-3}, \mathbf{C}_3 = \begin{bmatrix} 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 \end{bmatrix} \tag{4.30}$$

$$\mathbf{C}_{-4}, \mathbf{C}_4 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -3 & 0 \\ 0 & -3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{4.31}$$

Regarding the $\mathbf{F}$ vector, we have that it is just a column vector with entries $F_{-N}, \ldots, F_{-1}$, $F_1, \ldots, F_N$ from top to bottom. In fact, these entries are conjugate-symmetric with respect to the middle (i.e., $\overline{F_n} = F_{-n}$ for $1 \leq n \leq N$). Below we include an example for $N = 4$.

$$\mathbf{F} = \begin{bmatrix} F_{-4} & F_{-3} & F_{-2} & F_{-1} & F_1 & F_2 & F_3 & F_4 \end{bmatrix}^T \tag{4.32}$$

4.3.3 Some comments for higher dimension problems

If the dimension is higher than one ($d \geq 2$), the number of Fourier coefficients increases with respect to the 1D problem by a power equal to the concerned dimension. If we consider the Poisson's equation BVPs in a $[0,1]^d$ domain, we need to calculate $(2N)^d$ real coefficients to determine the solution (under the assumption that on each variable we have utilized a Fourier summation range precision of N).

In principle, because the only differences are the number of coefficients to determine, a possible alternative to solve these problems is to proceed as in the 1D problem but with powered sized matrices and vectors. Then, in a first approach, we could think on solving the general problem as an ordinary system of linear equations with a stiffness matrix of size $(2N)^d \times (2N)^d$, and input and output $(2N)^d$ sized vectors. The crux of the matter is then how we control what entry of the matrix represents what Fourier coefficient. The indexing is not as obvious as it was in the 1D case, where each entry was indexed by the pair row-column position of the matrix according to the results of the bilinear form when it was applied to the testing and training basis functions.

To deal with this indexing issue, it is convenient to introduce the concept of *tensor* [24, 35, 40]. Roughly speaking, and in what our topic concerns, a tensor is a generalization of

vectors and matrices structures. In the following subsection we describe briefly (and quite formally) the notion of tensor focusing on our particular problem in two dimensions.

4.3.4 Construction of a two dimensional tensor

Let V^1 and V^2 be two finite dimensional vector spaces with bases $\{e_i^1\}_i$ and $\{e_j^2\}_j$. Making abuse of notation, consider the vector coordinate representations of e_i^1 and e_j^2 for the aforementioned bases maintaining the same symbology (i.e., e_i^1 and e_j^2 are the canonical basis vectors, that is, their entries are the Kronecker deltas). Let us perform the products $T_{i,j} := e_i^1 (e_j^2)^T$ for all the possible pairs of canonical column vectors. Then, the assembled matrices $T_{i,j}$ are made up of zeros and a single one placed in the (i, j) row-column position. Take two vectors, $v^1 \in V^1$ and $v^2 \in V^2$, and calculate $(v^1)^T e_i^1 (e_j^2)^T v^2$ to get the scalar value of the i-th coordinate of v^1 multiplied by the j-th coordinate of v^2.

Formally, $e_i^1 \otimes e_j^2 := e_i^1 (e_j^2)^T$ is called the *outer product* between e_i^1 and e_j^2 and it may be performed for arbitrary (not necessarily canonical) vectors. The resulting $T_{i,j}$ matrix could be thought of as a mapping which is expected to be fed by two vectors, each of them belonging to its corresponding vector space, $v^1 \in V^1$ and $v^2 \in V^2$, so as to produce a scalar value, $T_{i,j}(v^1, v^2) := (v^1)^T e_i^1 (e_j^2)^T v^2$. It is straightforward to check that the obtained mapping $T_{i,j} : V^1 \times V^2 \longrightarrow \mathbf{R}$ is in fact a bilinear form. From now on, we prefer to call *tensors* to these bilinear functionals. The *tensor product* $V^1 \otimes V^2$ consists on the set of all the tensors made up as before by all the possible pairs of vectors of V^1 and V^2. Obviously, all these tensors are linear combinations of the $T_{i,j}$ tensors, and thus, $\{T_{i,j}\}_{i,j}$ conforms a basis of the vector space $V^1 \otimes V^2$. By construction it is easy to check that $\dim(V^1 \otimes V^2) = \dim(V^1) \dim(V^2)$. In some sense, we could say that the outer product mapping (which is bilinear), $\otimes : V^1 \times V^2 \longrightarrow V^1 \otimes V^2$, takes two elements in different spaces and performs a "dimensionalization" to produce a kind of "product element". It is important not to confuse cartesian products, $\times$ or $\oplus$, with tensor products, $\otimes$ (e.g., $\mathbf{R}^{15} \equiv \mathbf{R}^{3\times 5} \equiv \mathbf{R}^3 \otimes \mathbf{R}^5 \not\equiv \mathbf{R}^3 \times \mathbf{R}^5 \equiv \mathbf{R}^8$).

Now it is time to associate this special construction with the considered problem: let V^1 be the $2N$-dimensional vector space of functions expressed in Fourier summation for the (first) x_1 variable in the previously presented problem, and let V^2 be the one of the (second) x_2 variable. We have that $\{\exp(2\pi i\ n_1\ x_1) : -N \leq n_1 \leq N, n_1 \neq 0\}$ and $\{\exp(2\pi i\ n_2\ x_2) : -N \leq n_2 \leq N, n_2 \neq 0\}$ are bases for V^1 and V^2, respectively. Let us denote by e_{n_1} and e_{n_2} the previous bases functions in shortcut mode. Then, the $T_{n_1,n_2} :\equiv e_{n_1} \otimes e_{n_2}$ tensor represents the bilinear mapping that returns the coordinate corresponding to the $\exp(2\pi i\ (n_1 x_1 + n_2 x_2))$ basis term (coefficient W_{n_1,n_2} in (4.12)) in the $V^1 \otimes V^2$ space. In simple words, the constructed tensors allow us to connect each univariate-Fourier-space in the joint bivariate-Fourier-space.

In the context of vector spaces, the tensor product $V^1 \otimes V^2$ and the associated bilinear mapping $\otimes : V^1 \times V^2 \longrightarrow V^1 \otimes V^2$ are characterized up to isomorphism by a universal property regarding bilinear maps. Informally, $\otimes$ is the most general bilinear map out of $V^1 \times V^2$.

Theorem 4.1. *Let V^1, V^2 and W be vector spaces. Then, for any bilinear mapping $b : V^1 \times V^2 \longrightarrow W$, there exists a unique linear mapping $l : V^1 \otimes V^2 \longrightarrow W$ such that $b = l \circ \otimes$.*

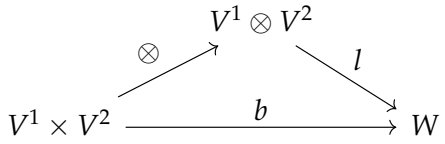

FIGURE 4.1: Theorem 4.1's representation

If $W = V^1 \otimes V^2$, this theorem aims to "justify" the following informal thinking about the problem: taking the elements in the univariate-Fourier-spaces (for x_1 and x_2) and solving a bilinear mapping, is equivalent to making a "dimensionalization" of the two univariate-Fourier-spaces and later solving the linear system in the powered space.

4.3.5 Example in general dimension. Methodology

Let us consider the Poisson's equation BVPs with null Dirichlet and Neumann boundary conditions in a $[0,1]^d$ domain. Let $V^1, V^2, \ldots, V^d$ be the $2N$-dimensional vector space of functions in Fourier summation for the $x_1, x_2, \ldots, x_d$ variables of the presented problem, respectively. Hence, we have that $\{\exp(2\pi i\ n_i\ x_i) : -N \leq n_i \leq N, n_i \neq 0\}$ conforms a basis of V^i for each $1 \leq i \leq d$.

Following a similar path as in the previous subsection, we are able to "dimensionalize" the spaces constructing the "product elements" (constructing kinds of d-multimatrices structures). We employ the outer product for constructing this space,

$$\otimes : V^1 \times V^2 \times \cdots \times V^d \longrightarrow V^1 \otimes V^2 \otimes \cdots \otimes V^d. \tag{4.33}$$

The basis tensors $T_{n_1,n_2,\ldots,n_d}$ (the multilinear forms that indicate the $n_1, n_2, \ldots, n_d$ indexed entry of the d-multimatrix) are given by

$$T_{n_1,n_2,\ldots,n_d} : v = ((v^1_{-N}, \ldots, v^1_N), (v^2_{-N}, \ldots, v^2_N), \ldots, (v^d_{-N}, \ldots, v^d_N)) \longmapsto \prod_{j=1}^{d} v^s_{n_s}. \tag{4.34}$$

We remark that to determine each element of each V^i in the cartesian product we need $2N$ real values (coordinates). Then, an element of $V^1 \otimes V^2 \otimes \cdots \otimes V^d$ needs $(2N)^d$ real values to be determined.

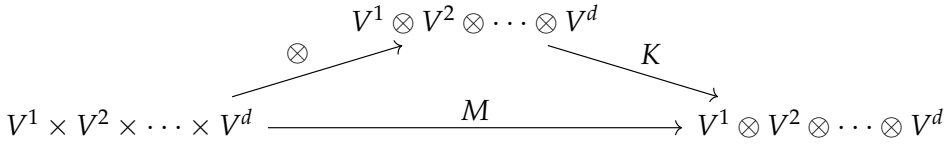

FIGURE 4.2: Problem representation

We now just want to solve the M multilinear mapping depicted in figure 4.2. To do so, we solve the $K \circ \otimes$ composition where K is the linear transformation already calculated in

(4.17). Hence, if we denote by $e_{(n_1,n_2,\ldots,n_d)} := T_{n_1,n_2,\ldots,n_d}(e_{n_1} \otimes e_{n_2} \otimes \cdots \otimes e_{n_d})$, where $e_{n_i} = \exp(2\pi i\, n_i\, x_i)$ are the basis functions of the V^i space for $1 \leq i \leq d$ and $-N \leq n_i \leq N$ with $n_i \neq 0$, we have that the linear transformation multimatrix $\mathbf{K} = [K_{(n_1,n_2,\ldots,n_d),(m_1,m_2,\ldots,m_d)}]$ has entries

$$\begin{aligned} K_{(n_1,n_2,\ldots,n_d),(m_1,m_2,\ldots,m_d)} &:= b\left(e_{(m_1,m_2,\ldots,m_d)}, \overline{e_{(n_1,n_2,\ldots,n_d)}}\right) = \\ &= 4\pi^2 \left(\sum_{s=1}^{d} \frac{n_s\, m_s}{L_s^2}\right) S_{(n_1-m_1,n_2-m_2,\ldots,n_d-m_d)}, \end{aligned} \tag{4.35}$$

where $S_{(n_1-m_1,n_2-m_2,\ldots,n_d-m_d)} = T_{n_1,n_2,\ldots,n_d}(\sigma)$, being σ the parameter function written in Fourier summation.

The "system" we want to solve is $\mathbf{KW} = \mathbf{F}$, with multivectors $\mathbf{W} = [W_{(m_1,m_2,\ldots,m_d)}]$ and $\mathbf{F} = [F_{(n_1,n_2,\ldots,n_d)}]$ such that

$$F_{(n_1,n_2,\ldots,n_d)} := F\left(\overline{e_{(n_1,n_2,\ldots,n_d)}}\right) = F_{n_1,n_2,\ldots,n_d}, \tag{4.36}$$

with $f_{n_1,n_2,\ldots,n_d} = T_{n_1,n_2,\ldots,n_d}(f)$, being f the source function written in Fourier summation.

The "system" is equivalently written in constraint terms by

$$\sum_{\substack{m_1=-N \\ m_1\neq 0}}^{N} \sum_{\substack{m_2=-N \\ m_2\neq 0}}^{N} \cdots \sum_{\substack{m_d=-N \\ m_d\neq 0}}^{N} K_{(n_1,n_2,\ldots,n_d),(m_1,m_2,\ldots,m_d)}\, W_{(m_1,m_2,\ldots,m_d)} = F_{(n_1,n_2,\ldots,n_d)}, \tag{4.37}$$

for all $-N \leq n_i \leq N, n_i \neq 0$ and $1 \leq i \leq d$.

As in the one dimensional example, we can split the $\mathbf{K}$ multimatrix as a sum of elemental matrices from which we can take out common factor the parameter function's Fourier coefficients:

$$\begin{aligned} \mathbf{K} &= 4\pi^2 \sum_{r_1=-N}^{N} \sum_{r_2=-N}^{N} \cdots \sum_{r_d=-N}^{N} \mathbf{K}_{(r_1,r_2,\ldots,r_d)} \\ &= 4\pi^2 \sum_{r_1=-N}^{N} \sum_{r_2=-N}^{N} \cdots \sum_{r_d=-N}^{N} S_{(r_1,r_2,\ldots,r_d)}\, \mathbf{C}_{(r_1,r_2,\ldots,r_d)}, \end{aligned} \tag{4.38}$$

with $\mathbf{C}_{(r_1,r_2,\ldots,r_d)} \in \mathbf{R}^{(2N)^d \times (2N)^d}$ for each $-N \leq r_1, r_2, \ldots, r_d \leq N$ being a completely determined multimatrix (with no parameters inside).

4.3.6 Rapid generation of stiffness matrices

Our BVP reads now as follows: find $\mathbf{W}$ of size $(2N)^d$ such that $\mathbf{KW} = \mathbf{F}$, where $\mathbf{W}$ and $\mathbf{F}$ are given by (4.35) and (4.36), respectively.

Moreover, we have that: (i) choices of σ are equivalent to choices of the $S_{(r_1,r_2,\ldots,r_d)}$ coefficients, and (ii) entries of $\mathbf{K}$ are already determined in an analytic way. Furthermore,

we have expressed the stiffness matrix as a sum of parameters coefficients times elemental matrices.

In addition, we consider a lower number of modes in the Fourier summation approximation of the σ parameter function in comparison to the solution function. Generally, parameter functions are piecewise constant functions because they represent physical coefficients that characterize the media. On the other side, solution functions usually have localized gradient peaks as a result of antenna sources. Material coefficients are well approximated with few Fourier modes except in the discontinuity points where they present the Gibbs effect. On the other hand, solutions require a significantly greater amount of Fourier modes. We select a 30% of material coefficient Fourier modes with respect to those employed in the solution. That is, if $N = 100$ (the solutions of the BVPs are represented by 100 modes in Fourier summation), we consider only 30 modes to approximate the material coefficients. Therefore, in the sum representation of the stiffness matrix (4.38), the only non-zero elemental matrices $\mathbf{C}_{(r_1,r_2,\ldots,r_d)}$ are those that satisfy $|r_k| \leq 30$.

In the above setting, we conclude:

- Each parameter σ corresponds to 60^d real coefficients, $S_{(r_1,r_2,\ldots,r_d)}$ such that $|r_k| \leq 30$ for each $1 \leq k \leq d$. For each coefficient, we have associated an elemental matrix $\mathbf{C}_{(r_1,r_2,\ldots,r_d)}$ which does not depend on the choice of the coefficients. Hence, if we precompute the elemental matrices and store them, the structure of the stiffness matrix is mostly set except for the componentwise products, $S_{(r_1,r_2,\ldots,r_d)}\ \mathbf{C}_{(r_1,r_2,\ldots,r_d)}$, and the final addition. The assembly of the $\mathbf{K}$ matrix is then faster than calculating all the entries of the matrix iteration by iteration.

- The sampling of the $S_{(r_1,r_2,\ldots,r_d)}$ coefficients needs to be both *fast solvable* and *sufficiently representative*

  The fast solvability refers to be able to find the solutions of the linear system rapidly in two possible ways: (i) the matrix has a form that can be fastly inverted (e.g., if we select the coefficients with $r_k \neq 0$ to be equal to zero, $\mathbf{K}$ is diagonal —in the $d = 1$ case, $\mathbf{K} = \mathbf{S}_0\ \mathbf{C}_0$—, and thus the solution is trivial), or (ii) previous computations provide useful information to solve more complex form matrices in a shorter time.

  The sufficient sampling representation means that the whole set of selected parameters coefficients is a representative set of the population of possible parameters coefficients.

The above statements are the starting point of the posterior research during the Ph.D. studies of Carlos Uriarte.

Chapter 5

Conclusions and Future Work

«Life is the art of drawing sufficient conclusions from insufficient premises.»

Samuel Butler

«I don't want happy-face conclusions. I want the truth.»

Elizabeth Warren

We have developed the current work in three main parts of content: (i) presentation and formalization of the research project and problem to be solved, (ii) technical and general introduction of the considered mathematical tool to be exploited, and finally, (iii) first proposal to solve the problem. The following two sections review the development performed in the last part and plans a future research work, respectively.

5.1 Review and conclusions

The development carried out in chapter 4 is theoretical and the obtained results are based on the following premises: (a) assumption of a Cartesian product domain, (b) Fourier summation representation of the functions involved in the problem, and (c) supposition that the parameter functions need an smaller proportion range of modes in Fourier summation than the solution function.

In the geoterrestrial framework of application, we justify premise (a) because most of the scenarios for mapping the subsoil deal with depth, length and/or width magnitudes variables, which are commonly represented in real Cartesian product domains. The same happens in premise (c), where the supposition is made in line with the nature of the piecewise constant form of parameter functions and the location of gradient peaks in the solution functions. Premise (b) allows to take advantage of the orthogonality of the Fourier system to calculate the analytic integrals.

From these hypotheses, we have obtained an explicit (analytic) expression of the entries of the stiffness matrix. This analytic expression avoids numerical integration when

building the stiffness matrix. Moreover, the stiffness matrix is represented as a linear sum of elemental matrices, each of them with an almost diagonal form. In addition, assumption (c) leads us to avoid calculating many of these elemental matrices.

In conclusion, we have managed to formulate the problem in a computational mathematics framework where FEM computations are need for creating DNN based inversion models. The analysis has led us to write a resolution methodology in very particular mathematical terms for the concerned problem: Fourier approximation based FEM (abbrev., Fourier-FEM). In consequence, the proposed methodology has some characteristics that, a priori, seem beneficial to the purpose of massive finite element computations: analytical expression of the stiffness matrix, and almost diagonal expression of it.

5.2 Future work

We plan now the future tasks to do in the already presented research project:

- **To study and propose change of coordinates systems for non-Cartesian 2D and 3D product domains for subsoil mapping applications.** The previously developed theory is consistent under assumption (a) at section 5.1. All the applications desired to be solved with the Fourier-FEM need to accomplish it. A field of interest is then how to model applications in these terms. The change of variable permits to solve these problems when the modeling has not been made in a Cartesian product domain. We want to find appropriate change of coordinates systems able to make those transformations. Moreover, alternative modelling techniques may be used to obtain Cartesian product domains in the concerned applications.

- **To study the parameters space in Fourier coefficients terms in 2D and 3D domains for subsoil mapping applications.** To be able to make a proper sampling of the problem, we need to first know the range the raw parameters have (domain), and then study their behaviour in such domain (parameter function's shape). Depending on the behavior, their corresponding Fourier coefficients have some behaviour or another. We want to know which is the domain for the Fourier coefficients. It varies from one applications to others.

- **To study and present parameters sampling options for the Fourier-FEM proposal in 2D and 3D domains for subsoil mapping applications.** The sampling needs to be both representative in the parameters population space and fast solvable in the sense explained in chapter 4. This part implies to know about both sampling techniques and (rapid) numerical linear systems solvability methods.

- **To study and propose fast solvers for massive computations of Fourier-FEM.** Once the sampling of the Fourier coefficients parameters space is established, we need to perform the simulations with high-performance. The fast solvability is going to be aimed in the following two lines: (i) to take advantage of the almost diagonal form

of the stiffness matrix, and (ii) to employ solver's first computations to extract useful information to solve more rapidly the posterior matrices which have a more complex structure.

- **To implement the Fourier-FEM algorithm.** Firstly, do it in a 1D problem and later generalize it to arbitrary dimensions employing convenient tensor libraries or modules. It is desirable to make the implementation in compiled languages (e.g., Fortrand, C or C++) instead of in interpreted languages (e.g., Python, Mathematica or MATLAB). Compiled languages assure a high and fast performance in the execution in detriment of a much more sophisticated coding compared to interpreted languages one. We propose to implement in Fortrand the computationally intensive routines and later wrap them with a Python interface. We recall that the research group in which this thesis has been developed has previously developed a Finite Element solver called *pFEM* (see [18] for more information).

- **To perform benchmarks of the presented methods.** We want to make comparisons among distinct proposed alternatives for sampling or solvavility criteria employing the Fourier-FEM.

Appendix A

Construction of a Finite Element space

To approximate the solution of the variational problem developed in section 3.1 we built a finite-dimensional subspace of $H^1([0,1])$ composed by piecewise-linear polynomials. In this appendix, we construct function spaces that are similar to that one, but which are defined on more general regions. We begin introducing the formal definition of a single *finite element*. This definition was originally given by Ciarlet in 1978 [8].

A.1 Finite Element

Let $(K, \mathcal{P}, \mathcal{N})$ be a 3-tuple (or triple) such that: (i) $K \subset \mathbf{R}^d$ is bounded and closed, with non-empty interior, and piecewise smooth boundary, (ii) $\mathcal{P}$ is a finite-dimensional space of functions on K, and (iii) $\mathcal{N} = \{N_0, N_1, \ldots, N_k\}$ is a basis of $\mathcal{P}'$. The set K is called the *element domain*, $\mathcal{P}$ is known as the space of *shape functions*, and the functionals belonging to $\mathcal{N}$ are called *nodal variables*. The system endowed by these three ingredients is called a *Finite Element* (abbrev., FE).

It is implicitly assumed that the nodal variables N_i lie in the dual space of some larger function space (e.g., a Sobolev space). Furthermore, the set $\{\phi_0, \phi_1, \ldots, \phi_k\}$ whose elements verify $N_i(\phi_j) = \delta_{ij}$ is a basis of $\mathcal{P}$ dual to $\mathcal{N}$ and it is called the *nodal basis* of $\mathcal{P}$.

An example of a one-dimensional FE is the following: let $K = [a, b] \subset \mathbf{R}$ with $a < b$, let $\mathcal{P}$ be the set of univariate polynomials with degree less than or equal to p, and let $\mathcal{N} = \{N_i\}_{i=0}^{p}$ be determined by $N_i(f) = f\left(a + \frac{(b-a)i}{k}\right)$ for each $f \in \mathcal{P}$ and each $i = 0, 1, \ldots, p$. Then, $(K, \mathcal{P}, \mathcal{N})$ is known as a *Lagrange interval*. In particular, if $K = [0,1]$, $p = 1, \mathcal{N} = \{N_0, N_1\}$ such that $N_0(f) = f(0)$ and $N_1(f) = f(1)$ for all $f \in \mathcal{P}$, then $\{\phi_0, \phi_1\}$ with $\phi_0(x) = 1 - x$ and $\phi_1(x) = x$ is a nodal basis of $(K, \mathcal{P}, \mathcal{N})$.

A.2 Examples of triangular FEs in two dimensions

In the previous section we provided a simple example of a FE in one dimension. We now construct two more examples of FEs in two dimensions with triangular shapes.

Let us denote by $\mathcal{P}_p^2$ to the space of bivariate polynomials with degree less than or equal to p, i.e., $\mathcal{P}_p^2 = \{\sum_{i=0}^{p-j}\sum_{j=0}^{p} A_{ij}x^i y^j : A_{ij} \in \mathbf{R}\}$. Then, it is easy to check that the dimension of $\mathcal{P}_p^2$ equals $d_p = (p+1)(p+2)/2$.

The following triangular FEs examples depend on the p parameter, nodal point positions and some directional derivatives criteria. In all the cases, they are of the form $(K, \mathcal{P}, \mathcal{N})$ with $K \subset \mathbf{R}^2$ being a triangle with vertices $\{v_1, v_2, v_3\}$ and edges $\{E_1 = [v_1, v_2], E_2 = [v_2, v_3], E_3 = [v_3, v_1]\}$, $\mathcal{P} = \mathcal{P}_p^2$ for some $p \in \mathbf{N}$, and the nodal variables $\mathcal{N} = \{N_1, N_2, \ldots, N_{d_p}\}$, $N_i : \mathcal{P}' \longrightarrow \mathbf{R}$, are determined by the laws described below.

- **Lagrange triangles ($p \geq 1$).** Let $\{n_i\}_{i=1}^{d_p}$ denote the set of nodes. Then: (i) $(n_1, n_2, n_3) = (v_1, v_2, v_3)$, (ii) nodes n_4 to n_{3p} lie in the three edges of K ($p-1$ per edge), and (iii) nodes n_{3p+1} to n_{d_p} are located in the interior of K. Overlappings are not allowed and the nodal variables are determined by $N_i(f) = f(n_i)$ for $i \in \{1, 2, \cdots, d_p\}$. In figure A.1 are depicted examples of representations of Lagrange Triangles for $p = 1, 2, 3$. "•" symbols indicate nodal variable evaluations at the point where the dots are located.

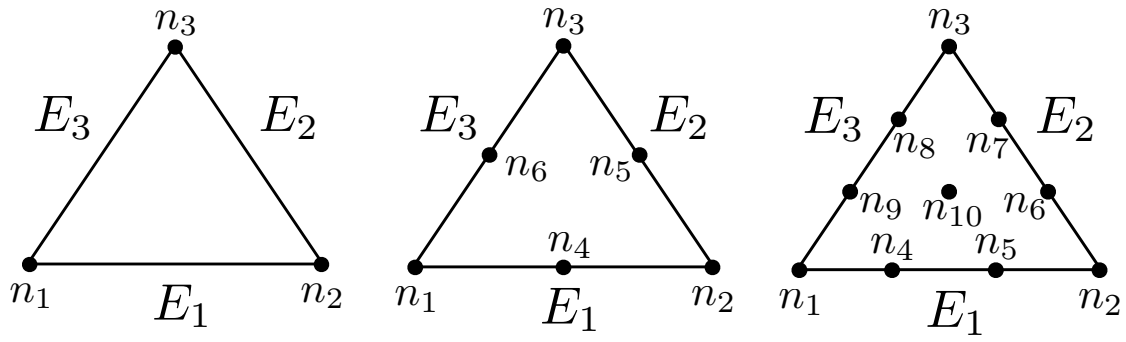

FIGURE A.1: Representations of Lagrange triangles for $p = 1, 2, 3$ from left to right, respectively

- **Hermite triangles ($p \geq 3$).** Let $\{n_i\}_{i=1}^{d_p-6}$ denote the set of nodes. Then: (i) $(n_1, n_2, n_3) = (v_1, v_2, v_3)$, (ii) nodes n_4 to n_{3p-6} are located in the interior of K, and (iii) nodes n_{3p-5} to n_{d_p-6} are placed in the edges of K ($p-3$ per edge). Overlappings are not allowed and the nodal variables are determined by: (i) $N_i(f) = f(n_i)$ for $i \in \{1, 2, \cdots, d_p - 6\}$, and (ii) $N_i(\partial_{E_j} f) = (\partial_{E_j} f)(n_i)$ for $i \in \{1, 2, 3\}$ where $\partial_{E_j} f$ denotes the E_j's directional derivative of f and j is chosen according to the edges of K that intersect in v_i (three evaluations per vertex —one normal and two of the directional derivatives—).

  See figure A.2 for a better comprehension of Hermite triangles structure. Red circles denote directional first derviatives (gradient) evaluations at the center of the circle.

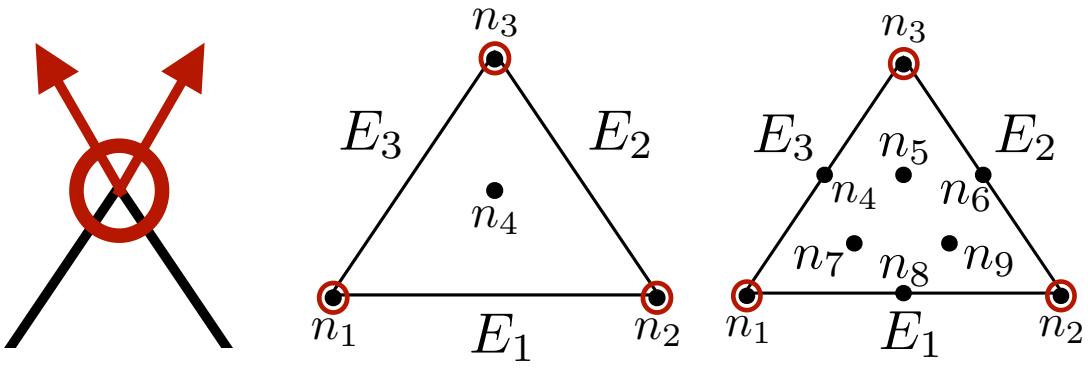

FIGURE A.2: Gradient evaluations depicted at left, and representations of Hermite triangles for $p = 3, 4$ at middle and right, respectively

A.3 The interpolant

Once we have introduced and examined a number of FEs, we join them together so as to construct subspaces of Sobolev spaces. In order to do it, we define the (local) *interpolant* of a FE (it was firstly introduced at subsection 3.1.3 in the particular case of Poisson's one-dimensional BVP).

Let $(K, \mathcal{P}, \mathcal{N})$ be a FE, let $\{\phi\}_{i=0}^{k}$ be a nodal basis of $\mathcal{P}$, and let m be the order of the highest partial derivative involved in the nodal variables of $\mathcal{N} = \{N_i : \mathcal{P}' \longrightarrow \mathbf{R} : 0 \leq i \leq k\}$. In what follows, we will refer to m as the *order* of $\mathcal{N}$. Then, the *interpolant* of a real-valued function $f \in C^m(K)$ is defined by

$$\mathcal{I}_K f := \sum_{i=0}^{k} N_i(f)\phi_i. \tag{A.1}$$

It is immediate to check that $\mathcal{I}_K$ is linear, idempotent (i.e., $\mathcal{I}_K^2 \equiv \mathcal{I}_K$ or $\mathcal{I}_K f = f$ for all $f \in \mathcal{P}$), and such that it verifies $N_i(\mathcal{I}_K f) = N_i(f)$ for all $0 \leq i \leq k$. The latter property has the interpretation that $\mathcal{I}_K f$ is the unique shape function which has the same nodal values as f.

We now formally define the concept of splitting a domain into subdomains: let $\Omega \subset \mathbf{R}^d$ be a domain and let $\{K^t\}_{t=1}^{T}$ be a finite collection of subsets of Ω such that: (i) int $K^{t_1} \cap$ int $K^{t_2} = \emptyset$ for all $t_1 \neq t_2$ and (ii) $\bigcup_{t=1}^{T} K^t = \overline{\Omega}$. Then, we say that $\{K^t\}_{t=1}^{T}$ is a *subdivision* of Ω. Furthermore, if for each $t \in \{1, 2, \ldots, T\}$ the triple $(K^t, \mathcal{P}^t, \mathcal{N}^t)$ conforms a FE (i.e., there exist a space of functions $\mathcal{P}^t$ and a collection of nodal variables $\mathcal{N}^t$ associated to the element domain K^t), we say that $\mathcal{G} = \{(K^t, \mathcal{P}^t, \mathcal{N}^t) : 0 \leq t \leq T\}$ is a *finite element grid* (or simply a *grid*) for Ω. In this setting, we call *order* of $\mathcal{G}$, and denote it by m, to the highest order of the collections of nodal variables (i.e., $m = \max_{1 \leq t \leq T}\{m_t : m_t$ is the order of $\mathcal{N}^t\}$).

Putting the previous ingredients together, we are able to generalize globally the previous (local) interpolant: let Ω be a domain and let $\mathcal{G} = \{(K^t, \mathcal{P}^t, \mathcal{N}^t)\}_{t=1}^{T}$ be a grid for Ω with order m. If $f : \Omega \longrightarrow \mathbf{R}$ such that $f \in C^m(\overline{\Omega})$, then the *global interpolant* of f with

respect to $\mathcal{G}$ is denoted by $\mathcal{I}_{\mathcal{G}} f$ and defined by $\mathcal{I}_{\mathcal{G}} f|_{K_t} = \mathcal{I}_{K_t} f$ for all $t \in \{1, 2, \ldots, T\}$.

Thereby, in the setting developed at chapter 3 for the Galerkin approximation, we may use a finite-dimensional space of piecewise polynomial functions as the testing space:

$$V^h = \{\mathcal{I}_{\mathcal{G}} f : f \in C^m(\overline{\Omega}) \text{ and } f|_{\partial\Omega_D} \equiv 0\}, \tag{A.2}$$

with $\mathcal{G}$ being a grid for Ω, m its order, and $\mathcal{I}_{\mathcal{G}}$ the global interpolant operator.

Without further assumptions on the subdivision, no continuity properties can be assured for the global interpolant. However, when it does, we say that a global interpolant has continuity order r whenever $\mathcal{I}_{\mathcal{G}} f \in C^r(\overline{\Omega})$ for all $f \in C^m(\overline{\Omega})$. Likewise, the space of piecewise polynomial functions $\{\mathcal{I}_{\mathcal{G}} f : f \in C^m(\overline{\Omega})\}$ is said to be a C^r *finite element space*.

We now formally present a particular kind of two-dimensional domain subdivision in triangular subdomains: if $\Omega \subset \mathbf{R}^2$ is a polygonal domain and $\mathcal{T} = \{(K^t, \mathcal{P}^t, \mathcal{N}^t)\}_{t=1}^T$ is a grid for Ω, then $\mathcal{T}$ is called a *triangulation* of Ω when each K^t is a triangle and no of its vertices lie in the interior of an edge of another triangle for $1 \leq t \leq T$.

The following theorem states a result according to the degree of continuity the space of interpolant functions have when a triangulation with the previous introduced triangles is performed in a polygonal domain.

Theorem A.1. *The Lagrange and Hermite elements are C^0. If $\Omega \subset \mathbf{R}^2$ is a polygonal domain and $\mathcal{T} = \{(K^t, \mathcal{P}^t, \mathcal{N}^t)\}_{t=1}^T$ is a triangulation of Ω made up of Lagrange or Hermite elements (all the elements are of the same kind in the triangulation), then it is possible to choose edge nodes in the triangles K^t such that the global interpolant is C^1 and such that the order of $\mathcal{G}$ is equal to $m = 0$ or $m = 1$, respectively. Furthermore, it suffices for each edge with vertices v_i^1 and v_i^2 to have $p - 1 - 2m$ nodes $\{\theta_i^j(v_i^1 - v_i^2) + v_i^1\}_j$ with the set of real parameters $\{\theta_i^j : 1 \leq j \leq p - 1 - 2m\}$ being symmetric around $1/2$.*

Moreover, under these hypotheses, $\mathcal{I}_{\mathcal{T}} f \in W_\infty^{r+1}(\Omega)$.

Proof. See [7, p. 81] □

In essence, this is the general structure to follow on a FEM based on the Galerkin's approximation theory: first we divide a domain into (possibly simpler) subdomains, then we solve in each subdomain the corresponding subproblem by employing the Galerkin's approximation method, and finally, we piece together all the subsolutions in a global solution through the global interpolant function.

Generally, polynomial functions are the most used to approximate the subsolutions elementwise. However, other kind of approximating shape functions may be considered, as it was the case with the Fourier system in chapter 4.

Bibliography

[1] R. A. Adams and J. Fournier. *Sobolev spaces*, volume 140. Elsevier, 2003.

[2] J. Alvarez-Aramberri and D. Pardo. Dimensionally adaptive hp-finite element simulation and inversion of 2d magnetotelluric measurements. *Journal of Computational Science*, 18:95–105, 2017.

[3] S. Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196, 1993.

[4] M. E. Baron. *The origins of the infinitesimal calculus*. Courier Corporation, 2003.

[5] B. Bogosel. Weak formulation for Laplace Equation with Robin boundary conditions. https://mathproblems123.wordpress.com/2012/10/22/weak-formulation-for-laplace-equation-with-robin-boundary-conditions/, 2012. [Last Accessed: 21th June 2019].

[6] B. A. Bolt, W. L. Horn, G. A. MacDonald, and R. F. Scott. *Geological Hazards: Earthquakes-tsunamis-volcanoes-avalanches-landslides-floods*. Springer Science & Business Media, 2013.

[7] S. Brenner and R. Scott. *The mathematical theory of finite element methods*, volume 15. Springer Science & Business Media, 2007.

[8] P. G. Ciarlet. *The finite element method for elliptic problems*, volume 40. SIAM, 2002.

[9] John B Conway. *A course in functional analysis*, volume 96. Springer Science & Business Media, 2013.

[10] N. Diamanti and A. Giannopoulos. Implementation of adi-fdtd subgrids in ground penetrating radar fdtd models. *Journal of Applied Geophysics*, 67(4):309–317, 2009.

[11] Y. V. Egorov and M. A. Shubin. *Foundations of the classical theory of partial differential equations*, volume 30. Springer Science & Business Media, 2013.

[12] C. L. Fefferman. Existence and smoothness of the Navier-Stokes equation. *The millennium prize problems*, 57:57–67, 2006.

[13] P. G. Fookes. Geology for engineers: the geological model, prediction and performance. *Quarterly Journal of Engineering Geology and Hydrogeology*, 30(4):293–424, 1997.

[14] A. Fournier. Review of machine-learning applications in exploration geophysics, 10 2017.

[15] I. B. Fridleifsson. Geothermal energy for the benefit of the people. *Renewable and sustainable energy reviews*, 5(3):299–312, 2001.

[16] M. W. Gardner and S. R. Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

[17] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[18] Mathmode Research Group. MATHMODE Website. https://sites.google.com/prod/view/mathmode, 2018. [Last Accessed: 5th September 2019].

[19] K. K. Gupta and J. L. Meek. A brief history of the beginning of the finite element method. *International journal for numerical methods in engineering*, 39(22):3761–3774, 1996.

[20] J. Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*, volume 37. Yale University Press, 1923.

[21] P. R. Halmos. *Measure theory*, volume 18. Springer, 2013.

[22] D. M. Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[23] R. Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier, 1992.

[24] L. Hogben. *Handbook of linear algebra*. Chapman and Hall/CRC, 2013.

[25] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[26] T. J. R. Hughes. *The finite element method: linear static and dynamic finite element analysis*. Courier Corporation, 2012.

[27] D. Hunter, H. Yu, M. S. Pukish III, J. Kolbusz, and B. M. Wilamowski. Selection of proper neural network sizes and architectures—a comparative study. *IEEE Transactions on Industrial Informatics*, 8(2):228–240, 2012.

[28] V. Isakov. *Inverse problems for partial differential equations*, volume 127. Springer, 2006.

[29] R. A. Jacobs. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295–307, 1988.

[30] K. Janocha and W. M. Czarnecki. On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*, 2017.

[31] V. Keilis-Borok and A. A. Soloviev. *Nonlinear dynamics of the lithosphere and earthquake prediction*. Springer Science & Business Media, 2013.

[32] K. Key and J. Ovall. A parallel goal-oriented adaptive finite element method for 2.5-d electromagnetic modelling. *Geophysical Journal International*, 186(1):137–154, 2011.

[33] Y. Kim and N. Nakata. Geophysical inversion versus machine learning in inverse problems. *The Leading Edge*, 37(12):894–901, 2018.

[34] T. W. Körner. *Fourier analysis*. Cambridge university press, 1989.

[35] J. M. Landsberg. Tensors: geometry and applications. *Representation theory*, 381(402):3, 2012.

[36] V. Mackevicius. *Integral and Measure: From Rather Simple to Rather Complex*. John Wiley & Sons, 2014.

[37] N. G. Meyers and J. Serrin. $H = W$, PYOC. In *Nat. Acad. Sci. USA*, volume 51, pages 1055–1056, 1964.

[38] M. N. Nabighian and M. W. Asten. Metalliferous mining geophysics—state of the art in the last decade of the 20th century and the beginning of the new millennium. *Geophysics*, 67(3):964–978, 2002.

[39] T. M. Nguyen. N-dimensional quasipolar coordinates-theory and application. 2125, 2014.

[40] University of Cambridge. What is a Tensor? https://www.doitpoms.ac.uk/tlplib/tensors/index.php, 2008. [Last Accessed: 25th August 2019].

[41] D. Pardo, V.M. Calo, C. Torres-Verdín, and M. J. Nam. Fourier series expansion in a non-orthogonal system of coordinates for the simulation of 3d-dc borehole resistivity measurements. *Computer Methods in Applied Mechanics and Engineering*, 197(21-24):1906–1925, 2008.

[42] D. Pardo, L. Demkowicz, C. Torres-Verdín, and L. Tabarovsky. A goal-oriented hp-adaptive finite element method with electromagnetic applications. part i: electrostatics. *International Journal for Numerical Methods in Engineering*, 65(8):1269–1309, 2006.

[43] L. Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

[44] H. L. Royden. *Real analysis*. Krishna Prakashan Media, 1968.

[45] D. W. Ruck, S. K. Rogers, and M. Kabrisky. Feature selection using a multilayer perceptron. *Journal of Neural Network Computing*, 2(2):40–48, 1990.

[46] W. Rudin. *Real and complex analysis*. Tata McGraw-hill education, 2006.

[47] L. Schwartz and Institut de mathématique (Strasbourg). *Théorie des distributions*, volume 2. Hermann Paris, 1957.

[48] M. Shahriari, D. Pardo, A. Picón, A. Galdrán, J. Del Ser, and C. Torres-Verdín. A Deep Learning Approach to the Inversion of Borehole Resistivity Measurements. *arXiv preprint arXiv:1810.04522*, 2018.

[49] G. D. Smith. *Numerical solution of partial differential equations: finite difference methods*. Oxford university press, 1985.

[50] Elias M Stein. *Singular integrals and differentiability properties of functions*, volume 2. Princeton university press, 1970.

[51] J. Stillwell. Mathematics and its history. *The Australian Mathem. Soc*, page 168, 2002.

[52] A. Tarantola. A strategy for nonlinear elastic inversion of seismic reflection data. *Geophysics*, 51(10):1893–1903, 1986.

[53] A. Tarantola. *Inverse problem theory and methods for model parameter estimation*, volume 89. siam, 2005.

[54] I. V. Tetko, D. J. Livingstone, and A. I. Luik. Neural network studies. 1. comparison of overfitting and overtraining. *Journal of chemical information and computer sciences*, 35(5):826–833, 1995.

[55] P. Vanapalli. Backprop is very simple. Who made it Complicated? https://github.com/Prakashvanapalli/TensorFlow/blob/master/Blogposts/Backpropogation_with_Images.ipynb, 2017. [Last Accessed: 4th June 2019].

[56] V. S. Vladimirov. Equations of mathematical physics. *Moscow Izdatel Nauka*, 1976.

[57] Curtis R Vogel. *Computational methods for inverse problems*, volume 23. Siam, 2002.

[58] E. Zauderer. *Partial differential equations of applied mathematics*, volume 71. John Wiley & Sons, 2011.

[59] O. C. Zienkiewicz, R. L. Taylor, P. Nithiarasu, and J. Z. Zhu. *The finite element method*, volume 3. McGraw-hill London, 1977.

[60] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.

[61] S. Zlotnik, P. Díez, D. Modesto, and A. Huerta. Proper generalized decomposition of a geometrically parametrized heat problem with geophysical applications. *International Journal for Numerical Methods in Engineering*, 103(10):737–758, 2015.