

OPTIMIZATION AND NUMERICAL ANALYSIS OF
PDE-CONSTRAINED OPTIMIZATION PROBLEMS WITH
APPLICATIONS TO MAXWELL'S EQUATIONS, BOUNDED
VARIATION AND NEURAL NETWORKS

by

Hugo Díaz

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Applied Mathematics

Spring 2023

© 2023 Hugo Díaz
All Rights Reserved

OPTIMIZATION AND NUMERICAL ANALYSIS OF
PDE-CONSTRAINED OPTIMIZATION PROBLEMS WITH
APPLICATIONS TO MAXWELL'S EQUATIONS, BOUNDED
VARIATION AND NEURAL NETWORKS

by

Hugo Díaz

Approved: _____
Mark Gockenbach, Ph.D.
Chair of the Department of Mathematical Sciences

Approved: _____
John A. Pelesko, Ph.D.
Dean of the College of Arts and Sciences

Approved: _____
Louis F. Rossi, Ph.D.
Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____
Peter Monk, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____
Harbir Antil, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____
Constantin Bacuta, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____
Armin Schikorra, Ph.D.
Member of dissertation committee

ACKNOWLEDGEMENTS

First, I want to thank Francisco-Javier Sayas and Gabriel Gatica for encouraging me to pursue a Ph.D. I would also like to express my gratitude to Prof. Dr. Harbir Antil for supervising my thesis and supporting me throughout the years. At the same time, I would like to thank Dr. Evelyn Herberg and Dr. Armin Schikorra for their invaluable contributions. I consider myself really lucky to have worked with them. My discussions with both of them about neural networks, fractional derivatives, and functional analysis taught me a lot. Lastly, but not least, I am also thankful for the help and encouragement I received from my family and friends.

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	ix
ABSTRACT	xi
Chapter	
1 INTRODUCTION	1
2 MAXWELL BOUNDARY VALUE PROBLEM	7
2.1 Motivation	7
2.2 Problem Setup	8
2.3 Notation and preliminary results	11
2.3.1 Tangential traces and Green’s identities for $H(\mathbf{curl}; \Omega)$	12
2.3.2 Well-posedness of the state equation	14
2.4 Reduced cost functional and its derivative	16
2.4.1 Wirtinger derivatives on complexified Hilbert spaces	18
2.4.2 The \mathbb{R} –linear derivative of the reduced cost functional and optimality conditions	22
2.4.3 Additional regularity of \mathbb{S}^* for Lipschitz polyhedra	25
2.5 Discrete Problem	28
2.5.1 Imposition of discrete boundary conditions	28
2.5.2 Discrete solution operator and cost functional	29
2.5.3 Particular choice of discrete control space	31
2.5.4 Discrete Adjoint equation and optimality conditions	36
2.5.5 Convergence of fully discrete scheme	37

2.5.6	Using $\ \operatorname{curl}_{\Gamma}\mathbf{z}\ _{0,\Gamma}^2$ as the regularization term	39
2.6	Numerical results	41
2.6.1	Code validation for Nédélec elements	41
2.6.2	Validation optimization routines	43
2.6.3	Convergence of optimization problem	43
3	AN OPTIMAL TIME VARIABLE LEARNING FRAMEWORK FOR DEEP NEURAL NETWORKS	45
3.1	Introduction	45
3.2	Preliminaries	48
3.2.1	Caputo fractional derivative	49
3.2.2	Deep Learning problem	50
3.3	Continuous DNNs	51
3.3.1	Ordinary Differential Equations and Neural Networks	52
3.3.2	Stability of continuous Fractional-DNN	54
3.4	Network architectures with fixed τ -parameter	56
3.5	Variable- τ framework for DNNs	58
3.5.1	ResNet with variable τ	59
3.5.2	Fractional-DNN with variable τ	60
3.6	Vanishing and exploding gradients	64
3.7	Numerical results	70
3.7.1	Maxwell's equations	71
4	NONLOCAL BOUNDED VARIATIONS WITH APPLICATIONS	78
4.1	Introduction	78
4.2	Fractional BV in the Riesz sense	81
4.3	Fractional BV in the Gagliardo sense	88
4.4	Image Denoising and Predual Problem	98
	BIBLIOGRAPHY	115
	Appendix	

A	DERIVATIVES OF THE LAGRANGIAN \mathcal{L}	127
A.0.1	Derivative with respect to $y^{[\ell]}$	127
A.0.2	Derivative with respect to $\tau^{[\ell]}$	129
B	SCALING IN L^P-NORMS AND STAR-SHAPED DOMAINS	131

LIST OF TABLES

3.1	Notation	49
3.2	Optimal learned τ variables for various DNN architectures with τ -variable framework with 6 layers and 2 layers. These are the same network architectures that are considered in Figure 3.3.	74

LIST OF FIGURES

1.1	Feedforward neural network	3
2.1	$F_+ \cup \mathbf{e} \cup F_-$, $\boldsymbol{\psi}_e _{F_+}$ and $(\mathbf{n}_{F_+} \times \boldsymbol{\psi}_e) _{F_+}$	32
2.2	Altitudes and outer normals of F	33
2.3	Affine transformation between \widehat{F} and F	36
2.4	Electrode	41
2.5	Log-log plot \mathcal{E}_h vs h , and coarsest mesh along with element-wise error.	43
2.6	<p>Left: Given a random direction $\boldsymbol{\xi}$, the panel shows the difference between $d^{\mathbb{R}}j_h(\mathbf{z}; \boldsymbol{\xi})$ and its finite difference approximation. As expected, we observe a linear rate of convergence. Right: We let $\alpha = 1e^{-3}$ and $\beta = 0$ in the cost functional $\mathcal{J}(\cdot)$. Let $\mathcal{J}_1(\mathbf{u}, \mathbf{z}) := \frac{1}{2}\ \mathbf{u} - \mathbf{u}_d\ ^2$ and $\mathcal{J}_2(\mathbf{z}) := \frac{\alpha}{2}\ \text{curl}_{\Gamma}\mathbf{z}\ _{L^2(\Gamma)}^2$. Moreover, let \mathbf{z} be the optimal control corresponding to the finest mesh. Then the three curves show $\mathcal{J}(\mathbf{u}_h, \mathbf{z}_h) - \mathcal{J}(\mathbf{u}, \mathbf{z}) /\mathcal{J}(\mathbf{u}, \mathbf{z})$, $\mathcal{J}_1(\mathbf{u}_h, \mathbf{z}_h) - \mathcal{J}_1(\mathbf{u}, \mathbf{z}) /\mathcal{J}_1(\mathbf{u}, \mathbf{z})$, and $\mathcal{J}_2(\mathbf{z}_h) - \mathcal{J}_2(\mathbf{z}) /\mathcal{J}_2(\mathbf{z})$ as $h \rightarrow 0$.</p>	44
3.1	The panel shows the mean squared error during training when the variable- τ framework is applied to a ResNet and a Fractional-DNN for an ill-posed 3D-Maxwell's equation. More details are available in Section 3.7.	46
3.2	Left: Optimal weights and biases for ResNet with variable τ with 5 hidden layers and 10 nodes each with bias ordering. Right: Reduced ResNet with 2 hidden layers, i.e., hidden layers 1 and 3 from the larger network. The color of the dots indicates the bias value, and the color of the lines indicates the magnitude of the weight.	73

3.3	Comparison between various DNN architectures and FEM. Top row: L^2 error between an exact solution and DNN approximation (6 hidden layers with a width of 50 each) or FEM approximation. B.O. indicates bias ordering. The left and right panels correspond to fixed and variable τ , respectively. Bottom row: The left panel shows the mean squared error during training of different DNNs with 2 hidden layers with a width of 50 nodes each. The right panel displays the L^2 error between an exact solution and a DNN approximation for the same DNNs and FEM.	75
3.4	Comparison of testing errors between ResNet, ResNet with τ -learning framework, Fractional-DNN and Fractional-DNN with τ -learning framework (2-50).	76
3.5	\mathbf{u} , \mathbf{u}^{NN} and pointwise error on Ω , at $x_3 = 0.5$ (x_1x_2 -plane).	76
B.1	Examples of star-shaped sets with discontinuous λ . Both sets are star-shaped with respect to the origin, and the first has even Lipschitz continuous boundary – however the conclusions of Lemma B.0.1 are not true.	133
B.2	Assuming that the ball $B(0, a)$ in the proof is actually equal to $B(0, 1)$ (which can always be obtained by scaling) the above figure explains the proof of Lemma B.0.2. Left: if $\lambda(x_k) < \lambda(\bar{x}) - \varepsilon$ and x_k is sufficiently close to \bar{x} then $\lambda(x_k)x_k$ must belong to the cone A . Right: if $\lambda(x_k) > \lambda(\bar{x}) + \varepsilon$ and x_k is sufficiently close to \bar{x} then \bar{x} must belong to A_k (using that the cone A_k has a minimal aperture that does not change and is determined by $B(0, a)$ as k changes)	136

ABSTRACT

The analysis and numerical discretization of a variety of problems related to electromagnetism, neural network architectures motivated by fractional time derivatives, and an image denoising problem based on some nonlocal differential operators are the focus of this thesis. Some of these optimization problems fall in the category of inverse problems and others are optimization problems with partial differential equations (PDEs) as constraints. A common thread linking them is the fact that they are usually ill-posed.

This thesis is divided into three chapters. In the first chapter, we consider a boundary control problem for Maxwell's equations in the frequency domain. The Wirtinger derivative is used in order to find the optimality conditions because the cost functional is not complex differentiable. The Rao-Wilton-Glisson basis and high-order Nédélec elements are used for the numerical discretization of the control and state equation, respectively. A modified version of the BFGS method is used for the numerical optimization.

In the second chapter, we consider a type of network architecture based on the discretization of a fractional time derivative. We consider a scaling factor for the activation functions, which is based on an adaptive time-stepping method for ODEs. This method may be used to remove unnecessary layers and help with the vanishing gradient problem. We also include several numerical experiments that support and illustrate our theoretical findings.

Finally, in the third chapter, we proposed two fractional bounded variation (BV) spaces based on the Riesz and Gagliardo gradients. We demonstrate that analogous properties of the BV space, such as lower semicontinuity and Sobolev embeddings, are still valid for the fractional case. However, the relationship with the space $W^{\alpha,1}$ is

different. This framework is used to create a fractional version of the total variation denoising model.

Chapter 1

INTRODUCTION

This thesis explores a variety of optimization problems with emphasis on optimization problems constrained by partial differential equations (PDEs) [13]. We hereby term these problems as PDECO problems. PDECO has several applications, including shape optimization, weather prediction, and inverse problems. It commonly uses the underlying physics or knowledge about the problem, which is typically linked to some PDE system, to solve control, design, and inverse problems. All of the problems discussed in this work may be formulated as the following constrained optimization problem:

$$\min_{\mathbf{u} \in \mathcal{U}} F(\mathbf{u}) + G(\Lambda \mathbf{u}), \quad (\mathcal{P})$$

for suitable operators F , G and Λ , where \mathcal{U} is a suitable Banach space. Some classic problems with this structure include:

- Noise removal: $\min_{u \in \mathcal{U}} |Du|(\Omega) + \frac{\alpha}{2} \|u - g\|^2$
- Obstacle problem: $\min_{u \in \mathcal{U}} \frac{1}{2} \int_{\Omega} |\nabla u|^2 dx - \int_{\Omega} f u dx + I_+(u)$
- Mossolov's problem: $\min_{u \in \mathcal{U}} \frac{\alpha}{2} \int_{\Omega} |\nabla u|^2 dx + \beta \int_{\Omega} |\nabla u| dx - \int_{\Omega} f u dx$
- Filtering-theory: $\min_{u \in \mathcal{U}} \frac{1}{2} \int_{\Omega} (\Delta u)^2 dx - \int_{\Omega} f u dx + I_+(u)$.

Here I_+ denotes an indicator function; for more details on these models, see [3, 62, 128]. Identifying the right operators F , G and Λ and the set \mathcal{U} simplifies the theoretical and numerical analysis of the problem. From a numerical optimization perspective, an appropriate splitting of the cost functional allows the use of algorithms like the

alternating direction method of multipliers (ADMM) or the primal-dual algorithm, cf. [29]. In the context of convex optimization, the concept of *dual* is connected to the Legendre-Fenchel transform, which is crucial for convex optimization. The dual problem, which is built on a family of perturbations, connects a minimization problem with a maximization problem. For example, the dual problem of (\mathcal{P}) is given by:

$$\sup_{\mathbf{p}^*} \left[-F^*(\Lambda^* \mathbf{p}^*) - G^*(-\mathbf{p}^*) \right]. \quad (\mathcal{P}^*)$$

In Chapter 4, we utilize a method known as *pre-dual* to study the theoretical properties of a new image denoising model that utilizes fractional derivatives. The existence and uniqueness of the solutions to the optimization problems under consideration in this study are established using the direct method in the calculus of variations and the method of Lagrange multipliers.

Recent advances in software, computer architecture, numerical simulations, and algorithms have made it possible to handle massive optimization problems. One application of such advances is modern artificial neural networks; see [32]. In this work, we took a different approach; we tried to simplify the numerical methods as much as possible by relying on some properties of the problem. In Chapter 2, for instance, we present a simple strategy based on the so-called Wirtinger derivative for dealing with optimization problems formulated on a complex Hilbert space that do not need rewriting as real-valued problems. This enables the acceleration of gradient-based approaches that rely on solving numerous linear systems, which may be expensive to solve otherwise. We utilize this strategy to solve a problem involving Maxwell's equations in the frequency domain, where we also employed high-order Nédélec elements to address the PDE discretization. In Chapter 3, we also use neural networks to approximate the solution of Maxwell's equations. This set of equations are crucial in the study of the electromagnetic phenomena. From 1856 through 1865, James Clerk Maxwell published a series of articles in which he generalized and unified all electrodynamic principles, see [113]. Oliver Heaviside later wrote them in modern vector calculus notation. They

relate the electric field E , the current density J , and the magnetic field B , which in their differential form are given by:

- Gauss' law for electricity: $\nabla \cdot E = \frac{\rho}{\varepsilon_0}$,
- Gauss' law for magnetism: $\nabla \cdot B = 0$,
- Faraday's law of induction: $\nabla \times E = -\frac{\partial B}{\partial t}$,
- Ampère's circuital law: $\nabla \times B = \frac{1}{\varepsilon_0 c^2} J + \frac{1}{c^2} \frac{\partial B}{\partial t}$,

see [115] for more details.

For the neural networks, we considered a *Feedforward architecture*, which is a type of architecture where the information flows only forward, from left to right; see Figure 1.1. This *causal* structure is shared with some discrete dynamical systems and

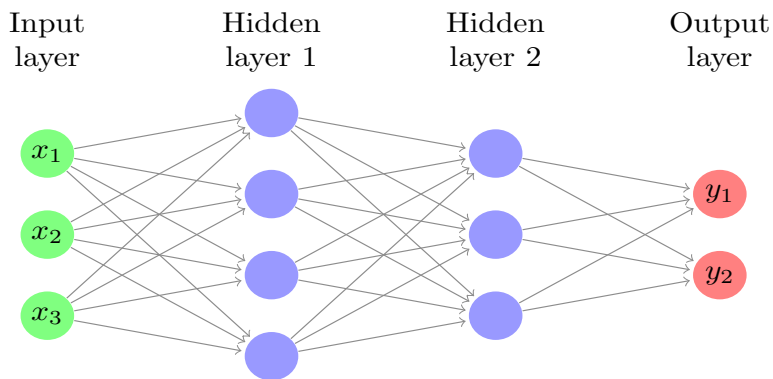


Figure 1.1: Feedforward neural network

has been studied in [47], where the authors address the connection between feedforward networks and the discretization of an ordinary differential equation. Consider, for instance,

$$y'(t) = \sigma(t, y(t)) \quad \text{for all } t \in (0, 1)$$

$$y(0) = y_0.$$

Euler's method approximates $y(\cdot)$ at $t_n = n \cdot \Delta t$ by y_n via the following recurrence relation:

$$y_n = y_{n-1} + \Delta t \cdot \sigma(t_{n-1}, y_{n-1}) \quad \text{with } n \in \mathbb{N}, \tag{1.0.1}$$

which resembles a ResNet’s architecture, cf. [77]. However, certain events, such as delayed effects or hysteresis cannot be modelled by the standard concept of derivatives, which are local operators. Fractional derivatives are a common technique to model such nonlocal phenomena. Following [118], we consider the left-sided Caputo fractional derivative of order γ , which is given by:

$$\partial_t^\gamma u(x, t_{k+1}) = \frac{1}{\Gamma(1-\gamma)} \int_0^{t_{k+1}} \frac{1}{(t_{k+1}-\tau)^\gamma} \partial_t u(x, \tau) d\tau, \quad (1.0.2)$$

where Γ is Euler’s Gamma function. And a finite difference scheme for (1.0.2) is given by:

$$\partial_t^\gamma u(\cdot, t_{k+1}) \approx \frac{1}{\Gamma(2-\gamma)} \sum_{j=0}^k a_j \cdot \left(\frac{u(\cdot, t_{k+1-j}) - u(\cdot, t_{k-j})}{(\Delta t)^\gamma} \right),$$

where $a_j = (j+1)^{1-\gamma} - (j)^{1-\gamma}$; see (3.5.4) for more details. It is worth mentioning that we consider an *adaptive step size* method in time, but in contrast with the standard theory of ODEs, a small time step will mean to us that we could ignore a given layer; see Figure 3.2. Note that this is not possible for a Feedforward architecture neural network without the additional connections between layers present in the Fractional Neural Network architecture.

We also investigate two types of spatial-fractional derivatives in Chapter 4; the Riesz and Gagliardo types. These fractional derivatives allow us to define the concepts of fractional gradient and divergence, and once we have both, we can describe first-order fractional systems and bounded variations. The Riesz type of fractional derivative is related to the well-known Riemann–Liouville integral, and it is based on operator *Riesz potential*: $I^\alpha f(x) := \mathcal{F}^{-1} (|\xi|^{-\alpha} \mathcal{F}f(\xi)) (x)$, then it is possible to define a notion of fractional gradient and divergence, namely:

$$\begin{aligned} D^\alpha f &:= DI^{1-\alpha} f, \\ \text{Div}_\alpha f &:= \text{div } I^{1-\alpha} f. \end{aligned}$$

In turn, the Gagliardo fractional derivative defines a more “local” type of fractional gradient and divergence:

$$\begin{aligned}
 (d_\alpha f)(x, y) &:= \frac{f(x) - f(y)}{|x - y|^\alpha}, \\
 (\operatorname{div}_\alpha F)(x) &:= - \int_{\mathbb{R}^n} \frac{F(x, y) - F(y, x)}{|x - y|^{n+\alpha}} dy.
 \end{aligned}$$

Having a notion of fractional divergence allow us to construct a fractional bounded variation minimization-based image denoising scheme.

Outline: In Chapter 2, we present a boundary control for Maxwell’s equations in the frequency domain. A surface type of curl operator is shown to be the appropriate regularization in order for the optimal control problem to be well-posed. Since, all underlying variables are assumed to be complex valued, the standard results on differentiability do not directly apply. Instead, we extend the notion of Wirtinger derivatives to complexified Hilbert spaces. Then, optimality conditions are derived and higher order boundary regularity of the adjoint variable is established. The state and adjoint variables are discretized using higher order Nédélec finite elements. The finite element space for controls is identified as a space which preserves the structure of the control regularization. Convergence of the fully discrete scheme is established. The theory is validated by numerical experiments, in some cases, motivated by realistic applications.

Later in Chapter 3, we investigate a type of neural network architecture that permits long-term memory via a discretization of the left-sided Caputo fractional derivative. The innovation of this approach comes in letting the discretization parameter (time-step size) change from layer to layer in an optimization framework, which must be learned. The suggested architecture is applicable to any current network, including ResNet, DenseNet, and Fractional-DNN. This approach is shown to help in overcoming the vanishing and exploding gradient issues. The stability of various existing continuous DNNs, such as the Fractional-DNN, is also investigated. The proposed method is used to approximate the solution of a 3D Maxwell’s equation.

Finally, in Chapter 4, we study the features of two fractional bounded variation (BV)-type spaces, which are motivated by problems involving jumps over lower-dimensional subsets and abrupt transitions across interfaces. One space is induced from the Riesz-fractional gradient, which Comi-Stefani recently studied in [52]; and the other is induced by the Gagliardo-type fractional gradient often used in Dirichlet forms and Peridynamics – this one is naturally related to the Caffarelli-Roquejoffre-Savin fractional perimeter, see [38]. As an application, novel image denoising models are proposed and their corresponding Fenchel pre-dual formulations are developed utilizing the features of these spaces. The latter requires a dense set of smooth functions with compact support. This density property is established for convex domains.

Parts of this thesis have appeared in the following articles:

- Boundary control of time-harmonic eddy current equations; with Harbir Antil, [7].
- An Optimal Time Variable Learning Framework for Deep Neural Networks; with Harbir Antil and Evelyn Herberg, [8].
- Nonlocal Bounded Variations with Applications; with Harbir Antil, Tian Jing and Armin Schikorra, [9].

Chapter 2

MAXWELL BOUNDARY VALUE PROBLEM

2.1 Motivation

In recent years, problems related to controllability and optimal control constrained by Maxwell's equations have gained a large amount of attention in both time and frequency domains; see, for instance, a series of works [18, 25, 40, 93, 95, 96, 97, 117, 145, 146, 152, 153, 154, 155]. Most of the existing literature focuses on the case where the control is in the interior, though the boundary control case can be found in some limited number of references [18, 95, 96, 97]. The novelty in this chapter is that it works in the frequency domain, whereas previous works were in the time domain. It also deals with appropriate tangential traces (control space) of $H(\text{curl})$ when the physical domain is only Lipschitz polyhedral, in contrast to earlier publications in which either no numerical method is given or the domain is assumed to be smooth. As a result, the method presented in those articles is rather impractical. There are no numerical examples given. In addition, we provide a full convergence analysis for the fully discrete scheme as well as a numerical implementation using finite elements. Furthermore, this method applies in a generic situation when all the variables are complex valued, and it introduces a framework for performing complex differentiation in Hilbert spaces.

Maxwell's equations with non-homogeneous boundary conditions in a non-smooth setting (Lipschitz polyhedral domains) is a relatively recent topic [33, 34, 143]. In addition, the underlying functional analytic framework is delicate. Nonetheless, non-homogeneous boundary conditions with non-smooth boundaries do occur in realistic

applications, such as microwave ovens, see also [26]. It is crucial to resolve the physical domain, for instance, using the finite element method, which permits a simple implementation and has a well-studied theoretical framework [115].

Motivated by [26], the articles [21, 22] introduced yet another boundary value problem for Maxwell equation (cf. (2.6.1)) where the boundary conditions are on certain electrodes. Notice that the non-homogeneous boundary conditions of the type considered here also arise in the scattering theory of electromagnetic fields. For instance, the incident and scattered fields denoted by \mathbf{E}^i and \mathbf{E}^s , respectively, satisfy the “boundary” condition: $\mathbf{E}^s \times \mathbf{n} = -\mathbf{E}^i \times \mathbf{n}$. Here \mathbf{n} is the outward unit normal. These works forms the motivation for us to study boundary optimal control of time-harmonic Maxwell’s equations.

2.2 Problem Setup

Let $\Omega \subset \mathbb{R}^3$ be a polyhedron with a Lipschitz continuous boundary denoted by Γ . Moreover, let the current density $\mathbf{j}_c \in L^2(\Omega; \mathbb{C}^3)$, a desired field vector \mathbf{u}_d in $L^2(\Omega; \mathbb{C}^3)$, and symmetric and positive definite functions $\boldsymbol{\kappa}$ and $\boldsymbol{\mu}$ in $L^\infty(\Omega; \mathbb{R}^{3 \times 3})$, $\omega \neq 0$, positive constants α and β , and a suitable lower semicontinuous convex function f . If (\mathbf{u}, \mathbf{z}) represents the state-control pair, then the goal of this work is to study the following boundary optimal control problem:

$$\min_{(\mathbf{u}, \mathbf{z}) \in U \times Z} \left\{ \mathcal{J}(\mathbf{u}, \mathbf{z}) := f(\mathbf{u}) + \frac{1}{2} \left(\alpha \|\text{curl}_\Gamma \mathbf{z}\|_{L^2(\Gamma)}^2 + \beta \|\mathbf{z}\|_{L^2(\Gamma)}^2 \right) \right\}, \quad (2.2.1a)$$

subject to the time-harmonic Maxwell’s equations as constraints

$$\begin{aligned} \text{curl}(\boldsymbol{\mu}^{-1} \text{curl} \mathbf{u}) + (i\omega) \boldsymbol{\kappa} \mathbf{u} &= \mathbf{j}_c \quad \text{in } \Omega, \\ \mathbf{u} \times \mathbf{n} &= \mathbf{z} \times \mathbf{n} \quad \text{on } \Gamma. \end{aligned} \quad (2.2.1b)$$

In (2.2.1a), curl_Γ denotes the scalar surface curl. For a precise definition of curl_Γ , the surface divergence div_Γ , the tangential gradient ∇_Γ , and the tangential vector curl_Γ when Ω is a Lipschitz polyhedron, see [33, 34].

To the best of our knowledge, this is the first work on boundary control of Maxwell’s equations with complete analysis and numerical implementation in the time

harmonic setting. Another significant difference between us and the existing literature on optimal control of Maxwell type equations is that we work in a complex setting without assuming a split between real and imaginary parts.

This, however, introduces additional challenges. For instance, even the standard quadratic cost functional

$$f(\mathbf{u}) := \frac{1}{2} \int_{\Omega} |\mathbf{u} - \mathbf{u}_d|^2 d\mathbf{x}, \quad (2.2.2)$$

is not differentiable in the (complex) Gâteaux sense, thus making most of the existing gradient-based optimization algorithms not directly applicable. However, provided that we can define an appropriate notion of derivatives, the complex structure leads to elegant analysis. Furthermore, because most modern programming languages can handle complex arithmetic, working in the complex field directly is advantageous. In the particular case of quadratic functionals of the form (2.2.2), the differentiability problem can be addressed via *directional derivatives* as in [145, Sec. 3.2], where they consider an optimal current problem with a complex control related to impressed currents (internal control), see also [40, Sec. 4.1]. Our approach allows us to study a considerably larger family of functionals. In the case of quadratic functions, this result is a natural extension to the real-valued case.

The *Wirtinger derivative* (1927) [149], which is a well-known concept in finite dimensions, inspired our idea of derivatives. Its origins may be traced back at least to 1899 in a paper by J.H. Poincaré on potentials [122, Théorème 8]. This notion of derivatives is motivated by the splitting of a function into its real and imaginary components. However, it allows one to work within the complex regime without using any such splitting. Wirtinger derivatives have been used in finite-dimensional optimization problems at least since the 1960s, in the works of Levinson, Mond, Hanson, and Kaul, cf. [102, 114, 90], and in the engineering community since the 1980s, see [30], with applications from *signal theory* to *Machine learning*, see [27]. We refer to [91, 94] for more details on Wirtinger derivatives.

In this work, we extend the notion of derivatives from finite dimensions to infinite dimensions with complex fields and rigorously derive the optimality conditions at the continuous level and identify continuous gradients. We discretize our state and adjoint variables using higher-order Nédélec elements, and for the control, we use the lowest-order Nédélec elements on each boundary face, which by construction have continuous tangential components. Next, we establish convergence of our numerical scheme. Numerical experiments confirm our theoretical findings. In particular, in our first experiment, motivated by [21, 22], we consider a realistic application with non-homogeneous boundary conditions where we first derive an explicit solution and next we validate our Nédélec finite element implementation against this explicit solution, the expected rate of convergence is observed. In the second experiment, we study the convergence of optimal control problem, see section 2.6 for more details.

Outline: In section 2.3, we introduce some notation and establish the well-posedness of the state equation. Section 2.4 first establishes existence of solution to the control problem. Next, section 2.4.1 introduces the notion of Wirtinger derivatives on complex Hilbert spaces and derives abstract optimality conditions. This is followed by a rigorous derivation of optimality conditions for our problem in section 2.4.2. Well-posedness of the adjoint equation is also established. Additional regularity for the adjoint equation and the optimality system are provided in section 2.4.3. Section 2.5 introduces the discrete optimal control problem. Details on imposing non-zero boundary conditions are provided in section 2.5.1. Moreover, section 2.5.2 is devoted to best approximation results for the state equations. The precise choice of the control space is discussed in section 2.5.3. Section 2.5.4 discusses the regularity of discrete adjoint equation and derives the discrete optimality system. Finally, in section 2.5.5, we provide convergence analysis of the optimal control problem. In Subsection 2.5.6, we establish that the lower order terms can be dropped, i.e., β in (2.2.1a) can be set to zero. Section 2.6 is devoted to numerical examples which confirm the theoretical results.

2.3 Notation and preliminary results

From now on, if X is a set of scalars, we use the notation \mathbf{X} to denote $(X)^3$, i.e., vectors. The term V^* will be used to refer to both the space of linear and conjugate linear functionals on a vector space V (see, for example, [120, p. 168]), this choice will be clear from the context. In what follows, we will need to identify the restriction of functions onto polygonal faces of Ω . For this purpose we use the notation $\Gamma = \bigcup \bar{\Gamma}_i$, where each Γ_i is an open subset of Γ such that its closure is a face of Γ and the Γ_i are pairwise disjoint. We use \mathbf{n} to denote the outward unit normal to Γ . Next, we define several Hilbert spaces, endowed with their standard inner products and norms:

$$\begin{aligned}
L^2(\Omega) &:= L^2(\Omega; \mathbb{C}), \\
H_0(\operatorname{div}; \Omega) &:= \{ \mathbf{v} \in H(\operatorname{div}; \Omega) : \gamma_{\mathbf{n}}(\mathbf{v}) := \mathbf{v} \cdot \mathbf{n} = 0 \text{ on } \Gamma \}, \\
H(\operatorname{curl}; \Omega) &:= \{ \mathbf{v} \in \mathbf{L}^2(\Omega) : \operatorname{curl} \mathbf{v} \in \mathbf{L}^2(\Omega) \}, \\
L_t^2(\Gamma) &:= \{ \mathbf{v} \in L^2(\Gamma; \mathbb{C}^3) : \mathbf{v} \cdot \mathbf{n} = 0 \text{ a.e. on } \Gamma \}, \\
H(\operatorname{curl}_{\Gamma}; \Gamma) &:= \{ \mathbf{v} \in L_t^2(\Gamma) : \operatorname{curl}_{\Gamma} \mathbf{v} \in L^2(\Gamma) \}, \\
H(\operatorname{div}_{\Gamma}; \Gamma) &:= \{ \mathbf{v} \in L_t^2(\Gamma) : \operatorname{div}_{\Gamma} \mathbf{v} \in L^2(\Gamma) \}, \\
H_-^{\frac{1}{2}}(\Gamma) &:= \left\{ \boldsymbol{\lambda} \in L_t^2(\Gamma) : \boldsymbol{\lambda}|_{\Gamma_i} \in \mathcal{H}^{\frac{1}{2}}(\Gamma_i), \text{ for each face } \Gamma_i \subset \Gamma \right\},
\end{aligned} \tag{2.3.1}$$

where all the differential operators are in the sense of distributions, cf. [69, Ch. I]. Unless stated otherwise, we will use the notation $(\cdot, \cdot)_{0, \Omega}$ and $\| \cdot \|_{0, \Omega}$ to denote the L^2 -inner product and norm (respectively), regardless of whether the functions are scalar-valued, vector-valued, etc. Similarly, we will use the notation $\langle \cdot, \cdot \rangle_{\Gamma}$ to denote the duality pairing of $H^{-\frac{1}{2}}(\Gamma)$ and $H^{\frac{1}{2}}(\Gamma)$.

Control space. We now define the set of *admissible controls* for our optimal control problem (2.2.1):

$$Z := \{ \mathbf{z} \in H(\operatorname{curl}_{\Gamma}; \Gamma) : \operatorname{curl}_{\Gamma} \mathbf{z} \in L_0^2(\Gamma) \}, \tag{2.3.2}$$

endowed with the norm on $H(\operatorname{curl}_{\Gamma}; \Gamma)$, given by

$$\| \mathbf{z} \|_{\operatorname{curl}_{\Gamma}} := \| \mathbf{z} \|_{L^2(\Gamma)} + \| \operatorname{curl}_{\Gamma} \mathbf{z} \|_{L^2(\Gamma)}, \tag{2.3.3}$$

where $L_0^2(\Gamma)$ is the space of L^2 -functions on Γ with zero mean. The zero mean condition is a natural restriction for the problem, related to the identity $\operatorname{div} \operatorname{curl} \mathbf{u} = 0$ for $\mathbf{u} \in H(\mathbf{curl}; \Omega)$; see for instance, [35, Corollary 5.4]. This condition also appears naturally in the finite element method setting, cf. [1, Sec. 2]. We further emphasize that the norm definition in (2.3.3) motivates the control regularization in (2.2.1a). In Subsection 2.5.6, we establish that the lower-order term in the regularization can be dropped. Whenever we write $a \lesssim b$ in what follows, we mean that $a \leq Cb$, where C is a positive non-essential constant and its value might change at each occurrence.

2.3.1 Tangential traces and Green's identities for $H(\mathbf{curl}; \Omega)$

We begin this section by defining the *tangential trace* of a function in $\mathbf{H}^1(\Omega)$, since from the boundary condition in (2.2.1b), it is clear that these are the traces that are being imposed by the control. The material in this subsection is known, and we refer the interested reader to [33] and [130, Chapter 16] for more details. Because we are considering Ω to be a polyhedron, the outer unit normal \mathbf{n} is well-defined almost everywhere on Γ , as well as along each edge of Γ one of the tangential traces is continuous when applied to smooth functions.

Definition 2.3.1. *The tangential traces of \mathbf{v} , defined from $\mathcal{H}^1(\Omega)$ onto $H_{\perp}^{\frac{1}{2}}(\Gamma)$, are given by*

$$\gamma_t \mathbf{v} := \gamma \mathbf{v} \times \mathbf{n}, \quad \text{and} \quad \gamma_T \mathbf{v} := \mathbf{n} \times (\gamma \mathbf{v} \times \mathbf{n}).$$

where γ denotes the standard restriction of \mathbf{v} on Γ in the trace sense.

We now define the Hilbert spaces $H_{\perp}^{\frac{1}{2}}(\Gamma)$ and $H_{\parallel}^{\frac{1}{2}}(\Gamma)$, see [33, Prop. 2.6], as the image of the maps γ_t and γ_T restricted to $\mathcal{H}^1(\Omega)$. Moreover, we have the following result

Lemma 2.3.2. *The following maps are linear, continuous and surjective*

$$\gamma_t : \mathcal{H}^1(\Omega) \mapsto H_{\perp}^{\frac{1}{2}}(\Gamma), \quad \text{and} \quad \gamma_T : \mathcal{H}^1(\Omega) \mapsto H_{\parallel}^{\frac{1}{2}}(\Gamma).$$

Proof. See [33, Proposition 2.7]. □

Definition 2.3.3. *The spaces*

$$(H_{\perp}^{-\frac{1}{2}}(\Gamma), \|\cdot\|_{\perp, -\frac{1}{2}, \Gamma}), \quad \text{and} \quad (H_{\parallel}^{-\frac{1}{2}}(\Gamma), \|\cdot\|_{\parallel, -\frac{1}{2}, \Gamma}) \quad (2.3.4)$$

are the dual spaces of $H_{\perp}^{\frac{1}{2}}(\Gamma)$ and $H_{\parallel}^{\frac{1}{2}}(\Gamma)$ (endowed with dual norms), respectively. In this case, $L_t^2(\Gamma)$ is taken as the pivot space.

However, Lemma 2.3.2 is not directly applicable in our setting. Instead of $\mathcal{H}^1(\Omega)$, the right function space for (2.2.1b) is $H(\mathbf{curl}; \Omega)$. Notice that $H(\mathbf{curl}; \Omega)$ is less regular than $\mathcal{H}^1(\Omega)$, but the dual space of its trace space is more delicate than $\mathbf{H}^{-\frac{1}{2}}(\Gamma)$.

To define weaker versions of γ_t and γ_T , we first note that for $\mathbf{v} \in \mathbf{C}^{\infty}(\overline{\Omega})$, the image of the maps $\gamma_t \mathbf{v}$ and $\gamma_T \mathbf{v}$ belong to $H_{\perp}^{-\frac{1}{2}}(\Gamma)$, according to Definition 2.3.1. Since no restrictions are imposed on the normal component of $\gamma \mathbf{u}$, those maps can be extended by density to elements of $H(\mathbf{curl}; \Omega)$.

To obtain a Green's identity when both functions are in $H(\mathbf{curl}; \Omega)$, the ranges of γ_t and γ_T acting on $H(\mathbf{curl}; \Omega)$ must be properly defined. We again refer to [33] and [130, Ch. 16] as well as [115, p. 58] for more details. Following the notation of [33], we define

$$\begin{aligned} H_{\parallel}^{-\frac{1}{2}}(\text{div}_{\Gamma}; \Gamma) &:= \left\{ \boldsymbol{\lambda} \in H_{\parallel}^{-\frac{1}{2}}(\Gamma) : \text{div}_{\Gamma} \boldsymbol{\lambda} \in H^{-\frac{1}{2}}(\Gamma) \right\}, \\ H_{\perp}^{-\frac{1}{2}}(\text{curl}_{\Gamma}; \Gamma) &:= \left\{ \boldsymbol{\lambda} \in H_{\perp}^{-\frac{1}{2}}(\Gamma) : \text{curl}_{\Gamma} \boldsymbol{\lambda} \in H^{-\frac{1}{2}}(\Gamma) \right\}, \end{aligned}$$

endowed with the norms

$$\begin{aligned} \|\boldsymbol{\lambda}\|_{H_{\parallel}^{-\frac{1}{2}}(\text{div}_{\Gamma}; \Gamma)} &:= \|\boldsymbol{\lambda}\|_{\parallel, -\frac{1}{2}, \Gamma} + \|\text{div}_{\Gamma} \boldsymbol{\lambda}\|_{-\frac{1}{2}, \Gamma}, \\ \|\boldsymbol{\lambda}\|_{H_{\perp}^{-\frac{1}{2}}(\text{curl}_{\Gamma}; \Gamma)} &:= \|\boldsymbol{\lambda}\|_{\perp, -\frac{1}{2}, \Gamma} + \|\text{curl}_{\Gamma} \boldsymbol{\lambda}\|_{-\frac{1}{2}, \Gamma}. \end{aligned} \quad (2.3.5)$$

Moreover, we have $H_{\perp}^{-\frac{1}{2}}(\text{curl}_{\Gamma}; \Gamma) := (H_{\parallel}^{-\frac{1}{2}}(\text{div}_{\Gamma}; \Gamma))^*$, where $L_t^2(\Gamma)$ is used as pivot, and the following holds:

Lemma 2.3.4. *The maps,*

$$\gamma_t : H(\mathbf{curl}; \Omega) \mapsto H_{\parallel}^{-\frac{1}{2}}(\text{div}_{\Gamma}; \Gamma), \quad \text{and} \quad \gamma_T : H(\mathbf{curl}; \Omega) \mapsto H_{\perp}^{-\frac{1}{2}}(\text{curl}_{\Gamma}; \Gamma)$$

are linear, continuous and surjective.

Proof. See [34, Thm. 5.4]. □

With these definitions we have the following Green's identity:

Theorem 2.3.5 ([115, Thm. 3.31]). *The space $H_{\parallel}^{-\frac{1}{2}}(\operatorname{div}_{\Gamma}; \Gamma)$ is a Hilbert space. The continuous linear mappings $\gamma_t: H(\mathbf{curl}; \Omega) \rightarrow H_{\parallel}^{-\frac{1}{2}}(\operatorname{div}_{\Gamma}; \Gamma)$ and $\gamma_T: H(\mathbf{curl}; \Omega) \rightarrow H_{\perp}^{-\frac{1}{2}}(\operatorname{curl}_{\Gamma}; \Gamma)$ are surjective, and for all $\mathbf{v}, \boldsymbol{\phi} \in H(\mathbf{curl}; \Omega)$*

$$(\mathbf{v}, \nabla \times \boldsymbol{\phi})_{0, \Omega} - (\nabla \times \mathbf{v}, \boldsymbol{\phi})_{0, \Omega} = \langle \gamma_t \mathbf{v}, \gamma_T \boldsymbol{\phi} \rangle_{\Gamma^*}, \quad (2.3.6)$$

where $\langle \cdot, \cdot \rangle_{\Gamma^*}$ is the duality pairing between $H_{\parallel}^{-\frac{1}{2}}(\operatorname{div}_{\Gamma}; \Gamma)$ and $H_{\perp}^{-\frac{1}{2}}(\operatorname{curl}_{\Gamma}; \Gamma)$.

Now, given $\mathbf{g} \in H_{\parallel}^{-\frac{1}{2}}(\operatorname{div}_{\Gamma}; \Gamma)$ we define

$$H_{\mathbf{g}}(\operatorname{curl}; \Omega) := \{ \mathbf{v} \in H(\mathbf{curl}; \Omega) : \gamma_t \mathbf{v} = \mathbf{g} \text{ on } \Gamma \}.$$

2.3.2 Well-posedness of the state equation

First, let us introduce the *sesquilinear form*

$a: H(\mathbf{curl}; \Omega) \times H(\mathbf{curl}; \Omega) \rightarrow \mathcal{C}$ given by

$$a(\mathbf{u}, \mathbf{v}) := \int_{\Omega} \boldsymbol{\mu}^{-1} \operatorname{curl} \mathbf{u} \cdot \operatorname{curl} \bar{\mathbf{v}} \, d\mathbf{x} + (i\omega) \int_{\Omega} \boldsymbol{\kappa} \mathbf{u} \cdot \bar{\mathbf{v}} \, d\mathbf{x}. \quad (2.3.7)$$

The following lemmas show the well-posedness of the state equation (2.2.1b).

Lemma 2.3.6. *Given $\mathbf{f} \in H(\mathbf{curl}; \Omega)^*$, the problem, find $\mathbf{u} \in H(\mathbf{curl}; \Omega)$ such that*

$$a(\mathbf{u}, \mathbf{v}) = \langle \mathbf{f}, \mathbf{v} \rangle \quad \forall \mathbf{v} \in H(\mathbf{curl}; \Omega), \quad (2.3.8)$$

is a well-posed problem in the sense of Hadamard, where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H(\mathbf{curl}; \Omega)$ and its dual.

Proof. The hypothesis on $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$ guarantees that $a(\cdot, \cdot)$ defines a sesquilinear, bounded and coercive form. The Lax-Milgram lemma implies that (2.3.8) has a unique solution; moreover, we have

$$c(\boldsymbol{\kappa}, \boldsymbol{\mu}; \omega) \|\mathbf{u}\|_{\operatorname{curl}, \Omega}^2 \leq |a(\mathbf{u}, \mathbf{u})| \lesssim |\langle \mathbf{f}, \mathbf{u} \rangle|, \quad (2.3.9)$$

where c is a positive constant that depends on ω and the eigenvalues of $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$. □

Lemma 2.3.7. *Given $\mathbf{z} \in H(\text{curl}_\Gamma; \Gamma)$ and $\mathbf{j}_c \in \mathbf{L}^2(\Omega)$, the problem, find $\mathbf{u} \in H_{\mathbf{z} \times \mathbf{n}}(\text{curl}; \Omega)$ such that*

$$a(\mathbf{u}, \mathbf{v}) = \int_{\Omega} \mathbf{j}_c \cdot \bar{\mathbf{v}} \, d\mathbf{x} \quad \forall \mathbf{v} \in H_0(\mathbf{curl}; \Omega), \quad (2.3.10)$$

is well-posed.

Proof. Let $\mathbf{z} \in H(\text{curl}_\Gamma; \Gamma)$ be given. Then $\mathbf{z} \times \mathbf{n} \in \gamma_t H(\mathbf{curl}; \Omega)$, which follows from the much more general results on weak rotations, see [130, Prop. 16.16 and Eq. (16.42)]. Therefore from Lemma 2.3.4, there exists $\mathbf{u}_z \in H(\mathbf{curl}; \Omega)$ such that $\gamma_t \mathbf{u}_z = \mathbf{z} \times \mathbf{n}$ and

$$\|\mathbf{u}_z\|_{\text{curl}, \Omega} \lesssim \|\mathbf{z} \times \mathbf{n}\|_{H_{\parallel}^{-\frac{1}{2}}(\text{div}_\Gamma; \Gamma)} \lesssim \|\mathbf{z}\|_{\text{curl}_\Gamma}, \quad (2.3.11)$$

where the last inequality follows because of the isometry $\|\mathbf{z} \times \mathbf{n}\|_{H_{\parallel}^{-\frac{1}{2}}(\text{div}_\Gamma; \Gamma)} = \|\mathbf{z}\|_{H_{\perp}^{-\frac{1}{2}}(\text{curl}_\Gamma; \Gamma)}$, see [130, p. 448], and

$$\|\mathbf{z}\|_{H_{\perp}^{-\frac{1}{2}}(\text{curl}_\Gamma; \Gamma)} \leq C \|\mathbf{z}\|_{\text{curl}_\Gamma},$$

which follows from the compact embedding of $L^2(\Gamma)$ into $H^{-\frac{1}{2}}(\Gamma)$.

Now, we look for $\mathbf{u}_0 \in H_0(\mathbf{curl}; \Omega)$ such that

$$a(\mathbf{u}_0, \mathbf{v}) = \int_{\Omega} \mathbf{j}_c \cdot \bar{\mathbf{v}} \, d\mathbf{x} - a(\mathbf{u}_z, \mathbf{v}) \quad \forall \mathbf{v} \in H_0(\mathbf{curl}; \Omega). \quad (2.3.12)$$

This problem is well-posed, according to Lemmas (2.3.6) and (2.3.11), and

$$\|\mathbf{u}_0\| \lesssim \|\mathbf{j}_c\|_{0, \Omega} + \|\mathbf{z}\|_{\text{curl}_\Gamma}.$$

Finally, $\mathbf{u} := \mathbf{u}_0 + \mathbf{u}_z$ is the unique solution to (2.3.10) and

$$\|\mathbf{u}\|_{\text{curl}, \Omega} \lesssim \|\mathbf{j}_c\|_{0, \Omega} + \|\mathbf{z}\|_{\text{curl}_\Gamma},$$

which finishes the proof. \square

From the previous analysis, \mathbf{u} depends on both \mathbf{z} and \mathbf{j}_c . The goal of the next section is to reduce the cost functional $\mathcal{J}(\mathbf{u}, \mathbf{z}) = \mathcal{J}(\mathbf{u}(\mathbf{z}; \mathbf{j}_c), \mathbf{z})$ to be only a function of \mathbf{z} , and then derive the optimality conditions.

2.4 Reduced cost functional and its derivative

For the remainder of the Chapter, we will assume that \mathbf{j}_c is given. By introducing the control-to-state map, we can obtain the so-called *reduced optimization problem*. The solution map is an affine transformation, and it is given by

$$\begin{aligned} \mathbb{S} : Z &\rightarrow H(\mathbf{curl}; \Omega) \hookrightarrow \mathbf{L}^2(\Omega) \\ \mathbf{z} &\mapsto \mathbb{S}\mathbf{z} := \mathbf{u}, \end{aligned} \quad (2.4.1)$$

where \mathbf{u} is the unique solution to (2.3.10) with right-hand-side \mathbf{j}_c and the boundary condition $\mathbf{z} \times \mathbf{n}$. The notation \hookrightarrow indicates the continuous embedding; as a result, we can consider $\mathbb{S} : Z \rightarrow \mathbf{L}^2(\Omega)$. The solution operator \mathbb{S} is an affine map, and it is common to split \mathbb{S} into two parts: the part that depends on \mathbf{z} and the one that depends on \mathbf{j}_c . We write

$$\mathbb{S}\mathbf{z} = \mathbb{S}_\Gamma \mathbf{z} + \mathbf{u}_\Omega,$$

where \mathbb{S}_Γ is the solution operator for the state equation with $\mathbf{j}_c \equiv \mathbf{0}$, and \mathbf{u}_Ω is the solution for the state equation when $\mathbf{z} \equiv \mathbf{0}$. By Lemma 2.3.7, \mathbb{S} is continuous, and there exists a $C = C(\boldsymbol{\mu}, \boldsymbol{\kappa}, \Omega) > 0$, such that

$$\|\mathbb{S}\mathbf{z}\|_{\mathbf{curl}, \Omega} = \|\mathbf{u}\|_{\mathbf{curl}, \Omega} \leq C (\|\mathbf{j}_c\|_{0, \Omega} + \|\mathbf{z}\|_{\mathbf{curl}_\Gamma}). \quad (2.4.2)$$

As a result, the *reduced-cost functional* $j(\mathbf{z}) := \mathcal{J}(\mathbb{S}\mathbf{z}, \mathbf{z})$ is also continuous, and it can be represented using the above-mentioned splitting as

$$\min_{\mathbf{z} \in Z} j(\mathbf{z}) = \min_{\mathbf{z} \in Z} \frac{1}{2} \int_{\Omega} |\mathbb{S}_\Gamma \mathbf{z} - \widehat{\mathbf{u}}_d|^2 d\mathbf{x} + \frac{\alpha}{2} \int_{\Gamma} |\mathbf{curl}_\Gamma \mathbf{z}|^2 dS + \frac{\beta}{2} \int_{\Gamma} |\mathbf{z}|^2 dS, \quad (2.4.3)$$

where $\widehat{\mathbf{u}}_d = \mathbf{u}_d - \mathbf{u}_\Omega$. Therefore, without loss of generality, in what follows we will consider $\mathbf{j}_c \equiv \mathbf{0}$, and therefore $\mathbb{S} = \mathbb{S}_\Gamma$. Notice that in this case, $\mathbf{u}_\Omega = 0$.

Our goal now is to discuss the existence and uniqueness of a solution to the reduced optimization problem

$$\min_{\mathbf{z} \in Z} j(\mathbf{z}) = \min_{\mathbf{z} \in Z} \mathcal{J}(\mathbb{S}\mathbf{z}, \mathbf{z}). \quad (2.4.4)$$

The proof follows immediately from the direct method of calculus of variations; we outline it for completeness.

Theorem 2.4.1 (existence and uniqueness). *The problem (2.4.4) has a unique solution $\bar{\mathbf{z}} \in Z$.*

Proof. Notice that $j(\cdot)$ is bounded below, therefore there exists an infimizing sequence $\{\mathbf{z}_n\}_{n=1}^\infty$ such that $\inf_{\mathbf{z} \in Z} j(\mathbf{z}) = \lim_{n \rightarrow \infty} j(\mathbf{z}_n)$. The previous limit and the definition of $j(\cdot)$ implies that $\{\mathbf{z}_n\}_{n=1}^\infty$ is a bounded sequence in Z . Notice that Z is closed subspace of a Hilbert space and is therefore a Hilbert space itself. Thus the boundedness of sequence $\{\mathbf{z}_n\}_{n=1}^\infty$ implies that there exists a subsequence (not relabeled) that converges to $\hat{\mathbf{z}}$ in Z . It then remains to show that $\hat{\mathbf{z}}$ is the minimizer of (2.4.4). This immediately follows from weak lower semicontinuity of $j(\cdot)$. The uniqueness is a direct consequence of the strict convexity of $j(\cdot)$. \square

As is usual, we want to apply a gradient-based method to locate the critical points of the cost functional $j(\cdot)$ in order to discover the optimal control \mathbf{z} of (2.4.3). Nonetheless, differentiability differs significantly between real and complex fields. However, Fréchet and Gâteaux differentiability on complex fields are quite similar, cf. [159, 160, 161]. Notice that, $|\mathbb{S}_\Gamma \mathbf{z} - \hat{\mathbf{u}}_d|^2$ in (2.4.3) is smooth, but as we will see below, it is not complex Gâteaux differentiable. To study (2.4.3), we require a weaker definition of $j(\cdot)$. In order to do that, we will propose a notion of derivative for $j(\cdot)$ that is more flexible than Gâteaux or Fréchet derivatives for complex spaces, and that it is strong enough to allow Taylor series expansions; this will allow us to characterize the critical points of $j(\cdot)$.

Before we begin our discussion of derivatives, let us define what it means for a complex-valued function to be \mathcal{C}^1 , but not necessarily analytic; see [99, (2.1)].

Definition 2.4.2 (\mathcal{C}^1 functions). *For a complex Banach space \mathcal{U} , suppose $\mathcal{D} \subset \mathcal{U}$ is open and $u : \mathcal{D} \rightarrow \mathcal{C}$ is a function. If the directional derivatives*

$$du(\mathbf{z}; \boldsymbol{\xi}) = \lim_{t \rightarrow 0} \frac{u(\mathbf{z} + t\boldsymbol{\xi}) - u(\mathbf{z})}{t}, \quad t \in \mathbb{R} \quad (2.4.5)$$

exist for all $\mathbf{z} \in \mathcal{D}$, $\boldsymbol{\xi} \in \mathcal{U}$, and $du : \mathcal{D} \times \mathcal{U} \rightarrow \mathcal{C}$ is continuous, we write $u \in \mathcal{C}^1(\mathcal{D})$.

Consider the function

$$z \mapsto z\bar{z} = |z|^2, \quad (2.4.6)$$

which is $\mathcal{C}^1(\mathcal{C})$ but not complex Fréchet differentiable, nor complex Gâteaux differentiable. This shows that complex differentiability is too restrictive, especially in the context of optimization. A weaker notion of “complex derivative” which has most of the required properties in optimization is the so-called *Wirtinger derivative*, cf. [149].

In the next section, we extend the concept of Wirtinger derivatives to spaces that are the complexification of a real Hilbert space; that generalizes case $f : \mathcal{C} \rightarrow \mathcal{C}$, cf. [147].

2.4.1 Wirtinger derivatives on complexified Hilbert spaces

Let $(\mathcal{U}, \|\cdot\|_{\mathcal{U}})$ be the complexification of a real Hilbert space $(H, (\cdot, \cdot)_H)$, and let $f : \mathcal{U} \rightarrow \mathcal{C}$ be a continuous function but not complex differentiable. We can extend f to a function on $\mathcal{U} \times \mathcal{U}$. Let $g : \mathcal{U} \times \mathcal{U} \rightarrow \mathcal{C}$ be a continuous extension of f such that

$$g(z, \bar{z}) = f(z) \quad \forall z \in \mathcal{U}.$$

Now, let us assume g is complex Fréchet differentiable, and define

$$\left. \frac{\partial f}{\partial z} \right|_{z=z_0} := \nabla g(z_0, \bar{z}_0)(\mathbf{e}_1), \quad \left. \frac{\partial f}{\partial \bar{z}} \right|_{z=z_0} := \nabla g(z_0, \bar{z}_0)(\mathbf{e}_2),$$

where the unit vectors \mathbf{e}_1 and \mathbf{e}_2 will give us the first and second components of ∇g . Even if the existence of a complex Fréchet differentiable extension g of f may look to be too restrictive, such an extension exists when f is continuous and f is analytic for z and \bar{z} ; separately, cf. [91, p. 2], these conditions hold for the function in (2.4.6), for instance. This follows from *Hartogs’ theorem* or one of its generalizations to infinite-dimensional spaces; see, for instance, [110, Thm. 3.2]. Now, for any z_0 and δz in \mathcal{U} ,

the following limit exists and the resulting expression is linear in $\delta \mathbf{z}$,

$$\begin{aligned}
d^{\mathbb{R}}f(\mathbf{z}_0; \delta \mathbf{z}) &:= \lim_{\substack{t \rightarrow 0 \\ \text{Im}(t)=0}} \frac{f(\mathbf{z}_0 + t\delta \mathbf{z}) - f(\mathbf{z}_0)}{t} \\
&= \lim_{\substack{t \rightarrow 0 \\ \text{Im}(t)=0}} \frac{g\left(\left(\mathbf{z}_0, \bar{\mathbf{z}}_0\right) + t\left(\delta \mathbf{z}, \overline{\delta \mathbf{z}}\right)\right) - g\left(\mathbf{z}_0, \bar{\mathbf{z}}_0\right)}{t} \\
&= \left(\left(\begin{array}{c} \delta \mathbf{z} \\ \overline{\delta \mathbf{z}} \end{array}\right), \nabla g(\mathbf{z}_0, \bar{\mathbf{z}}_0)\right)_{\mathcal{U} \times \mathcal{U}} \\
&= \left(\delta \mathbf{z}, \frac{\partial f}{\partial \mathbf{z}}(\mathbf{z}_0)\right)_{\mathcal{U}} + \left(\overline{\delta \mathbf{z}}, \frac{\partial f}{\partial \bar{\mathbf{z}}}(\mathbf{z}_0)\right)_{\mathcal{U}}.
\end{aligned} \tag{2.4.7}$$

Theorem 2.4.3. *Let \mathcal{U} be the complexification of a real Hilbert space H , and consider a continuous function $f : \mathcal{U} \mapsto \mathbb{R}, \mathbf{z} \mapsto f(\mathbf{z})$, such that f is analytic in \mathbf{z} and in $\bar{\mathbf{z}}$, separately, then $d^{\mathbb{R}}f(\mathbf{z}_0; \delta \mathbf{z})$ defined in (2.4.7) exists. Moreover,*

$$d^{\mathbb{R}}f(\mathbf{z}_0; \delta \mathbf{z}) = 2\text{Re}\left(\frac{\partial f}{\partial \mathbf{z}}(\mathbf{z}_0), \delta \mathbf{z}\right)_{\mathcal{U}} \quad \forall \mathbf{z}_0, \delta \mathbf{z} \in \mathcal{U}. \tag{2.4.8}$$

Proof. The existence of $d^{\mathbb{R}}f(\mathbf{z}_0; \delta \mathbf{z})$ follows from the previous analysis. On the other hand, if f is real-valued, then by definition $d^{\mathbb{R}}f(\mathbf{z}_0; \delta \mathbf{z})$ is also real-valued, yielding

$$\frac{\partial f}{\partial \mathbf{z}}(\mathbf{z}_0) = \overline{\frac{\partial f}{\partial \bar{\mathbf{z}}}(\mathbf{z}_0)}. \tag{2.4.9}$$

In order to prove this identity, consider the inner product on \mathcal{U} given by

$$(\mathbf{u}_1 + i\mathbf{v}_1, \mathbf{u}_2 + i\mathbf{v}_2)_{\mathcal{U}} := (\mathbf{u}_1, \mathbf{u}_2)_H + (\mathbf{v}_1, \mathbf{v}_2)_H + i((\mathbf{v}_1, \mathbf{u}_2)_H - (\mathbf{u}_1, \mathbf{v}_2)_H).$$

Now, we consider the splitting between real and imaginary parts

$$\delta \mathbf{z} = \delta^{Re} \mathbf{z} + i\delta^{Im} \mathbf{z}, \quad \frac{\partial f}{\partial \mathbf{z}}(\mathbf{z}_0) = f_z^{Re} + if_z^{Im}, \quad \text{and} \quad \frac{\partial f}{\partial \bar{\mathbf{z}}}(\mathbf{z}_0) = f_{\bar{z}}^{Re} + if_{\bar{z}}^{Im}.$$

In turn, from (2.4.7)

$$\begin{aligned}
d^{\mathbb{R}}f(\mathbf{z}_0; \delta \mathbf{z}) &= \left(\delta \mathbf{z}, \frac{\partial f}{\partial \mathbf{z}}(\mathbf{z}_0)\right)_{\mathcal{U}} + \left(\overline{\delta \mathbf{z}}, \frac{\partial f}{\partial \bar{\mathbf{z}}}(\mathbf{z}_0)\right)_{\mathcal{U}} \\
&= \left(\delta^{Re} \mathbf{z} + i\delta^{Im} \mathbf{z}, f_z^{Re} - if_z^{Im}\right)_{\mathcal{U}} + \left(\delta^{Re} \mathbf{z} - i\delta^{Im} \mathbf{z}, f_{\bar{z}}^{Re} - if_{\bar{z}}^{Im}\right)_{\mathcal{U}},
\end{aligned}$$

and because $\text{Im}\{d^{\mathbb{R}}f(\mathbf{z}_0, \delta\mathbf{z})\} = 0$, we have

$$\left(f_z^{Re}, \delta^{Im}\mathbf{z}\right)_H + \left(\delta^{Re}\mathbf{z}, f_z^{Im}\right)_H - \left(f_{\bar{z}}^{Re}, \delta^{Im}\mathbf{z}\right)_H + \left(\delta^{Re}\mathbf{z}, f_{\bar{z}}^{Im}\right)_H = 0.$$

In particular, if $\delta^{Re}\mathbf{z} = \mathbf{0}$ we get

$$\left(f_z^{Re}, \delta^{Im}\mathbf{z}\right)_H - \left(f_{\bar{z}}^{Re}, \delta^{Im}\mathbf{z}\right)_H = 0 \Leftrightarrow f_z^{Re} = f_{\bar{z}}^{Re}.$$

In turn, if $\delta^{Im}\mathbf{z} = \mathbf{0}$ we get

$$\left(\delta^{Re}\mathbf{z}, f_z^{Im}\right)_H + \left(\delta^{Re}\mathbf{z}, f_{\bar{z}}^{Im}\right)_H = 0 \Leftrightarrow f_z^{Im} = -f_{\bar{z}}^{Im}.$$

Thus,

$$\frac{\partial f}{\partial \mathbf{z}}(\mathbf{z}_0) = \overline{\left(\frac{\partial f}{\partial \bar{\mathbf{z}}}(\mathbf{z}_0)\right)}, \text{ and therefore } \left(\delta\mathbf{z}, \frac{\partial f}{\partial \mathbf{z}}(\mathbf{z}_0)\right)_u = \left(\overline{\delta\mathbf{z}}, \overline{\frac{\partial f}{\partial \bar{\mathbf{z}}}(\mathbf{z}_0)}\right)_u,$$

which concludes the proof. \square

Our next goal is to relate $d^{\mathbb{R}}f$ given in (2.4.8) with a gradient so that we can derive the first-order optimality conditions for problem (2.2.1a). In order to do that, we identify f with a real functional u , namely $u : H \times H \rightarrow \mathbb{R}$ that satisfies $f(\mathbf{z}) = f(\mathbf{x} + i\mathbf{y}) = u(\mathbf{x}, \mathbf{y})$ for all $\mathbf{z} = \mathbf{x} + i\mathbf{y}$ in \mathcal{U} . From the regularity of g (as defined above), we obtain

$$\begin{aligned} d^{\mathbb{R}}f(\mathbf{z}; \delta\mathbf{z}) &= \lim_{\substack{t \rightarrow 0 \\ t \in \mathbb{R}}} \frac{u(\mathbf{x} + t\delta\mathbf{x}, \mathbf{y} + t\delta\mathbf{y}) - u(\mathbf{x}, \mathbf{y})}{t} \\ &= \left(\nabla u, \begin{pmatrix} \delta\mathbf{x} \\ \delta\mathbf{y} \end{pmatrix} \right)_{H \times H} \\ &= \left(\frac{\partial u}{\partial \mathbf{x}}, \delta\mathbf{x} \right)_H + \left(\frac{\partial u}{\partial \mathbf{y}}, \delta\mathbf{y} \right)_H, \end{aligned} \tag{2.4.10}$$

where $\delta\mathbf{z} = \delta\mathbf{x} + i\delta\mathbf{y}$. Consequently, $f \in \mathcal{C}^1(\mathcal{U})$ (see Definition 2.4.2) and we have the following identities

$$u \left(\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} + \begin{pmatrix} \delta\mathbf{x} \\ \delta\mathbf{y} \end{pmatrix} \right) = u \left(\mathbf{x}, \mathbf{y} \right) + \left(\nabla u(\hat{\mathbf{x}}, \hat{\mathbf{y}}), \begin{pmatrix} \delta\mathbf{x} \\ \delta\mathbf{y} \end{pmatrix} \right), \tag{2.4.11}$$

$$f(\mathbf{z} + \delta\mathbf{z}) = f(\mathbf{z}) + d^{\mathbb{R}}f(\hat{\mathbf{z}}; \delta\mathbf{z}), \tag{2.4.12}$$

for some $\widehat{\mathbf{z}} = \widehat{\mathbf{x}} + i\widehat{\mathbf{y}}$ in the segment $[\mathbf{z}, \mathbf{z} + \delta\mathbf{z}]$. From now on, the expression $d^{\mathbb{R}}f(\mathbf{z}; \delta\mathbf{z})$ will be referred to as the \mathbb{R} -linear derivative of f at \mathbf{z} in the direction $\delta\mathbf{z}$. Many properties of f can be derived from its extension g or using the relation with u ; for example, f is (real) convex if and only if u is convex. Also, one can determine the directions of the steepest descent and stationary points of f . In fact, we have the following result:

Theorem 2.4.4 (Steepest descent). *Let f , g and u be as above, then the direction of steepest descent for f at $\mathbf{z}_0 = \mathbf{x}_0 + i\mathbf{y}_0$ is given by*

$$\delta\mathbf{z} = -\frac{\overline{\partial f}}{\partial \mathbf{z}}(\mathbf{z}_0) = -\frac{\partial f}{\partial \overline{\mathbf{z}}}(\mathbf{z}_0),$$

or equivalently

$$\begin{pmatrix} \delta\mathbf{x} \\ \delta\mathbf{y} \end{pmatrix} = -\nabla u(\mathbf{x}_0, \mathbf{y}_0).$$

Proof. The result follows from the two characterizations of $d^{\mathbb{R}}f$ given in (2.4.8) and (2.4.10), respectively. In the first case, we have also used (2.4.9). The proof is complete. \square

From the previous theorem, we obtain

$$\frac{\partial f}{\partial \mathbf{z}}(\mathbf{z}) = \frac{\partial u}{\partial \mathbf{x}}(\mathbf{x}, \mathbf{y}) + i\frac{\partial u}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y}), \quad \text{and} \quad \frac{\partial f}{\partial \overline{\mathbf{z}}}(\mathbf{z}) = \frac{\partial u}{\partial \mathbf{x}}(\mathbf{x}, \mathbf{y}) - i\frac{\partial u}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y}).$$

The identities, known as the *Wirtinger derivatives* (in finite dimensions) for a real-valued function f , are commonly written as

$$\frac{\partial f}{\partial \mathbf{z}} = \frac{\partial f}{\partial \mathbf{x}} + i\frac{\partial f}{\partial \mathbf{y}}, \tag{2.4.13}$$

$$\frac{\partial f}{\partial \overline{\mathbf{z}}} = \frac{\partial f}{\partial \mathbf{x}} - i\frac{\partial f}{\partial \mathbf{y}}. \tag{2.4.14}$$

Lemma 2.4.5. *Let f be as above. Then, $\mathbf{z}_0 = \mathbf{x}_0 + i\mathbf{y}_0$ is a stationary point of f if and only if*

$$\frac{\partial f}{\partial \mathbf{z}}(\mathbf{z}_0) = \mathbf{0} \quad \text{or} \quad \frac{\partial f}{\partial \overline{\mathbf{z}}}(\mathbf{z}_0) = \mathbf{0}.$$

Proof. It is enough to identify f with u , then $\nabla u(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$, and using (2.4.13) and (2.4.14) the proof is complete. \square

Finally, we present the following useful theorem, where $\text{Dom}(f)$ denotes the domain of f .

Theorem 2.4.6. *Let f be as above, and assume $f : \text{Dom}(f) \rightarrow \mathcal{C}$ with $\text{Dom}(f)$ convex. If \mathbf{z}_0 is an optimal point then*

$$d^{\mathbb{R}}f(\mathbf{z}_0; \mathbf{z} - \mathbf{z}_0) \geq 0, \quad \forall \mathbf{z} \in \text{Dom}(f). \quad (2.4.15)$$

In addition, if f is (real) convex then (2.4.15) is a sufficient condition.

Proof. This result follows directly from identifying f with u and applying well-known results for convex real-valued functions. \square

2.4.2 The \mathbb{R} -linear derivative of the reduced cost functional and optimality conditions

Since the map $\mathbf{z} \mapsto \bar{\mathbf{z}} \cdot \mathbf{z}$ is not (complex) Gâteaux differentiable, the same can be concluded for the reduced functional j defined in (2.4.3). The following lemma shows, however, that $d^{\mathbb{R}}j$ is well-defined.

Lemma 2.4.7. *Let $(X, \|\cdot\|_X)$ be a complex Banach space, $(\mathcal{U}, (\cdot, \cdot)_{\mathcal{U}})$ be a complex Hilbert space, $\tilde{\mathbf{u}} \in \mathcal{U}$, and $\tilde{\mathbb{S}} \in \mathcal{L}(X, \mathcal{U})$. Then, the convex function $f(\mathbf{z}) := \|\tilde{\mathbb{S}}\mathbf{z} - \tilde{\mathbf{u}}\|_{\mathcal{U}}^2$ has an \mathbb{R} -linear derivative given by*

$$d^{\mathbb{R}}f(\mathbf{z}; \mathbf{v}) = 2 \text{Re}(\tilde{\mathbb{S}}\mathbf{z} - \tilde{\mathbf{u}}, \tilde{\mathbb{S}}\mathbf{v})_{\mathcal{U}},$$

where $\|\mathbf{u}\|_{\mathcal{U}}^2 := (\mathbf{u}, \mathbf{u})_{\mathcal{U}}$ and $\text{Re}(a)$ denotes the real part of $a \in \mathcal{C}$.

Proof. We examine the difference quotient from the definition of $d^{\mathbb{R}}$, cf. 2.4.7, namely:

$$\begin{aligned} \frac{f(\mathbf{z} + t\mathbf{v}) - f(\mathbf{z})}{t} &= \frac{\|\tilde{\mathbb{S}}(\mathbf{z} + t\mathbf{v}) - \tilde{\mathbf{u}}\|_{\mathcal{U}}^2 - \|\tilde{\mathbb{S}}\mathbf{z} - \tilde{\mathbf{u}}\|_{\mathcal{U}}^2}{t} \\ &= \frac{(\|\tilde{\mathbb{S}}\mathbf{z} - \tilde{\mathbf{u}}\|_{\mathcal{U}}^2 + 2 \text{Re}(\tilde{\mathbb{S}}\mathbf{z} - \tilde{\mathbf{u}}, t\tilde{\mathbb{S}}\mathbf{v})_{\mathcal{U}} + \|t\tilde{\mathbb{S}}\mathbf{v}\|_X^2) - \|\tilde{\mathbb{S}}\mathbf{z} - \tilde{\mathbf{u}}\|_{\mathcal{U}}^2}{t} \\ &= \frac{2 \text{Re} \bar{t}(\tilde{\mathbb{S}}\mathbf{z} - \tilde{\mathbf{u}}, \tilde{\mathbb{S}}\mathbf{v})_{\mathcal{U}} + |t|^2 \|\tilde{\mathbb{S}}\mathbf{v}\|_{\mathcal{U}}^2}{t}. \end{aligned} \quad (2.4.16)$$

Considering $t \in \mathcal{C}$ such that $\text{Im}(t) = 0$, and then taking the limit as $t \rightarrow 0$ completes the proof. \square

Remark 2.4.8. From (2.4.16), we can also conclude that f is not complex Gâteaux differentiable.

In what follows we recall that $\mathbb{S} = \mathbb{S}_\Gamma$ and $\mathbf{u}_d = \widehat{\mathbf{u}}_d$.

Corollary 2.4.9. For \mathbf{z} and $\boldsymbol{\xi}$ in $H(\text{curl}_\Gamma; \Gamma)$, the \mathbb{R} -linear derivative of j at \mathbf{z} (given in (2.4.3)), in the direction $\boldsymbol{\xi}$ is given by

$$d^{\mathbb{R}}j(\mathbf{z}; \boldsymbol{\xi}) = \text{Re} \left\{ (\mathbb{S}\mathbf{z} - \mathbf{u}_d, \mathbb{S}\boldsymbol{\xi})_{0,\Omega} + \alpha(\text{curl}_\Gamma, \text{curl}_\Gamma \boldsymbol{\xi})_{0,\Gamma} + \beta(\mathbf{z}, \boldsymbol{\xi})_{0,\Gamma} \right\}, \quad (2.4.17)$$

where $(\cdot, \cdot)_{0,\Gamma}$ denotes the inner product in both $L^2(\Gamma)$ and $\mathbf{L}^2(\Gamma)$.

As usual, we will avoid computing the term $\mathbb{S}\boldsymbol{\xi}$ by introducing the adjoint of \mathbb{S} , denoted by \mathbb{S}^* . Since \mathbb{S} is bounded linear, $\mathbb{S}^* : L^2(\Omega) \rightarrow H(\text{curl}_\Gamma; \Gamma)^*$ is well defined. Then from (2.4.17), given $\boldsymbol{\xi} \in H(\text{curl}_\Gamma; \Gamma)$ we have that $d^{\mathbb{R}}j(\mathbf{z}; \boldsymbol{\xi})$ is given by

$$d^{\mathbb{R}}j(\mathbf{z}; \boldsymbol{\xi}) = \text{Re} \left\{ \langle \mathbb{S}^*(\mathbb{S}\mathbf{z} - \mathbf{u}_d), \boldsymbol{\xi} \rangle_{H(\text{curl}_\Gamma; \Gamma)^*, H(\text{curl}_\Gamma; \Gamma)} + \alpha(\text{curl}_\Gamma \mathbf{z}, \text{curl}_\Gamma \boldsymbol{\xi})_{0,\Gamma} + \beta(\mathbf{z}, \boldsymbol{\xi})_{0,\Gamma} \right\}. \quad (2.4.18)$$

We will make additional assumptions on $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$, so the duality pairing in (2.4.18) will become the inner product on $L_t^2(\Gamma)$. As is usual for optimal control of PDEs, \mathbb{S}^* is the solution operator for a problem similar to the state equation known as *adjoint state equation*. Given $\mathbf{f} \in \mathbf{L}^2(\Omega)$, find $\mathbf{w} \in H(\mathbf{curl}; \Omega)$ such that

$$\begin{aligned} \text{curl}(\boldsymbol{\mu}^{-1} \text{curl} \mathbf{w}) - (i\omega) \boldsymbol{\kappa} \mathbf{w} &= \mathbf{f} \quad \text{in } \Omega, \\ \mathbf{w} \times \mathbf{n} &= \mathbf{0} \quad \text{on } \Gamma. \end{aligned} \quad (2.4.19)$$

The weak formulation of this problem is: find $\mathbf{w} \in H_0(\mathbf{curl}; \Omega)$ such that

$$a^*(\mathbf{w}, \mathbf{v}) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in H_0(\mathbf{curl}; \Omega), \quad (2.4.20)$$

where

$$a^*(\mathbf{w}, \mathbf{v}) := \overline{a(\mathbf{v}, \mathbf{w})} \quad \forall \mathbf{w}, \mathbf{v} \in H(\mathbf{curl}; \Omega). \quad (2.4.21)$$

By the Lax-Milgram lemma (see Lemma 2.3.7), the above problem is well-posed. Now, given $\boldsymbol{\xi}$ in $H(\text{curl}_\Gamma; \Gamma)$ we set $\mathbf{u}_\xi = \mathbb{S}\boldsymbol{\xi}$, the unique solution to state equation, cf. (2.4.1), namely

$$\begin{aligned} a(\mathbf{u}_\xi, \mathbf{v}) &= 0 & \forall \mathbf{v} \in H_0(\mathbf{curl}; \Omega), \\ \gamma_t \mathbf{u}_\xi &= \boldsymbol{\xi} \times \mathbf{n} & \text{on } \Gamma. \end{aligned} \quad (2.4.22)$$

In turn, given $\mathbf{f} \in \mathbf{L}^2(\Omega)$ let \mathbf{w} be the solution for (2.4.20), and testing the first equation in (2.4.19) with $\overline{\mathbb{S}\boldsymbol{\xi}}$, and integrating by parts using (2.3.6) and the fact that $\boldsymbol{\mu}^{-1}\text{curl } \mathbf{w} \in H(\mathbf{curl}; \Omega)$ we arrive at

$$\begin{aligned} (\mathbf{f}, \mathbb{S}\boldsymbol{\xi})_{0,\Omega} &= -\langle \gamma_t(\boldsymbol{\mu}^{-1}\nabla \times \mathbf{w}), \gamma_T \mathbf{u}_\xi \rangle_{\Gamma^*} \\ &= -\langle \gamma_t(\boldsymbol{\mu}^{-1}\nabla \times \mathbf{w}), \boldsymbol{\xi} \rangle_{\Gamma^*}. \end{aligned}$$

Recall that $\langle \cdot, \cdot \rangle_{\Gamma^*}$ represents the duality pairing $H_{\parallel}^{-\frac{1}{2}}(\text{div}_\Gamma; \Gamma)$ and $H_{\perp}^{-\frac{1}{2}}(\text{curl}_\Gamma; \Gamma)$. Nevertheless, under some additional regularity on $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$ we will show that \mathbf{w} is smooth enough so this duality becomes an integral. Now, by setting $\mathbf{f} = \mathbf{u} - \mathbf{u}_d$, (2.4.19) becomes the *adjoint problem* for the state equation, and we have the following result.

Theorem 2.4.10. *The adjoint operator for \mathbb{S} is given by*

$$\begin{aligned} \mathbb{S}^* : \mathbf{L}^2(\Omega) &\rightarrow H_{\parallel}^{-\frac{1}{2}}(\text{div}_\Gamma; \Gamma) \\ \mathbf{f} &\mapsto \mathbb{S}^* \mathbf{f} = -\gamma_t(\boldsymbol{\mu}^{-1}\nabla \times \mathbf{w}), \end{aligned} \quad (2.4.23)$$

where $\mathbf{w} \in H_0(\mathbf{curl}; \Omega)$ solves

$$\begin{aligned} \text{curl}(\boldsymbol{\mu}^{-1}\text{curl } \mathbf{w}) - (i\omega)\boldsymbol{\kappa}\mathbf{w} &= \mathbf{f} & \text{in } \Omega \\ \mathbf{w} \times \mathbf{n} &= 0 & \text{on } \Gamma. \end{aligned} \quad (2.4.24)$$

Since, $H(\text{curl}_\Gamma; \Gamma) \hookrightarrow H_{\perp}^{-\frac{1}{2}}(\text{curl}_\Gamma; \Gamma)$, we can rewrite (2.4.17) as

$$d^{\mathbb{R}}j(\mathbf{z}; \boldsymbol{\xi}) = \text{Re} \left\{ \langle \mathbb{S}^*(\mathbb{S}\mathbf{z} - \mathbf{u}_d), \boldsymbol{\xi} \rangle_{\Gamma^*} + \alpha(\text{curl}_\Gamma \mathbf{z}, \text{curl}_\Gamma \boldsymbol{\xi})_{0,\Gamma} + \beta(\mathbf{z}, \boldsymbol{\xi})_{0,\Gamma} \right\}, \quad (2.4.25)$$

for all $\boldsymbol{\xi} \in H(\text{curl}_\Gamma; \Gamma)$.

Proof. The proof follows from the above analysis and the fact that $\boldsymbol{\mu}^{-1}\text{curl}\boldsymbol{w} \in H(\mathbf{curl}; \Omega)$, and therefore $\gamma_t(\boldsymbol{\mu}^{-1}\nabla \times \boldsymbol{w}) \in H_{\parallel}^{-\frac{1}{2}}(\text{div}_{\Gamma}; \Gamma)$, cf. Lemma 2.3.4. Finally, since $H(\text{curl}_{\Gamma}; \Gamma) \hookrightarrow H_{\perp}^{-\frac{1}{2}}(\text{curl}_{\Gamma}; \Gamma)$, therefore, $H_{\parallel}^{-\frac{1}{2}}(\text{div}_{\Gamma}; \Gamma) = \left(H_{\perp}^{-\frac{1}{2}}(\text{curl}_{\Gamma}; \Gamma)\right)^* \hookrightarrow H(\text{curl}_{\Gamma}; \Gamma)^*$. Thus, we can replace the $H(\text{curl}_{\Gamma}; \Gamma)^*-H(\text{curl}_{\Gamma}; \Gamma)$ duality pairing in (2.4.18) by the $\langle \cdot, \cdot \rangle_{\Gamma^*}$ pairing and the proof is complete. \square

2.4.3 Additional regularity of \mathbb{S}^* for Lipschitz polyhedra

The goal of this section is to show that we can obtain additional regularity for \mathbb{S}^* ; for instance, $\mathbb{S}^* : \mathbf{L}^2(\Omega) \rightarrow L_t^2(\Gamma)$, under suitable assumptions on $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$. These additional assumptions along with the regularity of the adjoint problem (2.4.24) will imply $\gamma(\boldsymbol{\mu}^{-1}\text{curl}\boldsymbol{w}) \in \mathcal{H}^{\sigma}(\Gamma)$, for some $\sigma > 0$. As a result, we can replace the duality pairing in (2.4.25) by the inner product in $\mathbf{L}^2(\Gamma)$. From now on, we will assume that $\boldsymbol{\kappa}$ and $\boldsymbol{\mu}$ are in $W^{1,\infty}(\Omega)$. In what follows, we give some technical results that are slight variations of the ones found in [152, Thm. 4.1]. We omit most of the proof details as they mostly follow from straightforward calculations.

Lemma 2.4.11. *If $\boldsymbol{\kappa} \in W^{1,\infty}(\Omega)$ such that $\boldsymbol{\kappa}(x) \geq \boldsymbol{\kappa}_0 > 0$ almost everywhere, and $\phi \in \mathcal{C}_0^{\infty}(\Omega)$, then $\boldsymbol{\kappa}^{-1}\phi \in H_0^1(\Omega)$ and*

$$\nabla\phi = (\boldsymbol{\kappa}^{-1}\nabla\boldsymbol{\kappa})\phi + \boldsymbol{\kappa}\nabla(\boldsymbol{\kappa}^{-1}\phi).$$

Proof. The proof follows immediately after using the product rule in $\nabla(\boldsymbol{\kappa}^{-1}\phi)$ and rearranging the resulting expression. \square

Using this result, we obtain the following two lemmas involving the product rule for divergence and curl:

Lemma 2.4.12. *Let $\boldsymbol{\kappa} \in W^{1,\infty}(\Omega)$ such that $\boldsymbol{\kappa}(x) \geq \boldsymbol{\kappa}_0 > 0$ almost everywhere, and $\boldsymbol{u} \in \mathbf{L}^2(\Omega)$ such that $\text{div}\boldsymbol{\kappa}\boldsymbol{u} = v \in L^2(\Omega)$, then $\boldsymbol{u} \in H(\text{div}; \Omega)$ and*

$$\text{div}(\boldsymbol{u}) = \boldsymbol{\kappa}^{-1}v - (\boldsymbol{\kappa}^{-1}\nabla\boldsymbol{\kappa}) \cdot \boldsymbol{u}. \tag{2.4.26}$$

Lemma 2.4.13. *Let $\zeta \in W^{1,\infty}(\Omega)$ such that $\zeta(x) \geq \zeta_0 > 0$ almost everywhere, and $\mathbf{u} \in \mathbf{L}^2(\Omega)$ such that $\text{curl} \zeta \mathbf{u} = \mathbf{v} \in \mathbf{L}^2(\Omega)$, then $\mathbf{u} \in H(\mathbf{curl}; \Omega)$ and*

$$\text{curl}(\mathbf{u}) = \zeta^{-1} \mathbf{v} + \mathbf{u} \times (\zeta^{-1} \nabla \zeta). \quad (2.4.27)$$

We will show that $\boldsymbol{\mu}^{-1} \nabla \times \mathbf{w}$ is smooth enough to have a well-defined and integrable trace, where \mathbf{w} is the solution for the adjoint problem (2.4.24). To do this, we will now show some results for the regularity of \mathbf{w} ,

Lemma 2.4.14. *If $\mathbf{w} \in H_0(\mathbf{curl}; \Omega)$, then $\text{curl} \mathbf{w} \in H_0(\text{div}; \Omega)$.*

Proof. It is clear that $\text{div}(\text{curl} \mathbf{w}) = 0$, then $\text{curl} \mathbf{w} \in H(\text{div}; \Omega)$ and therefore $\gamma_{\mathbf{n}}(\text{curl} \mathbf{w})$ belongs to $H^{-\frac{1}{2}}(\Gamma)$. Similarly; for each $v \in H^1(\Omega)$, $\nabla v \in H(\mathbf{curl}; \Omega)$, and

$$\begin{aligned} \langle \gamma_{\mathbf{n}}(\text{curl} \mathbf{w}), \gamma v \rangle_{-\frac{1}{2}, \frac{1}{2}; \Gamma} &= \int_{\Omega} v \text{div}(\text{curl} \mathbf{w}) d\mathbf{x} + \int_{\Omega} \text{curl} \mathbf{w} \cdot \nabla v d\mathbf{x} \\ &= 0 + \int_{\Omega} \mathbf{w} \cdot \text{curl}(\nabla v) d\mathbf{x} - \langle \gamma_t \mathbf{w}, \gamma_T(\nabla v) \rangle_{\Gamma^*} \\ &= 0. \end{aligned}$$

Thus, from the surjectivity of the trace map γ from $H^1(\Omega)$ onto $H^{\frac{1}{2}}(\Gamma)$ the proof concludes. \square

Theorem 2.4.15. *Let \mathbf{w} be the solution for the adjoint problem (2.4.24) with $\boldsymbol{\mu}, \boldsymbol{\kappa} \in W^{1,\infty}(\Omega)$, then*

$$\boldsymbol{\mu}^{-1} \text{curl} \mathbf{w} \in H(\mathbf{curl}; \Omega) \cap H_0(\text{div}; \Omega).$$

Proof. We will first invoke Lemma 2.4.12 with $\mathbf{u} = \mathbf{G} := \boldsymbol{\mu}^{-1} \text{curl} \mathbf{w}$ and $\boldsymbol{\kappa} = \boldsymbol{\mu}$. Notice that $\mathbf{G} \in \mathbf{L}^2(\Omega)$ and $\text{div} \boldsymbol{\mu} \mathbf{G} = \text{div} \text{curl} \mathbf{w} = 0 \in L^2(\Omega)$. Therefore, $\mathbf{G} \in H(\text{div}; \Omega)$ and

$$\text{div} \mathbf{G} = -(\boldsymbol{\mu}^{-1} \nabla \boldsymbol{\mu}) \cdot (\boldsymbol{\mu}^{-1} \text{curl} \mathbf{w}) \in L^2(\Omega). \quad (2.4.28)$$

In addition, we notice that $\text{curl} \mathbf{G} = (\mathbf{u} - \mathbf{u}_d) + i\boldsymbol{\omega} \boldsymbol{\kappa} \mathbf{w} \in \mathbf{L}^2(\Omega)$. Thus, $\mathbf{G} \in H(\mathbf{curl}; \Omega) \cap H(\text{div}; \Omega)$.

Finally, to complete the proof, we show that the normal trace of \mathbf{G} vanishes. From Lemma 2.4.11, $\boldsymbol{\mu}^{-1}v \in H^1(\Omega)$ for all $v \in H^1(\Omega)$, this along Lemma 2.4.14, and the Green's identity in $H(\text{div}; \Omega)$, cf. [66, Lemma 1.4], yields

$$\begin{aligned} 0 &= \langle \gamma_{\mathbf{n}}(\text{curl } \mathbf{w}), \gamma(\boldsymbol{\mu}^{-1}v) \rangle_{-\frac{1}{2}, \frac{1}{2}, \Gamma} = \int_{\Omega} \text{curl } \mathbf{w} \cdot \nabla(\boldsymbol{\mu}^{-1}v) dx + \int_{\Omega} (\boldsymbol{\mu}^{-1}v) \text{div}(\text{curl } \mathbf{w}) dx \\ &= \int_{\Omega} \boldsymbol{\mu}^{-1} \text{curl } \mathbf{w} \cdot \boldsymbol{\mu} \nabla(\boldsymbol{\mu}^{-1}v) dx \\ &= \int_{\Omega} \boldsymbol{\mu}^{-1} \text{curl } \mathbf{w} \cdot (\nabla v - \boldsymbol{\mu}^{-1} \nabla \boldsymbol{\mu} v) dx. \end{aligned}$$

Then using (2.4.28), we obtain that

$$\begin{aligned} 0 &= \langle \gamma_{\mathbf{n}}(\text{curl } \mathbf{w}), \gamma(\boldsymbol{\mu}^{-1}v) \rangle_{-\frac{1}{2}, \frac{1}{2}, \Gamma} = \int_{\Omega} \boldsymbol{\mu}^{-1} \text{curl } \mathbf{w} \cdot \nabla v dx + \int_{\Omega} v \text{div}(\boldsymbol{\mu}^{-1} \text{curl } \mathbf{w}) dx \\ &= \left\langle \gamma_{\mathbf{n}}(\boldsymbol{\mu}^{-1} \text{curl } \mathbf{w}), \gamma(v) \right\rangle_{-\frac{1}{2}, \frac{1}{2}, \Gamma}, \end{aligned}$$

which concludes the proof. \square

Corollary 2.4.16. *Let \mathbf{w} be the solution for the adjoint equation (2.4.24), then*

$$\gamma_t(\boldsymbol{\mu}^{-1} \nabla \times \mathbf{w}) \in L_t^2(\Gamma).$$

Proof. From [2, Thm. 4.4] $H(\mathbf{curl}; \Omega) \cap H_0(\text{div}; \Omega) \hookrightarrow \mathcal{H}^{\frac{1}{2} + \epsilon_*}(\Omega)$, for some $\epsilon_* > 0$. Thus, according to Theorem 2.4.15 we have $\boldsymbol{\mu}^{-1} \nabla \times \mathbf{w} \in \mathcal{H}^{\frac{1}{2} + \epsilon_*}(\Omega)$, then by trace theorem, $\gamma(\boldsymbol{\mu}^{-1} \nabla \times \mathbf{w}) \in \mathcal{H}^{\epsilon_*}(\Gamma)$. In particular, $\gamma_t(\boldsymbol{\mu}^{-1} \nabla \times \mathbf{w}) \in L_t^2(\Gamma)$. \square

As a result of Corollary 2.4.16, the duality pairing in (2.4.25) becomes an integral, namely

$$\begin{aligned} \langle \gamma_t(\boldsymbol{\mu}^{-1} \nabla \times \mathbf{w}), \boldsymbol{\xi} \rangle_{\Gamma^*} &= \langle \gamma(\boldsymbol{\mu}^{-1} \nabla \times \mathbf{w}) \times \mathbf{n}, \boldsymbol{\xi} \rangle_{\Gamma^*} \\ &= \int_{\Gamma} (\gamma(\boldsymbol{\mu}^{-1} \nabla \times \mathbf{w}) \times \mathbf{n}) \cdot \bar{\boldsymbol{\xi}} \, dS. \end{aligned}$$

Corollary 2.4.17. *The function $\widehat{\mathbf{z}} \in Z$ is an optimal control for (2.4.3), if and only if*

$$\begin{aligned} \widehat{\mathbf{u}} &= \mathbb{S} \widehat{\mathbf{z}}, \\ \widehat{\boldsymbol{\zeta}} &= \mathbb{S}^*(\widehat{\mathbf{u}} - \mathbf{u}_d), \\ \text{Re} \left\{ \left(\widehat{\boldsymbol{\zeta}}, \mathbf{z} - \widehat{\mathbf{z}} \right)_{0, \Gamma} + \alpha \left(\text{curl}_{\Gamma} \widehat{\mathbf{z}}, \text{curl}_{\Gamma}(\mathbf{z} - \widehat{\mathbf{z}}) \right)_{0, \Gamma} + \beta \left(\widehat{\mathbf{z}}, \mathbf{z} - \widehat{\mathbf{z}} \right)_{0, \Gamma} \right\} &= 0, \quad \forall \mathbf{z} \in Z. \end{aligned}$$

Proof. The result is a consequence of Theorems 2.4.6 and 2.4.10, and Corollary 2.4.16. \square

2.5 Discrete Problem

Let Ω be a connected polyhedral Lipschitz domain, and $\{\mathcal{T}_h\}_h$ be a family of shape regular simplicial triangulation for Ω . For a given mesh \mathcal{T}_h we denote the set of faces on the boundary by Γ_h , and the edges belonging to Γ_h by \mathcal{E}_Γ^h . We consider a conforming finite element space for $H(\mathbf{curl}; \Omega)$ denoted by \mathcal{N}_k^h , we utilize the basis given in [72]. More specifically, for $T \in \mathcal{T}_h$ we consider a local basis for $\mathcal{N}_k^h(T) := \mathcal{P}_k^3(T) \oplus (\mathbf{x} \times \tilde{\mathcal{P}}_k^3(T)) \subsetneq \mathcal{P}_{k+1}^3(T)$ where $\tilde{\mathcal{P}}_k$ is the set of homogeneous polynomials of degree k . Note that it is common to count these spaces in the index $k + 1$, so that the lowest order would be $k = 1$, but here that case will be $k = 0$. Thus, given $\mathbf{z} \in H(\mathbf{curl}_\Gamma; \Gamma)$ and $k \in \mathbb{N} \cup \{0\}$, we consider the semi-discrete state equation: find $\mathbf{u}_h \in \mathcal{N}_k^h$ such that

$$\begin{aligned} (\boldsymbol{\mu}^{-1} \mathbf{curl} \mathbf{u}_h, \mathbf{curl} \mathbf{v}_h)_{0,\Omega} + (i\omega) (\boldsymbol{\kappa} \mathbf{u}_h, \mathbf{v}_h)_{0,\Omega} &= 0 \quad \forall \mathbf{v}_h \in \mathcal{N}_{k0}^h, \\ \mathbf{u}_h \times \mathbf{n} &= \mathbf{z} \times \mathbf{n} \quad \text{on } \Gamma, \end{aligned} \tag{2.5.1}$$

where $\mathcal{N}_{k0}^h := \mathcal{N}_k^h \cap H_0(\mathbf{curl}; \Omega)$. Because \mathcal{N}_{k0}^h is a closed subspace of $H_0(\mathbf{curl}; \Omega)$ well-posedness of (2.5.1) follows from the continuous case, except for how the boundary condition $\mathbf{u}_h \times \mathbf{n} = \mathbf{z} \times \mathbf{n}$ is imposed. This can be done in several ways; for instance, with a $L^2(\Gamma_h)$ projection plus taking an average for the edge dofs. Nevertheless, for that approach, approximation and commutativity properties seem rather difficult to prove. We impose the Dirichlet condition in (2.5.1) using moments, which allows us to use the well-known approximation theory of Nédélec and Raviart-Thomas elements in 2D along the approximation theory described in [1].

2.5.1 Imposition of discrete boundary conditions

As we already mentioned, we impose the condition $\mathbf{u}_h \times \mathbf{n} = \mathbf{z} \times \mathbf{n}$ through a lifting defined by moments. Now, the natural strategy would be to use the 2D local moments for $H(\mathbf{div}; F)$, for each face F on Γ_h , which is a collection of linear integral

equations; see, for instance, [66, Sec. 3]. Nevertheless, because we consider $\mathbf{z} \times \mathbf{n}$ instead of just \mathbf{z} , imposing those moments turns out to be equivalent to imposing $\gamma_T \mathbf{u}_h = \mathbf{z}$ through the local moments for $H(\text{curl}, F)$, cf. [115, Sec. 5.5]. We now define a global lifting operator

$$\begin{aligned} \mathcal{L}_h : \mathcal{D}_{\mathcal{L}_h} \subseteq H(\text{curl}_\Gamma; \Gamma) &\mapsto \mathcal{N}_k^h \\ \mathbf{z} &\mapsto \mathcal{L}_h \mathbf{z} := \mathbf{u}_z^h, \end{aligned} \quad (2.5.2)$$

where $\mathcal{D}_{\mathcal{L}_h}$ denotes the domain of \mathcal{L}_h which is a finite subspace of $H(\text{curl}_\Gamma; \Gamma)$ and its elements have well-defined and continuous moments on Γ_h , \mathbf{u}_z^h is the unique element in \mathcal{N}_k^h which shares the local 2D Nédélec moments with \mathbf{z} for all the faces/edges in Γ_h , and it has zero interior moments. Because of the regularity needed for this lifting, we will choose the discrete control space Z^h so that the following holds

$$Z^h \subseteq \mathcal{D}_{\mathcal{L}_h} \subseteq \gamma_T \mathcal{N}_j^h, \quad \text{for some } j \in \{0, \dots, k\}. \quad (2.5.3)$$

2.5.2 Discrete solution operator and cost functional

Let $j \in \{0, \dots, k\}$ and define $Z^h := \mathbf{R}_h \cap Z$, where

$$\mathbf{R}_h := \left\{ \gamma_T \mathbf{u}_h : \mathbf{u}_h \in \mathcal{N}_j^h \right\}. \quad (2.5.4)$$

Now, let us introduce the discrete optimization problem,

$$\min_{\mathbf{u}_h, \mathbf{z}_h} \mathcal{J}(\mathbf{u}_h, \mathbf{z}_h) := \frac{1}{2} \int_{\Omega} |\mathbf{u}_h - \mathbf{u}_d|^2 d\mathbf{x} + \frac{\alpha}{2} \int_{\Gamma} |\text{curl}_\Gamma \mathbf{z}_h|^2 dS + \frac{\beta}{2} \int_{\Gamma} |\mathbf{z}_h|^2 dS, \quad (2.5.5)$$

where \mathbf{u}_h is the unique solution to (2.5.1), for $\mathbf{z} = \mathbf{z}_h$. As in the continuous case, we reduce this problem with the help of a (discrete) solution operator

$$\begin{aligned} \mathbb{S}_h : Z^h \subseteq \mathbf{R}_h &\mapsto \mathcal{N}_k^h \\ \mathbf{z}_h &\mapsto \mathbb{S}_h \mathbf{z}_h =: \mathbf{u}_h. \end{aligned}$$

To prove that \mathbb{S}_h is bounded, with constant independent of h , let us consider the following lemma.

Lemma 2.5.1. *Let \mathbf{u} and \mathbf{u}_h be the solutions for the continuous and discrete state equations, (2.2.1b) and (2.5.1), respectively, with respective boundary data given by $\mathbf{z} \times \mathbf{n}$ and $\mathbf{z}_h \times \mathbf{n}$. Then, the following holds*

$$\|\mathbf{u} - \mathbf{u}_h\|_{\text{curl}, \Omega} \leq C \left(\min_{\mathbf{v}_h \in \mathcal{N}_k^h} \|\mathbf{u} - \mathbf{v}_h\|_{\text{curl}, \Omega} + \|\mathbf{z} - \mathbf{z}_h\|_{H_{\perp}^{-\frac{1}{2}}(\text{curl}_{\Gamma}; \Gamma)} \right), \quad (2.5.6)$$

where C only depends on the shape regularity of \mathcal{T}_h , Ω and the eigenvalues of $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$.

Proof. From [1, Corollary 4.4], we have

$$\|\mathbf{u} - \mathbf{u}_h\|_{\text{curl}, \Omega} \leq C \left(\min_{\mathbf{v}_h \in \mathcal{N}_k^h} \|\mathbf{u} - \mathbf{v}_h\|_{\text{curl}, \Omega} + \|\mathbf{z} \times \mathbf{n} - \mathbf{z}_h \times \mathbf{n}\|_{H_{\parallel}^{-\frac{1}{2}}(\text{div}_{\Gamma}; \Gamma)} \right)$$

Finally, from the well-known isometry for weak rotations between $H_{\perp}^{-\frac{1}{2}}(\text{curl}_{\Gamma}; \Gamma)$ and $H_{\parallel}^{-\frac{1}{2}}(\text{div}_{\Gamma}; \Gamma)$, see [130, p. 448], the proof concludes. \square

A useful result related to (2.5.6) is the existence of a uniformly bounded operator

$$\begin{aligned} L_h : \mathbf{R}_h &\mapsto \mathcal{N}_k^h \\ \mathbf{z}_h &\mapsto \mathbf{u}_{\mathbf{z}_h}, \end{aligned}$$

such that $\gamma_{\Gamma}(\mathbf{u}_{\mathbf{z}_h}) = \mathbf{z}_h$, and $\|\mathbf{u}_{\mathbf{z}_h}\|_{\text{curl}, \Omega} \leq C\|\mathbf{z}_h\|_{\text{curl}_{\Gamma}}$, where C is independent of h , see [1] and references within. Thus, \mathbb{S}_h is linear and bounded independently of h . We now introduce the reduced discrete cost functional

$$\begin{aligned} j_h : Z^h &\mapsto \mathbb{R}, \\ \mathbf{z}_h &\mapsto j_h(\mathbf{z}_h) := \mathcal{J}(\mathbb{S}_h(\mathbf{z}_h), \mathbf{z}_h). \end{aligned} \quad (2.5.7)$$

It is clear that j_h is continuous and bounded independently of h . Now, under the same arguments as for the continuous case, we have

$$\begin{aligned} d^{\mathbb{R}} j_h(\mathbf{z}_h, \boldsymbol{\xi}_h) &= \text{Re}\{(\mathbb{S}_h \mathbf{z}_h - \mathbf{u}_d, \mathbb{S}_h \boldsymbol{\xi}_h)_{0, \Omega} + \alpha(\text{curl}_{\Gamma} \mathbf{z}_h, \text{curl}_{\Gamma} \boldsymbol{\xi}_h)_{0, \Gamma} + \beta(\mathbf{z}_h, \boldsymbol{\xi}_h)_{0, \Gamma}\} \\ &= \text{Re}\{\langle \mathbb{S}_h^*(\mathbb{S}_h \mathbf{z}_h - \mathbf{u}_d), \boldsymbol{\xi}_h \rangle_{\Gamma^*} + \alpha(\text{curl}_{\Gamma} \mathbf{z}_h, \text{curl}_{\Gamma} \boldsymbol{\xi}_h)_{0, \Gamma} + \beta(\mathbf{z}_h, \boldsymbol{\xi}_h)_{0, \Gamma}\} \end{aligned} \quad (2.5.8)$$

for all \mathbf{z}_h , and $\boldsymbol{\xi}_h$ in Z^h . Where the action of the discrete adjoint operator \mathbb{S}_h^* will be defined later, cf. (2.5.19).

2.5.3 Particular choice of discrete control space

For simplicity and because most of the theory for Nédélec elements focuses on this case, we consider the lowest-order space for our discrete control \mathbf{z}_h , along with a higher-order approximation for \mathbf{u}_h . As usual, we just construct the real-valued spaces. It turns out that it is easier to inherit properties for \mathbf{z}_h from the space of $\mathbf{z}_h \times \mathbf{n}$, which is related to the Raviart-Thomas space. In fact, given $F \in \Gamma_h$ we recall that the local Raviart-Thomas space $\mathcal{RT}_k(F)$ is just a $\pi/2$ -rotation of Nédélec's space $\mathcal{N}_k(F)$, see [49]. This idea can be easily generalized to the piecewise linear manifold Γ_h . In order to do that, we first introduce the Raviart-Thomas space for Γ_h

$$\mathcal{RT}_k(\Gamma_h) := \left\{ \mathbf{q} \in L_t^2(\Gamma) : \forall F \in \Gamma_h, \mathbf{q}|_F \in \mathcal{RT}_k(F) \text{ and } \forall \mathbf{e} \in \mathcal{E}_\Gamma^h, [\mathbf{q} \cdot \boldsymbol{\nu}]_e = 0 \right\}, \quad (2.5.9)$$

where $[\mathbf{q} \cdot \boldsymbol{\nu}]_e = 0$ denotes the continuity of $\mathbf{q} \cdot \boldsymbol{\nu}$ across the edge \mathbf{e} , and $\boldsymbol{\nu}$ is the 2D “normal vector” on \mathbf{e} , this will be clarified later, see (2.5.13). In the same manner, the space for the discrete control \mathbf{z}_h will be a subspace of the Nédélec space for Γ_h , which is given by

$$\mathcal{N}_k(\Gamma_h) := \left\{ \mathbf{q} \in L_t^2(\Gamma) : \forall F \in \Gamma_h, \mathbf{q}|_F \in \mathcal{N}_k(F) \text{ and } \forall \mathbf{e} \in \mathcal{E}_\Gamma^h, [\mathbf{q}]_e \cdot \mathbf{t}_e = 0 \right\}, \quad (2.5.10)$$

where \mathbf{t}_e is a unitary vector along the edge \mathbf{e} . Then, it follows that $\mathcal{RT}_k(\Gamma_h)$ is a $\pi/2$ -rotation of $\mathcal{N}_k(\Gamma_h)$ but just facewise, which can be compactly represented as $\mathcal{RT}_k(\Gamma_h) = \mathcal{N}_k(\Gamma_h) \times \mathbf{n}$. Now, we show that for the lowest order case, $k = 0$, the identity $\mathcal{RT}_k(\Gamma_h) = \mathcal{N}_k(\Gamma_h) \times \mathbf{n}$ can be reduced to an identity between elements of a particular choice of bases. To show that, notice that $\mathcal{RT}_0(\Gamma_h)$ is given by

$$\left\{ \mathbf{q} \in L_t^2(\Gamma) : \forall F \in \Gamma_h, \exists \mathbf{a} \in \mathbb{R}^3, \exists b \in \mathbb{R}, \mathbf{q}|_F = \mathbf{a} + b\mathbf{x} \text{ and } \forall \mathbf{e} \in \mathcal{E}_\Gamma^h, [\mathbf{q} \cdot \boldsymbol{\nu}]_e = 0 \right\}.$$

For a global basis for this space we consider the Rao-Wilton-Glisson basis [123]. From now on, we assume that all the faces on Γ_h are oriented counterclockwise in terms of its vertices, and the normal vector of a face points outward. Now, let us consider an

edge $\mathbf{e} \in \mathcal{E}_\Gamma^h$ and the two faces, F_+ and F_- on Γ_h , which share \mathbf{e} . Therefore, without loss of generality, we assume that F_- has \mathbf{e} with a negative orientation. We now want to find $\boldsymbol{\psi}_e$ in $\mathcal{RT}_0(\Gamma_h) \cap H_-^{\frac{1}{2}}(\Gamma)$ such that $\boldsymbol{\psi}_e \cdot \boldsymbol{\nu}|_{F_+} + \boldsymbol{\psi}_e \cdot \boldsymbol{\nu}|_{F_-}$ vanishes point-wise on \mathcal{E}_Γ^h . Here the main difference between this case and the 2D-case is that $\boldsymbol{\nu}|_{F_+} \neq -\boldsymbol{\nu}|_{F_-}$ in general. Nevertheless, “the definition for Raviart-Thomas basis” is the same for this case. Consider,

$$\boldsymbol{\psi}_e : \Gamma_h \mapsto \mathbb{R}^3,$$

$$\mathbf{x} \mapsto \boldsymbol{\psi}_e(\mathbf{x}) := \begin{cases} \frac{|\mathbf{e}|}{2|F_+|}(\mathbf{x} - \mathbf{v}_e^+) & : \mathbf{x} \in F_+, \\ -\frac{|\mathbf{e}|}{2|F_-|}(\mathbf{x} - \mathbf{v}_e^-) & : \mathbf{x} \in F_-, \\ \mathbf{0} & : \text{elsewhere,} \end{cases} \quad (2.5.11)$$

where \mathbf{v}_e^\pm is the vertex opposite to \mathbf{e} in F_\pm , see Figure 2.1, $\boldsymbol{\psi}_e$ was scaled for visualization purposes.

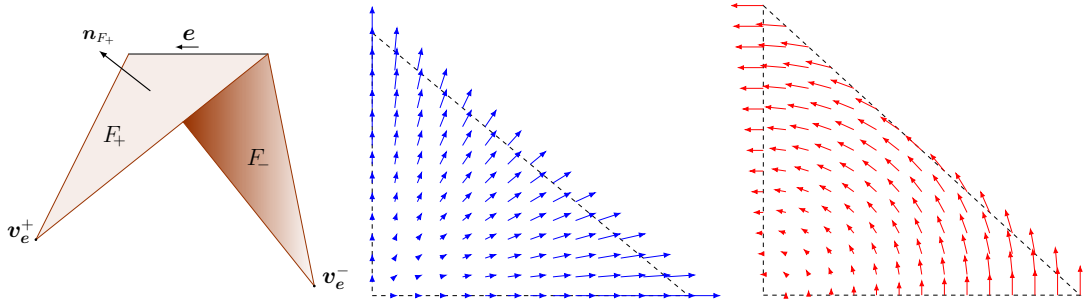


Figure 2.1: $F_+ \cup \mathbf{e} \cup F_-$, $\boldsymbol{\psi}_e|_{F_+}$ and $(\mathbf{n}_{F_+} \times \boldsymbol{\psi}_e)|_{F_+}$

The functions $\boldsymbol{\psi}_e$ have many good properties. For instance, it has a constant normal component on each edge of $\overline{F_+}$ and $\overline{F_-}$. To show this, we follow [16, Sec. 4]. Given $F \in \Gamma_h$, let us assume that $F = \text{conv}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, and define \mathbf{e}_j to be the edge opposite to the vertex \mathbf{v}_j . Then, by some basic properties of triangles and orthogonal projections in \mathbb{R}^2 , see Figure 2.2, yields

$$(\mathbf{x} - \mathbf{v}_j) \cdot \boldsymbol{\nu}_j = h_j = 2 \frac{|F|}{|\mathbf{e}_j|} \quad \forall \mathbf{x} \in \mathbf{e}_j, \quad \forall j \in \{1, 2, 3\}. \quad (2.5.12)$$

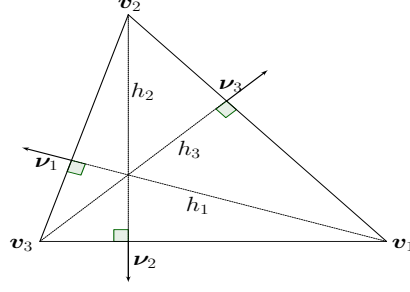


Figure 2.2: Altitudes and outer normals of F .

Now, we give a formal definition for the “normal” vector $\boldsymbol{\nu}$, see [33, section 2],

$$\boldsymbol{\nu} = \begin{cases} \mathbf{t}_e \times \mathbf{n}_{F_+} & \text{on } e \cap \overline{F_+}, \\ \mathbf{t}_e \times \mathbf{n}_{F_-} & \text{on } e \cap \overline{F_-}, \end{cases} \quad (2.5.13)$$

where \mathbf{t}_e is a unitary vector along e , following its orientation, and \mathbf{n}_{F_\pm} is the unitary outer normal vector to F_\pm . Note that, $\boldsymbol{\nu}|_{e \cap F_\pm} = \pm \boldsymbol{\nu}_{F_\pm}$, therefore $\boldsymbol{\nu}|_{e \cap F_-}$ points inward. With this definition we have the following result

Lemma 2.5.2. *Let $\boldsymbol{\psi}_e$ defined as in (2.5.11), then*

$$\begin{aligned} (\boldsymbol{\psi}_e \cdot \boldsymbol{\nu})(\mathbf{x}) &= \begin{cases} 1 & : \mathbf{x} \in e, \\ 0 & : \mathbf{x} \in \mathcal{E}_\Gamma^h \setminus \{e\}, \end{cases} \\ \operatorname{div}_\Gamma \boldsymbol{\psi}_e &= \begin{cases} \pm \frac{|e|}{|F_\pm|} & \text{in } F_\pm, \\ 0 & \text{elsewhere,} \end{cases} \\ \int_\Gamma \operatorname{div}_\Gamma \boldsymbol{\psi}_e dS &= 0. \end{aligned}$$

Proof. See [16, Lemma 4.1]. □

Now, we will study a basis for $\mathcal{N}_0(\Gamma_h)$, cf. (2.5.10). To simplify the notation, from now on we assume that e is the edge between the vertices $[\mathbf{x}_\ell, \mathbf{x}_m]$, following that

orientation. For the lowest order case, in 2D and 3D, there are only edge related functions of the form, $\tilde{\phi}_e = \lambda_\ell \nabla \lambda_m - \lambda_m \nabla \lambda_\ell$. This motivates us to consider the collection of functions $\{\phi_e\}_{e \in \mathcal{E}_\Gamma^h}$, given by

$$\begin{aligned} \phi_e &: \Gamma_h \mapsto \mathbb{R}^3, \\ \mathbf{x} \mapsto \phi_e(\mathbf{x}) &:= \begin{cases} |e| (\lambda_\ell \nabla_\Gamma \lambda_m - \lambda_m \nabla_\Gamma \lambda_\ell)|_{F_\pm} & : \mathbf{x} \in F_\pm, \\ \mathbf{0} & : \text{elsewhere.} \end{cases} \end{aligned} \quad (2.5.14)$$

Note that $\phi_e = \mathbf{n}_{F_\pm} \times \psi_e$, see Figure 2.1. Based on this observation, we have the following result

Lemma 2.5.3. *Let $e \in \mathcal{E}_\Gamma^h$ then*

$$\begin{aligned} \phi_e \cdot \mathbf{t}_{\hat{e}} &= \begin{cases} 1 & e = \hat{e}, \\ 0 & e \neq \hat{e}, \end{cases} \\ \phi_e \times \mathbf{n} &= \psi_e, \\ \text{curl}_\Gamma \phi_e &= \begin{cases} \pm \frac{|e|}{|F_\pm|} & \text{in } F_\pm, \\ 0 & \text{elsewhere,} \end{cases} \\ \int_\Gamma \text{curl}_\Gamma \phi_e dS &= 0. \end{aligned}$$

Proof. It follows from $\nabla_\Gamma \lambda_\ell \cdot \mathbf{t}_e = \nabla \lambda_\ell \cdot \mathbf{t}_e$, the definition of $\boldsymbol{\nu}$, cf. (2.5.13), $\phi_e \times \mathbf{n}$ having the same edge moments as ψ_e , the unisolvence of $\mathcal{RT}_0(F)$, and the identity $\text{curl}_\Gamma \phi_e = \text{div}_\Gamma \phi_e \times \mathbf{n}$. \square

We now show how to compute the terms that appear in the stabilization term in the cost functional and its derivative for the discrete setting. Our analysis only involves the length of the boundary edges, their local orientation, and the barycentric coordinates $\hat{\lambda}_1, \hat{\lambda}_2$ and $\hat{\lambda}_3$ on the 2D reference face $\hat{F} = \text{conv}\{(0,0)^T, (1,0)^T, (0,1)^T\}$.

Proposition 2.5.4. *Let $F \in \Gamma_h$ with edges $(\mathbf{e}_a, \mathbf{e}_b, \mathbf{e}_c)$, $\mathbf{z} \in Z^h$ with $\mathbf{z}|_F = \beta_a \boldsymbol{\phi}_{\mathbf{e}_a} + \beta_b \boldsymbol{\phi}_{\mathbf{e}_b} + \beta_c \boldsymbol{\phi}_{\mathbf{e}_c}$. Then, for each $i, j \in \{a, b, c\}$, yields*

$$\langle \beta_i \operatorname{curl}_\Gamma \boldsymbol{\phi}_{\mathbf{e}_i}, \beta_j \operatorname{curl}_\Gamma \boldsymbol{\phi}_{\mathbf{e}_j} \rangle_F = \beta_i \bar{\beta}_j \int_F \operatorname{curl}_\Gamma \boldsymbol{\phi}_{\mathbf{e}_i} \operatorname{curl}_\Gamma \boldsymbol{\phi}_{\mathbf{e}_j} dS = \beta_i \bar{\beta}_j \frac{|\mathbf{e}_i| |\mathbf{e}_j|}{|F|}, \quad (2.5.15)$$

$$\langle \beta_i \boldsymbol{\phi}_{\mathbf{e}_i}, \beta_j \boldsymbol{\phi}_{\mathbf{e}_j} \rangle_F = \beta_i \bar{\beta}_j \int_F \boldsymbol{\phi}_i \cdot \boldsymbol{\phi}_j dS = \beta_i \bar{\beta}_j |\mathbf{e}_i| |\mathbf{e}_j| \frac{|F|}{2} \int_{\hat{F}} (\hat{\boldsymbol{\phi}}_\ell)^t \mathbb{B}_F \hat{\boldsymbol{\phi}}_m dS, \quad (2.5.16)$$

where $\ell, m \in \{1, 2, 3\}$,

$$\mathbb{B}_F := \frac{1}{4|F|^2} \begin{bmatrix} |\mathbf{e}_a|^2 & \frac{|\mathbf{e}_b|^2 - (|\mathbf{e}_a|^2 + |\mathbf{e}_c|^2)}{2} \\ \frac{|\mathbf{e}_b|^2 - (|\mathbf{e}_a|^2 + |\mathbf{e}_c|^2)}{2} & |\mathbf{e}_c|^2 \end{bmatrix},$$

$$\hat{\boldsymbol{\phi}}_1 := -(\hat{\lambda}_1 + \hat{\lambda}_2) \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \hat{\lambda}_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \hat{\boldsymbol{\phi}}_2 := \hat{\lambda}_3 \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \hat{\lambda}_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \text{ and}$$

$$\hat{\boldsymbol{\phi}}_3 := \hat{\lambda}_3 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + (\hat{\lambda}_1 + \hat{\lambda}_3) \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Proof. First of all, because we are dealing with tangent fields, it is enough to show the 2D case. Let us consider a face $F = \operatorname{conv}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, where $\mathbf{v}_j = (x_j, y_j)^T$ for $j \in \{1, 2, 3\}$, and the reference face $\hat{F} = \operatorname{conv}\{(0, 0)^T, (1, 0)^T, (0, 1)^T\}$ with associated barycentric coordinates $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3$, and the affine map

$$\eta_F : \hat{F} \mapsto F,$$

$$\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} \mapsto \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} + \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}.$$

Let us assume the edges of F have length $|\mathbf{e}_a|, |\mathbf{e}_b|$ and $|\mathbf{e}_c|$, see Figure 2.3. And, as usual, define

$$B_F = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}. \quad (2.5.17)$$

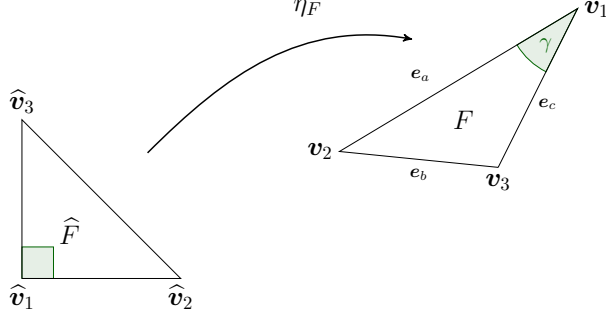


Figure 2.3: Affine transformation between \widehat{F} and F

It straightforward to show

$$\det(B_F) = |F|/|\widehat{F}| = 2|F|,$$

$$B_F^{-1} = \frac{1}{2|F|} \begin{pmatrix} y_3 - y_1 & x_1 - x_3 \\ y_1 - y_2 & x_2 - x_1 \end{pmatrix},$$

$$B_F^{-1} B_F^{-T} = \frac{1}{4|F|^2} \begin{pmatrix} |\mathbf{e}_a|^2 & \frac{|\mathbf{e}_b|^2 - (|\mathbf{e}_a|^2 + |\mathbf{e}_c|^2)}{2} \\ \frac{|\mathbf{e}_b|^2 - (|\mathbf{e}_a|^2 + |\mathbf{e}_c|^2)}{2} & |\mathbf{e}_c|^2 \end{pmatrix},$$

where we have used $\overrightarrow{\mathbf{v}_1 \mathbf{v}_2} \cdot \overrightarrow{\mathbf{v}_1 \mathbf{v}_3} = |\mathbf{e}_a| |\mathbf{e}_c| \cos(\gamma)$, see Figure 2.3, and the law of cosines

$$|\mathbf{e}_b|^2 = |\mathbf{e}_a|^2 + |\mathbf{e}_c|^2 - 2|\mathbf{e}_a| |\mathbf{e}_c| \cos(\gamma).$$

In practice we replace $|\mathbf{e}_j|$ by $\pm|\mathbf{e}_j|$, according to the orientation of the edge on F . The rest of the proof follows from the fact that the elements of our basis are up to a constant, the length of the edge, are the same as the standard lowest order Nédélec basis of the first kind. \square

A different approach to compute the above mentioned quantities can be done in terms of $\boldsymbol{\phi}_e \times \mathbf{n}$, for the 2D case, this can be found in [16, sec. 4.2-4.3].

2.5.4 Discrete Adjoint equation and optimality conditions

Because for $\mathbf{w}_h \in \mathcal{N}_k^h$, $\nabla \times \mathbf{w}_h \notin H(\mathbf{curl}; \Omega)$, we cannot apply the same strategy as for the continuous adjoint \mathbb{S}^* ; cf. Theorem 2.4.10. To overcome this problem, we first

consider a discretization of the continuous adjoint problem, cf. (2.4.24): find $\mathbf{w}_h \in \mathcal{N}_{k0}^h$ such that,

$$a^*(\mathbf{w}_h, \mathbf{v}_h) = (\mathbb{S}_h \mathbf{z}_h - \mathbf{u}_d, \mathbf{v}_h)_{0,\Omega} \quad \forall \mathbf{v}_h \in \mathcal{N}_{k0}^h, \quad (2.5.18)$$

and as in the continuous case, for $\boldsymbol{\xi}_h$ and \mathbf{z}_h in Z_h we want to simplify the term

$$(\mathbb{S}_h \mathbf{z}_h - \mathbf{u}_d, \mathbb{S}_h \boldsymbol{\xi}_h)_{0,\Omega}$$

which appears in the definition of $d^{\mathbb{R}} j_h$, so it does not require to compute/assemble the term $\mathbb{S}_h \boldsymbol{\xi}_h$, for each feasible direction $\boldsymbol{\xi}_h$. To do that, we apply the following splitting

$$\mathbb{S}_h \boldsymbol{\xi}_h = \mathbb{S}_{h0} \boldsymbol{\xi}_h + \mathcal{L}_h \boldsymbol{\xi}_h,$$

where $\mathbb{S}_{h0} \boldsymbol{\xi}_h \in \mathcal{N}_{k0}^h$, and \mathcal{L}_h is the lifting operator defined in (2.5.2). Thus, we define

$$\begin{aligned} \langle \mathbb{S}_h^*(\mathbb{S} \mathbf{z} - \mathbf{u}_d), \boldsymbol{\xi} \rangle_{\Gamma^*} &:= (\mathbb{S}_h \mathbf{z} - \mathbf{u}_d, \mathbb{S}_h \boldsymbol{\xi})_{0,\Omega} \\ &= (\mathbb{S}_h \mathbf{z} - \mathbf{u}_d, \mathbb{S}_{h0} \boldsymbol{\xi})_{0,\Omega} + (\mathbb{S}_h \mathbf{z} - \mathbf{u}_d, \mathcal{L}_h \boldsymbol{\xi})_{0,\Omega} \\ &\stackrel{(2.5.18)}{=} a^*(\mathbf{w}_h, \mathbb{S}_{h0} \boldsymbol{\xi}) + (\mathbb{S}_h \mathbf{z} - \mathbf{u}_d, \mathcal{L}_h \boldsymbol{\xi})_{0,\Omega} \\ &\stackrel{(2.4.21)}{=} \overline{a(\mathbb{S}_{h0} \boldsymbol{\xi}, \mathbf{w}_h)} + (\mathbb{S}_h \mathbf{z} - \mathbf{u}_d, \mathcal{L}_h \boldsymbol{\xi})_{0,\Omega} \\ &= -\overline{a(\mathcal{L}_h \boldsymbol{\xi}, \mathbf{w}_h)} + (\mathbb{S}_h \mathbf{z} - \mathbf{u}_d, \mathcal{L}_h \boldsymbol{\xi})_{0,\Omega}, \end{aligned} \quad (2.5.19)$$

where the last identity follows from (2.5.1), taking $\mathbf{v}_h = \mathbf{w}_h$. Note that the support of $\mathcal{L}_h \boldsymbol{\xi}_h$ is contained just in a small h -neighborhood of Γ_h .

Theorem 2.5.5. *The first-order optimality condition (2.4.15) implies in the discrete setting that: $\bar{\mathbf{z}}_h \in Z^h$ is an optimal control of (2.5.7) if and only if*

$$\bar{\mathbf{u}}_h = \mathbb{S}_h \bar{\mathbf{z}}_h,$$

$$\operatorname{Re} \left\{ \langle \mathbb{S}_h^*(\bar{\mathbf{u}}_h - \mathbf{u}_d), \mathbf{z}_h - \bar{\mathbf{z}}_h \rangle_{\Gamma^*} + \alpha (\operatorname{curl}_{\Gamma} \bar{\mathbf{z}}_h, \operatorname{curl}_{\Gamma} (\mathbf{z}_h - \bar{\mathbf{z}}_h))_{0,\Gamma} + \beta (\bar{\mathbf{z}}_h, \mathbf{z}_h - \bar{\mathbf{z}})_{0,\Gamma} \right\} = 0, \quad \forall \mathbf{z}_h \in Z^h.$$

2.5.5 Convergence of fully discrete scheme

Theorem 2.5.6. *Let $\{\bar{\mathbf{z}}_h\}_h$ be the family of discrete optimal controls related to $\{\mathcal{T}_h\}_h$ then*

(a) $\{\bar{\mathbf{z}}_h\}_h$ is uniformly bounded,

(b) there exists a subsequence $\{\bar{\mathbf{z}}_{\tilde{h}}\}_{\tilde{h}}$ of $\{\bar{\mathbf{z}}_h\}_h$ such that

$$\|\bar{\mathbf{z}}_{\tilde{h}} - \bar{\mathbf{z}}\|_{\text{curl}\Gamma} \rightarrow 0, \quad \text{as } \tilde{h} \rightarrow 0,$$

where $\bar{\mathbf{z}}$ is the unique solution to the continuous optimization problem (2.4.3).

Proof. The first part of the proof follows from

$$j_h(\bar{\mathbf{z}}_h) \leq j_h(\mathbf{0}) \Rightarrow \frac{\min\{\alpha, \beta\}}{2} \|\bar{\mathbf{z}}_h\|_{\text{curl}\Gamma}^2 \leq \frac{\alpha}{2} \|\text{curl}\Gamma \bar{\mathbf{z}}_h\|_{0,\Gamma}^2 + \frac{\beta}{2} \|\bar{\mathbf{z}}_h\|_{0,\Gamma}^2 \leq \|\mathbf{u}_d\|_{0,\Omega}^2. \quad (2.5.20)$$

Since $H(\text{curl}\Gamma; \Gamma)$ is a Hilbert space, $\{\bar{\mathbf{z}}_h\}_h$ has a weakly convergent subsequence i.e., there exists $\tilde{\mathbf{z}} \in H(\text{curl}\Gamma; \Gamma)$, such that

$$\bar{\mathbf{z}}_{\tilde{h}} \rightharpoonup \tilde{\mathbf{z}}, \quad \text{as } \tilde{h} \rightarrow 0. \quad (2.5.21)$$

Now, since the injection of $H(\text{curl}\Gamma; \Gamma)$ into $H_{\perp}^{-\frac{1}{2}}(\text{curl}\Gamma; \Gamma)$ is compact, therefore

$$\|\bar{\mathbf{z}}_{\tilde{h}} - \tilde{\mathbf{z}}\|_{H_{\perp}^{-\frac{1}{2}}(\text{curl}\Gamma; \Gamma)} \xrightarrow{\tilde{h}} 0, \quad \text{when } \bar{\mathbf{z}}_h \rightharpoonup \tilde{\mathbf{z}} \text{ in } H(\text{curl}\Gamma; \Gamma). \quad (2.5.22)$$

In turn, from (2.5.6) we get $\|\mathbb{S}_{\tilde{h}} \bar{\mathbf{z}}_{\tilde{h}} - \mathbb{S} \tilde{\mathbf{z}}\|_{\text{curl}, \Omega} \xrightarrow{\tilde{h}} 0$. In fact,

$$\|\mathbb{S}_{\tilde{h}} \bar{\mathbf{z}}_{\tilde{h}} - \mathbb{S} \tilde{\mathbf{z}}\|_{\text{curl}, \Omega} \lesssim \text{dist}(\mathbb{S} \tilde{\mathbf{z}}, \mathcal{N}_k^h) + \|\bar{\mathbf{z}}_{\tilde{h}} - \tilde{\mathbf{z}}\|_{H_{\perp}^{-\frac{1}{2}}(\text{curl}\Gamma; \Gamma)} \xrightarrow{\tilde{h}} 0.$$

Now, we will show convergence of the solutions for the discrete adjoint problem, cf. (2.5.18). In order to do that, let us consider $\tilde{\mathbf{w}} \in H_0(\mathbf{curl}; \Omega)$ to be the unique solution of

$$a^*(\tilde{\mathbf{w}}, \mathbf{v}) = (\mathbb{S} \tilde{\mathbf{z}} - \mathbf{u}_d, \mathbf{v}) \quad \forall \mathbf{v} \in H_0(\mathbf{curl}; \Omega),$$

and functionals ℓ and ℓ_h given by

$$\begin{aligned} \ell(\mathbf{v}) &= (\mathbb{S} \tilde{\mathbf{z}} - \mathbf{u}_d, \mathbf{v})_{0,\Omega} \quad \forall \mathbf{v} \in H_0(\mathbf{curl}; \Omega), \\ \ell_h(\mathbf{v}_h) &= (\mathbb{S}_h \bar{\mathbf{z}}_h - \mathbf{u}_d, \mathbf{v}_h)_{0,\Omega} \quad \forall \mathbf{v}_h \in \mathcal{N}_{k0}^h. \end{aligned}$$

Thus, from Strang's first lemma we obtain that

$$\begin{aligned} \|\tilde{\mathbf{w}} - \bar{\mathbf{w}}_{\tilde{h}}\|_{\text{curl},\Omega} &\lesssim \inf_{\mathbf{v}_{\tilde{h}} \in \mathcal{N}_{k_0}^{\tilde{h}}} \|\tilde{\mathbf{w}} - \mathbf{v}_{\tilde{h}}\|_{\text{curl},\Omega} + \sup_{\mathbf{v}_{\tilde{h}} \in \mathcal{N}_{k_0}^{\tilde{h}}} \left| \frac{\ell(\mathbf{v}_{\tilde{h}}) - \ell_h(\mathbf{v}_{\tilde{h}})}{\|\mathbf{v}_{\tilde{h}}\|_{\text{curl},\Omega}} \right| \\ &\lesssim \inf_{\mathbf{v}_{\tilde{h}} \in \mathcal{N}_{k_0}^{\tilde{h}}} \|\tilde{\mathbf{w}} - \mathbf{v}_{\tilde{h}}\|_{\text{curl},\Omega} + \|\mathbb{S}\tilde{\mathbf{z}} - \mathbb{S}_{\tilde{h}}\bar{\mathbf{z}}_{\tilde{h}}\|_{\text{curl},\Omega} \xrightarrow{\tilde{h}} 0. \end{aligned}$$

On the other hand, it is clear that for each $\mathbf{p} \in Z$ we have that

$$\alpha(\text{curl}_{\Gamma}\bar{\mathbf{z}}_h, \text{curl}_{\Gamma}\mathbf{p})_{0,\Gamma} + \beta(\bar{\mathbf{z}}_h, \mathbf{p})_{0,\Gamma} \xrightarrow{\tilde{h}} \alpha(\text{curl}_{\Gamma}\tilde{\mathbf{z}}, \text{curl}_{\Gamma}\mathbf{p})_{0,\Gamma} + \beta(\tilde{\mathbf{z}}, \mathbf{p})_{0,\Gamma}.$$

In turn, $\{Z^h\}_h$ is dense in $H(\text{curl}_{\Gamma}; \Gamma)$; moreover, $\{Z^h\}_h$ is dense in $H_{\perp}^{-\frac{1}{2}}(\text{curl}_{\Gamma}; \Gamma)$ which follows from [36, Corollary 5]. Thus, for a given $\mathbf{p} \in Z$ we define \mathbf{p}_h as its best approximation in Z^h , then

$$\begin{aligned} \left\langle \mathbb{S}_{\tilde{h}}^*(\mathbb{S}_{\tilde{h}}\bar{\mathbf{z}}_{\tilde{h}} - \mathbf{u}_d), \mathbf{p}_{\tilde{h}} \right\rangle_{\Gamma^*} &= (\mathbb{S}_{\tilde{h}}\bar{\mathbf{z}}_{\tilde{h}} - \mathbf{u}_d, \mathbb{S}_{\tilde{h}}\mathbf{p}_{\tilde{h}})_{0,\Omega} \\ &= (\mathbb{S}_{\tilde{h}}\bar{\mathbf{z}}_{\tilde{h}} - \mathbf{u}_d, \mathbb{S}_{\tilde{h}}\mathbf{p}_{\tilde{h}} - \mathbb{S}\mathbf{p})_{0,\Omega} + (\mathbb{S}_{\tilde{h}}\bar{\mathbf{z}}_{\tilde{h}} - \mathbf{u}_d, \mathbb{S}\mathbf{p})_{0,\Omega}. \end{aligned}$$

Thus, because $\mathbb{S}_h\bar{\mathbf{z}}_h$ is bounded (independently of h) we conclude

$$\left\langle \mathbb{S}_{\tilde{h}}^*(\mathbb{S}_{\tilde{h}}\bar{\mathbf{z}}_{\tilde{h}} - \mathbf{u}_d), \mathbf{p}_{\tilde{h}} \right\rangle_{\Gamma^*} \xrightarrow{\tilde{h}} \left\langle \mathbb{S}^*(\mathbb{S}\tilde{\mathbf{z}} - \mathbf{u}_d), \mathbf{p} \right\rangle_{\Gamma^*}.$$

Finally,

$$\text{Re} \left\{ \left\langle \mathbb{S}^*(\mathbb{S}\tilde{\mathbf{z}} - \mathbf{u}_d), \mathbf{p} \right\rangle_{\Gamma^*} + \alpha(\text{curl}_{\Gamma}\tilde{\mathbf{z}}, \text{curl}_{\Gamma}\mathbf{p})_{0,\Gamma} + \beta(\tilde{\mathbf{z}}, \mathbf{p})_{0,\Gamma} \right\} = 0 \quad \forall \mathbf{p} \in Z,$$

and by uniqueness of local minimum we conclude $(\tilde{\mathbf{z}}, \mathbb{S}\tilde{\mathbf{z}}, \tilde{\mathbf{w}}) = (\bar{\mathbf{z}}, \bar{\mathbf{u}}, \bar{\mathbf{w}})$. \square

2.5.6 Using $\|\text{curl}_{\Gamma}\mathbf{z}\|_{0,\Gamma}^2$ as the regularization term

In this section we study the problem

$$\min_{\mathbf{z} \in Z^2} \tilde{j}(\mathbf{z}) = \min_{\mathbf{z} \in Z} \frac{1}{2} \int_{\Omega} |\mathbb{S}_{\Gamma}\mathbf{z} - \hat{\mathbf{u}}_d|^2 d\mathbf{x} + \frac{\alpha}{2} \int_{\Gamma} |\text{curl}_{\Gamma}\mathbf{z}|^2 dS, \quad (2.5.23)$$

where

$$Z^2 := \{\mathbf{z} \in H(\text{curl}_{\Gamma}; \Gamma) : \text{curl}_{\Gamma}\mathbf{z} \in L_0^2(\Gamma)\} \cap \ker(\text{curl}_{\Gamma})^{\perp}. \quad (2.5.24)$$

The continuous analysis for this problem follows from the previous case, given that

$$\|\mathbf{z}\|_{0,\Gamma}^2 + \|\operatorname{curl}_\Gamma \mathbf{z}\|_{0,\Gamma}^2 \lesssim \|\operatorname{curl}_\Gamma \mathbf{z}\|_{0,\Gamma}^2 \quad \forall \mathbf{z} \in Z^2, \quad (2.5.25)$$

which follows from the open mapping theorem [31, Corollary 2.7], and the following result

Lemma 2.5.7. *The operators*

$$\operatorname{div}_\Gamma : L_t^2(\Gamma) \mapsto H^{-1}(\Gamma)/\mathbb{R}, \text{ and } \operatorname{curl}_\Gamma : L_t^2(\Gamma) \mapsto H^{-1}(\Gamma)/\mathbb{R}$$

are linear, continuous and surjective. Moreover,

$$\ker(\operatorname{div}_\Gamma) = \{\mathbf{curl}_\Gamma \varphi : \varphi \in H^1(\Gamma)\},$$

$$\ker(\operatorname{curl}_\Gamma) = \{\nabla_\Gamma \varphi : \varphi \in H^1(\Gamma)\}.$$

Proof. See Definition 2.3, Remark 3.2 and Proposition 3.1 in [34]. □

In turn, for the discrete case we have

Proposition 2.5.8. *The set of discrete admissible controls, Z^h , satisfies $Z^h \subset Z^2$.*

Proof. From Lemma 2.5.3 it is clear that

$$Z^h \subset \{\mathbf{z} \in H(\operatorname{curl}_\Gamma; \Gamma) : \operatorname{curl}_\Gamma \mathbf{z} \in L_0^2(\Gamma)\}.$$

On the other hand, we have the following Hodge decomposition for $L_t^2(\Gamma)$, cf. [34, Thm. 3.4],

$$L_t^2(\Gamma) = \nabla_\Gamma H^1(\Gamma) \oplus \mathbf{curl}_\Gamma H^1(\Gamma).$$

Thanks to Lemma 2.5.7, to conclude the proof we only need to show

$$\operatorname{div}_\Gamma Z^h = \{0\}.$$

In order to do that, let $\mathbf{r}(\mathbf{x})$ be the position vector associated with the point \mathbf{x} , which satisfies $\operatorname{curl} \mathbf{r} \equiv \theta$. Then, given an edge $\mathbf{e} \in \mathcal{E}_\Gamma^h$ and faces F_+ and F_- in Γ_h such that $\mathbf{e} = \overline{F_-} \cap \overline{F_+}$, cf. Figure 2.1, we have

$$\operatorname{div}_\Gamma \phi_e|_{F_\pm} = \operatorname{curl}_\Gamma \psi_e|_{F_\pm} = \pm \frac{|\mathbf{e}|}{|F_\pm|} \operatorname{curl}_\Gamma ((\mathbf{r} - \mathbf{r}(\mathbf{v}_e^\pm))) = \pm \frac{|\mathbf{e}|}{|F_\pm|} (\operatorname{curl}(\mathbf{r} - \mathbf{r}(\mathbf{v}_e^\pm))) \cdot \mathbf{n}_F = 0,$$

which concludes the proof. □

2.6 Numerical results

In this section we test our codes, developed in MATLAB[®], for dealing with higher order Nédélec spaces with nonhomogeneous Dirichlet boundary conditions. For optimization, we use a complex version of the BFGS algorithm [138].

2.6.1 Code validation for Nédélec elements

To test our code, we consider the problem involving electrodes given in [21, 22] but with a simpler Dirichlet boundary condition. Let us denote \mathbf{H} , \mathbf{E} , and \mathbf{J} as the complex amplitude of the magnetic field, the electric field, and the current density on a bounded domain Ω , Figure 2.4. And let us consider the eddy current problem: find $(\mathbf{H}, \mathbf{E}, \mathbf{J})$ such that

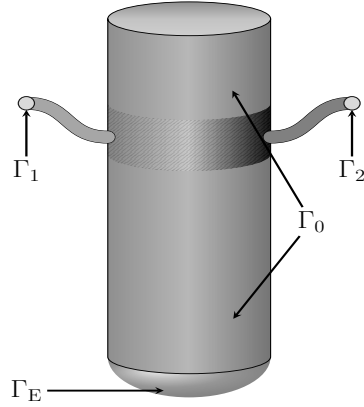


Figure 2.4: Electrode

$$\begin{aligned} \operatorname{curl} \mathbf{H} &= \mathbf{J} & \text{in } \Omega, \\ i\omega\mu\mathbf{H} + \operatorname{curl} \mathbf{E} &= \mathbf{0} & \text{in } \Omega, \\ \mathbf{J} &= \sigma\mathbf{E} & \text{in } \Omega, \end{aligned}$$

subject to:

$$\begin{aligned} \mathbf{E} \times \mathbf{n} &= \mathbf{0} & \text{on } \Gamma_E, & (2.6.1) \\ \int_{\Gamma_n} \mathbf{J} \cdot \mathbf{n} &= \iota_n, & n \in \{1, \dots, N\}, \\ \mathbf{E} \times \mathbf{n} &= \mathbf{0} & \text{on } \bigcup_{k=1}^N \Gamma_k, \\ \mathbf{J} \cdot \mathbf{n} &= 0 & \text{on } \Gamma_0, \\ \mathbf{H} \cdot \mathbf{n} &= 0 & \text{on } \partial\Omega. \end{aligned}$$

In the particular case that Ω is a cylinder of radius R and height L , $N = 1$, and the partition for the boundary Γ is given by

$$\Gamma_1 = \{(x, y, z) : z = L, x^2 + y^2 < R\}, \Gamma_E = \{(x, y, z) : z = 0, x^2 + y^2 < R\}, \Gamma_0 = \Gamma \setminus \overline{(\Gamma_1 \cup \Gamma_E)}.$$

Then, as shown in [21, 22], it is possible to find an analytic solution for (2.6.1) given by

$$\mathbf{H}(x, y, z) = \frac{\iota_1}{2\pi R} \frac{I_1(\gamma r)}{I_1(\gamma R)} \mathbf{e}_\theta, \quad \mathbf{E}(x, y, z) = \frac{\iota_1 \gamma}{2\pi R \sigma} \frac{I_0(\gamma r)}{I_1(\gamma R)} \mathbf{e}_z, \quad \text{and } \mathbf{J} = \sigma \mathbf{E}, \quad (2.6.2)$$

where I_ν is the modified Bessel function of the first kind of order ν , $\gamma := \sqrt{i\omega\mu\sigma}$, $r := \sqrt{x^2 + y^2}$, $\mathbf{e}_\theta := r^{-1}(-y, x, 0)$, and $\mathbf{e}_z := (0, 0, 1)$. Now, we approximate \mathbf{H} with \mathcal{N}_2^h which satisfies $\mathcal{P}_2^3 \subsetneq \mathcal{N}_2^h \subsetneq \mathcal{P}_3^3$. Also, we set all the parameters equal to 1 except for $R = \frac{1}{2}$, and then we consider the problem: find $\mathbf{H}_h \in \mathcal{N}_2^h$ such that

$$a(\mathbf{H}_h, \mathbf{v}_h) = \mathbf{0} \quad \forall \mathbf{v}_h \in \mathcal{N}_2^h \cap H_0(\mathbf{curl}; \Omega),$$

$$\mathbf{H}_h \times \mathbf{n} = \mathbf{H} \times \mathbf{n} \quad \text{on } \Gamma_h.$$

The domain Ω was approximated with a family of meshes generated with Gmsh [67]. Figure 2.5 shows a log-log plot for the error $\mathcal{E}_h := \|\mathbf{H} - \mathbf{H}_h\|_{\mathbf{curl}, \Omega}$, along with the coarsest mesh considered, where the color represents the element-wise error $\|\mathbf{H} - \mathbf{H}_h\|_{\mathbf{curl}, K}$.

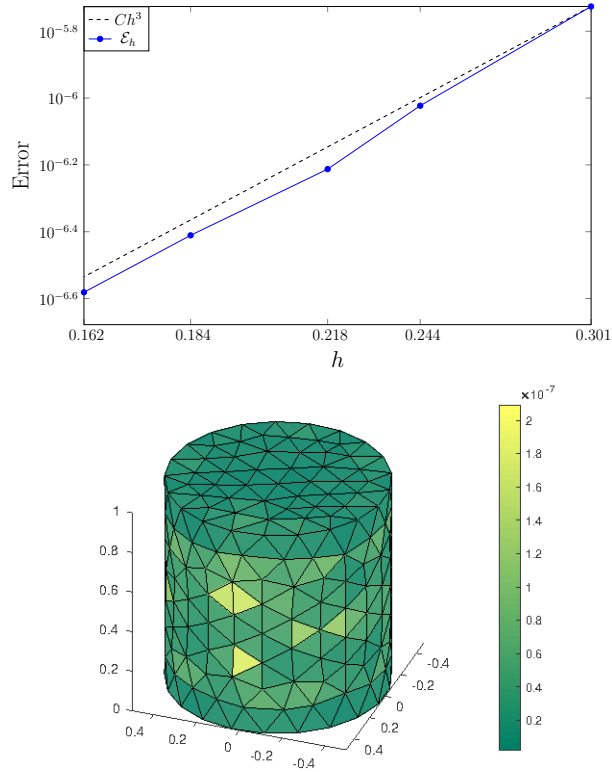


Figure 2.5: Log-log plot \mathcal{E}_h vs h , and coarsest mesh along with element-wise error.

2.6.2 Validation optimization routines

We devote this section to testing our codes related to the minimization problem. We start by testing our equivalent expression for $d^{\mathbb{R}}j_h$, cf. (2.5.8). The main difficulty is the term that involves \mathbb{S}_h^* , which can be computed as in (2.5.19). Figure 2.6 (left) shows the the difference between $d^{\mathbb{R}}j_h(\mathbf{z}; \boldsymbol{\xi})$ and its approximation using a finite difference quotient. Here $\boldsymbol{\xi}$ denotes the random direction. We see a linear rate of convergence until the round-off error kicks in, as expected.

2.6.3 Convergence of optimization problem

It is difficult to devise an exact solution to the optimal control problem in order to explicitly show the application of Theorem 2.5.6. Instead, we show the convergence of the cost functional $\mathcal{J}(\mathbf{u}_h, \mathbf{z}_h)$ to $\mathcal{J}(\mathbf{u}, \mathbf{z})$ as $h \rightarrow 0$. Here (\mathbf{u}, \mathbf{z}) is the optimal control corresponding to a mesh obtained after 6 refinements. The optimization problem is

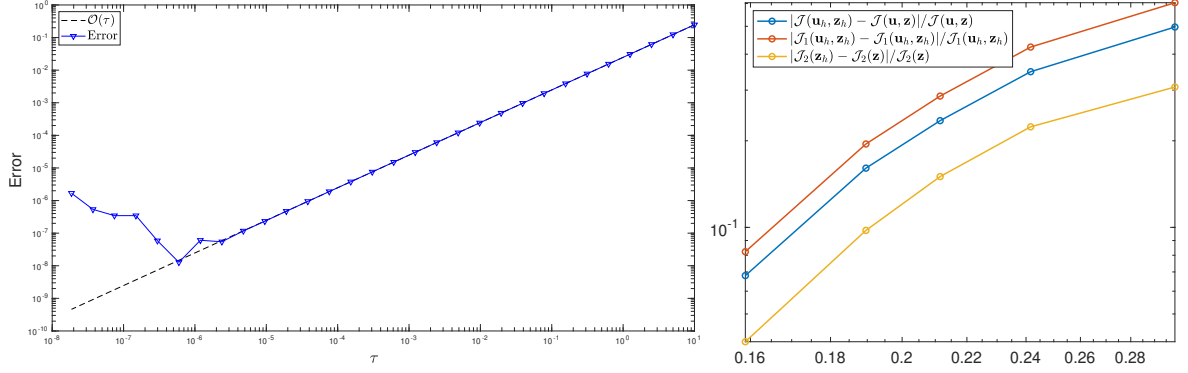


Figure 2.6: **Left:** Given a random direction $\boldsymbol{\xi}$, the panel shows the difference between $d^{\text{R}}j_h(\mathbf{z}; \boldsymbol{\xi})$ and its finite difference approximation. As expected, we observe a linear rate of convergence. **Right:** We let $\alpha = 1e^{-3}$ and $\beta = 0$ in the cost functional $\mathcal{J}(\cdot)$. Let $\mathcal{J}_1(\mathbf{u}, \mathbf{z}) := \frac{1}{2}\|\mathbf{u} - \mathbf{u}_d\|^2$ and $\mathcal{J}_2(\mathbf{z}) := \frac{\alpha}{2}\|\text{curl}_{\Gamma}\mathbf{z}\|_{L^2(\Gamma)}^2$. Moreover, let \mathbf{z} be the optimal control corresponding to the finest mesh. Then the three curves show $|\mathcal{J}(\mathbf{u}_h, \mathbf{z}_h) - \mathcal{J}(\mathbf{u}, \mathbf{z})|/\mathcal{J}(\mathbf{u}, \mathbf{z})$, $|\mathcal{J}_1(\mathbf{u}_h, \mathbf{z}_h) - \mathcal{J}_1(\mathbf{u}, \mathbf{z})|/\mathcal{J}_1(\mathbf{u}, \mathbf{z})$, and $|\mathcal{J}_2(\mathbf{z}_h) - \mathcal{J}_2(\mathbf{z})|/\mathcal{J}_2(\mathbf{z})$ as $h \rightarrow 0$.

solved using the BFGS method mentioned above with a stopping tolerance of 10^{-9} . We let $\mathbf{u}_d = \mathbf{H}$, cf. (2.6.2) and let $\alpha = 10^{-3}$ and $\beta = 0$. Let $\mathcal{J}_1(\mathbf{u}, \mathbf{z}) := \frac{1}{2}\|\mathbf{u} - \mathbf{u}_d\|^2$ and $\mathcal{J}_2(\mathbf{z}) := \frac{\alpha}{2}\|\text{curl}_{\Gamma}\mathbf{z}\|_{L^2(\Gamma)}^2$. Figure 2.6 (right) shows the errors $|\mathcal{J}(\mathbf{u}_h, \mathbf{z}_h) - \mathcal{J}(\mathbf{u}, \mathbf{z})|/\mathcal{J}(\mathbf{u}, \mathbf{z})$, $|\mathcal{J}_1(\mathbf{u}_h, \mathbf{z}_h) - \mathcal{J}_1(\mathbf{u}, \mathbf{z})|/\mathcal{J}_1(\mathbf{u}, \mathbf{z})$, and $|\mathcal{J}_2(\mathbf{z}_h) - \mathcal{J}_2(\mathbf{z})|/\mathcal{J}_2(\mathbf{z})$ as $h \rightarrow 0$. The expected convergence is observed.

Chapter 3

AN OPTIMAL TIME VARIABLE LEARNING FRAMEWORK FOR DEEP NEURAL NETWORKS

3.1 Introduction

Consider a network architecture, for example, the residual neural network (ResNet)

$$y^{[\ell]} = y^{[\ell-1]} + \tau \sigma(y^{[\ell-1]}, \theta^{[\ell-1]}), \quad (3.1.1)$$

which contains a parameter τ and an activation function σ . It is also possible to consider other architectures, such as feed-forward networks, etc. The above neural network can be understood as the time discretization of a non-linear ordinary differential equation (ODE). The feature vector $y^{[\ell]}$ is computed by forward propagation from the previous layers' feature vector $y^{[\ell-1]}$ and network parameters, which are collected in $\theta^{[\ell-1]}$. In most of the existing literature, τ is a given fixed constant. The main novelty of this work lies in

replacing τ by learning variables $\tau^{[\ell]}$

and the treatment of these $\tau^{[\ell]}$. These variables can be understood as the “time step-sizes” in Deep Neural Networks (DNNs). This work considers them as optimization variables, not as hyperparameters. Furthermore, the parameters $\tau^{[\ell]}$ are allowed to differ from layer to layer, i.e., the time grid can be non-uniform. Notice that this τ -variable framework can be applied to any of the existing networks of type (3.1.1). The deep learning optimization problem will now also learn optimal parameters $\tau^{[\ell]}$ in addition to the standard DNN parameters. As will be illustrated throughout the chapter, the presented approach is not just a scaling of the activation function σ by $\tau^{[\ell]}$.

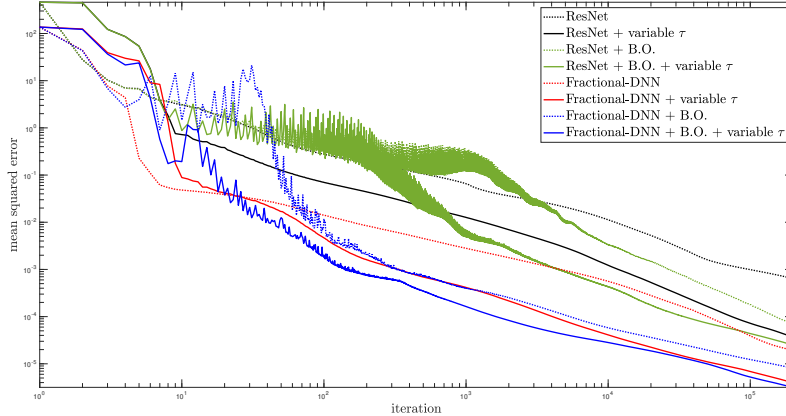


Figure 3.1: The panel shows the mean squared error during training when the variable- τ framework is applied to a ResNet and a Fractional-DNN for an ill-posed 3D-Maxwell’s equation. More details are available in Section 3.7.

This will become evident when applying the proposed framework to Fractional-DNNs, where $\tau^{[\ell]}$ enters in multiple ways; see Remark 3.5.1.

This proposed approach presents multiple advantages, including:

- The proposed framework leads to *optimal adaptive time discretizations of DNNs*, such as ResNets and Fractional-DNNs, tailored to the optimization/learning problem.
- The proposed framework can further help (as is rigorously established) overcome the *vanishing and exploding gradient* problems in networks such as ResNets and Fractional-DNNs. Notice, that motivation behind introducing ResNets [77] was vanishing gradients and Fractional-DNNs was vanishing and exploding gradients [12].
- If $\tau^{[\ell]}$ for any layer ℓ is close to zero, then the associated layer is redundant and can be deleted without sacrificing the accuracy. Thus leading to small yet accurate DNNs.
- Variable $\tau^{[\ell]}$ helps improve the training error decay, see Figure 3.1.

The main idea of approximate parameterized PDEs and solving the inverse problems using Fractional-DNNs has been recently introduced in [10]. The present approach not only introduces the aforementioned variable time step framework, but for the first time, to the best of our knowledge, also applies DNNs, such as ResNets and Fractional-DNNs, to ill-posed problems such as Maxwell’s equations with Gauss’s law. A comparison of

these standard DNNs with their time-step variable versions has also been carried out. The problem is ill-posed in the following sense: Nédélec finite elements are traditionally used to discretize Maxwell’s equations. However, they are curl-conforming, and thus Gauss’s law (divergence condition) cannot be directly imposed [48]. The DNNs are shown to generalize well on unseen data and physical domains.

Deep learning is a nascent field of research with many exciting applications, for example, imaging science [12, 77, 88, 124, 150], biomedical applications [45, 75, 98], satellite imagery, remote sensing [24, 140, 157], segmentation [126], and gaming [136]. Recently, this topic is starting to receive significant attention from mathematicians [55, 61]. Especially, the fact that DNNs of type (3.1.1) can be viewed as optimization problems constrained by discrete dynamical systems [12, 20, 43, 73, 74, 107, 134] and partial differential equations [106, 129]. In these settings, the state-of-the-art is to consider a uniform time grid, where the time step-size τ is chosen before running the optimization algorithm to identify weights, i.e. τ is a hyperparameter.

The articles [76, 87] suggest to consider $\tau^{[\ell]}$ as hyperparameters for physics-informed neural networks and ResNets, respectively. In both cases, the hyperparameters $\tau^{[\ell]}$ are seen as scalings applied, outside [76] and inside [87], the activation function. In [87], a global $\tau = \tau^{[\ell]}$ for all ℓ is considered leading to training error improvement and more accurate solutions, which will also hold true for our more general setting. In [76], a sequence of $\tau^{[\ell]}$, fulfilling a probabilistic condition, is chosen to avoid exploding gradients. In contrast, we let the optimization algorithm learn $\tau^{[\ell]}$ and provides deterministic arguments to overcome the vanishing and exploding gradient problems.

Outline: This Chapter is organized as follows: Section 3.2 introduces some basic preliminary results and notation. We introduce definitions of fractional derivatives and state a generic DNN. We also describe an extension of this DNN to include a recently introduced bias ordering idea from [6], which offers multiple advantages such as narrowing the parameter search space. This is followed by Section 3.3, where the relation between continuous DNNs and dynamical systems is considered. Special attention is

given to two continuous versions of DNNs: ResNet (DNN with a standard time derivative) and Fractional-DNN (DNN with a fractional time derivative). Stability results for these two DNNs are also provided. Notice that Fractional-DNNs have been recently introduced in [12] for classification and further extended in [10] to inverse problems with PDEs. They have multiple advantages over ResNets as they can incorporate memory into the network due to the nonlocal nature of fractional derivatives, and due to the low regularity requirements of fractional derivatives, they can be applied to non-smooth functions. As stated above, one of the main motivations behind introducing Fractional-DNN was to overcome vanishing and exploding gradients. The article [12] provides numerical evidence of overcoming the vanishing gradient problem.

Next, in Section 3.4 we state selected DNN architectures and compare them for a fixed τ . The new framework with variable $\tau^{[\ell]}$ is applied to the architectures of Section 3.4 in Section 3.5. Notice that our approach is broad and can be applied to any network architecture of type (3.1.1), and it is independent of the choice of the loss functional. Clearly, the proposed approach inherits all the positive aspects of these existing networks. Additionally, the new framework is rigorously shown to overcome vanishing and exploding gradient issues (cf. Section 3.6).

Finally, in Section 3.7, we illustrate the efficacy of our approach with the help of an ill-posed 3D-Maxwell's equation. The numerical examples validate the above-mentioned advantages of the proposed framework. In particular, stability and network reduction.

3.2 Preliminaries

The goal of this section is to introduce the relevant notation and abstract optimization problems arising while training the DNNs. The content of this section is well known [12, 10, 6].

Symbol	Description
$L \in \mathbb{N}$	Number of network layers (i.e. network depth)
$N \in \mathbb{N}$	Number of distinct data samples
n_ℓ	Number of nodes in layer ℓ
$y^{[\ell]} \in \mathbb{R}^{n_\ell}$	Feature vector in layer ℓ
σ	Activation function
$W^{[\ell]} \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$	Weights in layer ℓ
$b^{[\ell]} \in \mathbb{R}^{n_{\ell+1}}$	Biases in layer ℓ
$\tau^{[\ell]} \in \mathbb{R}$	Time step-size in layer ℓ
$\theta^{[\ell]} \in \mathbb{R}^{n_{\ell+1}(n_\ell+1)+1}$	Vector of all variables (weights, biases and time step-size) in layer ℓ
P_j^ℓ	Projection matrix from layer j onto layer ℓ
ϕ	Adjoint variables
\mathcal{F}	Network represented as a function
f_ℓ	Layer function
J	Loss function
$\lambda_1, \lambda_2, \beta \in \mathbb{R}$	Regularization parameters
\mathcal{L}	Lagrangian
$\{u, S(u)\}$	Input / Output pair of training data
$\Gamma(\cdot)$	Euler's Gamma function

Table 3.1: Notation

3.2.1 Caputo fractional derivative

In preparation for the Fractional-DNN architecture, we next introduce the left and right Caputo fractional derivatives for absolutely continuous functions and refer to [11, Definitions 2.1 and 2.4, and Proposition 2.3] and [92, (2.4.17) and (2.4.18)] for details.

Definition 3.2.1. (*Left Caputo Fractional Derivative*) Let $y \in W^{1,1}([0, T]; X)$, with X denoting a Banach space. The left Caputo fractional derivative of order $\gamma \in (0, 1)$ is given by

$$\partial_t^\gamma y(t) = c_\gamma \int_0^t \frac{y'(r)}{(t-r)^\gamma} dr, \quad (3.2.1)$$

where $c_\gamma := \frac{1}{\Gamma(1-\gamma)}$ and $\Gamma(\cdot)$ is Euler's Gamma function.

Definition 3.2.2. (*Right Caputo Fractional Derivative*) Let $y \in W^{1,1}([0, T]; X)$, with X denoting a Banach space. The right Caputo fractional derivative of order $\gamma \in (0, 1)$

is given by

$$\partial_{T-t}^\gamma y(t) = -c_\gamma \int_t^T \frac{y'(r)}{(r-t)^\gamma} dr.$$

Next, we introduce the general deep learning problem as an optimization problem with DNN constraints.

3.2.2 Deep Learning problem

Consider a neural network architecture with an input layer of dimension n_0 , $L-1$ hidden layers of dimension n_ℓ for $\ell = 1, \dots, L-1$ and an output layer of dimension n_L . Then we can represent this network as a function

$$\mathcal{F} = f_{L-1} \circ f_{L-2} \circ \dots \circ f_0, \quad (3.2.2)$$

where $\{f_\ell\}_{\ell=0}^{L-1}$ are the layer functions. These layer functions are parameterized by weight matrices $W^{[\ell]} \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$ and bias vectors $b^{[\ell]} \in \mathbb{R}^{n_{\ell+1}}$. The definition of f_ℓ depends on the network architecture. We will introduce different options in Section 3.4.

The weights and biases are identified during a training process that requires solving an optimization problem. Let $\{u^{(i)}, S(u^{(i)})\}_{i=1}^N$ (input/output pairs) denote the training data. Then the goal is to match the output of the DNN with the data points $S(u^{(i)})$. This is accomplished by minimizing a loss functional J and the resulting optimization problem is given by:

$$\begin{aligned} & \min_{\{W^{[\ell]}\}_{\ell=0}^{L-1}, \{b^{[\ell]}\}_{\ell=0}^{L-2}} J \left(\{(y^{[L](i)}, S(u^{(i)}))\}_i, \{W^{[\ell]}\}_\ell, \{b^{[\ell]}\}_\ell \right) \\ & \text{subject to} \quad y^{[L](i)} = \mathcal{F} \left(u^{(i)}; (\{W^{[\ell]}\}_\ell, \{b^{[\ell]}\}_\ell) \right) \quad i = 1, \dots, N. \end{aligned} \quad (3.2.3)$$

One standard choice for the loss function is the mean squared error:

$$J := \frac{1}{2N} \sum_{i=1}^N \|y^{[L](i)} - S(u^{(i)})\|_2^2.$$

Another example is the Cross-entropy loss, see [71] for more examples. The choice of J is dictated by the application. For our study, it is not relevant which of these options is chosen since our focus is on DNNs, represented by the operator \mathcal{F} .

It is also common to add regularization, for example, ℓ^1 and ℓ^2 regularizations on weights and biases

$$J_{\lambda_1} = J + \frac{\lambda_1}{2} \sum_{\ell=0}^{L-1} \left(\|W^{[\ell]}\|_2^2 + \|W^{[\ell]}\|_1 \right) + \frac{\lambda_1}{2} \sum_{\ell=0}^{L-2} \left(\|b^{[\ell]}\|_2^2 + \|b^{[\ell]}\|_1 \right), \quad (3.2.4)$$

where $\lambda_1 > 0$ is the regularization parameter.

Before we continue, we emphasize that recently, the article [6] has extended the above generic network (3.2.3) by incorporating ordering among the bias vector components in each layer. The main idea is that in each layer ℓ , with $\ell = 0, \dots, L-2$, one enforces

$$b_j^{[\ell]} \leq b_{j+1}^{[\ell]}, \quad j = 1, \dots, n_{\ell+1} - 1, \quad (3.2.5)$$

where the subscript j indicates the bias vector component. This approach offers multiple advantages, as highlighted in [6]. Our numerical examples further provides a comparison between with and without bias ordering framework. Notice that the bias ordering is implemented using a penalty framework; see [6] for details.

Next, we provide a mathematical background behind learning the time step-sizes $\tau^{[\ell]}$. To start, we discuss a link between some DNNs and dynamical systems.

3.3 Continuous DNNs

In this section, we study the continuous structure of multiple DNNs; cf. (3.1.1) and (3.2.2). This section mainly focuses on the stability of these architectures, which will follow from a connection with dynamical systems. For other approaches, we refer to [15, 76] and the references therein. Since we are primarily interested in stability results, we start with a basic remark on a (finite) network with a Lipschitz activation function, like ReLU. The finite composition of Lipschitz functions is also a Lipschitz function and therefore differentiable almost everywhere by Rademacher's theorem. In the following, we show some historical connections between neural networks and dynamical systems, and later on we consider continuous Fractional-DNNs which enable memory into the DNNs.

3.3.1 Ordinary Differential Equations and Neural Networks

The relation between Neural Networks (NNs) and differential equations is not new. In fact, in the late '80s, in [121] the following model was considered for the activity of j -th neuron:

$$\frac{dy_j}{dt} = -\alpha y_j + \beta \sigma \left(\sum_k w_{jk} y_k \right) + b_j, \quad (3.3.1)$$

where α and β are (given) positive constants, the weights $\{w_{jk}\}$ represent the connection strength between the k -th and j -th neurons, and b_j represents a bias. Note that most modern neural network architectures are related to stationary solutions of the ODE above. In the present work, we restrict our focus to a different connection with dynamical systems (cf. (3.1.1)), which is more recent, mainly because it has been tested more thoroughly. Also, because we are interested in optimal control, we will primarily focus on the ideas presented in [12], [47], and [129]. Nevertheless, for completeness, we also mention some other related works: [60] and [137] from a dynamical systems point of view, [142] for universal maps with memory, [103] for problems in the frequency domain, [65] for a Runge-Kutta based NN, PINNs [39] for PDE-related problems, and SINDy [89] for data-driven model discovery. Also, when ReLU is considered as the activation function, a DNN is a high-dimensional, piecewise linear function. Therefore, some techniques from the Finite Element and the Monte Carlo methods can be used for its analysis, cf. [82].

Motivated by ResNets [77], the authors in [47] relate DNNs of type (3.1.1) to a recurrence relation obtained when numerically solving a system of ODEs. For instance, for given $f : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ and $y_0 \in \mathbb{R}$, consider the problem: Find a function y , such that:

$$\begin{aligned} y'(t) &= f(t, y(t)), & \text{in } (0, T), \\ y(0) &= y_0. \end{aligned} \quad (3.3.2)$$

A solution for this system can be approximated, under mild assumptions on y and f , by the Euler method:

$$y(t + \tau) - y(t) = \int_t^{t+\tau} y'(s)ds = \int_t^{t+\tau} f(s, y(s))ds \approx \tau \cdot f(t, y(t)),$$

where we have (formally) used the fundamental theorem of calculus and a left Riemann-sum approximation. Namely, we can understand the layers of a DNN as samples from a continuous system that evolves from the input to the output. Here, the first and last layers are special cases due to common upsampling and downsampling techniques.

It is worth mentioning, that in [47, B.2] the authors consider $f = f(y(t), t, \theta)$, where θ represents the parameters of the DNN. Namely, θ is independent of t and therefore their ODE system is limited (essentially) to autonomous systems. Thus, DNNs generated with the method given in [47] are smooth by construction. This property allows one to use well-known results in the theory of dynamical systems, control theory, adaptive ODE solvers, among others. An interesting application where smooth trajectories are desired is when self-intersecting trajectories or surfaces are not allowed, as in shape optimization (manifold surfaces), cf. [125, 151]. It is clear that the additional smoothness also limits the usability of the model [59]. Obviously, the architecture of a neural network must match its purpose, i.e., the given data and desired application case.

Besides the networks of type (3.3.2), the present work also focuses on problems where the underlying system depends on its history in a nonlocal way. The latter is most commonly found in systems with *hysteresis* or delayed effects. Note that the derivative in (3.3.2) is a local operator; this follows from its pointwise limit definition. Therefore, based on [10, 11, 12], we consider a fractional derivative based approach. As pointed out in [10, 12], this serves two main purposes: it acts as a global operator (memory effect), and the order of a differential equation is allowed to be less than 1, which reduces the smoothness of the system.

3.3.2 Stability of continuous Fractional-DNN

As mentioned before, ResNet-like architectures can be connected to a classical ODE system, and therefore we can apply the well-known theories to analyze the properties of the DNN [47, 129]. A commonly desired property is continuous dependence on the data. In the context of machine learning, this means that input variables that are “close” should produce outputs of the DNN that are also “close”. Of course, in the context of real-life applications, the notion of distance is not always known, nor is the right dimension or the smoothness/regularity of the system underneath. Following [10, 12], we consider a DNN architecture that can be related to a different notion of derivative, the so-called fractional derivative, see Section 3.2.1, and here we show a stability result for this notion of derivative with respect to the initial data. In order to do so, we consider Ω to be an open, bounded and connected subset of \mathbb{R}^d , define $E := (L^2(\Omega), \|\cdot\|_\Omega)$, and let $f : \text{Dom}(f) \subseteq [0, \infty) \times E \mapsto E$. To establish the stability of continuous Fractional-DNN, we consider a dynamical system for y . Notice that similar structure holds for the continuous Fractional-DNN (cf. (3.5.6))

$$\partial_t^\gamma y = f(t, y), \quad \text{with} \quad y(0) = y_0, \quad (3.3.3)$$

where $y_0 \in E$, and f satisfies the standard assumptions:

$$\left\{ \begin{array}{l} \text{There exist positive constants } T \text{ and } r \text{ such that } f \text{ restricted to } [0, T] \times B_r(y_0) \\ \text{is continuous, bounded and Lipschitz with respect to the second argument.} \end{array} \right. \quad (\text{H})$$

The last hypothesis implies there exists $L > 0$ such that

$$\|f(t, y_1) - f(t, y_2)\|_\Omega \leq L\|y_1 - y_2\|_\Omega \quad \forall t \in [0, T], \text{ and } \forall y_1, y_2 \in B_r(y_0).$$

Let us remark that E can be replaced by any other space with the Radon-Nikodym property, but based on the most common loss functions, we restrict the analysis to $L^2(\Omega)$. From [11] we have the following result connecting the strong and generalized Caputo derivatives

Lemma 3.3.1. *Let $\gamma \in (0, 1)$, and $T > 0$. If $y \in W^{1,1}((0, T); E)$ then the following equality holds in the $L^1((0, T); E)$ -sense*

$$\partial_t^\gamma y(t) = D_t^\gamma(y - y(0))(t), \quad (3.3.4)$$

for a.e. $t \in (0, T]$, where D_t^γ denotes the Left Riemann-Liouville fractional derivative, cf. [11, Definition 2.2].

Proof. The proof follows from [11, Proposition 2.3] and the fact that every reflexive Banach space has the Radon-Nikodym property. \square

We write the Left Caputo derivative in the generalized Caputo derivative form (3.3.4), because several of the well-known results, which hold for standard ODEs, also have their counterparts in the generalized Caputo derivative setting. For instance, the solution operator for a non-autonomous fractional ODE can be represented in terms of a Volterra integral; cf. [57, Theorem 2.1]. By using this integral representation, the next proposition shows the stability of the fractional ODE (3.3.3) with respect to its initial value when the solution is smooth enough.

Proposition 3.3.2. *Given $y_0 \in E$, and f that satisfies (H). If $y_\alpha, y_\beta \in W^{1,1}((0, T); E)$ solve (3.3.3) with initial conditions $y_{\alpha,0}$ and $y_{\beta,0}$ both in $B_r(y_0)$. Then,*

$$\|y_\alpha - y_\beta\|_{L^1(0,T;E)} \leq C \|y_{\alpha,0} - y_{\beta,0}\|_\Omega, \quad (3.3.5)$$

where $C = C(\gamma, T, L) > 0$.

Proof. By Lemma 3.3.1, and because E is reflexive and therefore has the Radon-Nikodym property, we can recast (3.3.3) as (3.3.4), with $y(0) \in \{y_{\alpha,0}, y_{\beta,0}\}$, and represent each solution in terms of a nonlinear Volterra integral, cf. [57, Lemma 2.1]. Namely, if y_α, y_β represent the solutions for (3.3.3) with corresponding initial conditions $y_{\alpha,0}, y_{\beta,0}$, then for $t \in [0, T]$, and a.e. $x \in \Omega$

$$\begin{aligned} y_\alpha(x, t) &= y_{\alpha,0}(x) + c_\gamma \int_0^t \frac{1}{(t-\tau)^\gamma} f(\tau, y_\alpha(x, \tau)) d\tau, \\ y_\beta(x, t) &= y_{\beta,0}(x) + c_\gamma \int_0^t \frac{1}{(t-\tau)^\gamma} f(\tau, y_\beta(x, \tau)) d\tau, \end{aligned}$$

where we recall $c_\gamma = \frac{1}{\Gamma(1-\gamma)}$. Then,

$$\begin{aligned} \|y_\alpha(\cdot, t) - y_\beta(\cdot, t)\|_\Omega &\leq \|y_{\alpha,0} - y_{\beta,0}\|_\Omega + c_\gamma \int_0^t \frac{1}{(t-\tau)^\gamma} \|f(\tau, y_\alpha) - f(\tau, y_\beta)\|_\Omega d\tau \\ &\leq \|y_{\alpha,0} - y_{\beta,0}\|_\Omega + c_\gamma \int_0^t \frac{1}{(t-\tau)^\gamma} L \|y_\alpha(\cdot, \tau) - y_\beta(\cdot, \tau)\|_\Omega d\tau \end{aligned}$$

Finally, from Gronwall's inequality in its integral form

$$\begin{aligned} \|y_\alpha(\cdot, t) - y_\beta(\cdot, t)\|_\Omega &\leq \|y_{\alpha,0} - y_{\beta,0}\|_\Omega \exp\left(\frac{L}{\Gamma(1-\gamma)} \int_0^t \frac{1}{(t-\tau)^\gamma} d\tau\right) \\ &= \|y_{\alpha,0} - y_{\beta,0}\|_\Omega \exp\left(\frac{L}{\Gamma(1-\gamma)} \frac{t^{1-\gamma}}{1-\gamma}\right), \end{aligned}$$

and integrating over $(0, T)$ concludes the proof. \square

Remark 3.3.3. *It is important to point out that the previous results assume the regularity $W^{1,1}((0, T); E)$, but for most problems in Machine Learning the regularity of solutions is still an open question. Even at the “discrete level”, the regularity depends on the data, DNN architecture, optimization algorithm, loss function, among others factors. Another difficulty is that DNNs can have a different number of neurons in each layer, i.e., the space E can change over time.*

Finally, and as mentioned before, a DNN with Lipschitz activation functions defines a locally Lipschitzian operator. Later in Section 3.6, we will explore how the activation function and weights affect locally the gradient of a DNN and therefore the Lipschitz constant, and we will study the vanishing and exploding gradients problem of various DNNs under the variable- τ framework that is introduced in Section 3.5.

3.4 Network architectures with fixed τ -parameter

Let us begin by stressing that, in general, the proposed framework with variable τ can be applied to any DNN. We will illustrate our ideas using three representative DNNs. Subsequently, we will describe their strengths and weaknesses.

The first network architecture is ResNet [77]. As described in the previous section, see (3.3.2), this network arises after adding an identity map to a standard

feedforward network. This leads to connectivity between the adjacent layers. In order to connect all layers and additionally be able to approximate non-smooth functions, we refer to DenseNet [84] and Fractional-DNN [12]. We remark that there also exist other approaches that attempt to induce multilayer connections, e.g., Highway Net [139], AdaNet [53], ResNetPlus [46], etc.

DenseNet is an ad-hoc method that uses the feature maps of all preceding layers as inputs into all subsequent layers. Meanwhile, Fractional-DNN can be viewed as a time-discretization of a fractional in time non-linear ODE of type (3.3.3), connecting all layers in a mathematically rigorous manner. Both approaches, DenseNet and Fractional-DNN, improve the vanishing gradient effect issue due to the memory effect they incorporate. Furthermore, Fractional-DNN allows approximation of non-smooth functions and thus can also potentially help with exploding gradients.

We recall the **ResNet** [77] with uniform time-steps τ , cf. (3.1.1). The feature vector $y^{[\ell]} \in \mathbb{R}^{n_\ell}$ in layer $\ell = 1, \dots, L$ is computed by forward propagation in the following way

$$y^{[\ell]} = P_{\ell-1}^\ell y^{[\ell-1]} + \tau \sigma \left(W^{[\ell-1]} y^{[\ell-1]} + b^{[\ell-1]} \right), \quad \ell = 1, \dots, L,$$

where $P_0^1 = 0 \in \mathbb{R}^{n_0 \times n_1}$ and $y^{[0]} = u$. Here, σ is a nonlinear activation function; for instance, ReLU [71], τ is the fixed time-step length, and u is the input data. Notice that, if all the layers are the same size, then $P_{\ell-1}^\ell$ equals an identity matrix. In general, $P_{\ell-1}^\ell$ will allow layers to have different sizes, i.e.,

$$\dim(P_{\ell-1}^\ell y^{[\ell-1]}) = \dim(y^{[\ell]}).$$

While ResNet provides connectivity between adjacent layers, we are also interested in fully connected networks, such as DenseNet [84].

In a **DenseNet**, the connection through all layers is achieved by the following forward propagation:

$$y^{[\ell]} = \sum_{i=0}^{\ell-1} P_i^\ell y^{[i]} + \sigma \left(W^{[\ell-1]} y^{[\ell-1]} + b^{[\ell-1]} \right), \quad \ell = 1, \dots, L,$$

where $y^{[0]} = u$ is the input data.

Notice that this method does not contain a time step-size parameter. It is possible to artificially add the parameter τ before the activation function σ . The connection of the resulting expression to a dynamical system remains unclear. Instead of DenseNet, we focus on Fractional-DNN. In addition to connecting all layers, the Fractional-DNN can be understood as a time discretization of a dynamical system of type (3.3.3). Hence, learning the time step sizes is a meaningful task in this setup.

We recall the **Fractional-DNN**, with uniform time-steps τ , from [12, 10]. It corresponds to a time discretization of a system of type (3.3.3). The forward propagation for the Fractional-DNN is given by

$$y^{[\ell]} = P_{\ell-1}^{\ell} y^{[\ell-1]} - \sum_{j=1}^{\ell-1} a_{\ell-j} (P_j^{\ell} y^{[j]} - P_{j-1}^{\ell} y^{[j-1]}) + \tau^{\gamma} \Gamma(2 - \gamma) \sigma \left(W^{[\ell-1]} y^{[\ell-1]} + b^{[\ell-1]} \right),$$

where $\ell = 1, \dots, L$, $y^{[0]} = u$, and P_j^{ℓ} as before. Moreover

$$a_{\ell-j} := (\ell + 1 - j)^{1-\gamma} - (\ell - j)^{1-\gamma}.$$

Remark 3.4.1. *In Section 3.5.2 below, we will consider a Fractional-DNN with variable τ , i.e., $\tau^{[\ell]}$ for $\ell = 1, \dots, L - 1$. In this case, the coefficients $a_{\ell-j}$ will depend on $\tau^{[j]}, \dots, \tau^{[\ell]}$.*

We conclude this section by emphasizing that depending on the number of DNN outputs, the last layer may have a different size, which can be captured via

$$y^{[L]} = W^{[L-1]} y^{[L-1]}.$$

For the remainder of the Chapter, we will assume such a setup for the last layer.

3.5 Variable- τ framework for DNNs

Instead of a fixed τ , we propose to use a different $\tau^{[\ell]}$ for each layer, which is learned during the training process. This allows us to optimize the “time step-sizes” $\tau^{[\ell]}$, resulting in what can be viewed as an adaptive time discretization of the ODE

tailored to the optimization (learning) problem. The resulting optimization problem is given by (cf. 3.2.3)

$$\begin{aligned} & \min_{\{W^{[\ell]}\}_{\ell=0}^{L-1}, \{b^{[\ell]}\}_{\ell=0}^{L-2}, \{\tau^{[\ell]}\}_{\ell=0}^{L-2}} J_\lambda \left(\{(y^{[L](i)}, S(u^{(i)}))\}_i, \{W^{[\ell]}\}_\ell, \{b^{[\ell]}\}_\ell, \{\tau^{[\ell]}\}_\ell \right) \\ & \text{subject to } y^{[L](i)} = \mathcal{F} \left(u^{(i)}; (\{W^{[\ell]}\}_\ell, \{b^{[\ell]}\}_\ell, \{\tau^{[\ell]}\}_\ell) \right) \quad i = 1, \dots, N. \end{aligned} \quad (3.5.1)$$

Constraints on $\tau^{[\ell]}$, for instance, non-negativity can be easily incorporated. Recall from (3.2.4) that J_{λ_1} contains the regularization for the weights $W^{[\ell]}$ and $b^{[\ell]}$. Additional regularization on $\tau^{[\ell]}$ can be easily introduced as

$$J_\lambda := J_{\lambda_1} + \frac{\lambda_2}{2} \sum_{\ell=0}^{L-2} \left(\|\tau^{[\ell]}\|_2^2 + \|\tau^{[\ell]}\|_1 \right),$$

where $\lambda = \lambda(\lambda_1, \lambda_2)$.

We apply the τ -variable framework to the ResNet and the Fractional-DNN discussed above.

3.5.1 ResNet with variable τ

Consider (3.5.1) with \mathcal{F} denoting the ResNet with variable τ

$$\begin{aligned} y^{[\ell]} &= P_{\ell-1}^\ell y^{[\ell-1]} + \tau^{[\ell-1]} \sigma(W^{[\ell-1]} y^{[\ell-1]} + b^{[\ell-1]}), \quad \ell = 1, \dots, L-1, \\ y^{[L]} &= W^{[L-1]} y^{[L-1]}. \end{aligned} \quad (3.5.2)$$

For simplicity of notation, we write J instead of J_λ and collect $W^{[\ell]}$, $b^{[\ell]}$, and $\tau^{[\ell]}$ for all ℓ into one vector θ and denote the adjoint variables by ϕ .

We introduce the Lagrangian functional to derive the optimality system, following the approach of [12],

$$\begin{aligned} \mathcal{L}(y, \theta, \phi) &= J(\theta) - \sum_{\ell=1}^{L-1} \left\langle y^{[\ell]} - P_{\ell-1}^\ell y^{[\ell-1]} - \tau^{[\ell-1]} \sigma(W^{[\ell-1]} y^{[\ell-1]} + b^{[\ell-1]}), \phi^{[\ell]} \right\rangle \\ &\quad - \left\langle y^{[L]} - W^{[L-1]} y^{[L-1]}, \phi^{[L]} \right\rangle. \end{aligned}$$

Setting the variation of \mathcal{L} with respect to ϕ equals zero, we recover the *state equation* (3.5.2). Similarly, setting the variation of \mathcal{L} with respect to y equals zero, we arrive at the *adjoint system*

$$\begin{aligned}\phi^{[\ell]} &= (P_\ell^{\ell+1})^\top \phi^{[\ell+1]} + \tau^{[\ell]} (W^{[\ell]})^\top \left(\phi^{[\ell+1]} \odot \sigma'(W^{[\ell]} y^{[\ell]} + b^{[\ell]}) \right), & \ell = L-2, \dots, 1, \\ \phi^{[L-1]} &= (W^{[L-1]})^\top \phi^{[L]}, \\ \phi^{[L]} &= \partial_{y^{[L]}} J(\theta) = y^{[L]} - S(u),\end{aligned}$$

where the last equality is due to the specific choice of the least-squares loss function. It will be different, for example, in the case of the cross-entropy softmax.

Since we will be solving the above optimization problem using a gradient-based method, we also need to evaluate the derivatives with respect to θ :

$$\begin{aligned}\partial_{W^{[L-1]}} \mathcal{L} &= \phi^{[L]} (y^{[L-1]})^\top + \partial_{W^{[L-1]}} J(\theta), \\ \partial_{W^{[\ell]}} \mathcal{L} &= y^{[\ell]} \left(\phi^{[\ell+1]} \odot \tau^{[\ell]} \sigma'(W^{[\ell]} y^{[\ell]} + b^{[\ell]}) \right)^\top + \partial_{W^{[\ell]}} J(\theta), & \ell = 0, \dots, L-2, \\ \partial_{b^{[\ell]}} \mathcal{L} &= (\phi^{[\ell+1]})^\top \tau^{[\ell]} \sigma'(W^{[\ell]} y^{[\ell]} + b^{[\ell]}) + \partial_{b^{[\ell]}} J(\theta), & \ell = 0, \dots, L-2, \\ \partial_{\tau^{[\ell]}} \mathcal{L} &= \left\langle \sigma(W^{[\ell]} y^{[\ell]} + b^{[\ell]}), \phi^{[\ell+1]} \right\rangle + \partial_{\tau^{[\ell]}} J(\theta), & \ell = 0, \dots, L-2.\end{aligned}$$

Next, we state the Fractional-DNN [12] but now with variable $\tau^{[\ell]}$. Recall that, in contrast to a ResNet, the Fractional-DNN allows connectivity between all the layers.

3.5.2 Fractional-DNN with variable τ

Consider a time-discretization $t_0 \leq t_1 \leq \dots \leq t_L$ with $L \in \mathbb{N}$ and set $I_\ell := (t_\ell, t_{\ell+1}]$ and $\tau^{[\ell]} = t_{\ell+1} - t_\ell$ for $0 \leq \ell \leq L-1$. Throughout, we will assume that $\tau^{[\ell]} > 0$ to justify division by $\tau^{[\ell]}$.

We generalize the numerical scheme introduced in [105, 104] to a non-uniform time discretization and obtain the discrete approximation of the left-sided Caputo

fractional derivative of order $\gamma \in (0, 1)$. For $0 \leq \ell \leq L - 1$, we have that:

$$\begin{aligned} \partial_t^\gamma y(x, t_{\ell+1}) &= c_\gamma \int_0^{t_{\ell+1}} \frac{\partial_t y(x, t)}{(t_{\ell+1} - t)^\gamma} dt = c_\gamma \sum_{j=0}^{\ell} \int_{I_j} \frac{\partial_t y(x, t)}{(t_{\ell+1} - t)^\gamma} dt \\ &= c_\gamma \sum_{j=0}^{\ell} \frac{y(x, t_{j+1}) - y(x, t_j)}{\tau^{[j]}} \int_{I_j} \frac{1}{(t_{\ell+1} - t)^\gamma} dt + r_\gamma^{\ell+1} \end{aligned} \quad (3.5.3)$$

where we have used the finite difference approximation. Here $r_\gamma^{\ell+1}$ denotes the remainder from the Taylor formula, which can be estimated as described in [118, Section 3.2.1]. After carrying out the integration in (3.5.3), we arrive at

$$\begin{aligned} \partial_t^\gamma y(x, t_{\ell+1}) &= c_\gamma \sum_{j=0}^{\ell} \frac{y(x, t_{j+1}) - y(x, t_j)}{\tau^{[j]}} \frac{1}{(1-\gamma)} \left(\left(\sum_{i=j}^{\ell} \tau^{[i]} \right)^{1-\gamma} - \left(\sum_{i=j+1}^{\ell} \tau^{[i]} \right)^{1-\gamma} \right) + r_\gamma^{\ell+1} \\ &= c_{\gamma-1} \sum_{j=0}^{\ell} \frac{1}{\tau^{[j]}} \left(\left(\sum_{i=j}^{\ell} \tau^{[i]} \right)^{1-\gamma} - \left(\sum_{i=j+1}^{\ell} \tau^{[i]} \right)^{1-\gamma} \right) (y(x, t_{j+1}) - y(x, t_j)) + r_\gamma^{\ell+1}. \end{aligned} \quad (3.5.4)$$

Analogously, we obtain the approximation of the right-sided Caputo fractional derivative of order $\gamma \in (0, 1)$ for $0 \leq \ell \leq L - 1$:

$$\partial_{T-t}^\gamma y(x, t_\ell) = -c_{\gamma-1} \sum_{j=\ell}^{L-1} \frac{1}{\tau^{[j]}} \left(\left(\sum_{i=\ell}^j \tau^{[i]} \right)^{1-\gamma} - \left(\sum_{i=\ell}^{j-1} \tau^{[i]} \right)^{1-\gamma} \right) (y(x, t_{j+1}) - y(x, t_j)) + r_\gamma^\ell. \quad (3.5.5)$$

Before we apply the above discretization to the Fractional-DNN formulation, we consider two generic nonlinear ODEs of type (3.3.3) (cf. e.g. [12, Section 4.1]) with $\gamma \in (0, 1)$:

$$\begin{aligned} \partial_t^\gamma y(x, t) &= f(t, y(x, t)), & y(x, 0) &= y_0, \\ \partial_{T-t}^\gamma y(x, t) &= f(t, y(x, t)), & y(x, T) &= y_T. \end{aligned} \quad (3.5.6)$$

This also links back to Subsection 3.3.2, where the stability of the above continuous DNN was discussed. Here, we will move on to formulate the discrete version.

Using the discretizations from (3.5.4) and (3.5.5) in (3.5.6), for $0 \leq \ell \leq L - 1$, we arrive at

$$\begin{aligned} y(x, t_{\ell+1}) &= y(x, t_\ell) - \sum_{j=0}^{\ell-1} a_{\ell,j} (y(x, t_{j+1}) - y(x, t_j)) + (\tau^{[\ell]})^\gamma c_{\gamma-1}^{-1} f(t_\ell, y(x, t_\ell)), \\ y(x, t_\ell) &= y(x, t_{\ell+1}) + \sum_{j=\ell+1}^{L-1} b_{j,\ell} (y(x, t_{j+1}) - y(x, t_j)) + (\tau^{[\ell]})^\gamma c_{\gamma-1}^{-1} f(t_\ell, y(x, t_\ell)), \end{aligned}$$

with

$$a_{\ell,j} := \frac{(\tau^{[\ell]})^\gamma}{\tau^{[j]}} \left(\left(\sum_{i=j}^{\ell} \tau^{[i]} \right)^{1-\gamma} - \left(\sum_{i=j+1}^{\ell} \tau^{[i]} \right)^{1-\gamma} \right),$$

$$b_{j,\ell} := \frac{(\tau^{[\ell]})^\gamma}{\tau^{[j]}} \left(\left(\sum_{i=\ell}^j \tau^{[i]} \right)^{1-\gamma} - \left(\sum_{i=\ell}^{j-1} \tau^{[i]} \right)^{1-\gamma} \right).$$

Notice that the uniform case, $\tau^{[\ell]} = \tau$ for all ℓ , considered throughout the literature, is a special case of the above setting. Furthermore, in the uniform setting, $a_{j,\ell} = b_{j,\ell}$, which may not be true in the aforementioned general scenario.

After these preparations, we are ready to apply the τ -variable framework to Fractional-DNN. Here, we take into account that the feature vectors $y^{[\ell]}$ may have different sizes across the layers. Thus, as in case of τ -variable ResNet, we introduce projection matrices P_j^ℓ for $j = 0, \dots, \ell - 1$ and $\ell = 1, \dots, L - 1$ with $\dim(P_j^\ell y^{[j]}) = \dim(y^{[\ell]})$. The resulting Fractional DNN with variable τ is

$$y^{[\ell]} = P_{\ell-1}^\ell y^{[\ell-1]} - \sum_{j=0}^{\ell-2} a_{\ell-1,j} (P_{j+1}^\ell y^{[j+1]} - P_j^\ell y^{[j]})$$

$$+ (\tau^{[\ell-1]})^\gamma c_{\gamma-1}^{-1} \sigma(W^{[\ell-1]} y^{[\ell-1]} + b^{[\ell-1]}), \quad \ell = 1, \dots, L - 1 \quad (3.5.7)$$

$$y^{[L]} = W^{[L-1]} y^{[L-1]}.$$

Remark 3.5.1. *Before we proceed further, we stress that the τ -variable framework is not merely a scaling of the activation by $\tau^{[\ell]}$. Indeed, in (3.5.7) the scaling in front of σ is not simply $\tau^{[\ell]}$, but is $(\tau^{[\ell-1]})^\gamma c_{\gamma-1}^{-1}$. Furthermore, $a_{\ell-1,j}$ also contains $\tau^{[j]}, \dots, \tau^{[\ell-1]}$, which makes the impact of the time step sizes much more complex than just the scaling of σ .*

As in the ResNet case, we next derive the optimality conditions. This requires introducing the Lagrangian formulation as before. In this fractional derivative setting, we observe a subtle issue. It is well known that there are two approaches to derive the optimality conditions: optimize-then-discretize and discretize-then-optimize [13, 81]. Below, in the τ -variable fractional setting, we observe that the two approaches do not

coincide. It is not difficult to see that in the first case, optimize-then-discretize, we obtain the following adjoint equation:

$$\begin{aligned}
\phi^{[\ell]} &= (P_\ell^{\ell+1})^\top \phi^{[\ell+1]} + \sum_{j=\ell+1}^{L-2} b_{j,\ell} ((P_\ell^{j+1})^\top \phi^{[j+1]} - (P_\ell^j)^\top \phi^{[j]}) \\
&\quad + (\tau^{[\ell]})^\gamma c_{\gamma-1}^{-1} \left[(W^{[\ell]})^\top \left(\phi^{[\ell+1]} \odot \sigma'(W^{[\ell]} y^{[\ell]} + b^{[\ell]}) \right) \right], \quad \ell = L-2, \dots, 1, \\
\phi^{[L-1]} &= \left(W^{[L-1]} \right)^\top \phi^{[L]}, \\
\phi^{[L]} &= \partial_{y^{[L]}} J(\theta).
\end{aligned} \tag{3.5.8}$$

Next, we derive the adjoint equations for the second approach, i.e., discretize-then-optimize. We begin by introducing the Lagrangian

$$\begin{aligned}
\mathcal{L}(y, \theta, \phi) &= J(\theta) - \sum_{\ell=1}^{L-1} \langle y^{[\ell]} - P_{\ell-1}^\ell y^{[\ell-1]} + \sum_{j=0}^{\ell-2} a_{\ell-1,j} (P_{j+1}^\ell y^{[j+1]} - P_j^\ell y^{[j]}) \\
&\quad - (\tau^{[\ell-1]})^\gamma c_{\gamma-1}^{-1} \sigma(W^{[\ell-1]} y^{[\ell-1]} + b^{[\ell-1]}), \phi^{[\ell]} \rangle \\
&\quad - \langle y^{[L]} - W^{[L-1]} y^{[L-1]}, \phi^{[L]} \rangle.
\end{aligned}$$

Setting the variation of \mathcal{L} with respect to ϕ equal zero, we obtain the state equation (3.5.7). To derive the adjoint equation, we calculate the variation of \mathcal{L} with respect to $y^{[\ell]}$ for every $\ell = 1, \dots, L$. A detailed calculation can be found in Appendix A.0.1. Setting this variation equal to zero, we arrive at the following adjoint system

$$\begin{aligned}
\phi^{[\ell]} &= (1 - a_{\ell,\ell-1}) (P_\ell^{\ell+1})^\top \phi^{[\ell+1]} + \sum_{j=\ell+2}^{L-1} (a_{j-1,\ell} - a_{j-1,\ell-1}) (P_\ell^j)^\top \phi^{[j]} \\
&\quad + (\tau^{[\ell]})^\gamma c_{\gamma-1}^{-1} \left[(W^{[\ell]})^\top \left(\phi^{[\ell+1]} \odot \sigma'(W^{[\ell]} y^{[\ell]} + b^{[\ell]}) \right) \right], \quad \ell = L-2, \dots, 1, \\
\phi^{[L-1]} &= \left(W^{[L-1]} \right)^\top \phi^{[L]}, \\
\phi^{[L]} &= \partial_{y^{[L]}} J(\theta).
\end{aligned} \tag{3.5.9}$$

Below, we collect all summands that contain factors $b_{j,\ell}$ in (3.5.8) on the left side and all summands that contain factors $a_{j,\ell}$ in (3.5.9) on the right side. We see that the two

adjoint equations given in (3.5.8) and (3.5.9) differ in the following term:

$$\sum_{j=\ell+1}^{L-2} b_{j,\ell}((P_\ell^{j+1})^\top \phi^{[j+1]} - (P_\ell^j)^\top \phi^{[j]}) \neq \sum_{j=\ell+1}^{L-2} a_{j,\ell}(P_\ell^{j+1})^\top \phi^{[j+1]} - \sum_{j=\ell+1}^{L-1} a_{j-1,\ell-1}(P_\ell^j)^\top \phi^{[j]}$$

In our computations, we have implemented the discretize-then-optimize approach, i.e., (3.5.9). Finally, we compute the derivative with respect to θ :

$$\begin{aligned} \partial_{W^{[L-1]}} \mathcal{L} &= \phi^{[L]}(y^{[L-1]})^\top + \partial_{W^{[L-1]}} J(\theta), \\ \partial_{W^{[\ell]}} \mathcal{L} &= y^{[\ell]} \left(\phi^{[\ell+1]} \odot (\tau^{[\ell]})^\gamma c_{\gamma-1}^{-1} \sigma'(W^{[\ell]} y^{[\ell]} + b^{[\ell]}) \right)^\top + \partial_{W^{[\ell]}} J(\theta), \quad \ell = 0, \dots, L-2, \\ \partial_{b^{[\ell]}} \mathcal{L} &= (\phi^{[\ell+1]})^\top (\tau^{[\ell]})^\gamma c_{\gamma-1}^{-1} \sigma'(W^{[\ell]} y^{[\ell]} + b^{[\ell]}) + \partial_{b^{[\ell]}} J(\theta), \quad \ell = 0, \dots, L-2, \\ \partial_{\tau^{[\ell]}} \mathcal{L} &= - \sum_{k=\ell}^{L-2} \sum_{j=0}^{\min\{k-1,\ell\}} \partial_{\tau^{[k]}}(a_{k,j}) \left\langle P_{j+1}^{k+1} y^{[j+1]} - P_j^{k+1} y^{[j]}, \phi^{[k+1]} \right\rangle \\ &\quad + \left\langle \gamma (\tau^{[\ell]})^{\gamma-1} c_{\gamma-1}^{-1} \sigma(W^{[\ell]} y^{[\ell]} + b^{[\ell]}), \phi^{[\ell+1]} \right\rangle + \partial_{\tau^{[\ell]}} J(\theta), \quad \ell = 0, \dots, L-2. \end{aligned}$$

Details on the computation of $\partial_{\tau^{[\ell]}} \mathcal{L}$ can be found in Appendix A.0.2. Next, we examine the impact of variable τ on the stability of networks such as ResNets and Fractional-DNNs.

3.6 Vanishing and exploding gradients

It is well known that optimization problems with DNN constraints can suffer from vanishing and exploding gradients, see e.g. [19, 70]. In this section, we analyze the structure of the derivatives for several network architectures such as feedforward network, ResNet, DenseNet, Fractional-DNN, and the consequences of application of τ -variable framework on these networks. We will identify various conditions to help overcome the aforementioned challenges.

For simplicity of the notation, we define the abbreviation $a^{[\ell]} := \sigma(W^{[\ell]} y^{[\ell]} + b^{[\ell]})$ and omit the projection matrices $P_{\ell-1}^\ell$, i.e., $n_\ell = n$ for all layers ℓ . While the following result may not be new for the standard case with $\tau^{[\ell]} = \tau \in \mathbb{R}$ for all ℓ , to the best of our knowledge, this is new for variable $\tau^{[\ell]}$.

Theorem 3.6.1 (Feedforward Network and ResNet). *Consider the feedforward network and ResNet with τ -variable framework*

$$y^{[\ell]} = \tau^{[\ell-1]} a^{[\ell-1]}, \quad \ell = 1, \dots, L-1,$$

$$y^{[\ell]} = y^{[\ell-1]} + \tau^{[\ell-1]} a^{[\ell-1]}, \quad \ell = 1, \dots, L-1.$$

Let $\theta^{[k]} = (W^{[k]}(\cdot), b^{[k]}, \tau^{[k]})^\top$ be the parameters associated with layer k for $k = 0, \dots, L-2$. Then the respective derivatives take the form

$$d_{\theta^{[j]}} y^{[\ell]} = \prod_{i=\ell-1}^{j+1} \left(\tau^{[i]} d_{y^{[i]}} a^{[i]} \right) \partial_{\theta^{[j]}} (\tau^{[j]} a^{[j]}), \quad (3.6.1)$$

$$d_{\theta^{[j]}} y^{[\ell]} = \prod_{i=\ell-1}^{j+1} \left(\mathbb{I} + \tau^{[i]} d_{y^{[i]}} a^{[i]} \right) \partial_{\theta^{[j]}} (\tau^{[j]} a^{[j]}), \quad (3.6.2)$$

for all $\ell = 1, \dots, L-1$ and $j = 0, \dots, \ell-1$.

Proof. For the feedforward neural network we can compute with chain rule

$$\begin{aligned} d_{\theta^{[j]}} y^{[\ell]} &= d_{\theta^{[j]}} \tau^{[\ell-1]} a^{[\ell-1]} \\ &= \tau^{[\ell-1]} d_{y^{[\ell-1]}} a^{[\ell-1]} \cdot d_{\theta^{[j]}} y^{[\ell-1]} \\ &= \tau^{[\ell-1]} d_{y^{[\ell-1]}} a^{[\ell-1]} \cdot d_{\theta^{[j]}} \tau^{[\ell-2]} a^{[\ell-2]}. \end{aligned}$$

where \cdot denotes the standard matrix multiplication. By iterating we arrive at

$$d_{\theta^{[j]}} y^{[\ell]} = \prod_{i=\ell-1}^{j+1} \left(\tau^{[i]} d_{y^{[i]}} a^{[i]} \right) d_{\theta^{[j]}} y^{[j+1]} = \prod_{i=\ell-1}^{j+1} \left(\tau^{[i]} d_{y^{[i]}} a^{[i]} \right) \partial_{\theta^{[j]}} (\tau^{[j]} a^{[j]}).$$

Similarly, for the ResNet, we obtain

$$\begin{aligned} d_{\theta^{[j]}} y^{[\ell]} &= d_{\theta^{[j]}} (y^{[\ell-1]} + \tau^{[\ell-1]} a^{[\ell-1]}) \\ &= d_{\theta^{[j]}} y^{[\ell-1]} + \tau^{[\ell-1]} d_{y^{[\ell-1]}} a^{[\ell-1]} \cdot d_{\theta^{[j]}} y^{[\ell-1]} \\ &= (\mathbb{I} + \tau^{[\ell-1]} d_{y^{[\ell-1]}} a^{[\ell-1]}) d_{\theta^{[j]}} y^{[\ell-1]}, \end{aligned}$$

where $\mathbb{I} \in \mathbb{R}^{n \times n}$ denotes the identity matrix, with $n = n_\ell$ constant throughout all layers ℓ . By iterating we arrive at

$$d_{\theta^{[j]}} y^{[\ell]} = \prod_{i=\ell-1}^{j+1} \left(\mathbb{I} + \tau^{[i]} d_{y^{[i]}} a^{[i]} \right) d_{\theta^{[j]}} y^{[j+1]} = \prod_{i=\ell-1}^{j+1} \left(\mathbb{I} + \tau^{[i]} d_{y^{[i]}} a^{[i]} \right) \partial_{\theta^{[j]}} (\tau^{[j]} a^{[j]}),$$

where we use that $d_{\theta^{[j]}} y^{[j]} = 0$. This concludes the proof. \square

Remark 3.6.2. *Since σ is applied componentwise, special caution needs to be exercised when deriving $\mathbf{d}_{y^{[i]}}a^{[i]}$. Let $r_j^{[i]}$ be the j th row of $W^{[i]}y^{[i]} + b^{[i]}$, namely $\sum_{m=1}^n W_{j,m}^{[i]}y_m^{[i]} + b_j^{[i]}$ for $j \in \{1, \dots, n\}$. Then, with a slight abuse of notation, it holds*

$$\mathbf{d}_{y^{[i]}}a^{[i]} = \mathbf{d}_{y^{[i]}}\sigma(W^{[i]}y^{[i]} + b^{[i]}) = \mathbf{d}_{y^{[i]}}\left(\sigma(r_j^{[i]})\right)_{j=1}^n = \text{diag}(\sigma'(r_1^{[i]}), \dots, \sigma'(r_n^{[i]})) \cdot W^{[i]},$$

where $\sigma'(r_j^{[i]})$ is the one-dimensional derivative of σ at $r_j^{[i]}$. Furthermore, for the partial derivative $\partial_{\theta^{[j]}}(\tau^{[j]}a^{[j]})$, we recall $\theta^{[j]} = (W^{[j]}(:, b^{[j]}, \tau^{[j]})^\top \in \mathbb{R}^N$, with $N = n^2 + n + 1$ and the fact that $a^{[j]}$ depends on $W^{[j]}$ and $b^{[j]}$, but not $\tau^{[j]}$. Consequently, we see

$$\partial_{\theta^{[j]}}(\tau^{[j]}a^{[j]}) = \begin{pmatrix} \tau^{[j]}\partial_{W^{[j]}(:, \cdot)}a^{[j]} & \tau^{[j]}\partial_{b^{[j]}}a^{[j]} & a^{[j]} \end{pmatrix} \in \mathbb{R}^{n \times N}.$$

As pointed out above, the standard feedforward neural network, where $\tau^{[\ell]} = 1$ for all ℓ , can suffer from vanishing and exploding gradients, which can be a challenge for optimization with deep networks [19, 70]. Consider the structure of the derivatives in (3.6.1) with $\tau^{[\ell]} = 1$ for all ℓ , e.g. for the final hidden layer with $\ell = L - 1$,

$$\mathbf{d}_{\theta^{[j]}}y^{[L-1]} = \prod_{i=L-2}^{j+1} \left(\mathbf{d}_{y^{[i]}}a^{[i]} \right) \partial_{\theta^{[j]}}a^{[j]}.$$

Especially in the one-dimensional case, it is obvious that if the partial derivatives $\mathbf{d}_{y^{[i]}}a^{[i]}$ are smaller than one, the product will tend to 0 as the number of layers L increases, which leads to vanishing gradients. On the other hand, if the partial derivatives $\mathbf{d}_{y^{[i]}}a^{[i]}$ are larger than 1, the product will tend to ∞ as the number of layers L increases, which leads to exploding gradients. The feedforward neural network with variable τ , can potentially help overcome both problems, since now we have flexibility with respect to $\tau^{[\ell]}$. But one needs to be careful as if the gradient components are really small, then $\tau^{[\ell]}$ needs to be really large to compensate, which could lead to ill-conditioning issues.

A more appropriate approach is the standard ResNet with $\tau^{[\ell]} = \tau \in \mathbb{R}$ for all ℓ . It is known to be stable with respect to vanishing gradients. Recalling the gradient from (3.6.2)

$$\mathbf{d}_{\theta^{[j]}}y^{[L-1]} = \prod_{i=L-2}^{j+1} \left(\mathbb{I} + \tau^{[i]}\mathbf{d}_{y^{[i]}}a^{[i]} \right) \partial_{\theta^{[j]}}(\tau^{[j]}a^{[j]}),$$

it becomes clear that this stability is achieved by the added identity \mathbb{I} in every part of the product. Hence, even if the Jacobians $d_{y^{[i]}}a^{[i]}$ vanish, the product still contains the identity matrices. This advantage carries over to the τ -variable framework.

Furthermore, the introduction of $\tau^{[\ell]}$ in ResNet allows us to tackle the exploding gradients problem. This property has also been discussed using probabilistic bounds in [76]. Our approach is deterministic. The standard ResNet architecture does not have this property. Appropriate small $\tau^{[\ell]}$ can prevent the product from exploding with growing number of layers. However, choosing $\tau^{[\ell]}$ too small may lead to the vanishing gradient problem again, as we will illustrate in the following simple example. Recall that we do not tune $\tau^{[\ell]}$ by hand, but let the optimization find it.

Example 3.6.3. *Consider the ResNet architecture with variable τ in one dimension, i.e. one node per layer. We have*

$$\begin{aligned} d_{\theta^{[1]}}y^{[2]} &= \partial_{\theta^{[1]}}(\tau^{[1]}a^{[1]}), \\ d_{\theta^{[0]}}y^{[2]} &= (1 + \tau^{[1]}d_{y^{[1]}}a^{[1]}) \partial_{\theta^{[0]}}(\tau^{[0]}a^{[0]}). \end{aligned}$$

Assume that $d_{y^{[1]}}a^{[1]}$ is large, so that it leads to a large $d_{\theta^{[0]}}y^{[2]}$. This problem can be overcome if $\tau^{[1]}$ attains a small value. However, this may lead to $d_{\theta^{[1]}}y^{[2]}$ being accordingly small in its first two components, i.e., the derivatives by the weights and biases. Consequently, fixing one potential exploding gradient problem can cause another gradient to vanish. However, as emphasized earlier, we do not tune $\tau^{[\ell]}$ by hand but let the optimization find optimal values.

We also analyze the respective derivatives in the DenseNet architecture with variable τ . Finding a closed form for $d_{\theta^{[l]}}y^{[\ell]}$ is not so easy for this network architecture, but we can derive a recursive relation in terms of lower-order terms.

Theorem 3.6.4. *Consider the DenseNet with τ -variable framework*

$$y^{[\ell]} = \sum_{k=0}^{\ell-1} y^{[k]} + \tau^{[\ell-1]}a^{[\ell-1]}, \quad \ell = 1, \dots, L-1.$$

Then the derivatives can be recursively written as

$$\begin{aligned} d_{\theta^{[j]}} y^{[i]} &= d_{\theta^{[j]}} \sum_{k=j+1}^{i-2} y^{[k]} + (\mathbb{I} + \tau^{[i-1]} d_{y^{[i-1]}} a^{[i-1]}) d_{\theta^{[j]}} y^{[i-1]} \quad i = \ell, \dots, j+2, \\ d_{\theta^{[j]}} y^{[j+1]} &= \partial_{\theta^{[j]}} (\tau^{[j]} a^{[j]}). \end{aligned}$$

Proof. For $i = \ell, \dots, j+2$ we employ the chain rule of differentiation and use that $d_{\theta^{[j]}} y^{[k]} = 0$ for $k < j+1$ to arrive at

$$\begin{aligned} d_{\theta^{[j]}} y^{[i]} &= d_{\theta^{[j]}} \left(\sum_{k=0}^{i-1} y^{[k]} + \tau^{[i-1]} a^{[i-1]} \right) \\ &= d_{\theta^{[j]}} \sum_{k=j+1}^{i-2} y^{[k]} + d_{\theta^{[j]}} y^{[i-1]} + \tau^{[i-1]} d_{\theta^{[j]}} a^{[i-1]} \\ &= d_{\theta^{[j]}} \sum_{k=j+1}^{i-2} y^{[k]} + (\mathbb{I} + \tau^{[i-1]} d_{y^{[i-1]}} a^{[i-1]}) d_{\theta^{[j]}} y^{[i-1]}. \end{aligned}$$

The case $i = j+1$ is special since the chain rule does not need to be applied here. It simply holds

$$d_{\theta^{[j]}} y^{[j+1]} = d_{\theta^{[j]}} \left(\sum_{k=0}^j y^{[k]} + \tau^{[j]} a^{[j]} \right) = \partial_{\theta^{[j]}} (\tau^{[j]} a^{[j]}),$$

because $k < j+1$. The proof is complete. \square

Remark 3.6.5. In Theorem 3.6.4, we can successively insert the expressions for the lower-order terms in the higher-order terms, so that finally $d_{\theta^{[j]}} y^{[\ell]}$ depends only on $\partial_{\theta^{[j]}} (\tau^{[j]} a^{[j]})$ and $d_{y^{[i]}} a^{[i]}$ for $i = j+1, \dots, \ell-1$. Furthermore, we see that every next lower order term enters with a factor $(\mathbb{I} + \tau^{[i]} d_{y^{[i]}} a^{[i]})$, so that one can overcome the vanishing gradients problem in a DenseNet (both fixed and variable τ cases). To discuss the exploding gradients problem we consider for example the derivative $d_{\theta^{[j]}} y^{[L-1]}$, where one summand will be $\prod_{i=j+1}^{L-2} (\tau^{[i]} d_{y^{[i]}} a^{[i]}) \partial_{\theta^{[j]}} (\tau^{[j]} a^{[j]})$. In the standard one-dimensional DenseNet architecture with $\tau^{[\ell]} = 1$ for all ℓ , we see that the above product tends to ∞ with a growing number of layers L if $d_{y^{[i]}} a^{[i]} > 1$ for all i . The τ -variable architecture can help deal with this problem; see also Example 3.6.8.

Similarly to the above cases, we can express derivatives of Fractional-DNN architecture, with variable τ , in terms of lower-order terms.

Theorem 3.6.6. *Consider the Fractional-DNN with τ -variable framework*

$$y^{[\ell]} = y^{[\ell-1]} - \sum_{k=0}^{\ell-2} a_{\ell,k} (y^{[k+1]} - y^{[k]}) + (\tau^{[\ell-1]})^\gamma c_{\gamma-1}^{-1} a^{[\ell-1]}, \quad \ell = 1, \dots, L-1.$$

Then the derivatives can be recursively written as

$$\begin{aligned} d_{\theta^{[j]}} y^{[i]} &= d_{\theta^{[j]}} \sum_{k=j+1}^{i-2} (a_{i,k} - a_{i,k-1}) y^{[k]} + \left((1 - a_{i,i-2}) \mathbb{I} + (\tau^{[i-1]})^\gamma c_{\gamma-1}^{-1} d_{y^{[i-1]}} a^{[i-1]} \right) d_{\theta^{[j]}} y^{[i-1]}, \\ & \quad i = \ell, \dots, j+2, \end{aligned}$$

$$d_{\theta^{[j]}} y^{[j+1]} = c_{\gamma-1}^{-1} \partial_{\theta^{[j]}} ((\tau^{[j]})^\gamma a^{[j]}).$$

Proof. First of all, we rewrite the forward propagation for $\ell = 1, \dots, L-1$ in Fractional-DNN

$$\begin{aligned} y^{[\ell]} &= y^{[\ell-1]} - \sum_{k=0}^{\ell-2} a_{\ell,k} (y^{[k+1]} - y^{[k]}) + (\tau^{[\ell-1]})^\gamma c_{\gamma-1}^{-1} a^{[\ell-1]} \\ &= a_{\ell,0} y^{[0]} + \sum_{k=1}^{\ell-2} (a_{\ell,k} - a_{\ell,k-1}) y^{[k]} + (1 - a_{\ell,\ell-2}) y^{[\ell-1]} + (\tau^{[\ell-1]})^\gamma c_{\gamma-1}^{-1} a^{[\ell-1]}. \end{aligned}$$

Then for $i = \ell, \dots, j+2$ we use chain rule and $d_{\theta^{[j]}} y^{[k]} = 0$ for $k < j+1$ to obtain

$$\begin{aligned} d_{\theta^{[j]}} y^{[i]} &= d_{\theta^{[j]}} \left(a_{i,0} y^{[0]} + \sum_{k=1}^{i-2} (a_{i,k} - a_{i,k-1}) y^{[k]} + (1 - a_{i,i-2}) y^{[i-1]} + (\tau^{[i-1]})^\gamma c_{\gamma-1}^{-1} a^{[i-1]} \right) \\ &= d_{\theta^{[j]}} \sum_{k=j+1}^{i-2} (a_{i,k} - a_{i,k-1}) y^{[k]} + \left((1 - a_{i,i-2}) \mathbb{I} + (\tau^{[i-1]})^\gamma c_{\gamma-1}^{-1} d_{y^{[i-1]}} a^{[i-1]} \right) d_{\theta^{[j]}} y^{[i-1]}. \end{aligned}$$

Finally, for $i = j+1$, we exploit again $d_{\theta^{[j]}} y^{[k]} = 0$ for $k < j+1$, and derive

$$\begin{aligned} d_{\theta^{[j]}} y^{[j+1]} &= d_{\theta^{[j]}} \left(a_{j+1,0} y^{[0]} + \sum_{k=1}^{j-1} (a_{j+1,k} - a_{j+1,k-1}) y^{[k]} + (1 - a_{j+1,j-1}) y^{[j]} + (\tau^{[j]})^\gamma c_{\gamma-1}^{-1} a^{[j]} \right) \\ &= c_{\gamma-1}^{-1} \partial_{\theta^{[j]}} ((\tau^{[j]})^\gamma a^{[j]}). \end{aligned}$$

This completes the proof. □

Remark 3.6.7. *Again, the lower order term representations can be successively inserted into the higher order terms until we arrive at $d_{\theta^{[j]}}y^{[j]}$ depending only on $\partial_{\theta^{[j]}}((\tau^{[j]})^\gamma a^{[j]})$ and $d_{y^{[i]}}a^{[i]}$ for $i = j + 1, \dots, \ell - 1$. Here, the next lower order term enters with a factor $((1 - a_{i,i-2})\mathbb{I} + (\tau^{[i-1]})^\gamma c_{\gamma-1}^{-1} d_{y^{[i-1]}}a^{[i-1]})$, which allows us to overcome the vanishing gradient problem in Fractional-DNNs. Furthermore, the multiplication by $(\tau^{[i-1]})^\gamma$ in this factor can help us to deal with exploding gradients. This is similar to ResNet with variable τ .*

Example 3.6.8. *To get an idea of how different network architectures influence the derivatives, the derivative $d_{\theta^{[0]}}y^{[3]}$ is displayed here for the four different options that have been considered in this section, i.e., feedforward neural network, ResNet, DenseNet, and Fractional DNN:*

$$\begin{aligned} d_{\theta^{[0]}}y^{[3]} &= \tau^{[2]}d_{y^{[2]}}a^{[2]} \cdot \tau^{[1]}d_{y^{[1]}}a^{[1]} \cdot \partial_{\theta^{[0]}}(\tau^{[0]}a^{[0]}), \\ d_{\theta^{[0]}}y^{[3]} &= \left(\mathbb{I} + \tau^{[1]}d_{y^{[1]}}a^{[1]} + \tau^{[2]}d_{y^{[2]}}a^{[2]} + \tau^{[1]}\tau^{[2]}d_{y^{[2]}}a^{[2]} \cdot d_{y^{[1]}}a^{[1]} \right) \partial_{\theta^{[0]}}(\tau^{[0]}a^{[0]}), \\ d_{\theta^{[0]}}y^{[3]} &= \left(2\mathbb{I} + \tau^{[1]}d_{y^{[1]}}a^{[1]} + \tau^{[2]}d_{y^{[2]}}a^{[2]} + \tau^{[1]}\tau^{[2]}d_{y^{[2]}}a^{[2]} \cdot d_{y^{[1]}}a^{[1]} \right) \partial_{\theta^{[0]}}(\tau^{[0]}a^{[0]}), \\ d_{\theta^{[0]}}y^{[3]} &= \left\{ (1 - a_{2,0} - a_{3,0} + a_{2,0}a_{3,1})\mathbb{I} + (1 - a_{3,1})(\tau^{[1]})^\gamma c_{\gamma-1}^{-1} d_{y^{[1]}}a^{[1]} \right. \\ &\quad \left. + (1 - a_{2,0})(\tau^{[2]})^\gamma c_{\gamma-1}^{-1} d_{y^{[2]}}a^{[2]} + (\tau^{[1]})^\gamma (\tau^{[2]})^\gamma c_{\gamma-1}^{-2} d_{y^{[2]}}a^{[2]} \cdot d_{y^{[1]}}a^{[1]} \right\} c_{\gamma-1}^{-1} \partial_{\theta^{[0]}}((\tau^{[0]})^\gamma a^{[0]}). \end{aligned}$$

In conclusion, ResNet, DenseNet and Fractional-DNN have a visible additive structure in the derivatives, which helps with the vanishing gradients problem. Furthermore, the parameters $\tau^{[j]}$ can help overcome both vanishing and exploding gradients.

3.7 Numerical results

In this section, we apply the τ -variable framework to a ResNet and a Fractional DNN with and without bias ordering (3.2.5). A thorough comparison is carried out in the context of an ill-posed 3D parametrized Maxwell's equation with Gauss's law. This problem is ill-posed because the standard Nédélec finite element is only curl conforming

and cannot directly impose the Gauss's law. In all cases, we apply the smoothed version of standard $\text{ReLU}(y) = \max\{0, y\}$ as the activation function

$$\text{smoothReLU}(y) = \begin{cases} \max\{0, y\}, & \text{if } |y| > \eta \\ \frac{1}{4\eta}y^2 + 0.5y + 0.25\eta, & \text{if } y \in [-\eta, \eta]. \end{cases}$$

We have found that $\eta = 10^{-4}$ is a robust choice for the examples under consideration. Notice, that one can also use other activation functions which can differ from layer to layer.

3.7.1 Maxwell's equations

Our findings suggests that the τ -variable framework outperforms the standard approach (with fixed τ) for deeper networks; see Figure 3.1. This is expected, since the effect of variable $\tau^{[\ell]}$ will be more prominent when more layers (and consequently more time-step parameters $\tau^{[\ell]}$) are present. On the other hand, for shallow networks, the τ -variable framework provides comparatively less improvements; see Figure 3.3(c). Nevertheless, the τ -variable framework applied to ResNet yields error improvements compared to a standard ResNet for model extrapolation, see Figure 3.4. These results are also comparable to the approximation obtained with the finite element method (FEM) using the lowest-order Nédélec space (see Figure 3.3(d)).

Consider the Maxwell-Dirac equations; our goal is to learn $\mathbf{u} : \Omega \subset \mathbb{R}^3 \mapsto \mathbb{R}^3$ that satisfies

$$\begin{aligned} \text{curl}(\boldsymbol{\mu}^{-1}\text{curl}\mathbf{u}) &= \mathbf{f} & \text{in } \Omega, \\ \text{div}(\boldsymbol{\varepsilon}\mathbf{u}) &= \rho & \text{in } \Omega, \\ \mathbf{u} \times \mathbf{n} &= \mathbf{g} & \text{on } \partial\Omega, \end{aligned} \tag{3.7.1}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\varepsilon}$ are positive definite symmetric tensors in $L^\infty(\Omega)^3$, $\mathbf{f} \in L^2(\Omega)^3$, $\rho \in L^2(\Omega)$ and $\mathbf{g} \in H_{\parallel}^{-\frac{1}{2}}(\text{div}_\Gamma; \partial\Omega)$. This problem is particularly difficult at the discrete level due to its divergence-related constraints and requires rather tailored algorithms to

deal with it, see for instance [48]. Therefore, an interesting question is to approximate the map:

$$(\mathbf{x}, \mathbf{f}(\mathbf{x}), \boldsymbol{\mu}(\mathbf{x}), \rho(\mathbf{x})) \mapsto \mathbf{u}(\mathbf{x}),$$

that can lead to a reasonable and noise-robust approximation of the solution \mathbf{u} to (3.7.1), note that we ignore the boundary data \mathbf{g} . This approach is similar to a surrogate model where its output could be used as an initial guess by an iterative method like the domain decomposition method or in the reduced basis method [141]. Nevertheless, those problems are beyond the scope of the present study, and thus will be studied in future works.

This learning problem is challenging because the solutions to (3.7.1) can have discontinuities, while most neural networks, except for those with the Heaviside activation function, lead to continuous approximations. Also, it is still not clear how to incorporate the geometry (domain Ω) of the problem in a meaningful way. Thus, we consider an example with a known smooth solution and we compare it with an approximation obtained by various DNNs and by the lowest order Nédélec space of the first kind, cf. [116], denoted by $\mathcal{N}_0(\Omega)$. Here, we consider the basis proposed in [72]. In order to do that, let us consider $\boldsymbol{\varepsilon} = \mathbb{I}_{3 \times 3}$, and for a smooth $\varphi : \Omega \mapsto \mathbb{R}^+$ we define $\boldsymbol{\mu}^{-1}(\mathbf{x}) = \varphi(\mathbf{x})\mathbb{I}_{3 \times 3}$, then $\text{curl}(\boldsymbol{\mu}^{-1}\text{curl}\mathbf{u}) = \nabla\varphi \times \text{curl}\mathbf{u} + \varphi\text{curl}(\text{curl}\mathbf{u})$.

Thus, if we consider

$$\begin{aligned} \mathbf{u} : \Omega &\mapsto \mathbb{R}^3, & (x_1, x_2, x_3) &\mapsto I_1(r(x_1, x_2, x_3))\mathbf{e}_\theta, \\ \varphi : \Omega &\mapsto \mathbb{R}, & (x_1, x_2, x_3) &\mapsto \frac{1}{2}(x_1^2 + x_2^2 + 1), \end{aligned} \tag{3.7.2}$$

where Ω is the cylinder $\{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1^2 + x_2^2 \leq 1 \text{ and } x_3 \in [0, 1]\}$, I_ν is the modified Bessel functions of the first kind of order ν , $r(x_1, x_2, x_3) = \sqrt{x_1^2 + x_2^2}$, and $\mathbf{e}_\theta(x_1, x_2, x_3) = (x_1^2 + x_2^2)^{-\frac{1}{2}}(-x_2, x_1, 0)$, we obtain:

$$\begin{aligned} \text{curl}\mathbf{u} &= I_0(r)\mathbf{e}_z, & \text{curl}(\text{curl}\mathbf{u}) &= -\mathbf{u}, & \text{div}\mathbf{u} &= 0, \text{ and} \\ \text{curl}(\boldsymbol{\mu}^{-1}\text{curl}\mathbf{u}) &= -rI_0(r)\mathbf{e}_\theta - \varphi\mathbf{u} =: \mathbf{f}. \end{aligned}$$

Because \mathbf{u} is divergence free, we consider the reduced map $(\mathbf{x}, \mathbf{f}(\mathbf{x}), \varphi(\mathbf{x})) \mapsto \mathbf{u}(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, x_3)$, and $\mathbf{u}(\mathbf{x}) = (\mathbf{u}_1(\mathbf{x}), \mathbf{u}_2(\mathbf{x}), \mathbf{u}_3(\mathbf{x}))$. In order to generate the input/output data for the DNNs, we consider points $\{\mathbf{x}_i\}_{i=1}^N \subset \Omega$ randomly chosen from Ω obtained with Matlab's function `unifrnd` along with `philox` as the random number algorithm. Here, we set $N = 12,000$. Then, $\{(\mathbf{x}_i, f(\mathbf{x}_i), \varphi(\mathbf{x}_i))\}_{i=1}^N$ and $\{\mathbf{u}(\mathbf{x}_i)\}_{i=1}^N$ can be utilized as input/output data. We are now ready to train and compare several DNNs.

Neural Network size & network reduction

The bigger the network that we use for training, the bigger the computational time and the memory requirements. We employ the τ -variable framework and start with a ResNet architecture with five hidden layers with ten nodes each. As target functional, we implement the mean squared error with no regularization, i.e., $\lambda_1 = \lambda_2 = 0$. Additionally, we consider bias ordering (B.O.) with a fixed Moreau-Yosida parameter $\beta = 10$. After 1000 steepest descent steps we observe the following result: The relative error in the Euclidean norm on the test set is 0.07, and we see $\tau^{[1]}, \tau^{[3]}$ and $\tau^{[4]}$ are approximately 0. Recalling the ResNet structure with variable τ , (3.5.2), it is obvious that e.g. from $\tau^{[1]} \approx 0$ we can deduce $y^{[2]} \approx P_1^2 y^{[1]}$. Consequently, we delete the hidden layers 2,4 and 5, cf. Figure 3.2. The reduced network with 2 hidden layers achieves the same relative error on the test set, i.e. 0.07.

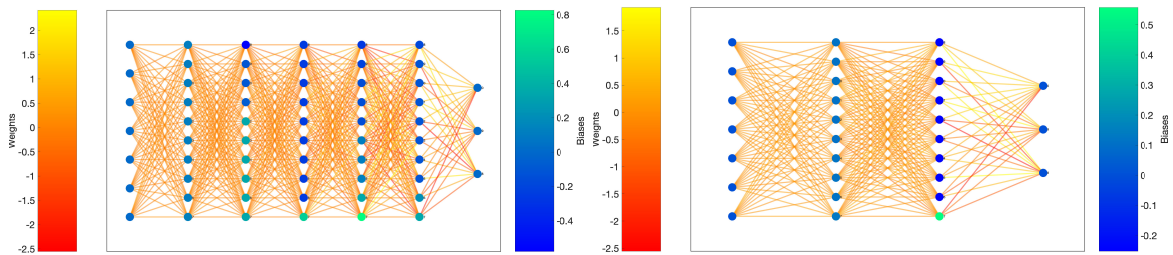


Figure 3.2: Left: Optimal weights and biases for ResNet with variable τ with 5 hidden layers and 10 nodes each with bias ordering. Right: Reduced ResNet with 2 hidden layers, i.e., hidden layers 1 and 3 from the larger network. The color of the dots indicates the bias value, and the color of the lines indicates the magnitude of the weight.

While the same relative error is obtained with the reduced network, we aim at achieving even better results, therefore we next consider a larger network size with 6 hidden layers and 50 nodes in each layer (6-50). The proposed variable- τ approach seems to always outperform its constant τ counterpart, see Figure 3.1. As stated before, this is expected because the variable- τ framework and also Fractional-DNN have a bigger impact for deeper architectures with more hidden layers. Let us remark that the curves in Figure 3.1 are not monotone because we have only plotted the mean squared error term. In case of the entire J we do observe monotone behavior as expected. Even though we see in Table 3.2 that $\tau^{[\ell]} > 0$ for all ℓ in this setup, motivated by the reduction in the previous architecture of 5 hidden layers and 10 nodes per layer instead of (6-50), we also consider a network with 2 hidden layers with 50 nodes per layer (2-50), which yields better results; cf. Figure 3.3.

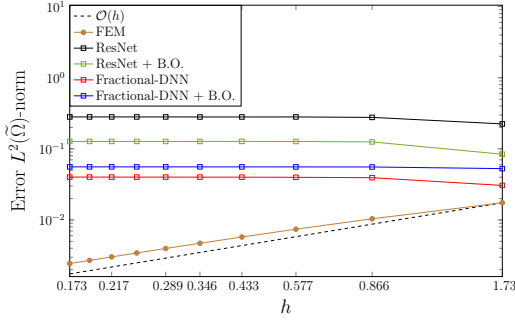
Results: 6 layers-50 nodes vs. 2 layers-50 nodes

From now on, $\mathbf{u}^{NN}(\mathbf{x})$ will denote the approximation of \mathbf{u} obtained with a neural network at a point \mathbf{x} , the specific architecture will be clear from the context. Note that, $\mathbf{u}^{NN}(\mathbf{x}) = (\mathbf{u}_1^{NN}(\mathbf{x}), \mathbf{u}_2^{NN}(\mathbf{x}), \mathbf{u}_3^{NN}(\mathbf{x})) \in \mathbb{R}^3$.

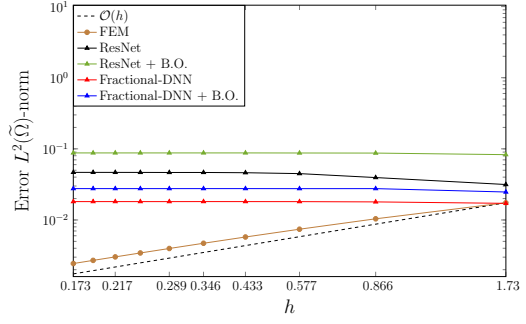
6-50	$\tau^{[0]}$	$\tau^{[1]}$	$\tau^{[2]}$	$\tau^{[3]}$	$\tau^{[4]}$	$\tau^{[5]}$	2-50	$\tau^{[0]}$	$\tau^{[1]}$
ResNet	0.92	0.95	0.99	0.95	0.92	0.88		0.84	0.87
ResNet + B.O.	0.70	0.94	1.00	0.96	0.86	0.70		0.28	0.51
Fractional-DNN	0.59	0.70	0.53	0.34	0.28	0.30		0.94	0.93
Fractional-DNN + B.O.	0.67	0.80	0.78	0.68	0.44	0.04		0.85	0.95

Table 3.2: Optimal learned τ variables for various DNN architectures with τ -variable framework with 6 layers and 2 layers. These are the same network architectures that are considered in Figure 3.3.

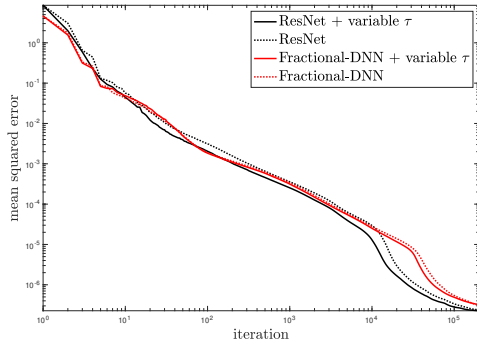
We compare the neural network results with an approximation obtained with the FEM, see Figures 3.3(a), 3.3(b), and 3.3(d). To do that, we consider the unit cube $(0, 1)^3$, denoted by $\tilde{\Omega}$, as a domain. The unit cube is considered, to test how well the Neural Network performs for unseen data, and to test its extrapolation properties. Recall that the training data has been generated on Ω , which is cylindrical. For the



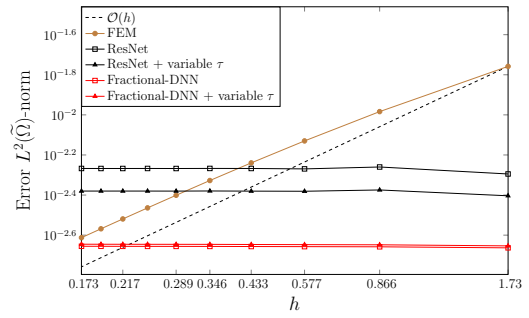
(a) Fixed τ – 6 layers and 50 nodes



(b) Variable τ – 6 layers and 50 nodes



(c) MSE plot – 2 layers and 50 nodes



(d) Comparison – 2 layers and 50 nodes

Figure 3.3: Comparison between various DNN architectures and FEM. **Top row:** L^2 error between an exact solution and DNN approximation (6 hidden layers with a width of 50 each) or FEM approximation. B.O. indicates bias ordering. The left and right panels correspond to fixed and variable τ , respectively. **Bottom row:** The left panel shows the mean squared error during training of different DNNs with 2 hidden layers with a width of 50 nodes each. The right panel displays the L^2 error between an exact solution and a DNN approximation for the same DNNs and FEM.

FEM, we consider 10 uniform refinements of the unit cube $(0, 1)^3$ and denote by h the mesh size of each one. Then, we compute $\|\mathbf{u} - \mathbf{u}^{NN}\|_{\tilde{\Omega}}$ and $\|\mathbf{u} - \mathbf{u}_h\|_{\tilde{\Omega}}$, where $\|\cdot\|_{\tilde{\Omega}}$ denotes the $L^2(\tilde{\Omega})^3$ -norm and \mathbf{u}_h denotes the best approximation of \mathbf{u} into $\mathcal{N}_0(\tilde{\Omega})$, with respect to the $H(\text{curl}; \tilde{\Omega})$ -norm.

In Figure 3.3(d), we observe that, for the DNN with 2 layers and for large h , the DNN approach gives a better approximation than the Lowest Order Nédélec space. We further notice that, as h gets smaller, \mathbf{u}^{NN} needs to be evaluated at more points in $\tilde{\Omega} \setminus \Omega$ and it is not obvious if the DNN approximation will remain stable. However,

Figure 3.3(a), 3.3(b), and 3.3(d) show that the DNN approximation remains stable.

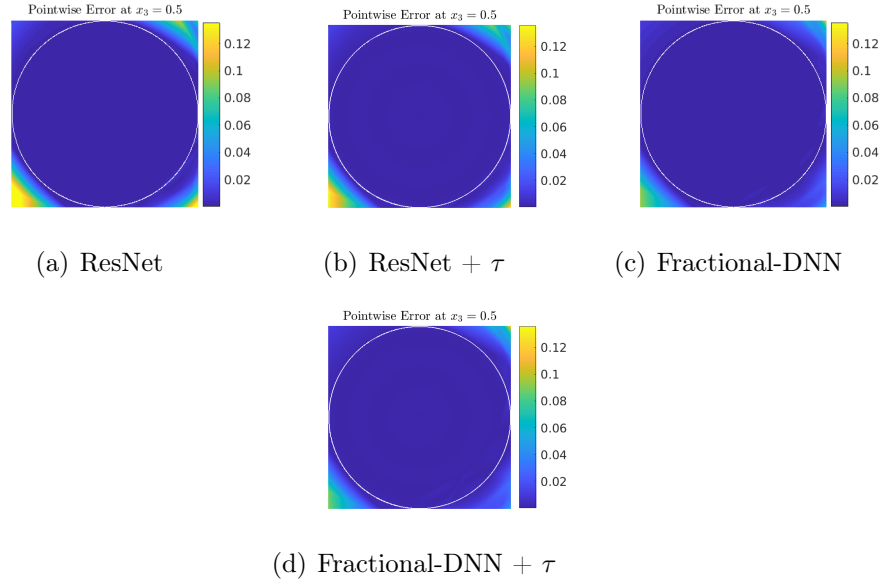


Figure 3.4: Comparison of testing errors between ResNet, ResNet with τ -learning framework, Fractional-DNN and Fractional-DNN with τ -learning framework (2-50).

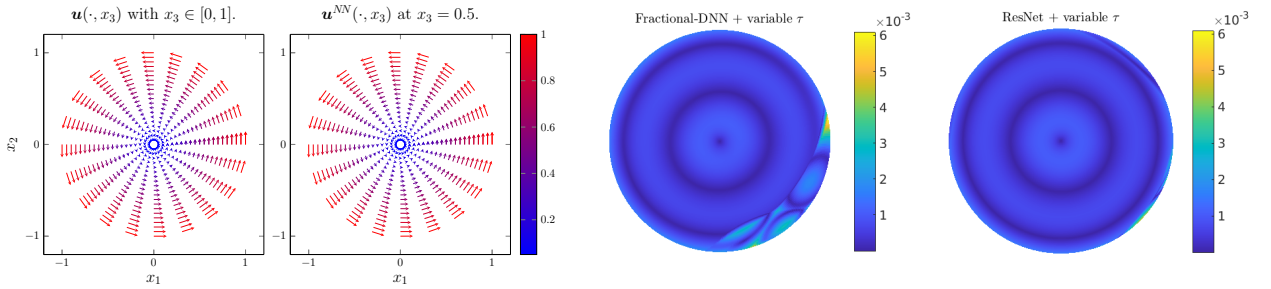


Figure 3.5: \mathbf{u} , \mathbf{u}^{NN} and pointwise error on Ω , at $x_3 = 0.5$ (x_1x_2 -plane).

There are several ways to measure how well a neural network performs. For instance, in the case of 2 layers and 50 nodes, the training error is slightly better with the ResNet-based architectures, cf. Figure 3.3(c), while a smaller error on unseen data is achieved with Fractional-DNNs, cf. Figure 3.3(d). Furthermore, when we plot the error on the square $(-1, 1)^2 \times \{0.5\}$, we see that Fractional-DNNs, cf. Figure 3.4(c) and 3.4(d), extrapolate better than ResNets, cf. Figure 3.4(a) and 3.4(b). In the ResNet setting, employing the τ -learning framework (Figure 3.4(b)) yields to slightly better

results than fixing τ (Figure 3.4(a)). Additionally, from its definition, we know $\mathbf{u}_3 \equiv 0$, cf. (3.7.2). Hence, we present quiver plots of the first two components of $\mathbf{u}(\cdot, x_3)$ for any x_3 , and $\mathbf{u}^{NN}(\cdot, 0.5)$, in Figure 3.5. The two plots seem to coincide. Therefore, we further present pointwise errors restricted to Ω , where the pointwise error is measured in the (\mathbb{R}^2) Euclidean norm. For Fractional-DNN with variable τ (2-50) and no bias ordering inside Ω , we observe that $-0.0051 \leq \mathbf{u}_3^{NN}(\mathbf{x}) \leq 0.0077$, i.e., the constraint violation is of the order of approximation error. Meanwhile, ResNet with variable τ achieves better results in this test case.

Besides improving the training, the approximation could be further improved if we knew a priori some qualitative properties of the exact solution. Then, they could be forced into the loss functional, similarly, as it is done with PINNs; cf. [39]. It is important to mention that several other numerical examples were considered to test the robustness of our τ -variable framework. For instance, we considered problems where the standard Neural ODEs (cf. [47]) struggle to obtain good approximations, as pointed out in [59]. We obtained similar results to the ones presented here. For the sake of brevity, those results have been excluded.

Conclusion

A time-variable learning framework for DNNs has been presented, which can be used to almost any DNN. However, from a mathematical perspective, DNN architectures which can be related to dynamical systems are of interest, since learning τ then corresponds to optimal adaptive time stepping. Consequently, special emphasis has been put on applying the τ -variable framework to ResNet and Fractional-DNN. The τ -variable framework is argued to overcome vanishing and exploding gradient challenges. The numerical results suggest that DNNs with τ -variable framework outperform their counterparts with fixed τ for deep architectures and enjoy an improved training error decay. Moreover, this method has the potential to identify redundant layers so that the network size can be reduced while maintaining the quality of the prediction.

Chapter 4

NONLOCAL BOUNDED VARIATIONS WITH APPLICATIONS

4.1 Introduction

Fractional calculus and nonlocal operators have emerged as natural tools for studying numerous phenomena in science and engineering in recent years. Fractional operators differ from their classical counterparts in various ways, including the fact that they need less smoothness and are nonlocal in nature. Such flexibilities have led to multiple successes of fractional derivative-based models in practical applications. For instance, magnetotellurics in geophysics [148], viscoelastic models [109], quantum spin chains and harmonic maps [54, 100, 5], deep neural networks [12], repulsive curves [156], etc.

A fundamental concept in inverse problems, such as image denoising, is the use of regularization. The article [4] introduced the fractional Laplacian as a regularizer in image denoising as an alternative to well-known approaches such as total-variation regularization. Subsequently, this model has been successfully used by various authors in imaging science as it provides a behavior that is closer to total variation based approaches [85], yet it is simple to implement in practice. The current approach is motivated by these observations. We also refer to [68] for a different (discrete) nonlocal regularization in imaging.

Fundamental developments are being made in fractional calculus. In fact, now there are notions of fractional divergence and gradient. For example, the aforementioned fractional Laplacian, can be obtained by the composition of fractional divergence and fractional gradient. This is similar to the classical integer-order setting. Such discoveries are not only fueling further developments in analysis but are also leading to

new application areas or improving the existing ones. Inspired by image denoising, the goal of this work is to study the fundamental properties of the space of (nonlocal) fractional bounded variation. Based on such fractional order spaces, we introduce novel image denoising models, and we derive Fenchel dual formulations [62, chapter III] for these. Notice that such formulations are critical in deriving efficient numerical methods in the classical setting. To motivate the analytical tools developed in this work, the remainder of this section provides a detailed discussion of new image denoising models.

Total variation minimization is a well-established method for solving image denoising problems [3, 127, 128]. Let $u_N : \Omega \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ denote a continuous representation of an image (possibly noisy). Given a regularization parameter $\beta > 0$, a standard image denoising problem amounts to finding u solving

$$\arg \min_{u \in \mathcal{X}} \left\{ \beta |Du|_{\mathcal{X}} + \frac{1}{2} \|u - u_N\|_{L^2(\Omega)}^2 \right\}, \quad (4.1.1)$$

where the space \mathcal{X} is chosen in conjunction with the norm $|\cdot|_{\mathcal{X}}$ such that Du is well defined at least in a distributional sense, and u can be piecewise smooth. One of the most popular spaces used in practice is the space of functions with *bounded variation* (BV), defined by

$$\text{BV}(\Omega) = \{u \in L^1(\Omega) : \text{Var}(u; \Omega) < \infty\}.$$

Namely, a function u in $L^1(\Omega, \mathbb{R})$ is said to have bounded variation if and only if

$$\text{Var}(u; \Omega) := \sup \left\{ \int_{\mathbb{R}^n} u(x) \text{Div} \Phi(x) dx : \Phi \in C_c^1(\Omega, \mathbb{R}^n), \|\Phi\|_{L^\infty(\Omega)} \leq 1 \right\} < \infty.$$

If the variation $\text{Var}(u; \Omega)$ is finite, one can show that its distributional derivative Du is a Radon measure and $\text{Var}(u; \Omega) = |Du|(\Omega)$, see [14, Ch. 10]. It is well known that $\text{BV}(\Omega)$ preserves edges better than $W^{1,1}(\Omega)$ in a noisy images while retaining several of its properties. For instance, it is a Banach space; it is lower semi-continuous on $L^1(\Omega)$; Sobolev inequalities; etc.

In this work, we are interested in the fractional version of the problem (4.1.1). For this, we first need to decide on the notion of fractional *BV*. We accomplish this by

replacing the derivative in the preceding definition with a suitable fractional derivative. Alas, there are many different, yet natural, fractional operators that are considered extensions of the usual gradient – and each one induces its own BV -space.

We will consider the two most popular notions. Firstly, we will consider the space BV^α , which we refer to as *Riesz-type*. The study of BV^α was initiated by Comi-Stefani in [52], see Section 4.2. It relies on the notion of what is sometimes referred to as *Riesz gradient* D^α , which is simply the usual gradient combined with a regularizing Riesz potential.

The other type of fractional BV we consider will be denoted by bv^α and is referred to as *Gagliardo-type*; see Section 4.3. We are not aware whether this has been considered in the literature prior to this work. The notion of a fractional derivative is what we will refer to as the Gagliardo-type derivative, which is considered in various aspects of mathematics, e.g., Dirichlet forms [80], peridynamics [58], and harmonic analysis [111]. This Gagliardo-type bv^α is naturally related to the most popular notion of a fractional perimeter defined by Caffarelli–Roquejoffre–Savin [38]. Indeed, we will show in Theorem 4.3.4 that bv^α coincides with the Gagliardo-Sobolev space $W^{\alpha,1}$ – a maybe surprising feature of the case $\alpha < 1$, since this is false for $\alpha = 1$: indeed it is well-known that $W^{1,1} \neq BV$, see [64]. This is one of the main theoretical contributions of the current study.

We will introduce new types of variational models for image denoising based on these fractional BV notions. Namely, we study the fractional versions of (4.1.1),

$$\arg \min_{u \in \mathcal{X}} \left\{ \beta \text{Var}_\alpha(u; \Omega) + \frac{\gamma}{p} \|u - u_N\|_{L^p(\Omega)}^p \right\}. \quad (4.1.2)$$

A related model was studied by Bartels and one of the authors in [4], but working in fractional order Hilbert space $H^s(\Omega)$ rather than $\mathcal{X} = BV^\alpha(\Omega)$.

We emphasize that the numerical algorithms for solving problems of type (4.1.1) make extensive use of the Fenchel dual formulations [17, 42]. However, this requires dealing with the dual space of $BV(\Omega)$, whose full characterization is still unknown [144]. Instead, one proceeds by finding a predual problem to (4.1.1), i.e., a problem

whose Fenchel conjugate is (4.1.1); see, for instance, [37, 41, 78]. In this case, one does not need to deal with $\text{BV}(\Omega)^*$ but instead the closure in $L^p(\Omega)$ of the range of a divergence-like operator, which is the conjugate of $-D : \mathcal{X} \subset \text{BV}(\Omega) \rightarrow \mathcal{M}(\Omega, \mathbb{R}^n)$. We will derive a pre-dual problem corresponding to (4.1.2) in Section 4.4. Derivation of pre-dual requires density of smooth functions with compact support. This is highly non-trivial in general, even in the local case. We establish this result, provided that the domain Ω is convex. Such results are of interest by themselves, see Propositions 4.4.4 and 4.4.8.

4.2 Fractional BV in the Riesz sense

To begin, consider the fractional Laplacian and its inverse, the Riesz potential. Denote by \mathcal{F} and \mathcal{F}^{-1} the Fourier transform on \mathbb{R}^n . For $\alpha > 0$ the fractional Laplacian of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with differential order α , denoted by $|D|^\alpha f$, is given by

$$|D|^\alpha f(x) := \mathcal{F}^{-1} (|\xi|^\alpha \mathcal{F} f(\xi)) (x).$$

The notation $|D|^\alpha = (-\Delta)^{\frac{\alpha}{2}}$ is common, but we will mostly use the notation $|D|^\alpha$ in this chapter, since it states the order of derivatives more clearly. The definition above also makes sense when $\alpha < 0$. In that case, we call the operator Riesz potential. More precisely, for all $\alpha \in (0, n)$ we define

$$I^\alpha f(x) := \mathcal{F}^{-1} (|\xi|^{-\alpha} \mathcal{F} f(\xi)) (x).$$

It is then easy to see that $|D|^\alpha I^\alpha f = I^\alpha |D|^\alpha f = f$, at least for suitably smooth functions with decay at infinity, i.e. the fractional Laplacian and Riesz potential are inverses to each other. The fractional Laplacian $|D|^\alpha$ has no gradient structure. It does not converge to the gradient D when $\alpha \rightarrow 1$. Recently, many authors considered a fractional-order operator with a gradient structure. Although this operator may be traced as far back as [83], it has received increased interest in various applications since the works e.g., [52, 131, 132, 135]. It is defined very simply as the usual gradient of the Riesz potential

$$D^\alpha f := D I^{1-\alpha} f. \tag{4.2.1}$$

From its Fourier transform representation, it is easy to show that $D^\alpha \rightarrow D$ as $\alpha \rightarrow 1$. The fractional divergence Div_α is defined as

$$\text{Div}_\alpha f = \text{div } I^{1-\alpha} f.$$

Note that Div_α is the adjoint of $-D^\alpha$. In fact, the following integration-by-parts formula holds

$$\int_{\mathbb{R}^n} F \cdot D^\alpha g dx = - \int_{\mathbb{R}^n} \text{Div}_\alpha F g dx \quad \forall F \in C_c^\infty(\mathbb{R}^n, \mathbb{R}^n), \forall g \in C_c^\infty(\mathbb{R}^n), \quad (4.2.2)$$

which follows readily from the definition via the Fourier transform and Plancherel's theorem. We comment on the integral definition of the above operators. For any $\alpha \in (0, 1]$, we have

$$\begin{aligned} |D|^\alpha f(x) &= c_{1,\alpha} \int_{\mathbb{R}^n} \frac{f(x) - f(y)}{|x - y|^{n+\alpha}} dy, \\ D^\alpha f(x) &= c_{2,\alpha} \int_{\mathbb{R}^n} \frac{(f(x) - f(y))(x - y)}{|x - y|^{n+\alpha+1}} dy, \\ \text{Div}_\alpha F(x) &= c_{3,\alpha} \int_{\mathbb{R}^n} \frac{(F(x) - F(y)) \cdot (x - y)}{|x - y|^{n+\alpha+1}} dy, \end{aligned} \quad (4.2.3)$$

for some constants $c_{1,\alpha}$, $c_{2,\alpha}$ and $c_{3,\alpha}$, which can be found in the literature. Having the notion of a fractional gradient, we naturally obtain the notion of fractional BV spaces. Our definitions are very similar to [52] and different from other natural approaches as in [28] or an approach via a different type of nonlocal gradient and divergence as in [58, 111], which we will discuss later in Section 4.3. When employing the concept of fractional BV spaces in this work, most of the needed properties will follow the same basic principles as in conventional BV spaces. We give a derivation of the results that we were unable to find in the literature, and we provide references otherwise. Some of these results may already be known to experts.

To distinguish the resulting space from the one discussed in Section 4.3, we use the notations BV^α , Div_α and Var_α . In Section 4.3 we will use bv^α , div_α , and var_α instead.

Let $\alpha \in (0, 1]$ and $f \in L^1(\mathbb{R}^n)$, the variation of f is defined as

$$\text{Var}_\alpha(f; \mathbb{R}^n) := \sup \left\{ \int_{\mathbb{R}^n} f \text{Div}_\alpha \Phi \, dx : \Phi \in C_c^1(\mathbb{R}^n; \mathbb{R}^n), \|\Phi\|_{L^\infty(\mathbb{R}^n)} \leq 1 \right\}. \quad (4.2.4)$$

Let $\Omega \subseteq \mathbb{R}^n$. For any $f \in L^1(\Omega)$, we define

$$\text{Var}_\alpha(f; \Omega) := \text{Var}_\alpha(\chi_\Omega f; \mathbb{R}^n),$$

where $\chi_\Omega f$ is the extension of f by zero to \mathbb{R}^n . The integral

$$\int_{\mathbb{R}^n} f \text{Div}_\alpha \Phi \, dx$$

is well defined for all $f \in L^1(\mathbb{R}^n)$ and $\Phi \in C_c^1(\mathbb{R}^n, \mathbb{R}^n)$, which is a consequence of the following result.

Lemma 4.2.1. *Let $\Phi \in C_c^1(\mathbb{R}^n; \mathbb{R}^n)$, then for any $\alpha \in (0, 1]$ and any $p \in [1, \infty]$ we have*

$$\text{Div}_\alpha \Phi \in L^p(\mathbb{R}^n).$$

Proof. Fix $\Phi \in C_c^1(\mathbb{R}^n; \mathbb{R}^n)$. For $\alpha = 1$, we have $\text{Div}_\alpha \Phi \in C_c(\mathbb{R}^n) \subseteq L^p(\mathbb{R}^n)$ for all $p \in [1, \infty]$. For $\alpha < 1$, we have from (4.2.3) that

$$|\text{Div}_\alpha \Phi(x)| \lesssim_\alpha (2\|\Phi\|_{L^\infty(\mathbb{R}^n)} + \|\nabla \Phi\|_{L^\infty(\mathbb{R}^n)}) \int_{\mathbb{R}^n} \frac{\min\{1, |x-y|\}}{|x-y|^{n+\alpha}} dy.$$

Here \lesssim_α implies that the hidden constant depends on α (and any constant may depend on the dimension n). Since $\alpha < 1$, the following integral is finite and has the same value for every $x \in \mathbb{R}^n$, i.e.,

$$\int_{\mathbb{R}^n} \frac{\min\{1, |x-y|\}}{|x-y|^{n+\alpha}} dy \equiv C(n, \alpha) < \infty,$$

which implies that

$$\|\text{Div}_\alpha \Phi\|_{L^\infty(\mathbb{R}^n)} \lesssim_\alpha (\|\Phi\|_{L^\infty(\mathbb{R}^n)} + \|\nabla \Phi\|_{L^\infty(\mathbb{R}^n)}).$$

It remains to prove that $\text{Div}_\alpha \Phi \in L^1(\mathbb{R}^n)$. Once this is shown we conclude $\text{Div}_\alpha \Phi \in L^p(\mathbb{R}^n)$ for any $p \in [1, \infty]$ by interpolation. Taking $R \geq 1$ large enough, such that $\text{supp } \Phi \subset B(0, R/2)$, then for $x \in \mathbb{R}^n \setminus B(0, R)$ we have

$$|\text{Div}_\alpha \Phi(x)| \lesssim_\alpha \int_{B(0, R/2)} \frac{|\Phi(y)|}{|x-y|^{n+\alpha}} dy.$$

By Fubini's theorem, we have

$$\begin{aligned} \|\operatorname{Div}_\alpha \Phi\|_{L^1(\mathbb{R}^n \setminus B(0,R))} &\lesssim \int_{B(0,R/2)} |\Phi(y)| \left(\int_{\mathbb{R}^n \setminus B(0,R)} \frac{1}{|x-y|^{n+\alpha}} dx \right) dy \\ &\lesssim \|\Phi\|_{L^1(\mathbb{R}^n)} \sup_{y \in \mathbb{R}^n \setminus B(0,R/2)} \left(\int_{\{x:|x-y| \geq R/2\}} \frac{1}{|x-y|^{n+\alpha}} dx \right). \end{aligned}$$

Here we hide the constant by using \lesssim . Using the fact that

$$\int_{\{x:|x-y| \geq R/2\}} \frac{1}{|x-y|^{n+\alpha}} dx \lesssim_\alpha R^{-\alpha} < \infty,$$

we obtain

$$\|\operatorname{Div}_\alpha \Phi\|_{L^1(\mathbb{R}^n \setminus B(0,R))} \lesssim_\alpha \|\Phi\|_{L^1(\mathbb{R}^n)}.$$

On the complement $B(0, R)$, we have $\operatorname{Div}_\alpha \Phi \in L^\infty(B(0, R)) \subset L^1(B(0, R))$. Thus, we obtain that $\|\operatorname{Div}_\alpha \Phi\|_{L^1(\mathbb{R}^n)} < \infty$, which finishes the proof. \square

Now we are ready to define the first fractional BV space of this work, i.e., BV^α ; see also [52, 51, 50] where this space was considered first. This space inherits most of its properties from the gradient structure of the Riesz-derivative D^α , cf. (4.2.1).

Definition 4.2.2 (Riesz-type fractional BV). *For $\Omega \subset \mathbb{R}^n$, we define*

$$BV_{00}^\alpha(\Omega) := \{f \in L^1(\mathbb{R}^n) : f \equiv 0 \text{ on } \mathbb{R}^n \setminus \Omega, \operatorname{Var}_\alpha(f; \Omega) < \infty\}, \quad (4.2.5)$$

endowed with the norm

$$\|f\|_{BV^\alpha(\Omega)} := \|f\|_{L^1(\Omega)} + \operatorname{Var}_\alpha(f; \Omega).$$

In this chapter, we often identify $f \in L^1(\Omega)$ with its extension by zero $\chi_\Omega f \in L^1(\mathbb{R}^n)$. Observe that we do not need to assume any regularity of $\partial\Omega$ in the above (and following) definitions and results. The regularity of $\partial\Omega$ is only relevant for determining if constant functions in Ω belong to $BV^\alpha(\Omega)$. Namely $1 \in L^1(\Omega)$ belongs to $BV_{00}^\alpha(\Omega)$ (with the usual identification $1 \in L^1(\Omega)$ corresponds to $\chi_\Omega \in L^1(\mathbb{R}^n)$) if the α -Cacciopoli-perimeter of $\partial\Omega$ is finite. We refer to [52] for the definition of this perimeter. Essentially by definition we immediately obtain

Proposition 4.2.3. *The surface $\partial\Omega$ has finite α -Cacciopoli-perimeter, i.e. $\text{Per}_\alpha(\partial\Omega) < \infty$ if and only if $\text{Var}_\alpha(1; \Omega) < \infty$.*

Observe that the Cacciopoli-perimeter above is different from the more commonly used fractional perimeter introduced by Caffarelli-Roquejoffre-Savin [38]. The latter one is related to the fractional version of BV functions defined using the divergence as used in, e.g. [58, 111]. We shall discuss it in Section 4.3.

Next, we note that one can obtain the existence of the distributional derivative $D^\alpha f$ (which is a Radon measure) just like for BV , see [64, p.167, Theorem 1, Structure Theorem] If $f \in BV_{00}^\alpha(\Omega)$, then the mapping

$$C_c^1(\mathbb{R}^n; \mathbb{R}^n) \ni \Phi \mapsto \int_{\mathbb{R}^n} f \text{Div}_\alpha \Phi dx$$

extends to a linear functional on $(C_c(\mathbb{R}^n; \mathbb{R}^n), \|\cdot\|_{L^\infty(\mathbb{R}^n)})$. By the Riesz representation theorem [64, Section 1.8, Theorem 1], there exists a Radon measure μ on \mathbb{R}^n and a μ -measurable function $\sigma : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $|\sigma| = 1$ μ -a.e. and

$$\int_{\mathbb{R}^n} f \text{Div}_\alpha \Phi dx = - \int_{\mathbb{R}^n} \Phi \cdot \sigma d\mu.$$

Moreover, we have

$$|\mu(\mathbb{R}^n)| \leq \text{Var}_\alpha(f; \Omega).$$

The latter follows by the definition of the norm. By slight abuse of notation we will denote by $D^\alpha f$ both the distributional derivative and the measure $D^\alpha f := \sigma \llcorner \mu$ (where \llcorner denotes the concatenation of function and measure), whichever is applicable.

We now consider the approximation of $BV_{00}^\alpha(\Omega)$ functions by smooth functions. Since f is compactly supported, the convolution $f * \eta_\varepsilon$ is in $C_c^\infty(\mathbb{R}^n)$. Using the same argument as in [64, Theorem 5.2], we obtain the following result.

Proposition 4.2.4. *Let $\Omega \subset \mathbb{R}^n$ be open and bounded. For any $f \in BV_{00}^\alpha(\Omega)$ there exists $f_k \in C_c^\infty(\mathbb{R}^n)$ such that*

$$\|f_k - f\|_{L^1(\mathbb{R}^n)} + |\text{Var}_\alpha(f; \mathbb{R}^n) - \text{Var}_\alpha(f_k; \mathbb{R}^n)| \xrightarrow{k \rightarrow \infty} 0.$$

Equivalently, (since f vanishes outside of Ω),

$$\|f_k - f\|_{L^1(\mathbb{R}^n)} + |\mathrm{Var}_\alpha(f; \Omega) - \mathrm{Var}_\alpha(f_k; \mathbb{R}^n)| \xrightarrow{k \rightarrow \infty} 0.$$

We also have the following embedding theorem.

Proposition 4.2.5. *Let $\Omega \subset \mathbb{R}^n$ be open and bounded and $n \geq 2$. Then for all $p \in [1, \frac{n}{n-\alpha}]$ we have $BV_{00}^\alpha(\Omega) \subseteq L^p(\mathbb{R}^n)$ and*

$$\|f\|_{L^p(\mathbb{R}^n)} \leq C(n, p, \alpha) \|f\|_{BV_{00}^\alpha(\Omega)}.$$

If $n = 1$, then the same results hold for all $p \in [1, \frac{1}{1-\alpha})$.

Proof. Let f_k be the approximation of f as in Proposition 4.2.4. By the main result in [133], we have for all $p \in [1, \frac{n}{n-\alpha}]$

$$\|f_k\|_{L^p(\mathbb{R}^n)} \leq C (\|f_k\|_{L^1(\mathbb{R}^n)} + \|D^\alpha f_k\|_{L^1(\mathbb{R}^n)}),$$

since $f_k \in C_c^\infty(\mathbb{R}^n)$. Observe that by an integration-by-parts formula, since we already know $D^\alpha f_k \in L^1(\mathbb{R}^n, \mathbb{R}^n)$,

$$\|D^\alpha f_k\|_{L^1(\mathbb{R}^n, \mathbb{R}^n)} = \mathrm{Var}_\alpha(f_k; \mathbb{R}^n).$$

Since up to subsequences f_k converges to f almost everywhere we conclude from Fatou's lemma,

$$\begin{aligned} \|f\|_{L^p(\mathbb{R}^n)} &\leq \liminf_{k \rightarrow \infty} \|f_k\|_{L^p(\mathbb{R}^n)} \leq C \liminf_{k \rightarrow \infty} (\|f_k\|_{L^1(\mathbb{R}^n)} + \mathrm{Var}_\alpha(f_k; \mathbb{R}^n)) \\ &= C (\|f\|_{L^1(\mathbb{R}^n)} + \mathrm{Var}_\alpha(f; \mathbb{R}^n)), \end{aligned}$$

which concludes the proof. □

Using the duality definition of Var_α and the same argument as in [64, Theorem 5.2], we obtain the lower semicontinuity with respect to the so-called intermediate convergence; see Definition 10.1.3 and Remark 10.1.3 in [14] for details.

Proposition 4.2.6 (Lower semicontinuity). *Let $\Omega \subset \mathbb{R}^n$ be open and bounded. Assume $\{f_k\}_{k=1}^\infty \subset BV_{00}^\alpha(\Omega)$, and assume that $f \in L^1(\mathbb{R}^n)$ such that*

$$\|f_k - f\|_{L^1(\mathbb{R}^n)} \xrightarrow{k \rightarrow \infty} 0.$$

Then $f \in BV_{00}^\alpha(\Omega)$ and we have

$$\text{Var}_\alpha(f; \mathbb{R}^n) \leq \liminf_{k \rightarrow \infty} \text{Var}_\alpha(f_k; \mathbb{R}^n).$$

Or, equivalently,

$$\text{Var}_\alpha(f; \Omega) \leq \liminf_{k \rightarrow \infty} \text{Var}_\alpha(f_k; \Omega).$$

Corollary 4.2.7. *Let $\Omega \subset \mathbb{R}^n$ be bounded. Then $(BV_{00}^\alpha(\Omega), \|\cdot\|_{BV^\alpha(\Omega)})$ is a complete space.*

Proof. Let $\{f_k\}_{k=1}^\infty$ be a Cauchy sequence in $BV_{00}^\alpha(\Omega)$. Since f_k is Cauchy in $L^1(\mathbb{R}^n)$, there exists $f \in L^1(\mathbb{R}^n)$ with $f \equiv 0$ in $\mathbb{R}^n \setminus \Omega$, such that $f_k \rightarrow f$ in $L^1(\mathbb{R}^n)$. By Proposition 4.2.6, we find that $f \in BV_{00}^\alpha(\Omega)$. Using the lower semicontinuity of the variation still from Proposition 4.2.6, we obtain

$$\lim_{k \rightarrow \infty} \text{Var}_\alpha(f - f_k; \Omega) \leq \lim_{k \rightarrow \infty} \liminf_{\ell \rightarrow \infty} \text{Var}_\alpha(f_\ell - f_k; \Omega) = 0,$$

which completes the proof. □

Using the weak*-convergence of Radon measures, and the arguments of the standard Rellich-Kondrachov compactness, see [64, Theorem 5.2 & Theorem 5.5], we have the following result.

Proposition 4.2.8 (Weak compactness). *Let $\Omega \subset \mathbb{R}^n$ be open and bounded. Assume $\{f_k\}_{k=1}^\infty \subset BV_{00}^\alpha(\Omega)$ such that*

$$\sup_{k \geq 1} \|f_k\|_{BV^\alpha(\Omega)} < \infty.$$

Then there exists $f \in BV_{00}^\alpha(\Omega)$ such that

$$\text{Var}_\alpha(f; \mathbb{R}^n) \leq \liminf_{k \rightarrow \infty} \text{Var}_\alpha(f_k; \mathbb{R}^n),$$

or equivalently,

$$\mathrm{Var}_\alpha(f; \Omega) \leq \liminf_{k \rightarrow \infty} \mathrm{Var}_\alpha(f_k; \Omega),$$

and there is a subsequence $\{f_{k_i}\}_{i=1}^\infty$ such that for all $p \in \left[1, \frac{n}{n-\alpha}\right)$ we have

$$\|f_{k_i} - f\|_{L^p(\mathbb{R}^n)} \xrightarrow{i \rightarrow \infty} 0.$$

Lastly, as in the local case where we know that $H^{1,1}(\Omega)$ is a subspace of $BV(\Omega)$ (where $H^{1,1}(\Omega)$ is the space of functions $f \in L^1(\Omega)$ such that $Df \in L^1(\Omega)$), the corresponding result for the fractional situation holds as well.

Lemma 4.2.9. *Let $f \in H^{\alpha,1}(\mathbb{R}^n)$, i.e., $f \in L^1(\mathbb{R}^n)$ and $D^\alpha f \in L^1(\mathbb{R}^n; \mathbb{R}^n)$. Assume additionally that $f \equiv 0$ in $\mathbb{R}^n \setminus \Omega$. Then $f \in BV_{00}^\alpha(\Omega)$.*

Proof. We only need to show $\mathrm{Var}_\alpha(\chi_\Omega f; \mathbb{R}^n) < \infty$. For any $\Phi \in C_c^1(\mathbb{R}^n; \mathbb{R}^n)$ such that $\|\Phi\|_{L^\infty(\mathbb{R}^n)} \leq 1$, we have by Fubini's theorem

$$\int_{\mathbb{R}^n} \chi_\Omega f \mathrm{Div}_\alpha \Phi = - \int_{\mathbb{R}^n} D^\alpha f \cdot \Phi \leq \|\Phi\|_{L^\infty(\mathbb{R}^n)} \|D^\alpha f\|_{L^1(\mathbb{R}^n)} \leq \|D^\alpha f\|_{L^1(\mathbb{R}^n)},$$

which implies that $\mathrm{Var}_\alpha(\chi_\Omega f; \mathbb{R}^n) < \infty$. □

4.3 Fractional BV in the Gagliardo sense

The notion of fractional BV from Section 4.2 (as in [52]) is very similar to the usual BV , since it is essentially a lifting by the Riesz potential. In this section, we introduce another natural notion, which is denoted by bv^α . This notion recovers the fractional perimeter as defined by Caffarelli-Roquejoffre-Savin in [38]. We begin by introducing a different type of fractional divergence, as defined in [111]. We stress that related notions were known before [58] and are often utilized in the theory of Dirichlet forms, see [80].

A (nonlocal) vector-field F on \mathbb{R}^n is defined as an $\mathcal{L}^n \times \mathcal{L}^n$ -measurable map $F : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$, which is additionally antisymmetric, i.e., $F(x, y) = -F(y, x)$. As in [111] the set of such vector-fields is denoted by $\mathcal{M}(\bigwedge_{od} \mathbb{R}^n)$, where od stands for off-diagonal and (as in the theory of Dirichlet forms) \bigwedge_{od} stands for a sort of one-form

(we will not really use this aspect, we recommend the reader to take it as a purely notational choice).

We say that $F \in L^p(\Lambda_{od} \mathbb{R}^n)$ if $F \in \mathcal{M}(\Lambda_{od} \mathbb{R}^n)$ and

$$\|F\|_{L^p(\Lambda_{od} \mathbb{R}^n)} := \left(\int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \frac{|F(x, y)|^p}{|x - y|^n} dx dy \right)^{\frac{1}{p}} < \infty$$

for $p \in [1, \infty)$, and

$$\|F\|_{L^\infty(\Lambda_{od} \mathbb{R}^n)} := \operatorname{ess\,sup}_{x, y \in \mathbb{R}^n} |F(x, y)| < \infty$$

for $p = \infty$. For $\Omega \subset \mathbb{R}^n$, we say $F \in L^p_{00}(\Lambda_{od} \Omega)$ if $F \in L^p(\Lambda_{od} \mathbb{R}^n)$ and $F(x, y) = 0$ for \mathcal{L}^n -a.e. $x \in \mathbb{R}^n \setminus \Omega$ (and thus for a.e. $y \in \mathbb{R}^n \setminus \Omega$).

The (Gagliardo sense) fractional derivative d_α , which has similar properties to a function's gradient, takes an \mathcal{L}^n -measurable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ into a vector-field

$$(d_\alpha f)(x, y) := \frac{f(x) - f(y)}{|x - y|^\alpha}.$$

Let us remark that if one were to consider stability as $\alpha \rightarrow 1$, then it would make more sense to set

$$(d_\alpha f)(x, y) := (1 - \alpha) \frac{f(x) - f(y)}{|x - y|^\alpha}.$$

However, for the purpose of clarity, we shall not utilize this term.

The scalar product of two vectorfields F and G is given by:

$$(F \cdot G)(x) := \int_{\mathbb{R}^n} \frac{F(x, y)G(x, y)}{|x - y|^n} dy. \quad (4.3.1)$$

The fractional divergence $\operatorname{div}_\alpha$ is then the formal adjoint to $-d_\alpha$ with respect to the $L^2(\mathbb{R}^n)$ scalar product, i.e., for all $\varphi \in C_c^\infty(\mathbb{R}^n)$, we have

$$\int_{\mathbb{R}^n} \operatorname{div}_\alpha F \varphi dx := - \int_{\mathbb{R}^n} F \cdot d_\alpha \varphi dx = - \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \frac{F(x, y)(\varphi(x) - \varphi(y))}{|x - y|^{n+\alpha}} dy dx. \quad (4.3.2)$$

The multiplication of a scalar function $f(x)$ and a vector field $F(x, y)$ is defined as:

$$(fF)(x, y) := \frac{f(x) + f(y)}{2} F(x, y). \quad (4.3.3)$$

Using (4.3.2), we can obtain the integral formula of $\operatorname{div}_\alpha$. By antisymmetry $F(x, y) = -F(y, x)$ and the Fubini's theorem, we have

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^n} F(x, y) \frac{\varphi(x) - \varphi(y)}{|x - y|^{n+\alpha}} dy dx = 2 \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \frac{F(x, y)}{|x - y|^{n+\alpha}} dy \varphi(x) dx,$$

which enables us to give the integral definition of $\operatorname{div}_\alpha F$ by

$$(\operatorname{div}_\alpha F)(x) := -2 \int_{\mathbb{R}^n} \frac{F(x, y)}{|x - y|^{n+\alpha}} dy = - \int_{\mathbb{R}^n} \frac{F(x, y) - F(y, x)}{|x - y|^{n+\alpha}} dy.$$

In what follows, by yet another slight abuse of notation we are going to use this formulation even when $F(x, y) \neq -F(y, x)$:

$$(\operatorname{div}_\alpha F)(x) := - \int_{\mathbb{R}^n} \frac{F(x, y) - F(y, x)}{|x - y|^{n+\alpha}} dy. \quad (4.3.4)$$

It was shown in [111] how this fractional divergence naturally appears and leads to conservation laws and div-curl type results in the theory of fractional harmonic maps.

The Fourier transform can be used to verify that

$$(-\Delta)^\alpha f = -c \operatorname{div}_\alpha(d_\alpha f) \quad (4.3.5)$$

for some constant $c = c(n, \alpha)$.

With the fractional divergence $\operatorname{div}_\alpha$, we may define the fractional bounded variation in the Gagliardo sense.

Definition 4.3.1 (Gagliardo-type fractional BV). *Let $f \in L^1_{loc}(\mathbb{R}^n)$. For an open set $\Omega \subset \mathbb{R}^n$, we define*

$$\operatorname{var}_\alpha(f; \Omega) := \sup \left\{ \int_{\mathbb{R}^n} f \operatorname{div}_\alpha \Phi dx : \Phi \in C_c^\infty(\Omega \times \Omega), \|\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq 1 \right\}.$$

Observe that this is equivalent to

$$\operatorname{var}_\alpha(f; \Omega) = \sup \left\{ \int_{\Omega} f \operatorname{div}_\alpha \Phi dx : \Phi \in C_c^\infty(\Omega \times \Omega), \|\Phi\|_{L^\infty(\Omega \times \Omega)} \leq 1 \right\}.$$

We say that $f \in bv^\alpha(\Omega)$ if

$$\|f\|_{bv^\alpha(\Omega)} := \|f\|_{L^1(\Omega)} + \operatorname{var}_\alpha(f; \Omega) < \infty.$$

The notion $\text{var}_\alpha(f; \Omega)$ is well-defined by the following observations: first, in order to have consistency, we observe that

Lemma 4.3.2. *Let $\Phi \in C_c^1(\mathbb{R}^n \times \mathbb{R}^n)$, then for all $\alpha \in (0, 1)$ and all $p \in [1, \infty]$, we have*

$$\text{div}_\alpha \Phi \in L^p(\mathbb{R}^n).$$

Observe that we exclude the case $\alpha = 1$ since $\text{div}_\alpha \Phi$ is not well defined for $\alpha = 1$. A multiplication with $(1 - \alpha)$ would lead to a stable theory as $\alpha \rightarrow 1$.

Proof. Observe that by differentiability of Φ ,

$$|\Phi(x, y) - \Phi(y, x)| \leq |\Phi(x, y) - \Phi(x, x)| + |\Phi(x, x) - \Phi(y, x)| \leq 2\|D\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} |x - y|.$$

Then, using a similar argument as in Lemma 4.2.1, we have

$$\begin{aligned} |(\text{div}_\alpha \Phi)(x)| &\leq 2 \left(\|\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} + \|D\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \right) \int_{\mathbb{R}^n} \frac{\min\{1, |x - y|\}}{|x - y|^{n+\alpha}} dy \\ &\lesssim_\alpha \left(\|\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} + \|D\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \right), \end{aligned} \quad (4.3.6)$$

which implies that $\|\text{div}_\alpha \Phi\|_{L^\infty(\mathbb{R}^n)} < \infty$. It remains to prove that $\|\text{div}_\alpha \Phi\|_{L^1(\mathbb{R}^n)} < \infty$. Then the required result can be obtained using interpolation. Since Φ is compactly supported, we may suppose $\text{supp } \Phi \subseteq B(0, M) \times B(0, M)$ for some $M > 0$. Thus, we obtain

$$\begin{aligned} \|\text{div}_\alpha \Phi\|_{L^1(\mathbb{R}^n)} &= \int_{B(0, M)} \left| \int_{B(0, M)} \frac{\Phi(x, y) - \Phi(y, x)}{|x - y|^{n+\alpha}} dy \right| dx \\ &\lesssim \|D\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \int_{B(0, M)} \int_{B(0, M)} \frac{1}{|x - y|^{n+\alpha-1}} dy dx < \infty, \end{aligned} \quad (4.3.7)$$

which finishes the proof. \square

We introduce the definition of space $W^{\alpha, 1}(\Omega)$, see [111] for details.

Definition 4.3.3. *Let $\Omega \subseteq \mathbb{R}^n$ be an open set. A function f is in $W^{\alpha, 1}(\Omega)$ when $f \in L^1(\Omega)$ and*

$$[f]_{W^{\alpha, 1}(\Omega)} := \int_{\Omega} \int_{\Omega} \frac{|f(x) - f(y)|}{|x - y|^{n+\alpha}} dy dx < \infty. \quad (4.3.8)$$

The norm of $W^{\alpha,1}(\Omega)$ is defined as

$$\|f\|_{W^{\alpha,1}(\Omega)} := \|f\|_{L^1(\Omega)} + [f]_{W^{\alpha,1}(\Omega)}. \quad (4.3.9)$$

We now state our main theorem of this section, which is in strong contrast to the Riesz-type fractional BV functions, cf. Lemma 4.2.9. The fractional BV space bv^α is actually equivalent to $W^{\alpha,1}$, which makes this space more tractable and probably more attainable for numerical purposes.

Theorem 4.3.4. *Let $\alpha \in (0, 1)$. Let $\Omega \subseteq \mathbb{R}^n$ be any open set. Then $bv^\alpha(\Omega) = W^{\alpha,1}(\Omega)$. More precisely, for any $f \in L^1(\Omega)$ we have*

$$\text{var}_\alpha(f; \Omega) = [f]_{W^{\alpha,1}(\Omega)},$$

whenever one of the two sides are finite.

Remark 4.3.5. *It is well known that Theorem 4.3.4 is false for $\alpha = 1$: e.g. take any nonempty open and bounded set Ω with finite perimeter. Then $\chi_\Omega \notin W^{1,1}(\mathbb{R}^n)$ (e.g. because it is not continuous on almost all lines). However, we have $\chi_\Omega \in BV(\mathbb{R}^n)$. In that sense, Theorem 4.3.4 may be surprising at first. Let us mention that, although we are not aware of Theorem 4.3.4 in the literature, intuitively related observations have been made by people working with fractional perimeters.*

Remark 4.3.6. *An immediate corollary is that the fractional perimeter as defined by Caffarelli-Roquejoffre-Savin, [38], $\text{Per}_\alpha(\Omega; \mathbb{R}^n) = \text{var}_\alpha(\chi_\Omega; \mathbb{R}^n)$. Thus, the space bv^α is the naturally associated notion for a fractional BV space when working with that perimeter.*

We prove several lemmas before proving Theorem 4.3.4.

Lemma 4.3.7. *Suppose $f \in W^{\alpha,1}(\Omega)$, then we have $f \in bv^\alpha(\Omega)$ and $\text{var}_\alpha(f; \Omega) = [f]_{W^{\alpha,1}(\Omega)}$.*

Proof. Given any $\Phi \in C_c^1(\Omega \times \Omega, \mathbb{R})$. Without loss of generality, we may suppose $\text{supp } \Phi \subseteq K \times K$, while K is a compact subset of Ω . Then we obtain that also $\text{div}_\alpha \Phi = 0$ outside K .

We have from Fubini's theorem (since $f \in W^{\alpha,1}(\Omega)$, both sides converge absolutely)

$$\int_{\mathbb{R}^n} f \operatorname{div}_\alpha \Phi dx = - \int_\Omega \int_\Omega \frac{f(x) - f(y)}{|x - y|^\alpha} \Phi(x, y) \frac{dy dx}{|x - y|^n}. \quad (4.3.10)$$

Since $L^\infty(\Omega \times \Omega)$ is the dual of $L^1(\Omega \times \Omega)$, from (4.3.10) we obtain

$$\operatorname{var}_\alpha(f; \Omega) = [f]_{W^{\alpha,1}(\Omega)},$$

which completes the proof. \square

The lemma above has not yet proven Theorem 4.3.4: if we only know $f \in bv^\alpha(\Omega)$ we cannot yet apply Lemma 4.3.7. However, Lemma 4.3.7 does give us the direction $\operatorname{var}_\alpha(f; \Omega) \leq [f]_{W^{\alpha,1}(\Omega)}$ whenever the right-hand side is finite (because in that case we can indeed apply Lemma 4.3.7).

Next, we observe the following lower semi-continuity result.

Lemma 4.3.8. *Suppose $f_k \in bv^\alpha(\Omega)$ for all $k \in \mathbb{N}$ and $\|f_k - f\|_{L^1(\Omega)} \rightarrow 0$ as $k \rightarrow \infty$. Then we have*

$$\operatorname{var}_\alpha(f; \Omega) \leq \liminf_{k \rightarrow \infty} \operatorname{var}_\alpha(f_k; \Omega),$$

and in particular $f \in bv^\alpha(\Omega)$.

Proof. Consider any $\Phi \in C_c^1(\Omega \times \Omega)$ with $\|\Phi\|_{L^\infty(\Omega \times \Omega)} \leq 1$. Since $f_k \rightarrow f$ in $L^1(\Omega)$, and $\operatorname{div}_\alpha \Phi$ is bounded by Lemma 4.3.2, we have $\|f_k \operatorname{div}_\alpha \Phi - f \operatorname{div}_\alpha \Phi\|_{L^1(\Omega)} \rightarrow 0$. Thus, we have

$$\int_\Omega f \operatorname{div}_\alpha \Phi dx = \lim_{k \rightarrow \infty} \int_\Omega f_k \operatorname{div}_\alpha \Phi dx \leq \liminf_{k \rightarrow \infty} \operatorname{var}_\alpha(f_k; \Omega). \quad (4.3.11)$$

Taking the supremum over all admissible Φ , we have

$$\operatorname{var}_\alpha(f; \Omega) \leq \liminf_{k \rightarrow \infty} \operatorname{var}_\alpha(f_k; \Omega), \quad (4.3.12)$$

which completes the proof. \square

To prove $\operatorname{var}_\alpha(f; \Omega) \geq [f]_{W^{\alpha,1}(\Omega)}$ (whenever the left-hand side is finite), the last missing ingredient is the following recovery sequence result. In the following, we say that a set G is compactly contained in a set Ω , in symbols $G \subset\subset \Omega$, if G is bounded and $\overline{G} \subset \Omega$.

Lemma 4.3.9. *Let $\Omega \subseteq \mathbb{R}^n$ be any open set. Assume $f \in L^1(\Omega)$ with $\text{var}_\alpha(f; \Omega) < \infty$. Then for any open $G \subset\subset \Omega$ there exists $f_k \in C_c^\infty(\Omega)$, for all $k \in \mathbb{N}$, such that*

$$f_k \rightarrow f \quad \text{in } L^1(G)$$

and

$$\limsup_{k \rightarrow \infty} \text{var}_\alpha(f_k; G) \leq \text{var}_\alpha(f; \Omega).$$

Proof. Since $G \subset\subset \Omega$, there exist open sets U and V such that $G \subset\subset U \subset\subset V \subset\subset \Omega$. Pick $\zeta \in C_c^\infty(V)$ such that $\zeta = 1$ on U and $\zeta \leq 1$ in all of \mathbb{R}^n . Take $\varepsilon_0 > 0$ such that $B_\varepsilon(G) := \{z \in \mathbb{R}^n : \text{dist}(z, G) < \varepsilon\} \subset\subset U$ and $B_\varepsilon(V) \subset\subset \Omega$ for any $\varepsilon \in (0, \varepsilon_0)$. Let $\eta \in C_c^\infty(B(0, 1))$, $\int \eta = 1$, be the usual mollifier kernel and set $\eta_\varepsilon := \varepsilon^{-n} \eta(\cdot/\varepsilon)$. For $\varepsilon \in (0, \varepsilon_0)$, we define $f_\varepsilon := \eta_\varepsilon * (f\zeta)$, then $\text{supp} f_\varepsilon \subseteq \Omega$. Given any $\Phi \in C_c^1(G \times G)$ with $\|\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq 1$. Using (4.3.4), the Fubini's theorem and the substitution $x' = x - z$ and $y' = y - z$, we obtain

$$\begin{aligned} & \int_{\mathbb{R}^n} f_\varepsilon \text{div}_\alpha \Phi dx \\ &= \int_G \eta_\varepsilon * (f\zeta) \text{div}_\alpha \Phi dx \\ &= \int_G \left(\int_{B(0, \varepsilon)} \eta_\varepsilon(z) f(x-z) \zeta(x-z) dz \right) \left(- \int_G \frac{\Phi(x, y) - \Phi(y, x)}{|x-y|^{n+\alpha}} dy \right) dx \\ &= - \int_G \int_G \int_{B(0, \varepsilon)} f(x-z) \zeta(x-z) \eta_\varepsilon(z) \frac{\Phi(x, y) - \Phi(y, x)}{|x-y|^{n+\alpha}} dz dy dx \\ &= - \int_{B_\varepsilon(G)} \int_{B_\varepsilon(G)} \int_{B(0, \varepsilon)} f(x') \zeta(x') \eta_\varepsilon(z) \frac{\Phi(x'+z, y'+z) - \Phi(y'+z, x'+z)}{|x'-y'|^{n+\alpha}} dz dy' dx'. \end{aligned} \tag{4.3.13}$$

Notice that since $\eta_\varepsilon(-z) = \eta_\varepsilon(z)$, we have

$$(\eta_\varepsilon * \Phi)(x', y') := \int_{B(0, \varepsilon)} \eta_\varepsilon(z) (\Phi(x'+z, y'+z) - \Phi(y'+z, x'+z)) dz. \tag{4.3.14}$$

Thus, by (4.3.13) we have

$$\begin{aligned}
\int_{\mathbb{R}^n} f_\varepsilon \operatorname{div}_\alpha \Phi dx &= - \int_{B_\varepsilon(G)} \int_{B_\varepsilon(G)} f(x') \zeta(x') \frac{(\eta_\varepsilon * \Phi)(x', y')}{|x' - y'|^{n+\alpha}} dy' dx' \\
&= - \int_{B_\varepsilon(G)} \int_{B_\varepsilon(G)} f(x') \frac{1}{2} (\zeta(x') + \zeta(y')) \frac{(\eta_\varepsilon * \Phi)(x', y')}{|x' - y'|^{n+\alpha}} dy' dx' \\
&\quad - \int_{B_\varepsilon(G)} \int_{B_\varepsilon(G)} f(x') \frac{1}{2} (\zeta(x') - \zeta(y')) \frac{(\eta_\varepsilon * \Phi)(x', y')}{|x' - y'|^{n+\alpha}} dy' dx'.
\end{aligned}$$

Since $B_\varepsilon(G) \subset U$ and $\zeta \equiv 1$ in U , the second term vanishes. Setting

$$\Psi_\varepsilon(x', y') := \frac{1}{2} (\zeta(x') + \zeta(y')) (\eta_\varepsilon * \Phi)(x', y'),$$

we see that $\Psi_\varepsilon \in C_c^\infty(\Omega \times \Omega)$, $\Psi_\varepsilon(x', y') = -\Psi_\varepsilon(y', x')$, and

$$\int_{\mathbb{R}^n} f_\varepsilon \operatorname{div}_\alpha \Phi dx = \int_{\Omega} f \operatorname{div}_\alpha \Psi_\varepsilon dx'.$$

It is easy to check that

$$\left| \frac{1}{2} (\zeta(x') + \zeta(y')) (\eta_\varepsilon * \Phi)(x', y') \right| \leq |(\eta_\varepsilon * \Phi)(x', y')| \leq \|\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq 1.$$

Thus, we have shown that for any $\Phi \in C_c^\infty(G \times G)$ with $\|\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq 1$, and any $\varepsilon < \varepsilon_0$, there is

$$\int_{\mathbb{R}^n} f_\varepsilon \operatorname{div}_\alpha \Phi dx \leq \operatorname{var}_\alpha(f; \Omega).$$

Taking the supremum over such test-functions Φ we obtain

$$\sup_{\varepsilon \in (0, \varepsilon_0)} \operatorname{var}_\alpha(f_\varepsilon; G) \leq \operatorname{var}_\alpha(f; \Omega).$$

In particular,

$$\limsup_{\varepsilon \rightarrow 0} \operatorname{var}_\alpha(f_\varepsilon; G) \leq \operatorname{var}_\alpha(f; \Omega).$$

By usual mollifier arguments we have $f_\varepsilon \rightarrow \zeta f$ in $L^1(\mathbb{R}^n)$ as $\varepsilon \rightarrow 0$. Since $\zeta \equiv 1$ in G , we have $f_\varepsilon \rightarrow f$ in $L^1(G)$ as $\varepsilon \rightarrow 0$. \square

We now finish the proof of the main theorem.

Proof of Theorem 4.3.4. Let $f \in L^1(\Omega)$. In Lemma 4.3.7, we have proved $[f]_{W^{\alpha,1}(\Omega)} \geq \text{var}_\alpha(f; \Omega)$, whenever the left-hand side is finite. So we only need to establish $[f]_{W^{\alpha,1}(\Omega)} \leq \text{var}_\alpha(f; \Omega)$, whenever the right-hand side is finite.

Given any $G \subset\subset \Omega$, we can take a sequence $\{f_k\}_{k=1}^\infty$ as stated in Lemma 4.3.9. Since $f_k \in C_c^\infty(\Omega)$, we have $f_k \in W^{\alpha,1}(G)$, so Lemma 4.3.7 is applicable. Combining Lemma 4.3.7 and Lemma 4.3.9, we find

$$\limsup_{k \rightarrow \infty} [f_k]_{W^{\alpha,1}(G)} \leq \limsup_{k \rightarrow \infty} \text{var}_\alpha(f_k; G) \leq \text{var}_\alpha(f; \Omega).$$

Since $f_k \rightarrow f$ in $L^1(G)$, up to passing to a subsequence, we may assume that $f_k(x)$ converges to $f(x)$ a.e. in G . Using Fatou's lemma, we obtain

$$\int_G \int_G \frac{|f(x) - f(y)|}{|x - y|^{n+\alpha}} dy dx \leq \liminf_{k \rightarrow \infty} \int_G \int_G \frac{|f_k(x) - f_k(y)|}{|x - y|^{n+\alpha}} dy dx. \quad (4.3.15)$$

Thus, we obtain

$$[f]_{W^{\alpha,1}(G)} \leq \liminf_{k \rightarrow \infty} [f_k]_{W^{\alpha,1}(G)} \leq \text{var}_\alpha(f; \Omega). \quad (4.3.16)$$

Picking an increasing sequence of open sets $\{G_m\}$, such that $G_m \subset\subset \Omega$ and

$$\bigcup_{m=1}^\infty G_m = \Omega. \quad (4.3.17)$$

Applying the above argument to $G = G_m$, we have $[f]_{W^{\alpha,1}(G_m)} \leq \text{var}_\alpha(f; \Omega)$ for any $m \in \mathbb{N}$. Using Fatou's lemma again, we have

$$\begin{aligned} [f]_{W^{\alpha,1}(\Omega)} &\leq \liminf_{m \rightarrow \infty} \int_{G_m} \int_{G_m} \frac{|f(x) - f(y)|}{|x - y|^{n+\alpha}} dy dx \\ &\leq \liminf_{m \rightarrow \infty} [f]_{W^{\alpha,1}(G_m)} \leq \text{var}_\alpha(f; \Omega), \end{aligned} \quad (4.3.18)$$

which concludes the proof. \square

Using Theorem 4.3.4, we can easily obtain the following result.

Proposition 4.3.10 (Weak compactness). *Let $\Omega \subset \mathbb{R}^n$ be an open and bounded set with Lipschitz boundary. Assume that $\{f_k\}_{k=1}^\infty \subset bv^\alpha(\Omega)$ such that*

$$\sup_{k \in \mathbb{N}} \|f_k\|_{bv^\alpha(\Omega)} < \infty.$$

Then there exists $f \in bv^\alpha(\Omega)$ such that

$$\text{var}_\alpha(f; \Omega) \leq \liminf_{k \rightarrow \infty} \text{var}_\alpha(f_k; \Omega),$$

and there is a subsequence $\{f_{k_i}\}_{i=1}^\infty$, such that for all $p \in \left[1, \frac{n}{n-\alpha}\right)$

$$\|f_{k_i} - f\|_{L^p(\Omega)} \xrightarrow{i \rightarrow \infty} 0.$$

Proof. By Theorem 4.3.4, we have

$$\text{var}_\alpha(f_k, \Omega) = [f_k]_{W^{\alpha,1}(\Omega)}.$$

Since Ω is a Lipschitz domain, it is regular in the sense of [158]. Thus, by the main result of [158], we can find an extension $\tilde{f}_k \in W^{\alpha,1}(\mathbb{R}^n)$ with compact support, $\tilde{f}_k = f_k$ a.e. in Ω , such that

$$[\tilde{f}_k]_{W^{\alpha,1}(\mathbb{R}^n)} \lesssim [f_k]_{W^{\alpha,1}(\Omega)}.$$

From the usual Rellich theorem, we find a subsequence $(f_{k_i})_{i \in \mathbb{N}}$, such that for all $p \in \left[1, \frac{n}{n-\alpha}\right)$

$$\|f_{k_i} - f\|_{L^p(\Omega)} \xrightarrow{i \rightarrow \infty} 0.$$

see [56, Corollary 7.2]. In particular, in view of Lemma 4.3.8,

$$\text{var}_\alpha(f; \Omega) \leq \liminf_{k \rightarrow \infty} \text{var}_\alpha(f_k; \Omega).$$

□

Using Theorem 4.3.4, we also readily obtain the Sobolev embedding theorem, which can be proved using the extension theorem as in Proposition 4.3.10 above and then [108, Theorem 9].

Proposition 4.3.11. *Let $\Omega \subset \mathbb{R}^n$ be an open and bounded set with Lipschitz boundary. Then there exists a constant $C = C(n, \alpha) > 0$ such that for any $f \in bv^\alpha(\Omega)$,*

$$\|f\|_{L^{\frac{n}{n-\alpha}}(\Omega)} \leq C \text{var}_\alpha(f; \Omega).$$

We also obtain the following density result, which might be known to experts (observe that this density is not true for $\alpha = 1$, cf. [64, Theorem 5.3 and remark after]). Using the identification in Theorem 4.3.4, the extension property in [158], and the usual mollification in [56, Theorem 2.4.] or [112, Lemma 26], we have the following result.

Corollary 4.3.12. *Let $\alpha \in (0, 1)$. Let $\Omega \subset \mathbb{R}^n$ be any open and bounded set with Lipschitz boundary, then $\mathcal{C}^\infty(\bar{\Omega})$ is dense in $bv^\alpha(\Omega)$.*

Let us make a last remark about traces. For a classical BV function there is a trace, [14, Theorem 10.2.1]. However, this will not be true for $bv^\alpha(\Omega)$, since $W^{\alpha,1}(\Omega)$ does not have a reasonably defined trace. The typical approach is then the notion of a fat boundary trace, which we do not pursue here.

4.4 Image Denoising and Primal Problem

Let $\Omega \subset \mathbb{R}^n$ be a open and bounded set with a Lipschitz continuous boundary, $\alpha \in (0, 1)$, $p \in (1, \infty)$, $p^\infty := \frac{n}{n-\alpha}$, and $u_N \in L^p(\Omega)$. Based on the two fractional variations considered in this work, we consider the (primal) problems for some fixed positive parameters β and γ

$$\inf_{u \in L^p(\Omega)} \left\{ \frac{\gamma}{p} \|u - u_N\|_{L^p(\Omega)}^p + \beta \text{Var}_\alpha(\chi_\Omega u; \mathbb{R}^n) \right\}, \quad (\mathcal{P}_R)$$

$$\inf_{u \in L^p(\Omega)} \left\{ \frac{\gamma}{p} \|u - u_N\|_{L^p(\Omega)}^p + \beta \text{var}_\alpha(u; \Omega) \right\}. \quad (\mathcal{P}_G)$$

Note that the condition of u having bounded fractional variation is imposed implicitly, and it is also clear that both problems are strictly convex for $p > 1$. As a result, we use well-known results from *convex analysis*, cf. [62], to study the minimizers of problems (\mathcal{P}_R) and (\mathcal{P}_G) . The regularity theory of a related problem to \mathcal{P}_G was recently studied in [119, 23].

Convex Analysis and Optimization

As usual in convex optimization, we consider the so-called *dual problem*, which usually gives new insights about the structure of the primal problem. In this work, we

consider a different but related approach coined the *predual method*. Here we mainly follow the approach given in [37, 44, 78]. In order to introduce this method, we need some definitions, cf. [62, Ch. I]. Consider a Banach space V and its topological dual V^* , with duality pairing denoted by $\langle \cdot, \cdot \rangle_{V^*, V}$. Given $\mathcal{F} : V \rightarrow \overline{\mathbb{R}}$, its *Fenchel conjugate* is given by $\mathcal{F}^* : V^* \rightarrow \overline{\mathbb{R}}$,

$$u^* \mapsto \mathcal{F}^*(u^*) := \sup_{u \in V} \{ \langle u^*, u \rangle_{V^*, V} - \mathcal{F}(u) \}. \quad (4.4.1)$$

We denote by $\partial\mathcal{F}(u)$ the subdifferential map of \mathcal{F} at the point $u \in V$, see [62, Definition I.5.1]. The following characterization holds,

$$\begin{aligned} u^* \in \partial\mathcal{F}(u) \text{ if and only if } \mathcal{F}(u) \text{ is finite and} \\ \langle u^*, v - u \rangle_{V^*, V} + \mathcal{F}(u) \leq \mathcal{F}(v), \quad \forall v \in V. \end{aligned} \quad (4.4.2)$$

We now introduce a process known as *dualization* [62, Chs. III-IV]. Here we will focus on problems with the form:

$$\inf_{u \in V} \{ F(u) + G(\Lambda u) \}, \quad (\mathcal{Q})$$

where Y is a Hausdorff topological space with dual Y^* , $\Lambda \in \mathcal{L}(V, Y)$, with transpose $\Lambda^* \in \mathcal{L}(Y^*, V^*)$, and $F : V \rightarrow \overline{\mathbb{R}}$, $G : Y \rightarrow \overline{\mathbb{R}}$. We define the dual problem of (\mathcal{Q}) as

$$\sup_{v \in Y^*} -\Phi^*(0, v), \quad (\mathcal{Q}^*)$$

where $\Phi^* : V^* \times Y^* \rightarrow \overline{\mathbb{R}}$ is the Fenchel conjugate (dual) of $\Phi : V \times Y \rightarrow \overline{\mathbb{R}}$, $(u, p) \mapsto \Phi(u, p) := F(u) + G(p + \Lambda u)$, see (4.4.1). The next theorem gives conditions for the so-called *Fenchel's duality*; cf. [62, Theorem III.4.1] and [63, Pg. 130].

Theorem 4.4.1. *Assume V and Y are Banach spaces, F and G are convex and lower semicontinuous (l.s.c.), and there exists $v_0 \in V$ such that $F(v_0) < \infty$, $G(\Lambda v_0) < \infty$, and G is continuous at Λv_0 . Then, the problems (\mathcal{Q}) and (\mathcal{Q}^*) are related by:*

$$\begin{aligned} \inf_{u \in V} \{ F(u) + G(\Lambda u) \} &= \sup_{v \in Y^*} -\Phi^*(0, v) \\ &= \sup_{v \in Y^*} \{ -F^*(\Lambda^* v) - G^*(-v) \}, \end{aligned}$$

and there exists at least one solution to (\mathcal{Q}^*) . Moreover, if \bar{u} and \bar{v} are solutions for (\mathcal{Q}) and (\mathcal{Q}^*) , respectively, then

$$\begin{aligned}\Lambda^* \bar{v} &\in \partial F(\bar{u}), \\ -\bar{v} &\in \partial G(\Lambda \bar{u}).\end{aligned}\tag{4.4.3}$$

In general, there are several options for F, G and Λ in order to write a given problem as shown in (\mathcal{Q}) . Here we consider one that satisfies the hypothesis of Theorem 4.4.1 in a straightforward manner. We now show existence and characterization for minimizers of problems (\mathcal{P}_R) and (\mathcal{P}_G) .

Riesz-type

By Proposition 4.2.5, for $p \in [1, p^\infty]$, with $p^\infty := \frac{n}{n-\alpha}$, we can consider the problem (\mathcal{P}_R) defined on $L^p(\Omega)$ or $\text{BV}_{00}^\alpha(\Omega)$, cf. (4.2.5), interchangeably. The next lemma shows that the problem (\mathcal{P}_R) , related to the Riesz-type of fractional bounded variation, has a solution and for $p > 1$ it is unique.

Lemma 4.4.2. *For $p \in (1, p^\infty)$, the problem (\mathcal{P}_R) has a unique solution $\bar{u} \in \text{BV}_{00}^\alpha(\Omega)$*

Proof. Let $p \in [1, \infty)$, define $\mathcal{J}_R : (L^p(\Omega), \|\cdot\|_{L^p(\Omega)}) \rightarrow \overline{\mathbb{R}}$, given by

$$\mathcal{J}_R(u) := \frac{\gamma}{p} \|u - u_N\|_{L^p(\Omega)}^p + \beta \text{Var}_\alpha(\chi_\Omega u; \mathbb{R}^n).\tag{4.4.4}$$

It is clear that

$$0 \leq \inf_{u \in L^p(\Omega)} \mathcal{J}_R(u) \leq \frac{\gamma}{p} \|u_N\|_{L^p(\Omega)}^p.$$

Now, let $(u_k)_{k \in \mathbb{N}} \subseteq L^p(\Omega)$ be a minimizing sequence associated to the problem (\mathcal{P}_R) , then for each $k \in \mathbb{N}$

$$\begin{aligned}\|u_k\|_{L^p(\Omega)} &\leq \|u_k - u_N\|_{L^p(\Omega)} + \|u_N\|_{L^p(\Omega)} \leq 2\|u_N\|_{L^p(\Omega)}, \text{ and} \\ \text{Var}_\alpha(\chi_\Omega u_k; \mathbb{R}^n) &\leq \frac{\gamma}{p\beta} \|u_N\|_{L^p(\Omega)}^p.\end{aligned}$$

Then, for $p \in [1, p^\infty)$, Propositions 4.2.5 and 4.2.8 imply there exist $\bar{u} \in \text{BV}_{00}^\alpha(\Omega) \hookrightarrow L^p(\mathbb{R}^n)$ and a subsequence $\{u_{k_i}\}_{i \in \mathbb{N}}$ such that

$$\text{Var}_\alpha(\bar{u}; \mathbb{R}^n) \leq \liminf_{i \rightarrow \infty} \text{Var}_\alpha(u_{k_i}; \mathbb{R}^n) \quad \text{and} \quad \|u_{k_i} - u_N\|_{L^p(\Omega)}^p \xrightarrow{i \rightarrow \infty} \|\bar{u} - u_N\|_{L^p(\Omega)}^p.$$

Thus, the existence of a solution for (\mathcal{P}_R) follows from the fact that $\bar{u} = \chi_\Omega \bar{u}$, a.e., for the uniqueness it is enough to notice that \mathcal{J}_R , cf. (4.4.4), is a strictly convex functional for $p > 1$. In fact, if \bar{u}_1 and \bar{u}_2 were two different solutions to (\mathcal{P}_R) , then for $\lambda \in (0, 1)$,

$$\begin{aligned}
\mathcal{J}_R(\lambda \bar{u}_1 + (1 - \lambda) \bar{u}_2) &= \frac{\gamma}{p} \|\lambda \bar{u}_1 + (1 - \lambda) \bar{u}_2 - u_d\|_{L^p(\Omega)}^p + \beta \text{Var}_\alpha(\chi_\Omega(\lambda \bar{u}_1 + (1 - \lambda) \bar{u}_2); \mathbb{R}^n) \\
&= \frac{\gamma}{p} \|\lambda(\bar{u}_1 - u_d) + (1 - \lambda)(\bar{u}_2 - u_d)\|_{L^p(\Omega)}^p \\
&\quad + \beta \text{Var}_\alpha(\chi_\Omega(\lambda \bar{u}_1 + (1 - \lambda) \bar{u}_2); \mathbb{R}^n) \\
&< \frac{\gamma}{p} \lambda \|\bar{u}_1 - u_d\|_{L^p(\Omega)}^p + \frac{\gamma}{p} (1 - \lambda) \|\bar{u}_2 - u_d\|_{L^p(\Omega)}^p \\
&\quad + \beta \text{Var}_\alpha(\chi_\Omega(\lambda \bar{u}_1 + (1 - \lambda) \bar{u}_2); \mathbb{R}^n) \\
&\leq \lambda \mathcal{J}_R(\bar{u}_1) + (1 - \lambda) \mathcal{J}_R(\bar{u}_2).
\end{aligned}$$

Thus, $\bar{u}_1 = \bar{u}_2$ a.e. and by the definition of Var_α , cf. (4.2.4), the proof concludes. \square

Next, we will derive an expression for the predual of (\mathcal{P}_R) . In order to do that, we begin with the regularity for the “test functions” in (4.2.4). It is clear that if $u \in L^1(\mathbb{R}^n)$ and $\text{supp } u \subset \Omega$, then $\int_{\mathbb{R}^n} u \text{Div}_\alpha \Phi \, dx$ does not depend on $\text{Div}_\alpha \Phi|_{\Omega^c}$. This motivates us to define $\|\Phi\|_{X_{\text{Riesz}}} := \sqrt[q]{\|\Phi\|_{L^q(\mathbb{R}^n, \mathbb{R}^n)}^q + \|\text{Div}_\alpha \Phi\|_{L^q(\Omega)}^q}$ for $\Phi \in C_c^1(\mathbb{R}^n; \mathbb{R}^n)$, where $q := \frac{p}{p-1}$. We consider the space $X_{\text{Riesz}} := X_{\text{Riesz}}(\Omega, q, \alpha)$, given by

$$X_{\text{Riesz}} := \overline{C_c^1(\mathbb{R}^n; \mathbb{R}^n)}^{\|\cdot\|_{X_{\text{Riesz}}}}.$$

We also define an auxiliary problem:

$$\inf_{\Phi \in X_{\text{Riesz}}} \left\{ \frac{1}{q} \|\text{Div}_\alpha \Phi\|_{L^q(\Omega)}^q - \int_{\Omega} u_N (-\text{Div}_\alpha \Phi) + I_\beta(\Phi) \right\}, \quad (\mathcal{Q}_R)$$

where I_β denotes the convex indicator function defined as

$$I_\beta(\Phi) := \begin{cases} 0 & : \|\Phi\|_{L^\infty(\mathbb{R}^n)} \leq \beta, \\ +\infty & : \text{otherwise.} \end{cases}$$

We will establish that (\mathcal{Q}_R) is the pre-dual problem to (\mathcal{P}_R) , i.e., dual of (\mathcal{Q}_R) will be (\mathcal{P}_R) if Ω is convex.

We begin by noting that (\mathcal{Q}_R) fits in the abstract framework of (\mathcal{Q}) if we consider the spaces: $Y := (L^q(\Omega), \|\cdot\|_{L^q(\Omega)})$, $V := (X_{\text{Riesz}}, \|\cdot\|_{X_{\text{Riesz}}})$, and the operators:

$$\begin{aligned} G : Y &\rightarrow \overline{\mathbb{R}}, & G(v) &:= \frac{1}{q} \|v\|_{L^q(\Omega)}^q - \int_{\Omega} u_N v dx, \\ F : V &\rightarrow \overline{\mathbb{R}}, & F(\Phi) &:= I_{\beta}(\Phi), \\ \Lambda : V &\rightarrow Y, & \Lambda(\Phi) &:= (-\text{Div}_{\alpha} \Phi)|_{\Omega}. \end{aligned} \tag{4.4.5}$$

To compute the dual problem of (\mathcal{Q}_R) , we compute the Fenchel conjugate of F, G , and Λ , given in (4.4.5).

Proposition 4.4.3. *Let $\Omega \subseteq \mathbb{R}^n$ be open, bounded, convex. Let $Y := (L^q(\Omega), \|\cdot\|_{q,\Omega})$, $V := (X_{\text{Riesz}}, \|\cdot\|_{X_{\text{Riesz}}})$, and operators F, G and Λ be defined as in (4.4.5), then*

$$\begin{aligned} G^* : Y^* &\rightarrow \overline{\mathbb{R}}, & u &\mapsto \frac{1}{p} \|u + u_N\|_{L^p(\Omega)}^p, \\ F^* : V^* &\rightarrow \overline{\mathbb{R}}, & \Psi^* &\mapsto \sup_{\substack{\Phi \in V \\ \|\Phi\|_{L^\infty(\mathbb{R}^n)} \leq \beta}} \langle \Psi^*, \Phi \rangle_{V^*, V}, \\ F^* \circ \Lambda^* : Y^* &\rightarrow \overline{\mathbb{R}}, & u &\mapsto \beta \text{Var}_{\alpha}(\chi_{\Omega} u; \mathbb{R}^n). \end{aligned}$$

Proposition 4.4.3, while in principle looking very similar to the arguments in [78, Section 2], contains a serious nuance. Observe that [78] does not consider test-functions with the natural restriction $\|\Phi\|_{L^\infty(\mathbb{R}^n)} \leq \beta$, but rather component-wise control $|\Phi^i| \leq \beta$, $i = 1, \dots, n$ (leading to a nonstandard BV -space) – which is critical in their argument to compute the predual.

Instead, we show here that such unnatural restrictions are unnecessary for bounded, open, convex sets Ω . The main property we use is the controllable distance of rescaled sets; cf. Lemma B.0.1 and Lemma B.0.2. The main novelty is contained in the next proposition. Notice that such a result is even critical to prove the result in [78, 79], where $\alpha = 1$, for the natural BV -space.

Proposition 4.4.4. *If Ω is convex, then for all $\Phi \in X_{Riesz}$,*

$$I_\beta(\Phi) = \tilde{I}_\beta(\Phi),$$

where

$$\tilde{I}_\beta(\Phi) := \begin{cases} 0 & : \text{if there exists } \Psi_k \in C_c^\infty(\mathbb{R}^n; \mathbb{R}^n), \Psi_k \rightarrow \Phi \text{ in } X_{Riesz} \\ & \text{such that } \sup_k \|\Psi_k\|_{L^\infty(\mathbb{R}^n)} \leq \beta, \\ +\infty & : \text{otherwise.} \end{cases}$$

Proof. We first observe that

$$\tilde{I}_\beta(\Phi) = 0 \Rightarrow I_\beta(\Phi) = 0. \quad (4.4.6)$$

Indeed, if $\tilde{I}_\beta(\Phi) = 0$ then there exists a sequence $\Psi_k \in C_c^\infty(\mathbb{R}^n; \mathbb{R}^n)$ with $\|\Psi_k\|_{L^\infty(\mathbb{R}^n)} \leq \beta$, such that $\Psi_k \rightarrow \Phi$ in X_{Riesz} . In particular, we have $\|\Psi_k - \Phi\|_{L^q(\mathbb{R}^n)} \xrightarrow{k \rightarrow \infty} 0$. Then there exists a subsequence, still denoted by Ψ_k , such that Ψ_k converges a.e. to Φ , which implies that $|\Phi(x)| \leq \beta$ a.e. in \mathbb{R}^n , i.e. $\|\Phi\|_{L^\infty(\mathbb{R}^n)} \leq \beta$. By the definition of I_β , we have $I_\beta(\Phi) = 0$, which proves (4.4.6).

From (4.4.6) we conclude $\tilde{I}_\beta(\Phi) \geq I_\beta(\Phi)$. It remains to prove $\tilde{I}_\beta(\Phi) \leq I_\beta(\Phi)$. If the right-hand side is $+\infty$ then there is nothing to show. Thus, we only need to show

$$I_\beta(\Phi) = 0 \Rightarrow \tilde{I}_\beta(\Phi) = 0. \quad (4.4.7)$$

Suppose that $\Phi \in X_{Riesz}$ and $I_\beta(\Phi) = 0$. In order to establish (4.4.7), we need to show

$$\forall \varepsilon > 0 \exists \Theta \in C_c^\infty(\mathbb{R}^n; \mathbb{R}^n), \quad \|\Theta\|_{L^\infty(\mathbb{R}^n)} \leq \beta, \quad \|\Theta - \Phi\|_{X_{Riesz}} < \varepsilon. \quad (4.4.8)$$

We proceed in several steps.

Step 1: We first show that

$$\forall \varepsilon > 0 \exists \Theta_1 \in X_{Riesz}, \quad \text{supp } \Theta_1 \subset\subset \mathbb{R}^n, \quad \|\Theta_1\|_{L^\infty(\mathbb{R}^n)} \leq \beta, \quad \|\Theta_1 - \Phi\|_{X_{Riesz}} < \varepsilon. \quad (4.4.9)$$

For $m \in \mathbb{N}$, we choose a smooth cut-off function $0 \leq \zeta_m \leq 1$, such that $\zeta_m = 1$ when $|x| < m$; $\zeta_m = 0$ when $|x| > 2m$; and $|\nabla \zeta_m| \lesssim \frac{1}{m}$. For sufficiently large m , we set

$$\Theta_1 := \zeta_m \Phi.$$

It is clear that $\|\Theta_1\|_{L^\infty(\mathbb{R}^n)} \leq \|\zeta_m \Phi\|_{L^\infty(\mathbb{R}^n)} \leq \|\Phi\|_{L^\infty(\mathbb{R}^n)} \leq \beta$. Thus, we only need to show

$$\|\Phi \zeta_m - \Phi\|_{L^q(\mathbb{R}^n)} + \|\operatorname{Div}_\alpha(\Phi \zeta_m) - \operatorname{Div}_\alpha \Phi\|_{L^q(\Omega)} \rightarrow 0.$$

Since $|\Phi \zeta_m| \leq |\Phi|$ and $\zeta_m \rightarrow 1$ almost everywhere on \mathbb{R}^n , using the Lebesgue dominated convergence theorem, we have $\|\Phi \zeta_m - \Phi\|_{L^q(\mathbb{R}^n)} \rightarrow 0$. Since m is sufficiently large, we may assume without loss of generality that $\zeta_m(x) \equiv 1$ when $\operatorname{dist}(x, \Omega) \leq 1$. Since $\Phi \in X_{\text{Riesz}}$, there exists a sequence $\Phi_k \in C_c^\infty(\mathbb{R}^n)$, such that $\|\Phi_k - \Phi\|_{X_{\text{Riesz}}} \xrightarrow{k \rightarrow \infty} 0$ and

$$\int_{\mathbb{R}^n} \Phi_k \cdot D^\alpha \varphi = - \int_{\Omega} \operatorname{Div}_\alpha(\Phi_k) \varphi, \quad \forall \varphi \in C_c^\infty(\Omega).$$

By the definition of X_{Riesz} -convergence, we can take the limits of both sides, which implies

$$\int_{\mathbb{R}^n} \Phi \cdot D^\alpha \varphi = - \int_{\Omega} \operatorname{Div}_\alpha(\Phi) \varphi, \quad \forall \varphi \in C_c^\infty(\Omega).$$

Now we claim that $\operatorname{Div}_\alpha(\zeta_m \Phi) \in L^q(\Omega)$. Indeed, let $\varphi \in C_c^\infty(\Omega)$, then

$$\begin{aligned} \int_{\mathbb{R}^n} \Phi \zeta_m \cdot D^\alpha \varphi &= \int_{\mathbb{R}^n} \Phi \cdot D^\alpha(\zeta_m \varphi) + \int_{\mathbb{R}^n} \Phi \cdot (\zeta_m D^\alpha(\varphi) - D^\alpha(\zeta_m \varphi)) \\ &= \int_{\mathbb{R}^n} \Phi \cdot D^\alpha \varphi + \int_{\mathbb{R}^n} \Phi \cdot (\zeta_m D^\alpha(\varphi) - D^\alpha(\zeta_m \varphi)). \end{aligned}$$

In the last step we used that $\zeta_m \varphi = \varphi$ by the definition of ζ_m and the support of φ . Using e.g. the Coifman-McIntosh-Meyer commutator estimate (e.g., see [101, Theorem 6.1] or [86, Theorem 3.2.1]), we have

$$\|\zeta_m D^\alpha(\varphi) - D^\alpha(\zeta_m \varphi)\|_{L^{q'}(\mathbb{R}^n)} \lesssim [\zeta_m]_{\text{Lip}} \|I^{1-\alpha} \varphi\|_{L^{q'}(\mathbb{R}^n)},$$

where $I^{1-\alpha}$ denotes the Riesz potential and $q' = \frac{q}{q-1}$. Since φ has compact support in the bounded set Ω , we have by Sobolev-Poincaré inequality

$$\|I^{1-\alpha} \varphi\|_{L^{q'}(\mathbb{R}^n)} \lesssim_{\Omega} \|\varphi\|_{L^{q'}(\Omega)},$$

which follows from the usual blow-up argument used for the classical Poincaré inequality. That is, we have shown that for any $\varphi \in C_c^\infty(\Omega)$,

$$\left| \int_{\mathbb{R}^n} \operatorname{Div}_\alpha(\Phi\zeta_m - \Phi) \varphi \right| \equiv \left| \int_{\mathbb{R}^n} (\Phi\zeta_m - \Phi) \cdot D^\alpha \varphi \right| \leq C(\Omega, q) \|\Phi\|_{L^q(\mathbb{R}^n)} \|\varphi\|_{L^{q'}(\mathbb{R}^n)} [\zeta_m]_{\operatorname{Lip}}.$$

Observe that $[\zeta_m]_{\operatorname{Lip}} \lesssim \frac{1}{m}$, so we have shown by duality that

$$\|\operatorname{Div}_\alpha(\Phi\zeta_m - \Phi)\|_{L^q(\Omega)} \lesssim \frac{1}{m} \|\Phi\|_{L^q(\mathbb{R}^n)} \xrightarrow{m \rightarrow \infty} 0,$$

which establishes (4.4.9).

Step 2: By translation we may assume Ω is convex and $0 \in \Omega$. For $\rho > 1$, we set

$$\Omega_\rho := \rho\Omega = \{\rho x : x \in \Omega\}. \quad (4.4.10)$$

Then, from Lemma B.0.1 and Lemma B.0.2, we have $\Omega \subset\subset \Omega_\rho$ for $\rho > 1$.

In this step, we show that

$$\begin{aligned} \forall \varepsilon > 0 \exists \Theta_2 \in X_{\operatorname{Riesz}}, \rho > 1, \operatorname{supp} \Theta_2 \subset\subset \mathbb{R}^n, \operatorname{Div}_\alpha \Theta_2 \in L^q(\Omega_\rho), \\ \|\Theta_2\|_{L^\infty(\mathbb{R}^n)} \leq \beta, \quad \|\Theta_2 - \Phi\|_{X_{\operatorname{Riesz}}} < \varepsilon. \end{aligned} \quad (4.4.11)$$

By the results in Step 1, we only need to show (4.4.11) with Φ replaced by Θ_1 . We let $\Psi := \Theta_1$ for convenience. For $\rho > 1$,

$$\Psi_\rho(x) := \Psi(x/\rho).$$

Then we have

$$\|\Psi_\rho\|_{L^\infty(\mathbb{R}^n)} = \|\Psi\|_{L^\infty(\mathbb{R}^n)} \leq \beta.$$

Moreover, in view of Lemma B.0.4, we have

$$\|\Psi_\rho - \Psi\|_{L^q(\mathbb{R}^n)} \xrightarrow{\rho \rightarrow 1^+} 0.$$

It remains to show $\operatorname{Div}_\alpha \Psi_\rho \in L^q(\Omega_\rho)$ and

$$\|\operatorname{Div}_\alpha \Psi_\rho - \operatorname{Div}_\alpha \Psi\|_{L^q(\Omega)} \xrightarrow{\rho \rightarrow 1^+} 0.$$

We first observe that

$$\int_{\mathbb{R}^n} \Psi_\rho \cdot D^\alpha \varphi \, dx = \rho^{n-\alpha} \int_{\mathbb{R}^n} \Psi \cdot D^\alpha (\varphi(\rho \cdot)) \, dx.$$

Thus, from $\varphi(\cdot) \in C_c^\infty(\Omega_\rho)$ we have $\varphi(\rho \cdot) \in C_c^\infty(\Omega)$. Then we conclude

$$\operatorname{Div}_\alpha \Psi_\rho(x) = \rho^{-\alpha} (\operatorname{Div}_\alpha \Psi)(x/\rho) \quad \text{for a.e. } x \in \Omega_\rho.$$

In particular $\operatorname{Div}_\alpha \Psi_\rho \in L^q(\Omega_\rho)$. We now have

$$\begin{aligned} & \| \operatorname{Div}_\alpha \Psi_\rho - \operatorname{Div}_\alpha \Psi \|_{L^q(\Omega)} \\ & \leq \| \operatorname{Div}_\alpha \Psi_\rho - \rho^{-\alpha} \operatorname{Div}_\alpha \Psi \|_{L^q(\Omega)} + \| \rho^{-\alpha} \operatorname{Div}_\alpha \Psi - \operatorname{Div}_\alpha \Psi \|_{L^q(\Omega)} \\ & = \rho^{-\alpha} \| (\operatorname{Div}_\alpha \Psi)(\cdot/\rho) - (\operatorname{Div}_\alpha \Psi)(\cdot) \|_{L^q(\Omega)} + (1 - \rho^{-\alpha}) \| \operatorname{Div}_\alpha \Psi \|_{L^q(\Omega)} \\ & \xrightarrow{\rho \rightarrow 1^+} 0, \end{aligned}$$

where we have used Lemma B.0.4 for the first term and $\operatorname{Div}_\alpha \Psi \in L^q(\Omega)$ for the second term. This implies that (4.4.11) is satisfied, by considering Ψ_ρ for $\rho > 1$ close enough to 1.

Step 3: Conclusion Given $\varepsilon > 0$, we take $\Psi := \Theta_2$ and pick $\rho > 1$ such that (4.4.11) is satisfied for $\frac{\varepsilon}{2}$ instead of ε . Since Ω is convex, by Lemma B.0.1 and Lemma B.0.2, there exists $D > 0$ such that $\operatorname{dist}(\Omega, \partial\Omega_\rho) > D$. We let $\delta_0 := \frac{D}{100}$ and choose $\delta \in (0, \delta_0)$. Let $\eta \in C_c^\infty(B(0, 1))$ be the usual symmetric mollifier kernel, and set

$$\Psi_\delta := \eta_\delta * \Psi.$$

Since $\operatorname{supp} \Psi \subset\subset \mathbb{R}^n$, we have $\Psi_\delta \in C_c^\infty(\mathbb{R}^n)$ and

$$\| \Psi_\delta \|_{L^\infty(\mathbb{R}^n)} \leq \| \Psi \|_{L^\infty(\mathbb{R}^n)} \leq \beta$$

We also have by usual mollification

$$\| \Psi_\delta - \Psi \|_{L^q(\mathbb{R}^n)} \xrightarrow{\delta \rightarrow 0} 0.$$

Lastly, for $\varphi \in C_c^\infty(\Omega)$ we have by Fubini's theorem

$$\int_{\mathbb{R}^n} \Psi_\delta \cdot D^\alpha \varphi = \int_{\mathbb{R}^n} \Psi \cdot D^\alpha (\varphi * \eta_\delta).$$

Observe that $\varphi \in C_c^\infty(\Omega)$ implies that $\varphi * \eta_\delta \in C_c^\infty(B(\Omega, \delta)) \subset C_c^\infty(\Omega_\rho)$. Thus, we have

$$\int_{\mathbb{R}^n} \Psi_\delta \cdot D^\alpha \varphi = \int_{\mathbb{R}^n} \operatorname{Div}_\alpha \Psi(\varphi * \eta_\delta) \quad \forall \varphi \in C_c^\infty(\Omega).$$

Since $\operatorname{Div}_\alpha \Psi \in L^q(\Omega_\rho)$, we conclude that

$$\operatorname{Div}_\alpha \Psi_\delta = (\operatorname{Div}_\alpha \Psi) * \eta_\delta \quad \text{in } \Omega.$$

Then, since $\operatorname{Div}_\alpha \Psi \in L^q(\Omega_\rho)$ we have that $(\operatorname{Div}_\alpha \Psi) * \eta_\delta$ converges to $\operatorname{Div}_\alpha \Psi$ in $L^q(\Omega)$ as $\delta \rightarrow 0$, i.e.

$$\|\operatorname{Div}_\alpha \Psi_\delta - \operatorname{Div}_\alpha \Psi\|_{L^q(\Omega)} \xrightarrow{\delta \rightarrow 0^+} 0.$$

Thus, we conclude that

$$\|\Psi_\delta - \Psi\|_{X_{\operatorname{Riesz}}} \xrightarrow{\delta \rightarrow 0^+} 0.$$

Now by choosing $\delta > 0$ sufficiently small, we have

$$\|\Psi_\delta - \Phi\|_{X_{\operatorname{Riesz}}} \leq \|\Psi_\delta - \Psi\|_{X_{\operatorname{Riesz}}} + \|\Psi - \Phi\|_{X_{\operatorname{Riesz}}} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Letting $\Theta := \Psi_\delta$, we have shown (4.4.8), which implies (4.4.7). Therefore, we have proved $\tilde{I}_\beta(\Phi) \leq I_\beta(\Phi)$, which completes the proof. \square

With the help of Proposition 4.4.4, we can now continue with the optimizing problem.

Proof of Proposition 4.4.3. For G^* the procedure is standard, cf. [62, Ch. I], and follows from (4.4.1),

$$G^* : Y^* \rightarrow \overline{\mathbb{R}},$$

$$G^*(u) = \sup_{v \in L^q(\Omega)} \left\{ \int_{\Omega} v u dx - G(v) \right\} = \frac{1}{p} \|u + u_N\|_{L^p(\Omega)}^p.$$

As for F^* , we follow [78, Section 2], with the crucial adaptation of using Proposition 4.4.4 in the last step

$$F^* : V^* \rightarrow \overline{\mathbb{R}},$$

$$F^*(\Psi^*) = \sup_{\Phi \in V} \left\{ \langle \Psi^*, \Phi \rangle_{V^*, V} - F(\Phi) \right\} = \sup_{\Phi \in V} \left\{ \langle \Psi^*, \Phi \rangle_{X^*, X} - I_\beta(\Phi) \right\}$$

$$= \sup_{\Phi \in V} \left\{ \langle \Psi^*, \Phi \rangle_{X^*, X} - \tilde{I}_\beta(\Phi) \right\} = \sup_{\substack{\Phi \in V \cap C_c^\infty(\mathbb{R}^n) \\ \|\Phi\|_{L^\infty(\mathbb{R}^n)} \leq \beta}} \langle \Psi^*, \Phi \rangle_{V^*, V}.$$

The condition in the last line that we can assume $\Phi \in C_c^\infty(\mathbb{R}^n)$ is the crucial point of Proposition 4.4.4, and the only place where convexity of Ω appears. Finally, by definition we have $\Lambda^* : Y^* \rightarrow V^*$. Therefore,

$$\begin{aligned} F^*(\Lambda^*u) &= \sup_{\substack{\Phi \in V \cap C_c^\infty(\mathbb{R}^n) \\ \|\Phi\|_{L^\infty(\mathbb{R}^n)} \leq \beta}} \langle \Lambda^*u, \Phi \rangle_{V^*,V} = \sup_{\substack{\Phi \in V \cap C_c^\infty(\mathbb{R}^n) \\ \|\Phi\|_{L^\infty(\mathbb{R}^n)} \leq \beta}} \langle u, \Lambda\Phi \rangle_{Y^*,Y} \\ &= \beta \sup_{\substack{\Phi \in C_c^\infty(\mathbb{R}^n, \mathbb{R}^n) \\ \|\Phi\|_{L^\infty(\mathbb{R}^n)} \leq 1}} \int_{\mathbb{R}^n} \chi_\Omega u (-\text{Div}_\alpha \Phi) dx = \beta \text{Var}_\alpha(\chi_\Omega u; \mathbb{R}^n), \end{aligned}$$

which concludes the proof. \square

From Theorem 4.4.1, we have the following result.

Corollary 4.4.5. *If Ω is convex, the problems (\mathcal{P}_R) and (\mathcal{Q}_R) are related by*

$$\begin{aligned} \min_{\Phi \in X_{Riesz}} \left\{ \frac{1}{q} \| -\text{Div}_\alpha \Phi \|_{L^q(\Omega)}^q - \int_{\Omega} u_N (-\text{Div}_\alpha \Phi) dx + I_\beta(\Phi) \right\} \\ = - \min_{u \in L^p(\Omega)} \left\{ \frac{\gamma}{p} \|u - u_N\|_{L^p(\Omega)}^p + \beta \text{Var}_\alpha(\chi_\Omega u; \mathbb{R}^n) \right\}. \end{aligned}$$

It is important to mention that the (predual) problem (\mathcal{Q}_R) has at least one solution. Moreover, we have the following results for the optimality conditions, cf. (4.4.3),

Lemma 4.4.6. *Let \bar{u} be the unique solution for (\mathcal{P}_R) and let $\bar{\Phi}$ be any solution for (\mathcal{Q}_R) . Then we have*

$$\Lambda^*\bar{u} \in \partial F(\bar{\Phi}) \Leftrightarrow \langle \Lambda^*\bar{u}, \mathbf{v} - \bar{\Phi} \rangle \leq 0 \quad \forall \mathbf{v} \in X_{Riesz}, \quad (4.4.12)$$

$$-\bar{u} \in \partial G(\Lambda\bar{\Phi}) \Leftrightarrow -\bar{u} = - \left| \text{Div}_\alpha \bar{\Phi} \right|^{q-2} \text{Div}_\alpha \bar{\Phi} - u_N. \quad (4.4.13)$$

Proof. It is clear that (4.4.12) follows from (4.4.2). On the other hand, if G is Gâteaux differentiable at $u \in Y$, then $\partial G(u) = \{G'(u)\}$, cf. [62, Prop. I.5.3]. In turn, the following property about the duality map, it is also well known:

$$\begin{aligned} \partial \| \cdot \|_{L^q(\Omega)}^q : L^q(\Omega) &\rightarrow L^p(\Omega) \\ u &\mapsto \{q|u|^{q-2}u\}, \end{aligned}$$

which proves (4.4.13) and finishes the proof. \square

Gagliardo-Type

Next, we focus on the Gagliardo case. We refer to [119] where they studied a related problem. As in the case of Riesz, we begin by establishing the existence and uniqueness of solution to (\mathcal{P}_G) .

Lemma 4.4.7. *For $p \in (1, p^\infty)$, the problem (\mathcal{P}_G) has a unique solution $\bar{u} \in bv_\alpha(\Omega) \cap L^p(\Omega)$.*

Proof. The proof is similar to the Riesz case in Lemma 4.4.2 after using Proposition 4.3.10. \square

Now, we characterize the minimizers of (\mathcal{P}_G) using the predual strategy as discussed in the Riesz case. Note that u does not need to be extended by zero outside Ω . As a result, our approach is largely motivated by [37, Section 2]. We now study the predual problem associated to (\mathcal{P}_G) . In a similar way as in the Riesz case, we consider the spaces

$$X_{\text{Gagliardo}} = \overline{\{\Phi : \Phi \in C_c^1(\Omega \times \Omega), \Phi(x, y) = -\Phi(y, x)\}}^{\|\cdot\|_{X_{\text{Gagliardo}}}}, \quad (4.4.14)$$

where

$$\|\Phi\|_{X_{\text{Gagliardo}}} := \sqrt[q]{\|\Phi\|_{L^q(\Lambda_{od} \Omega)}^q + \|\operatorname{div}_\alpha \Phi\|_{L^q(\Omega)}^q},$$

which is well defined because of Lemma 4.3.2; note that we may equivalently assume $\Phi \equiv 0$ in $(\mathbb{R}^n \times \mathbb{R}^n) \setminus (\Omega \times \Omega)$ and set

$$\|\Phi\|_{X_{\text{Gagliardo}}} := \sqrt[q]{\|\Phi\|_{L^q(\Lambda_{od} \mathbb{R}^n)}^q + \|\operatorname{div}_\alpha \Phi\|_{L^q(\mathbb{R}^n)}^q}.$$

As in the Riesz case, we will use the indicator function I_β for some $\beta > 0$. For $\Phi \in X_{\text{Gagliardo}}$, we define

$$I_\beta(\Phi) := \begin{cases} 0 & : \|\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq \beta, \\ +\infty & : \text{otherwise.} \end{cases}$$

As in the Riesz case, our main novelty is that we are able to pass from I_β to a new \tilde{I}_β which has better approximation properties.

Proposition 4.4.8. *If Ω is convex then for all $\Phi \in X_{\text{Gagliardo}}$,*

$$I_\beta(\Phi) = \tilde{I}_\beta(\Phi),$$

where

$$\tilde{I}_\beta(\Phi) := \begin{cases} 0 & : \text{if there exists } \Psi_k \in C_c^\infty(\Omega \times \Omega), \\ & \Psi_k \rightarrow \Phi \text{ in } X_{\text{Gagliardo}} \text{ such that } \sup_k \|\Psi_k\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq \beta, \\ +\infty & : \text{otherwise.} \end{cases}$$

Proof. We may assume without loss of generality that Ω is convex and $0 \in \Omega$. First, we establish that $I_\beta(\Phi) \leq \tilde{I}_\beta(\Phi)$ for all $\Phi \in X_{\text{Gagliardo}}$. The case $\tilde{I}_\beta(\Phi) = \infty$ is trivial. Suppose that $\tilde{I}_\beta(\Phi) = 0$, then there exist $\Psi_k \in C_c^\infty(\Omega \times \Omega)$ with $\Psi_k(x, y) = -\Psi_k(y, x)$ and $\sup_k \|\Psi_k\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq \beta$, such that $\Psi_k \rightarrow \Phi$ in $X_{\text{Gagliardo}}$. From the $L^q(\wedge^{od} \mathbb{R}^n)$ -convergence of Ψ_k , we can find a subsequence, still denoted by Ψ_k , such that

$$\frac{|\Psi_k(x, y) - \Phi(x, y)|}{|x - y|^{\frac{n}{q} + s}} \xrightarrow{k \rightarrow \infty} 0 \quad \text{for } \mathcal{L}^{2n}\text{-a.e. } (x, y) \in \mathbb{R}^{2n},$$

which in particular implies

$$|\Psi_k(x, y) - \Phi(x, y)| \xrightarrow{k \rightarrow \infty} 0 \quad \text{for } \mathcal{L}^{2n}\text{-a.e. } (x, y) \in \mathbb{R}^{2n}.$$

Thus, we have

$$|\Phi(x, y)| \leq \beta \quad \text{for } \mathcal{L}^{2n}\text{-a.e. } (x, y) \in \mathbb{R}^{2n},$$

which implies that $I_\beta(\Phi) = 0$ and proves that $I_\beta(\Phi) \leq \tilde{I}_\beta(\Phi)$ for all $\Phi \in X_{\text{Gagliardo}}$.

Now we to prove the opposite direction, i.e. $I_\beta(\Phi) \geq \tilde{I}_\beta(\Phi)$ for all $\Phi \in X_{\text{Gagliardo}}$. If $I_\beta(\Phi) = \infty$ there is nothing to show, so we actually need to show

$$I_\beta(\Phi) = 0 \quad \Rightarrow \quad \tilde{I}_\beta(\Phi) = 0.$$

Assuming that $\Phi \in X_{\text{Gagliardo}}$ satisfies $\|\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq \beta$, we prove that

$$\forall \varepsilon > 0 \exists \Theta \in C_c^\infty(\Omega \times \Omega), \quad \|\Theta\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq \beta, \quad \|\Theta - \Phi\|_{X_{\text{Gagliardo}}} < \varepsilon. \quad (4.4.15)$$

Step 1: In contrast to the Riesz case, we scale the functions inwards for the Gagliardo case, which ensures that the mollification produces a function still in $C_c^\infty(\Omega)$.

Using again the notation

$$\Omega_\rho := \rho\Omega = \{\rho x : x \in \Omega\}$$

in (4.4.10) with $\rho < 1$. Since Ω is convex and $0 \in \Omega$, we have that $\Omega_\rho \subset\subset \Omega$ for any $\rho < 1$. We prove that

$$\forall \varepsilon > 0 \exists \Theta_1 \in X_{\text{Gagliardo}}, \exists \rho < 1, \text{supp } \Theta_1 \subset \Omega_\rho \times \Omega_\rho, \|\Theta_1\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq \beta, \|\Theta_1 - \Phi\|_{X_{\text{Gagliardo}}} < \varepsilon. \quad (4.4.16)$$

For $\rho < 1$, we define

$$\Phi_\rho(x, y) := \Phi\left(\frac{x}{\rho}, \frac{y}{\rho}\right).$$

Then we have $\text{supp } \Phi_\rho \subset \Omega_\rho \times \Omega_\rho$ and $\|\Phi_\rho\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} = \|\Phi\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq \beta$. So in order to establish (4.4.16) we need to show

$$\|\Phi_\rho - \Phi\|_{X_{\text{Gagliardo}}} \xrightarrow{\rho \rightarrow 0} 0. \quad (4.4.17)$$

We first observe that

$$\frac{\Phi_\rho(x, y)}{|x - y|^{\frac{n}{q}}} = \rho^{-\frac{n}{q}} \frac{\Phi(x/\rho, y/\rho)}{|x/\rho - y/\rho|^{\frac{n}{q}}}.$$

So we have

$$\begin{aligned} \|\Phi_\rho - \Phi\|_{L^q(\Lambda_{od} \mathbb{R}^n)} &\leq \left| \rho^{-\frac{n}{q}} - 1 \right| \left\| \frac{\Phi(x/\rho, y/\rho)}{|x/\rho - y/\rho|^{\frac{n}{q}}} \right\|_{L^q(\mathbb{R}^n \times \mathbb{R}^n)} + \left\| \frac{\Phi(x/\rho, y/\rho)}{|x/\rho - y/\rho|^{\frac{n}{q}}} - \frac{\Phi(x, y)}{|x - y|^{\frac{n}{q}}} \right\|_{L^q(\mathbb{R}^n \times \mathbb{R}^n)} \\ &= \left| \rho^{-\frac{n}{q}} - 1 \right| \rho^{\frac{2n}{q}} \left\| \frac{\Phi(x, y)}{|x - y|^{\frac{n}{q}}} \right\|_{L^q(\mathbb{R}^n \times \mathbb{R}^n)} + \left\| \frac{\Phi(x/\rho, y/\rho)}{|x/\rho - y/\rho|^{\frac{n}{q}}} - \frac{\Phi(x, y)}{|x - y|^{\frac{n}{q}}} \right\|_{L^q(\mathbb{R}^n \times \mathbb{R}^n)} \\ &= \left| \rho^{-\frac{n}{q}} - 1 \right| \rho^{\frac{2n}{q}} \|\Phi\|_{L^q(\Lambda_{od} \mathbb{R}^n)} + \left\| \frac{\Phi(x/\rho, y/\rho)}{|x/\rho - y/\rho|^{\frac{n}{q}}} - \frac{\Phi(x, y)}{|x - y|^{\frac{n}{q}}} \right\|_{L^q(\mathbb{R}^n \times \mathbb{R}^n)} \\ &\xrightarrow{\rho \rightarrow 1} 0, \end{aligned}$$

where for the first term we use that $\|\Phi\|_{L^q(\Lambda_{od} \mathbb{R}^n)} = \|\Phi\|_{L^q(\Lambda_{od} \Omega)} < \infty$, for the second term we use Lemma B.0.4 in $\mathbb{R}^n \times \mathbb{R}^n$. Moreover, a direct computation from (4.3.4) yields

$$\text{div}_\alpha \Phi_\rho(x) = \rho^{-\alpha} (\text{div}_\alpha \Phi)(x/\rho) \quad \text{a.e. } x \in \mathbb{R}^n.$$

So again with Lemma B.0.4 we have

$$\begin{aligned} \|\operatorname{div}_\alpha \Phi_\rho - \operatorname{div}_\alpha \Phi\|_{L^q(\mathbb{R}^n)} &\leq |\rho^{-\alpha} - 1| \|(\operatorname{div}_\alpha \Phi)(\cdot/\rho)\|_{L^q(\mathbb{R}^n)} + \|(\operatorname{div}_\alpha \Phi)(\cdot/\rho) - \operatorname{div}_\alpha \Phi\|_{L^q(\mathbb{R}^n)} \\ &= |\rho^{-\alpha} - 1| \rho^{\frac{n}{q}} \|(\operatorname{div}_\alpha \Phi)\|_{L^q(\mathbb{R}^n)} + \|(\operatorname{div}_\alpha \Phi)(\cdot/\rho) - \operatorname{div}_\alpha \Phi\|_{L^q(\mathbb{R}^n)} \\ &\xrightarrow{\rho \rightarrow 1} 0 \end{aligned}$$

where we used crucially that by the support of Φ in $\Omega \times \Omega$ we have

$$\|\operatorname{div}_\alpha \Phi\|_{L^q(\mathbb{R}^n)} = \|\operatorname{div}_\alpha \Phi\|_{L^q(\Omega)} < \infty.$$

This establishes (4.4.17) and thus (4.4.16) is proven.

Step 2: Let Θ_1 and $\rho \in (0, 1)$ be from Step 1. Set $\delta_0 := \frac{D}{100}$ where $D := \operatorname{dist}(\Omega_\rho, \partial\Omega) > 0$. Let $\eta \in C_c^\infty(B(0, 1))$ be the usual symmetric mollifier, i.e. $\eta \geq 0$ and $\int \eta = 1$. For $\delta \in (0, \delta_0)$, we define $\eta_\delta(x) := \eta(x/\delta)/\delta^n$. Using the notation from (4.3.14), we define

$$\Psi_\delta(x, y) := (\eta_\delta * \Theta_1)(x, y).$$

Then $\Psi_\delta \in C_c^\infty(B(\Omega_\rho \times \Omega_\rho, \delta)) \subset C_c^\infty(\Omega \times \Omega)$ and $\|\Psi_\delta\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq \|\Theta_1\|_{L^\infty(\mathbb{R}^n \times \mathbb{R}^n)} \leq \beta$.

Notice that

$$\frac{(\eta_\delta * \Theta_1)(x, y)}{|x - y|^{n/q}} = (\eta_\delta * \Xi)(x, y) \quad (4.4.18)$$

where $\Xi(x', y') := \Theta_1(x', y')/|x' - y'|^{n/q}$. By the definition of $\Theta_1 \in L^q(\wedge_{od} \mathbb{R}^n)$, we have $\Xi \in L^q(\mathbb{R}^n \times \mathbb{R}^n)$. Thus, we have

$$\|\Psi_\delta - \Theta_1\|_{L^q(\wedge_{od}^1 \mathbb{R}^n)} \xrightarrow{\delta \rightarrow 0} 0. \quad (4.4.19)$$

For any $x \in \mathbb{R}^n$, by letting $y' = y - z$, we obtain

$$\begin{aligned} (\operatorname{div}_\alpha \Psi_\delta)(x) &= (\operatorname{div}_\alpha (\eta_\delta * \Theta_1))(x) = - \int_{\mathbb{R}^n} \frac{(\eta_\delta * \Theta_1)(x, y) - (\eta_\delta * \Theta_1)(y, x)}{|x - y|^{n+\alpha}} dy \\ &= - \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} \eta_\delta(z) (\Theta_1(x - z, y - z) - \Theta_1(y - z, x - z)) dz \right) \frac{dy}{|x - y|^{n+\alpha}} \\ &= \int_{\mathbb{R}^n} \eta_\delta(z) \left(- \int_{\mathbb{R}^n} \frac{\Theta_1(x - z, y') - \Theta_1(y', x - z)}{|(x - z) - y'|^{n+\alpha}} dy' \right) dz \\ &= \int_{\mathbb{R}^n} \eta_\delta(z) (\operatorname{div}_\alpha \Theta_1)(x - z) dz = (\eta_\delta * (\operatorname{div}_\alpha \Theta_1))(x). \end{aligned} \quad (4.4.20)$$

Thus, we have

$$\|\operatorname{div}_\alpha \Psi_\delta - \operatorname{div}_\alpha \Theta_1\|_{L^q(\mathbb{R}^n)} = \|\eta_\delta * (\operatorname{div}_\alpha \Theta_1) - \operatorname{div}_\alpha \Theta_1\|_{L^q(\mathbb{R}^n)} \rightarrow 0 \quad (4.4.21)$$

as $\delta \rightarrow 0^+$. Using (4.4.19) and (4.4.21), for a sufficiently small δ , the function $\Theta := \Psi_\delta$ satisfies the requirements in (4.4.15), which completes the proof. \square

Now we continue with the optimizing problem. We set $V = (X_{\text{Gagliardo}}, \|\cdot\|_{X_{\text{Gagliardo}}})$, $Y = (L^q(\Omega), \|\cdot\|_{L^q(\Omega)})$ and the operators

$$\begin{aligned} G : Y &\rightarrow \overline{\mathbb{R}}, & G(v) &:= \frac{1}{q} \|v\|_{L^q(\Omega)}^q - \int_{\Omega} u_N v dx, \\ F : V &\rightarrow \overline{\mathbb{R}}, & F(\Phi) &:= I_\beta(\Phi), \\ \Lambda : V &\rightarrow Y, & \Lambda(\Phi) &:= -\operatorname{div}_\alpha \Phi. \end{aligned} \quad (4.4.22)$$

Similarly as in Lemma 4.4.3, we have

Corollary 4.4.9. *Let Ω be open, bounded, and convex set, $V = (X_{\text{Gagliardo}}, \|\cdot\|_{X_{\text{Gagliardo}}})$, $Y = (L^q(\Omega), \|\cdot\|_{L^q(\Omega)})$, and let F, G and Λ defined as in (4.4.22), then for all $u \in L^p(\Omega)$*

$$\begin{aligned} G^*(-u) &= \frac{1}{p} \|u - u_N\|_{L^p(\Omega)}^p, & \text{and} \\ F^*(\Lambda^* u) &= \beta \operatorname{var}_\alpha(u; \Omega). \end{aligned}$$

This motivates us to consider the problem

$$\inf_{\Phi \in X_{\text{Gagliardo}}} \left\{ \frac{1}{q} \|-\operatorname{div}_\alpha \Phi\|_{L^q(\Omega)}^q - \int_{\Omega} u_N (-\operatorname{div}_\alpha \Phi) dx + I_\beta(\Phi) \right\}. \quad (\mathcal{Q}_G)$$

We have the following result (see, Corollary 4.4.5 for the Riesz case).

Corollary 4.4.10. *If Ω is an open, bounded, and convex set, the problems (\mathcal{P}_G) and (\mathcal{Q}_G) are related by*

$$\begin{aligned} &\min_{\Phi \in X_{\text{Gagliardo}}} \left\{ \frac{1}{q} \|-\operatorname{div}_\alpha \Phi\|_{L^q(\Omega)}^q - \int_{\Omega} u_N (-\operatorname{div}_\alpha \Phi) dx + I_\beta(\Phi) \right\} \\ &= - \min_{u \in L^p(\Omega)} \left\{ \frac{\gamma}{p} \|u - u_N\|_{L^p(\Omega)}^p + \beta \operatorname{var}_\alpha(u; \Omega) \right\}. \end{aligned}$$

Finally, we have the following optimality conditions as consequences of Theorem 4.4.1.

Corollary 4.4.11. *Let \bar{u} be the unique solution to (\mathcal{P}_G) and let $\bar{\Phi}$ be any solution to (\mathcal{Q}_G) , then*

$$\Lambda^*\bar{u} \in \partial F(\bar{\Phi}) \Leftrightarrow \langle \Lambda^*\bar{u}, \Psi - \bar{\Phi} \rangle \leq 0 \quad \forall \Psi \in X_{Gagliardo}, \text{ and} \quad (4.4.23)$$

$$-\bar{u} \in \partial G(\Lambda\bar{\Phi}) \Leftrightarrow -\bar{u} = -\left| \operatorname{div}_\alpha \bar{\Phi} \right|^{q-2} \operatorname{div}_\alpha \bar{\Phi} - u_N. \quad (4.4.24)$$

BIBLIOGRAPHY

- [1] M. Ainsworth, J. Guzmán, and F.-J. Sayas. Discrete extension operators for mixed finite element spaces on locally refined meshes. *Math. Comp.*, 85(302):2639–2650, 2016.
- [2] A. Alonso and A. Valli. An optimal domain decomposition preconditioner for low-frequency time-harmonic Maxwell equations. *Math. Comp.*, 68(226):607–631, 1999.
- [3] L. Álvarez, P.-L. Lions, and J.-M. Morel. Image selective smoothing and edge detection by nonlinear diffusion. ii. *SIAM Journal on Numerical Analysis*, 29:845–866, 1992.
- [4] H. Antil and S. Bartels. Spectral Approximation of Fractional PDEs in Image Processing and Phase Field Modeling. *Comput. Methods Appl. Math.*, 17(4):661–678, 2017.
- [5] H. Antil, S. Bartels, and A. Schikorra. Approximation of fractional harmonic maps. *IMA Journal of Numerical Analysis*, 07 2022. drac029.
- [6] H. Antil, T. S. Brown, R. Löhner, F. Togashi, and D. Verma. Deep neural nets with fixed bias configuration. *arXiv preprint arXiv:2107.01308*, 2021.
- [7] H. Antil and H. Díaz. Boundary control of time-harmonic eddy current equations. *arXiv preprint arXiv:2209.15129*, 2022.
- [8] H. Antil, H. Díaz, and E. Herberg. An optimal time variable learning framework for deep neural networks. *arXiv preprint arXiv:2204.08528*, 2022.
- [9] H. Antil, H. Díaz, T. Jing, and A. Schikorra. Nonlocal bounded variations with applications. *arXiv preprint arXiv:2208.11746*, 2022.
- [10] H. Antil, H. C. Elman, A. Onwunta, and D. Verma. Novel deep neural networks for solving bayesian statistical inverse. *arXiv preprint arXiv:2102.03974*, 2021.
- [11] H. Antil, C. G. Gal, and M. Warma. A unified framework for optimal control of fractional in time subdiffusive semilinear pdes. *Discrete and Continuous Dynamical Systems - Series S*, 10 2021.

- [12] H. Antil, R. Khatri, R. Löhner, and D. Verma. Fractional deep neural network via constrained optimization. *Machine Learning: Science and Technology*, 2(1):015003, 2020.
- [13] H. Antil, D. P. Kouri, M.-D. Lacasse, and D. Ridzal, editors. *Frontiers in PDE-constrained optimization*, volume 163 of *The IMA Volumes in Mathematics and its Applications*. Springer, New York, 2018. Papers based on the workshop held at the Institute for Mathematics and its Applications, Minneapolis, MN, June 6–10, 2016.
- [14] H. Attouch, G. Buttazzo, and G. Michaille. *Variational Analysis in Sobolev and BV Spaces*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2014.
- [15] T. Avant and K. A. Morgansen. Analytical bounds on the local lipschitz constants of relu networks. *ArXiv*, abs/2104.14672, 2021.
- [16] C. Bahriawati and C. Carstensen. Three MATLAB implementations of the lowest-order Raviart-Thomas MFEM with a posteriori error control. *Comput. Methods Appl. Math.*, 5(4):333–361, 2005.
- [17] S. Bartels. *Numerical methods for nonlinear partial differential equations*, volume 47 of *Springer Series in Computational Mathematics*. Springer, Cham, 2015.
- [18] M. Belishev and A. Glasman. Boundary control of the Maxwell dynamical system: lack of controllability by topological reasons. *ESAIM Control Optim. Calc. Var.*, 5:207–217, 2000.
- [19] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.
- [20] M. Benning, E. Celledoni, M. J. Ehrhardt, B. Owren, and C.-B. Schönlieb. Deep learning as optimal control problems: Models and numerical methods. *Journal of Computational Dynamics*, 6(2):171–198, 2019.
- [21] A. Bermúdez, R. Rodríguez, and P. Salgado. Numerical analysis of electric field formulations of the eddy current model. *Numer. Math.*, 102(2):181–201, 2005.
- [22] A. Bermúdez, R. Rodríguez, and P. Salgado. Numerical treatment of realistic boundary conditions for the eddy current problem in an electrode via Lagrange multipliers. *Math. Comp.*, 74(249):123–151, 2005.
- [23] K. Bessas. Fractional total variation denoising model with l^1 fidelity, 2021.
- [24] B. Bischke, P. Bhardwaj, A. Gautam, P. Helber, D. Borth, and A. Dengel. Detection of flooding events in social multimedia and satellite imagery using deep neural networks. In *MediaEval*, 2017.

- [25] V. Bommer and I. Yousept. Optimal control of the full time-dependent Maxwell equations. *ESAIM Math. Model. Numer. Anal.*, 50(1):237–261, 2016.
- [26] A. Bossavit. Most general non-local boundary conditions for the Maxwell equation in a bounded region. *COMPEL*, pages 239–245, 2000.
- [27] P. Bouboulis, K. Slavakis, and S. Theodoridis. Adaptive learning in complex reproducing kernel Hilbert spaces employing Wirtinger’s subgradients. *IEEE Transactions on Neural Networks*, 23:425–438, 03 2012.
- [28] C. Bourdarias, M. Gisclon, and S. Junca. Fractional BV spaces and applications to scalar conservation laws. *J. Hyperbolic Differ. Equ.*, 11(4):655–677, 2014.
- [29] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 01 2011.
- [30] D. H. Brandwood. A complex gradient operator and its application in adaptive array theory. *IEE Proceedings F: Communications Radar and Signal Processing*, 130(1):11–16, Feb. 1983.
- [31] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.
- [32] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.
- [33] A. Buffa and P. Ciarlet, Jr. On traces for functional spaces related to Maxwell’s equations. I. An integration by parts formula in Lipschitz polyhedra. *Math. Methods Appl. Sci.*, 24(1):9–30, 2001.
- [34] A. Buffa and P. Ciarlet, Jr. On traces for functional spaces related to Maxwell’s equations. II. Hodge decompositions on the boundary of Lipschitz polyhedra and applications. *Math. Methods Appl. Sci.*, 24(1):31–48, 2001.
- [35] A. Buffa, M. Costabel, and D. Sheen. On traces for $\mathbf{H}(\mathbf{curl}, \Omega)$ in Lipschitz domains. 276(2):845–867, 2002.
- [36] A. Buffa and R. Hiptmair. *Galerkin Boundary Element Methods for Electromagnetic Scattering*, pages 83–124. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

- [37] M. Burger, K. Frick, S. Osher, and O. Scherzer. Inverse total variation flow. *Multiscale Model. Simul.*, 6(2):365–395, 2007.
- [38] L. Caffarelli, J.-M. Roquejoffre, and O. Savin. Nonlocal minimal surfaces. *Comm. Pure Appl. Math.*, 63(9):1111–1144, 2010.
- [39] S. Cai, Z. Mao, Z. Wang, M. Yin, and G. E. Karniadakis. Physics-informed neural networks (pinns) for fluid mechanics: A review, 2021.
- [40] G. Caselli. Optimal control of an eddy current problem with a dipole source. *J. Math. Anal. Appl.*, 489(1):124152, 20, 2020.
- [41] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vision*, 20(1-2):89–97, 2004. Special issue on mathematics and image analysis.
- [42] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.
- [43] B. Chang, L. Meng, E. Haber, L. Ruthotto, D. Begert, and E. Holtham. Reversible architectures for arbitrarily deep residual neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [44] G. Chavent and K. Kunisch. Regularization of linear least squares problems by total bounded variation. *ESAIM Control Optim. Calc. Var.*, 2:359–376, 1997.
- [45] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. *NeuroImage*, 170:446–455, 2018. Segmenting the Brain.
- [46] K. Chen, K. Chen, Q. Wang, Z. He, J. Hu, and J. He. Short-term load forecasting with deep residual networks. *IEEE Transactions on Smart Grid*, 10(4):3943–3952, 2018.
- [47] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations, 2019.
- [48] P. Ciarlet, H. Wu, and J. Zou. Edge element methods for Maxwell’s equations with strong convergence for Gauss’ laws. *SIAM J. Numer. Anal.*, 52(2):779–807, 2014.
- [49] B. Cockburn and J. Gopalakrishnan. Incompressible finite elements via hybridization. II. The Stokes system in three space dimensions. *SIAM J. Numer. Anal.*, 43(4):1651–1672, 2005.
- [50] G. E. Comi, D. Spector, and G. Stefani. The fractional variation and the precise representative of $BV^{\alpha,p}$ functions. *Fractional Calculus and Applied Analysis*, 25(2):520–558, apr 2022.

- [51] G. E. Comi and G. Stefani. A distributional approach to fractional Sobolev spaces and fractional variation: asymptotics I. *preprint, arXiv.1910.13419*, 2019.
- [52] G. E. Comi and G. Stefani. A distributional approach to fractional Sobolev spaces and fractional variation: existence of blow-up. *J. Funct. Anal.*, 277(10):3373–3435, 2019.
- [53] C. Cortes, X. Gonzalvo, V. Kuznetsov, M. Mohri, and S. Yang. Adanet: Adaptive structural learning of artificial neural networks. In *International conference on machine learning*, pages 874–883. PMLR, 2017.
- [54] F. Da Lio and T. Rivière. Three-term commutator estimates and the regularity of $\frac{1}{2}$ -harmonic maps into spheres. *Anal. PDE*, 4(1):149–190, 2011.
- [55] R. DeVore, B. Hanin, and G. Petrova. Neural network approximation. *Acta Numer.*, 30:327–444, 2021.
- [56] E. Di Nezza, G. Palatucci, and E. Valdinoci. Hitchhiker’s guide to the fractional Sobolev spaces. *Bull. Sci. Math.*, 136(5):521–573, 2012.
- [57] K. Diethelm and N. J. Ford. Analysis of fractional differential equations. *J. Math. Anal. Appl.*, 265(2):229–248, 2002.
- [58] Q. Du, M. Gunzburger, R. B. Lehoucq, and K. Zhou. A nonlocal vector calculus, nonlocal volume-constrained problems, and nonlocal balance laws. *Math. Models Methods Appl. Sci.*, 23(3):493–540, 2013.
- [59] E. Dupont, A. Doucet, and Y. W. Teh. Augmented neural odes, 2019.
- [60] W. E. A proposal on machine learning via dynamical systems. *Commun. Math. Stat.*, 5(1):1–11, 2017.
- [61] W. E. Machine learning: Mathematical theory and scientific applications. *Notices of the American Mathematical Society*, 66(11):1813–1820, 2019.
- [62] I. Ekeland and R. Témam. *Convex Analysis and Variational Problems*. Society for Industrial and Applied Mathematics, 1999.
- [63] I. Ekeland and T. Turnbull. *Infinite-dimensional optimization and convexity*. Chicago Lectures in Mathematics. University of Chicago Press, Chicago, IL, 1983.
- [64] L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. Textbooks in Mathematics. CRC Press, Boca Raton, FL, revised edition, 2015.
- [65] R. M. Farber, A. S. Lapedes, R. Rico-Martínez, and I. G. Kevrekidis. Identification of continuous-time dynamical systems: Neural network based algorithms and parallel implementation. In R. F. Sincovec, D. E. Keyes, M. R. Leuze, L. R. Petzold, and D. A. Reed, editors, *PPSC*, pages 287–291. SIAM, 1993.

- [66] G. N. Gatica. *A simple introduction to the mixed finite element method*. Springer-Briefs in Mathematics. Springer, Cham, 2014. Theory and applications.
- [67] C. Geuzaine and J.-F. Remacle. Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities. *Internat. J. Numer. Methods Engrg.*, 79(11):1309–1331, 2009.
- [68] G. Gilboa and S. Osher. Nonlocal linear image regularization and supervised segmentation. *Multiscale Model. Simul.*, 6(2):595–630, 2007.
- [69] V. Girault and P.-A. Raviart. *Finite element approximation of the Navier-Stokes equations*, volume 749 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin-New York, 1979.
- [70] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.
- [71] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [72] J. Gopalakrishnan, L. E. García-Castillo, and L. F. Demkowicz. Nédélec spaces in affine coordinates. *Comput. Math. Appl.*, 49(7-8):1285–1294, 2005.
- [73] S. Gunther, L. Ruthotto, J. B. Schroder, E. C. Cyr, and N. R. Gauger. Layer-parallel training of deep residual neural networks. *SIAM Journal on Mathematics of Data Science*, 2(1):1–23, 2020.
- [74] E. Haber and L. Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1):014004, 2017.
- [75] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.
- [76] S. Hayou, E. Clerico, B. He, G. Deligiannidis, A. Doucet, and J. Rousseau. Stable resnet. In *International Conference on Artificial Intelligence and Statistics*, pages 1324–1332. PMLR, 2021.
- [77] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [78] M. Hintermüller and K. Kunisch. Total bounded variation regularization as a bilaterally constrained optimization problem. *SIAM J. Appl. Math.*, 64(4):1311–1333, 2004.

- [79] M. Hintermüller, C. N. Rautenberg, and S. Rösel. Density of convex intersections and applications. *Proc. A.*, 473(2205):20160919, 28, 2017.
- [80] M. Hinze. Magnetic energies and Feynman-Kac-Itô formulas for symmetric Markov processes. *Stoch. Anal. Appl.*, 33(6):1020–1049, 2015.
- [81] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.
- [82] Q. Hong, J. W. Siegel, and J. Xu. A priori analysis of stable neural network solutions to numerical pdes, 2021.
- [83] J. Horváth. On some composition formulas. *Proc. Amer. Math. Soc.*, 10:433–437, 1959.
- [84] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [85] J. A. Iglesias and G. Mercier. Geometric convergence in fractional laplacian regularization. *arXiv preprint arXiv:2201.13281*, 2022.
- [86] J. Ingmanns. Estimates for commutators of fractional differential operators via harmonic extension. *Master Thesis, arXiv:2012.12072*, 2020.
- [87] A. D. Jagtap, K. Kawaguchi, and G. E. Karniadakis. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *Journal of Computational Physics*, 404:109136, 2020.
- [88] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.
- [89] E. Kaiser, J. N. Kutz, and S. L. Brunton. Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proc. A.*, 474(2219):20180335, 25, 2018.
- [90] R. N. Kaul. Symmetric dual nonlinear programs in complex space. *J. Math. Anal. Appl.*, 33:140–148, 1971.
- [91] L. Kaup and B. Kaup. *Holomorphic functions of several variables*, volume 3 of *De Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 1983. An introduction to the fundamental theory, With the assistance of Gottfried Barthel, Translated from the German by Michael Bridgland.

- [92] A. A. Kilbas, H. M. Srivastava, and J. J. Trujillo. *Theory and applications of fractional differential equations*, volume 204 of *North-Holland Mathematics Studies*. Elsevier Science B.V., Amsterdam, 2006.
- [93] M. Kolmbauer and U. Langer. A robust preconditioned MinRes solver for distributed time-periodic eddy current optimal control problems. *SIAM J. Sci. Comput.*, 34(6):B785–B809, 2012.
- [94] K. Kreutz-Delgado. The complex gradient operator and the $\mathbb{C}\mathbb{R}$ -calculus. *arXiv preprint arXiv:0906.4835*, 2009.
- [95] J. E. Lagnese. Exact boundary controllability of Maxwell’s equations in a general region. *SIAM J. Control Optim.*, 27(2):374–388, 1989.
- [96] J. E. Lagnese. A singular perturbation problem in exact controllability of the Maxwell system. *ESAIM Control Optim. Calc. Var.*, 6:275–289, 2001.
- [97] J. E. Lagnese and G. Leugering. Time domain decomposition in final value optimal control of the Maxwell system. volume 8, pages 775–799. 2002. A tribute to J. L. Lions.
- [98] D. Lee, J. Yoo, S. Tak, and J. C. Ye. Deep residual learning for accelerated mri using magnitude and phase networks. *IEEE Transactions on Biomedical Engineering*, 65(9):1985–1995, 2018.
- [99] L. Lempert. The cauchy-riemann equations in infinite dimensions. *Journées équations aux dérivées partielles*, pages 1–8, 1998.
- [100] E. Lenzmann and A. Schikorra. On energy-critical half-wave maps into \mathbb{S}^2 . *Invent. Math.*, 213(1):1–82, 2018.
- [101] E. Lenzmann and A. Schikorra. Sharp commutator estimates via harmonic extensions. *Nonlinear Anal.*, 193:111375, 37, 2020.
- [102] N. Levinson. Linear programming in complex space. *J. Math. Anal. Appl.*, 14:44–62, 1966.
- [103] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations, 2021.
- [104] Y. Lin, X. Li, and C. Xu. Finite difference/spectral approximations for the fractional cable equation. *Math. Comp.*, 80(275):1369–1396, 2011.
- [105] Y. Lin and C. Xu. Finite difference/spectral approximations for the time-fractional diffusion equation. *J. Comput. Phys.*, 225(2):1533–1552, 2007.

- [106] H. Liu and P. Markowich. Selection dynamics for deep neural networks. *Journal of Differential Equations*, 269(12):11540–11574, 2020.
- [107] Y. Lu, A. Zhong, Q. Li, and B. Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *International Conference on Machine Learning*, pages 3276–3285. PMLR, 2018.
- [108] M. Ludwig. Anisotropic fractional perimeters. *J. Differential Geom.*, 96(1):77–93, 2014.
- [109] F. Mainardi. *Fractional calculus and waves in linear viscoelasticity*. Imperial College Press, London, 2010. An introduction to mathematical models.
- [110] M. C. Matos. Holomorphically bornological spaces and infinite-dimensional versions of Hartogs’ theorem. *J. London Math. Soc. (2)*, 17(2):363–368, 1978.
- [111] K. Mazowiecka and A. Schikorra. Fractional div-curl quantities and applications to nonlocal geometric equations. *J. Funct. Anal.*, 275(1):1–44, 2018.
- [112] P. Mironescu. Fine properties of functions: an introduction. Lecture, June 2005.
- [113] M. Mitolo and R. Araneo. A brief history of maxwell’s equations [history]. *IEEE Industry Applications Magazine*, 25(3):8–13, 2019.
- [114] B. Mond and M. Hanson. Symmetric duality for quadratic programming in complex space. *J. Math. Anal. Appl.*, 23:284–293, 1968.
- [115] P. Monk. *Finite element methods for Maxwell’s equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2003.
- [116] J.-C. Nédélec. Mixed finite elements in \mathbb{R}^3 . *Numer. Math.*, 35(3):315–341, 1980.
- [117] S. Nicaise, S. Stingelin, and F. Tröltzsch. On two optimal control problems for magnetic fields. *Comput. Methods Appl. Math.*, 14(4):555–573, 2014.
- [118] R. H. Nochetto, E. Otárola, and A. J. Salgado. A PDE approach to space-time fractional parabolic problems. *SIAM J. Numer. Anal.*, 54(2):848–873, 2016.
- [119] M. Novaga and F. Onoue. Local hölder regularity of minimizers for nonlocal variational problems, 2021.
- [120] J. T. Oden and L. F. Demkowicz. *Applied functional analysis*. Textbooks in Mathematics. CRC Press, Boca Raton, FL, 2018.
- [121] F. J. Pineda. Generalization of back-propagation to recurrent neural networks. *Phys. Rev. Lett.*, 59:2229–2232, 1987.
- [122] H. Poincaré. Sur les propriétés du potentiel et sur les fonctions abéliennes. *Acta mathematica*, 22:89–178, 1898.

- [123] S. Rao, D. Wilton, and A. Glisson. Electromagnetic scattering by surfaces of arbitrary shape. *IEEE Transactions on Antennas and Propagation*, 30(3):409–418, 1982.
- [124] W. Rawat and Z. Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.
- [125] W. Ring and B. Wirth. Optimization methods on Riemannian manifolds and their application to shape space. *SIAM J. Optim.*, 22(2):596–627, 2012.
- [126] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [127] L. I. Rudin. *Images, numerical analysis of singularities and shock filters*. PhD thesis, California Institute of Technology, 1987.
- [128] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.
- [129] L. Ruthotto and E. Haber. Deep neural networks motivated by partial differential equations. *Journal of Mathematical Imaging and Vision*, 62(3):352–364, 2020.
- [130] F.-J. Sayas, T. Brown, and M. Hassell. *Variational techniques for elliptic partial differential equations*. CRC Press, Boca Raton, FL, 2019. Theoretical tools and advanced applications.
- [131] A. Schikorra. L^p -gradient harmonic maps into spheres and $SO(N)$. *Differential Integral Equations*, 28(3-4):383–408, 2015.
- [132] A. Schikorra. ε -regularity for systems involving non-local, antisymmetric operators. *Calc. Var. Partial Differential Equations*, 54(4):3531–3570, 2015.
- [133] A. Schikorra, D. Spector, and J. Van Schaftingen. An L^1 -type estimate for Riesz potentials. *Rev. Mat. Iberoam.*, 33(1):291–303, 2017.
- [134] C.-B. Schoenlieb, M. Benning, M. Ehrhardt, B. Owren, and E. Celledoni. Research data supporting " deep learning as optimal control problems". 2019.
- [135] T.-T. Shieh and D. E. Spector. On a new class of fractional partial differential equations. *Adv. Calc. Var.*, 8(4):321–336, 2015.
- [136] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.
- [137] S. Sonoda and N. Murata. Transport analysis of infinitely deep neural network. *Journal of Machine Learning Research*, 20(2):1–52, 2019.

- [138] L. Sorber, M. Van Barel, and L. De Lathauwer. Unconstrained optimization of real functions in complex variables. *SIAM J. OPTIM*, 22, 2012.
- [139] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. *arXiv preprint arXiv:1507.06228*, 2015.
- [140] Y. Tai, J. Yang, and X. Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.
- [141] T. Tam Nhan Quyen, H. Antil, and H. Díaz. Optimal control of parameterized maxwell’s system: Reduced basis, convergence analysis, and a posteriori error estimates. *To appear: Math Control & Related Fields*, 2022.
- [142] V. E. Tarasov. Differential equations with fractional derivative and universal map with memory. *J. Phys. A*, 42(46):465102, 13, 2009.
- [143] L. Tartar. On the characterization of traces of a sobolev space used for Maxwell’s equation. *Proceedings of a meeting held in Bordeaux, in honour of Michel Artola*, 1997.
- [144] M. Torres. On the dual of BV . In *Contributions of Mexican mathematicians abroad in pure and applied mathematics. Second meeting “Matemáticos Mexicanos en el Mundo”, Centro de Investigación en Matemáticas, Guanajuato, Mexico, December 15–19, 2014*, pages 115–129. Providence, RI: American Mathematical Society (AMS); México: Sociedad Matemática Mexicana, 2018.
- [145] F. Tröltzsch and A. Valli. Optimal control of low-frequency electromagnetic fields in multiply connected conductors. *Optimization*, 65(9):1651–1673, 2016.
- [146] F. Tröltzsch and I. Yousept. PDE-constrained optimization of time-dependent 3D electromagnetic induction heating by alternating voltages. *ESAIM Math. Model. Numer. Anal.*, 46(4):709–729, 2012.
- [147] A. van den Bos. Complex gradient and hessian. *IEE Proceedings - Vision, Image and Signal Processing*, 141(6):380–383, 1994.
- [148] C. J. Weiss, B. G. van Bloemen Waanders, and H. Antil. Fractional operators applied to geophysical electromagnetics. *Geophysical Journal International*, 220(2):1242–1259, 2020.
- [149] W. Wirtinger. Zur formalen Theorie der Funktionen von mehr komplexen Veränderlichen. *Math. Ann.*, 97(1):357–375, 1927.
- [150] S. Wu, S. Zhong, and Y. Liu. Deep residual learning for image steganalysis. *Multimedia tools and applications*, 77(9):10437–10453, 2018.

- [151] G. Yang, X. Huang, Z. Hao, M.-Y. Liu, S. Belongie, and B. Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4541–4550, 2019.
- [152] I. Yousept. Finite element analysis of an optimal control problem in the coefficients of time-harmonic eddy current equations. *J. Optim. Theory Appl.*, 154(3):879–903, 2012.
- [153] I. Yousept. Optimal control of Maxwell’s equations with regularized state constraints. *Comput. Optim. Appl.*, 52(2):559–581, 2012.
- [154] I. Yousept. Optimal control of quasilinear $H(\text{curl})$ -elliptic partial differential equations in magnetostatic field problems. *SIAM J. Control Optim.*, 51(5):3624–3651, 2013.
- [155] I. Yousept. Optimal bilinear control of eddy current equations with grad-div regularization. *J. Numer. Math.*, 23(1):81–98, 2015.
- [156] C. Yu, H. Schumacher, and K. Crane. Repulsive curves. *ACM Trans. Graph.*, 40(2), 2021.
- [157] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei. Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8):4274–4288, 2018.
- [158] Y. Zhou. Fractional Sobolev extension and imbedding. *Trans. Amer. Math. Soc.*, 367(2):959–979, 2015.
- [159] M. A. Zorn. Characterization of analytic functions in Banach spaces. *Ann. of Math. (2)*, 46:585–593, 1945.
- [160] M. A. Zorn. Gâteaux differentiability and essential boundedness. *Duke Math. J.*, 12:579–583, 1945.
- [161] M. A. Zorn. Derivatives and Fréchet differentials. *Bull. Amer. Math. Soc.*, 52:133–137, 1946.

Appendix A

DERIVATIVES OF THE LAGRANGIAN \mathcal{L}

We provide detailed calculations of derivatives needed in Section 3.5.2. To this end we recall

$$\begin{aligned} \mathcal{L}(y, \theta, \phi) = & J(\theta) - \sum_{\ell=1}^{L-1} \langle y^{[\ell]} - P_{\ell-1}^{\ell} y^{[\ell-1]} + \sum_{j=0}^{\ell-2} a_{\ell-1,j} (P_{j+1}^{\ell} y^{[j+1]} - P_j^{\ell} y^{[j]}) \\ & - (\tau^{[\ell-1]})^{\gamma} c_{\gamma-1}^{-1} \sigma(W^{[\ell-1]} y^{[\ell-1]} + b^{[\ell-1]}), \phi^{[\ell]} \rangle \\ & - \langle y^{[L]} - W^{[L-1]} y^{[L-1]}, \phi^{[L]} \rangle \end{aligned}$$

A.0.1 Derivative with respect to $y^{[\ell]}$

Here, we calculate the variation of \mathcal{L} with respect to $y^{[\ell]}$ for $\ell = 1, \dots, L$:

$$\begin{aligned} \partial_{y^{[\ell]}} \mathcal{L}(y, \theta, \phi) = & \partial_{y^{[\ell]}} J(\theta) - \partial_{y^{[\ell]}} \left(\sum_{k=1}^{L-1} \langle y^{[k]} - P_{k-1}^k y^{[k-1]}, \phi^{[k]} \rangle \right) \\ & - \partial_{y^{[\ell]}} \left(\sum_{k=1}^{L-1} \left\langle \sum_{j=0}^{k-2} a_{k-1,j} (P_{j+1}^k y^{[j+1]} - P_j^k y^{[j]}), \phi^{[k]} \right\rangle \right) \\ & + \partial_{y^{[\ell]}} \left(\sum_{k=1}^{L-1} \left\langle (\tau^{[k-1]})^{\gamma} c_{\gamma-1}^{-1} \sigma(W^{[k-1]} y^{[k-1]} + b^{[k-1]}), \phi^{[k]} \right\rangle \right) \\ & - \partial_{y^{[\ell]}} \langle y^{[L]} - W^{[L-1]} y^{[L-1]}, \phi^{[L]} \rangle. \end{aligned}$$

From e.g. [12, Section 4.2.] we have most components of this expression already given. The main difference lies in the factors $a_{k-1,j}$, since the contained $\tau^{[j]}, \dots, \tau^{[k-1]}$ are

variable now. We only calculate the remaining unknown term, i.e. the second line in the above equation. First we rewrite the double sum in the following way:

$$\begin{aligned}
& - \sum_{k=1}^{L-1} \left\langle \sum_{j=0}^{k-2} a_{k-1,j} (P_{j+1}^k y^{[j+1]} - P_j^k y^{[j]}), \phi^{[k]} \right\rangle \\
& = \sum_{k=0}^{L-2} \left(\sum_{j=0}^{k-1} a_{k,j} \langle P_j^{k+1} y^{[j]}, \phi^{[k+1]} \rangle - \sum_{j=0}^{k-1} a_{k,j} \langle P_{j+1}^{k+1} y^{[j+1]}, \phi^{[k+1]} \rangle \right) \\
& = \sum_{k=0}^{L-2} \left(\sum_{j=0}^{k-1} a_{k,j} \langle P_j^{k+1} y^{[j]}, \phi^{[k+1]} \rangle - \sum_{j=1}^k a_{k,j-1} \langle P_j^{k+1} y^{[j]}, \phi^{[k+1]} \rangle \right).
\end{aligned}$$

Now we take the derivative with respect to $y^{[\ell]}$ and can apply the sum rule of differentiation:

$$\begin{aligned}
& \sum_{k=0}^{L-2} \sum_{j=0}^{k-1} a_{k,j} \partial_{y^{[\ell]}} \langle y^{[j]}, (P_j^{k+1})^\top \phi^{[k+1]} \rangle - \sum_{k=0}^{L-2} \sum_{j=1}^k a_{k,j-1} \partial_{y^{[\ell]}} \langle y^{[j]}, (P_j^{k+1})^\top \phi^{[k+1]} \rangle \\
& = \sum_{k=0, \ell \leq k-1}^{L-2} a_{k,\ell} (P_\ell^{k+1})^\top \phi^{[k+1]} - \sum_{k=0, \ell \leq k}^{L-2} a_{k,\ell-1} (P_\ell^{k+1})^\top \phi^{[k+1]} \\
& = \sum_{k=\ell+1}^{L-2} a_{k,\ell} (P_\ell^{k+1})^\top \phi^{[k+1]} - \sum_{k=\ell}^{L-2} a_{k,\ell-1} (P_\ell^{k+1})^\top \phi^{[k+1]}
\end{aligned}$$

For $\ell = L$ and $\ell = L - 1$ this derivative vanishes. For $\ell = 1, \dots, L - 2$ we can slightly rewrite to get

$$\begin{aligned}
& \sum_{k=\ell+1}^{L-2} a_{k,\ell} (P_\ell^{k+1})^\top \phi^{[k+1]} - \sum_{k=\ell}^{L-2} a_{k,\ell-1} (P_\ell^{k+1})^\top \phi^{[k+1]} \\
& = \sum_{k=\ell+2}^{L-1} (a_{k-1,\ell} - a_{k-1,\ell-1}) (P_\ell^k)^\top \phi^{[k]} - a_{\ell,\ell-1} (P_\ell^{\ell+1})^\top \phi^{[\ell+1]}.
\end{aligned}$$

Now, the derivative can be assembled.

A.0.2 Derivative with respect to $\tau^{[\ell]}$

Care must be observed as $a_{k,j}$ contains $\tau^{[j]}, \dots, \tau^{[k]}$. The derivative of $\mathcal{L}(y, \theta, \phi)$ with respect to $\tau^{[\ell]}$ for $\ell = 0, \dots, L-2$ therefore consists of the following terms:

$$\begin{aligned} \partial_{\tau^{[\ell]}} \mathcal{L}(y, \theta, \phi) &= \partial_{\tau^{[\ell]}} J(\theta) - \partial_{\tau^{[\ell]}} \left(\sum_{k=1}^{L-1} \left\langle \sum_{j=0}^{k-2} a_{k-1,j} (P_{j+1}^k y^{[j+1]} - P_j^k y^{[j]}), \phi^{[k]} \right\rangle \right) \\ &\quad + \partial_{\tau^{[\ell]}} \left(\sum_{k=1}^{L-1} \left\langle (\tau^{[k-1]})^\gamma c_{\gamma-1}^{-1} \sigma(W^{[k-1]} y^{[k-1]} + b^{[k-1]}), \phi^{[k]} \right\rangle \right). \end{aligned}$$

We make an index shift from $k-1$ to k , use the sum rule of differentiation and the fact that only $\tau^{[j]}, \dots, \tau^{[k]}$ are contained in $a_{k,j}$ to obtain

$$\begin{aligned} &\partial_{\tau^{[\ell]}} \left(- \sum_{k=1}^{L-1} \left\langle \sum_{j=0}^{k-2} a_{k-1,j} (P_{j+1}^k y^{[j+1]} - P_j^k y^{[j]}), \phi^{[k]} \right\rangle \right) \\ &= - \sum_{k=0}^{L-2} \sum_{j=0}^{k-1} \partial_{\tau^{[\ell]}}(a_{k,j}) \left\langle P_{j+1}^{k+1} y^{[j+1]} - P_j^{k+1} y^{[j]}, \phi^{[k+1]} \right\rangle \\ &= - \sum_{k=\ell}^{L-2} \sum_{j=0}^{\min\{k-1, \ell\}} \partial_{\tau^{[\ell]}}(a_{k,j}) \left\langle P_{j+1}^{k+1} y^{[j+1]} - P_j^{k+1} y^{[j]}, \phi^{[k+1]} \right\rangle. \end{aligned}$$

Using the above equality, we arrive at

$$\begin{aligned} \partial_{\tau^{[\ell]}} \mathcal{L}(y, \theta, \phi) &= \partial_{\tau^{[\ell]}} J(\theta) - \sum_{k=\ell}^{L-2} \sum_{j=0}^{\min\{k-1, \ell\}} \partial_{\tau^{[\ell]}}(a_{k,j}) \left\langle P_{j+1}^{k+1} y^{[j+1]} - P_j^{k+1} y^{[j]}, \phi^{[k+1]} \right\rangle \\ &\quad + \left\langle \gamma (\tau^{[\ell]})^{\gamma-1} c_{\gamma-1}^{-1} \sigma(W^{[\ell]} y^{[\ell]} + b^{[\ell]}), \phi^{[\ell+1]} \right\rangle. \end{aligned}$$

For implementations we may want to understand the double summation in more detail. In fact, it can be split up into 3 terms in the following way:

$$\begin{aligned}
& - \sum_{k=\ell}^{L-2} \sum_{j=0}^{\min\{k-1, \ell\}} \partial_{\tau^{[\ell]}}(a_{k,j}) \left\langle P_{j+1}^{k+1} y^{[j+1]} - P_j^{k+1} y^{[j]}, \phi^{[k+1]} \right\rangle \\
& = - \sum_{j=0}^{\ell-1} \partial_{\tau^{[\ell]}}(a_{\ell,j}) \left\langle P_{j+1}^{\ell+1} y^{[j+1]} - P_j^{\ell+1} y^{[j]}, \phi^{[\ell+1]} \right\rangle \\
& \quad - \sum_{k=\ell+1}^{L-2} \sum_{j=0}^{\ell-1} \partial_{\tau^{[\ell]}}(a_{k,j}) \left\langle P_{j+1}^{k+1} y^{[j+1]} - P_j^{k+1} y^{[j]}, \phi^{[k+1]} \right\rangle \\
& \quad - \sum_{k=\ell+1}^{L-2} \partial_{\tau^{[\ell]}}(a_{k,\ell}) \left\langle P_{\ell+1}^{k+1} y^{[\ell+1]} - P_{\ell}^{k+1} y^{[\ell]}, \phi^{[k+1]} \right\rangle.
\end{aligned}$$

Now for each of the above terms we can easily compute the contained derivative using basic differentiation rules. Employing the product rule for $j = 0, \dots, \ell - 1$, i.e. $j < \ell$, we get

$$\begin{aligned}
\partial_{\tau^{[\ell]}}(a_{\ell,j}) & = (1 - \gamma) \frac{(\tau^{[\ell]})^\gamma}{\tau^{[j]}} \left(\left(\sum_{i=j}^{\ell} \tau^{[i]} \right)^{-\gamma} - \left(\sum_{i=j+1}^{\ell} \tau^{[i]} \right)^{-\gamma} \right) \\
& \quad + \gamma \frac{(\tau^{[\ell]})^{\gamma-1}}{\tau^{[j]}} \left(\left(\sum_{i=j}^{\ell} \tau^{[i]} \right)^{1-\gamma} - \left(\sum_{i=j+1}^{\ell} \tau^{[i]} \right)^{1-\gamma} \right).
\end{aligned}$$

For $j = 0, \dots, \ell - 1$ and $k = \ell + 1, \dots, L - 2$, i.e. $j < \ell < k$, we have

$$\partial_{\tau^{[\ell]}}(a_{k,j}) = (1 - \gamma) \frac{(\tau^{[k]})^\gamma}{\tau^{[j]}} \left(\left(\sum_{i=j}^k \tau^{[i]} \right)^{-\gamma} - \left(\sum_{i=j+1}^k \tau^{[i]} \right)^{-\gamma} \right).$$

And finally, using the product rule again for $k = \ell + 1, \dots, L - 2$, i.e. $\ell < k$, we see

$$\partial_{\tau^{[\ell]}}(a_{k,\ell}) = \frac{(\tau^{[k]})^\gamma}{(\tau^{[\ell]})^2} \left(\left(\sum_{i=\ell+1}^k \tau^{[i]} \right)^{1-\gamma} - \left(\sum_{i=\ell}^k \tau^{[i]} \right)^{1-\gamma} \right) + (1 - \gamma) \frac{(\tau^{[k]})^\gamma}{\tau^{[\ell]}} \left(\sum_{i=\ell}^k \tau^{[i]} \right)^{-\gamma}$$

Appendix B

SCALING IN L^P -NORMS AND STAR-SHAPED DOMAINS

In this appendix we state and prove for the convenience of the reader some facts about star-shaped domains that are most likely well-known to experts.

Denote the $n - 1$ -dimensional unit sphere by $\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : |x| = 1\}$. For $x \in \mathbb{S}^{n-1}$.

Lemma B.0.1. *Assume $\lambda : \mathbb{S}^{n-1} \rightarrow (0, \infty)$ is continuous and consider*

$$\Omega = \left\{ x \in \mathbb{R}^n \setminus \{0\} : |x| < \lambda \left(\frac{x}{|x|} \right) \right\} \cup \{0\}.$$

For $\rho > 0$ set

$$\Omega_\rho := \{\rho x : x \in \Omega\}$$

then we have for any $\rho_1 < \rho_2$

$$\text{dist}(\Omega_{\rho_1}, \mathbb{R}^n \setminus \Omega_{\rho_2}) > 0.$$

Proof. We first observe

$$\partial\Omega = \left\{ x \in \mathbb{R}^n \setminus \{0\} : |x| = \lambda \left(\frac{x}{|x|} \right) \right\}. \quad (\text{B.0.1})$$

Indeed $\bar{x} \in \partial\Omega$. Since $0 \in \Omega$ and Ω is open by continuity of λ , we have $\bar{x} \neq 0$. Then there exists $0 \neq x_k \in \Omega$, $0 \neq y_k \in \mathbb{R}^n \setminus \Omega$ such that $\lim_k |x_k - \bar{x}| = \lim_k |y_k - \bar{x}| = 0$. We have

$$|x_k| < \lambda \left(\frac{x_k}{|x_k|} \right), \quad |y_k| \geq \lambda \left(\frac{y_k}{|y_k|} \right) \quad \forall k.$$

Since $x_k, y_k, \bar{x} \neq 0$ these expressions are continuous and passing to the limit as $k \rightarrow \infty$,

$$|\bar{x}| \leq \lambda \left(\frac{\bar{x}}{|\bar{x}|} \right), \quad |\bar{x}| \geq \lambda \left(\frac{\bar{x}}{|\bar{x}|} \right) \quad \forall k.$$

This implies $|\bar{x}| = \lambda\left(\frac{\bar{x}}{|\bar{x}|}\right)$ and thus we have established

$$\partial\Omega \subseteq \left\{ x \in \mathbb{R}^n \setminus \{0\} : |x| = \lambda\left(\frac{x}{|x|}\right) \right\}.$$

Now assume $\bar{x} \in \mathbb{R}^n \setminus \{0\}$ with $|\bar{x}| = \lambda\left(\frac{\bar{x}}{|\bar{x}|}\right)$. Then for $\mu > 0$ we have

$$|\mu\bar{x}| = \mu\lambda\left(\frac{\mu\bar{x}}{|\mu\bar{x}|}\right).$$

Thus, if $\mu > 1$ we have $\mu\bar{x} \notin \Omega$ and if $\mu < 1$ we have $\mu\bar{x} \in \Omega$. In particular,

$$x_k := \left(1 - \frac{1}{k}\right)\bar{x} \in \Omega, \quad y_k := \left(1 + \frac{1}{k}\right)\bar{x} \notin \Omega,$$

and $\lim_{k \rightarrow \infty} x_k = \lim_{k \rightarrow \infty} y_k = \bar{x}$, so $\bar{x} \in \overline{\Omega} \cap \overline{\mathbb{R}^n \setminus \Omega} = \partial\Omega$. This implies

$$\partial\Omega \supseteq \left\{ x \in \mathbb{R}^n \setminus \{0\} : |x| = \lambda\left(\frac{x}{|x|}\right) \right\}.$$

So (B.0.1) is established.

Next we observe

$$\Omega_\rho = \left\{ x \in \mathbb{R}^n \setminus \{0\} : |x| < \rho\lambda\left(\frac{x}{|x|}\right) \right\} \cup \{0\}.$$

In particular if $\rho_1 < \rho_2$ we have that

$$\Omega_{\rho_1} \cap (\mathbb{R}^n \setminus \Omega_{\rho_2}) = \left\{ x : |x| < \rho_1\lambda\left(\frac{x}{|x|}\right), \quad \text{and} \quad |x| \geq \rho_2\lambda\left(\frac{x}{|x|}\right) \right\} = \emptyset.$$

Since Ω_{ρ_1} and $(\mathbb{R}^n \setminus \Omega_{\rho_2})$ are disjoint, and Ω_{ρ_1} is bounded we conclude that

$$\begin{aligned} \text{dist}(\Omega_{\rho_1}, \mathbb{R}^n \setminus \Omega_{\rho_2}) &= \text{dist}(\partial\Omega_{\rho_1}, \partial\Omega_{\rho_2}) \\ &= \inf_{x, y \in \mathbb{R}^n} \left| \rho_1 \frac{x}{|x|} \lambda\left(\frac{x}{|x|}\right) - \rho_2 \frac{y}{|y|} \lambda\left(\frac{y}{|y|}\right) \right| \\ &= \inf_{x, y \in \mathbb{S}^{n-1}} \left| \rho_1 x \lambda(x) - \rho_2 y \lambda(y) \right|. \end{aligned}$$

Since $\lambda(\cdot)$ is continuous and \mathbb{S}^{n-1} is compact, this infimum is attained at some $\bar{x}, \bar{y} \in \mathbb{S}^{n-1}$,

$$\text{dist}(\Omega_{\rho_1}, \mathbb{R}^n \setminus \Omega_{\rho_2}) = \left| \rho_1 \bar{x} \lambda(\bar{x}) - \rho_2 \bar{y} \lambda(\bar{y}) \right|$$

We claim that $|\rho_1 \bar{x} \lambda(\bar{x}) - \rho_2 \bar{y} \lambda(\bar{y})| > 0$. Indeed if this was not the case we would have

$$\rho_1 \bar{x} \lambda(\bar{x}) = \rho_2 \bar{y} \lambda(\bar{y})$$

Since the scalar factors $\rho_1, \rho_2, \lambda(\bar{x}), \lambda(\bar{y})$ are all positive – and $|\bar{x}| = |\bar{y}| = 1$ this implies that $\bar{x} = \bar{y}$. Whence we would find

$$\rho_1 \lambda(\bar{x}) = \rho_2 \lambda(\bar{x}),$$

and thus – since $\lambda(\bar{x}) \in (0, \infty)$, $\rho_1 = \rho_2$ – a contradiction to $\rho_1 < \rho_2$. Thus we have established

$$\text{dist}(\Omega_{\rho_1}, \mathbb{R}^n \setminus \Omega_{\rho_2}) > 0.$$

□

In Lemma B.0.2, the continuity of λ is not guaranteed for generic star-shaped domain – even if their boundaries are Lipschitz. We provide two examples in Figure B.1. The first example is the union of an open disk and an open sector. The second is an open unit disk with a slit, i.e. the ray $\{(c + 1/2, c + 1/2) : c \geq 0\}$ is excluded from the disk.

However, the assumptions of the set Ω in Lemma B.0.1 are satisfied if Ω is star-shaped w.r.t to an *open neighborhood* of the origin – this can be obtained by a careful inspection of the proof below. We will focus on convexity here.

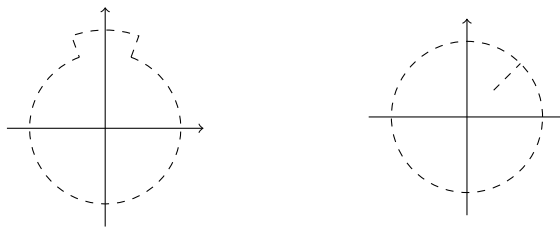


Figure B.1: Examples of star-shaped sets with discontinuous λ . Both sets are star-shaped with respect to the origin, and the first has even Lipschitz continuous boundary – however the conclusions of Lemma B.0.1 are not true.

Lemma B.0.2. *Let Ω be an open, bounded, convex set with $0 \in \Omega$, then there exists continuous $\lambda : \mathbb{S}^{n-1} \rightarrow (0, \infty)$ such that*

$$\Omega = \left\{ x \in \mathbb{R}^n \setminus \{0\} : |x| < \lambda \left(\frac{x}{|x|} \right) \right\} \cup \{0\}. \quad (\text{B.0.2})$$

In particular the results of Lemma B.0.1 are true.

Proof. For $x \in \mathbb{S}^{n-1}$, we define

$$\lambda(x) := \sup \{ r \geq 0 : rx \in \Omega \}. \quad (\text{B.0.3})$$

Since Ω is open and $0 \in \Omega$ there exists a ball $B(0, a) \subset \Omega$, and thus $\lambda(x) \geq r$ for all $x \in \mathbb{S}^{n-1}$. Since Ω is bounded there must be some $b > 0$ such that $\lambda(x) \leq b$ for all $x \in \mathbb{S}^{n-1}$.

We first establish (B.0.2). If $x \in \Omega$ then $|x| \frac{x}{|x|} \in \Omega$ and since $0 \in \Omega$ we have that $r \frac{x}{|x|} \in \Omega$ for all $r \in [0, |x|]$. Since Ω is open, there actually must be some $\delta > 0$ such that $r \frac{x}{|x|} \in \Omega$ for all $r \in [0, |x| + \delta]$. Thus $\lambda(x/|x|) \geq |x| + \delta > |x|$.

On the other hand if $x \in \mathbb{R}^n \setminus \{0\}$ and $|x| < \lambda(x/|x|)$, then by definition of $\lambda(\cdot)$ there must be some $r > |x|$ such that $rx/|x| \in \Omega$. Since $0 \in \Omega$ and Ω is convex we conclude that $x = |x|x/|x| \in \Omega$. Thus (B.0.2) is established.

It remains to prove the continuity of λ on \mathbb{S}^{n-1} . Given any $\bar{x} \in \mathbb{S}^{n-1}$, we let $\{x_k\}_{k=1}^\infty \subseteq \mathbb{S}^{n-1}$ be a sequence such that $x_k \xrightarrow{k \rightarrow \infty} \bar{x}$.

Recall that the open ball $B(0, a) \subset \Omega$. We denote the open cone from $\lambda(\bar{x})\bar{x}$ to $B(0, a)$ as

$$A := \{ \theta \lambda(\bar{x})\bar{x} + (1 - \theta)z : z \in B(0, a), \theta \in [0, 1) \}. \quad (\text{B.0.4})$$

Clearly, A is an open set. Also, whenever $\theta \in [0, 1)$ we have that $\theta \lambda(\bar{x})\bar{x} \in \Omega$, by convexity of Ω and definition of $\lambda(\cdot)$. Since z is taken from an open ball $B(0, a) \subset \Omega$ we conclude that $\theta \bar{x} + (1 - \theta)z \in \Omega$. That is we have $A \subset \Omega$.

Similarly, we define the open sets A_k by

$$A_k := \{ \theta \lambda(x_k)x_k + (1 - \theta)z : z \in B(0, a), \theta \in [0, 1) \} \subset \Omega \quad (\text{B.0.5})$$

Now we assume that there exists $\varepsilon > 0$ and a sequence $x_k \in \mathbb{S}^{n-1}$ converging to $\bar{x} \in \mathbb{S}^{n-1}$ such that $\lambda(x_k) \leq \lambda(\bar{x}) - \varepsilon$. Then $x_k \lambda(x_k) \subset A$ when k is sufficiently large, see Figure B.2. Thus we have lower semicontinuity of λ :

$$\lambda(\bar{x}) \leq \liminf_{\mathbb{S}^{n-1} \ni x \rightarrow \bar{x}} \lambda(x).$$

On the other hand, if there exists $\varepsilon > 0$ and a sequence $x_k \in \mathbb{S}^{n-1}$ converging to $\bar{x} \in \mathbb{S}^{n-1}$ such that $\lambda(x_k) \geq \lambda(\bar{x}) + \varepsilon$. Then we have that $\bar{x} \lambda(\bar{x}) \in A_k$ for all large k , see Figure B.2. Thus we have established upper semicontinuity of λ

$$\lambda(\bar{x}) \geq \limsup_{\mathbb{S}^{n-1} \ni x \rightarrow \bar{x}} \lambda(x).$$

Therefore, we have proved the continuity of λ . □

Remark B.0.3. *We leave the technical details to the reader, but observe that the lower semicontinuity of λ holds under the assumption that Ω is open and star-shaped. It is the upper semicontinuity of λ that requires the center of Ω containing an open neighborhood of the origin $B(0, a)$ (which in particular is a consequence of convexity and openness).*

Lemma B.0.4. *Let $\Omega \subseteq \mathbb{R}^n$ be an open domain star-shaped with respect to the origin. Fix $p \in [1, \infty)$, let $f \in L^p(\Omega)$ and set for $\rho > 1$*

$$f_\rho := f(\cdot/\rho).$$

Then

$$\|f_\rho - f\|_{L^p(\Omega)} \xrightarrow{\rho \rightarrow 1^+} 0.$$

Proof. Let $\varepsilon > 0$. Since $p \in [1, \infty)$ we have $C^0(\bar{\Omega})$ is dense in $L^p(\Omega)$, and thus there exists $g \in C_c^0(\mathbb{R}^n)$ with

$$\|f - g\|_{L^p(\Omega)} \leq \varepsilon.$$

Set for some $\rho \in (1, 2)$,

$$g_\rho := g(\cdot/\rho).$$

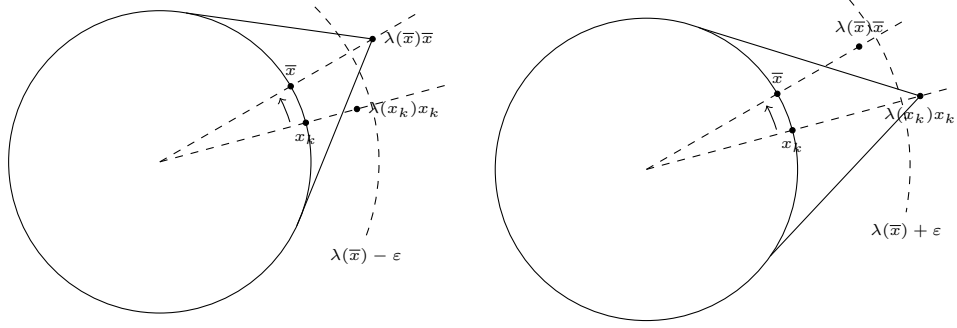


Figure B.2: Assuming that the ball $B(0, a)$ in the proof is actually equal to $B(0, 1)$ (which can always be obtained by scaling) the above figure explains the proof of Lemma B.0.2.

Left: if $\lambda(x_k) < \lambda(\bar{x}) - \varepsilon$ and x_k is sufficiently close to \bar{x} then $\lambda(x_k)x_k$ must belong to the cone A . Right: if $\lambda(x_k) > \lambda(\bar{x}) + \varepsilon$ and x_k is sufficiently close to \bar{x} then \bar{x} must belong to A_k (using that the cone A_k has a minimal aperture that does not change and is determined by $B(0, a)$ as k changes)

Since Ω is star-shaped with respect to the origin,

$$\|f_\rho - g_\rho\|_{L^p(\Omega)} = \rho^{\frac{n}{p}} \|f - g\|_{L^p(\frac{1}{\rho}\Omega)} \stackrel{\rho < 2}{\leq} 2^{\frac{n}{p}} \|f - g\|_{L^p(\Omega)} \leq 2^{\frac{n}{p}} \varepsilon.$$

Then we have for any $\rho > 1$,

$$\|f_\rho - f\|_{L^p(\Omega)} \leq \|f_\rho - g_\rho\|_{L^p(\Omega)} + \|f - g\|_{L^p(\Omega)} + \|g - g_\rho\|_{L^p(\Omega)} \leq 2^{\frac{n}{p}+1} \varepsilon + \|g - g_\rho\|_{L^p(\Omega)}. \quad (\text{B.0.6})$$

Take $R > 0$ such that

$$\text{supp } g \subset B(0, R/4).$$

Since g has compact support we can find such an R . Then g is uniformly continuous on $\overline{B(0, R)}$ and thus there exists some $\rho_0 \in (1, 2)$ such that

$$|g(x) - g(x/\rho)| \leq \frac{\varepsilon}{|B(0, R)|^{\frac{1}{p}}} \quad \forall x \in \overline{B(0, R)}, \forall \rho \in [1, \rho_0].$$

On the other hand if $x \notin B(0, R)$ then

$$g(x) = g(x/\rho) = 0 \quad \forall \rho \in [1, 2].$$

Thus we have

$$\|g - g_\rho\|_{L^\infty(\mathbb{R}^n)} < \frac{\varepsilon}{|B(0, R)|^{\frac{1}{p}}} \quad \forall \rho \in [1, \rho_0]$$

and thus

$$\|g - g_\rho\|_{L^p(\Omega)} = \|g - g_\rho\|_{L^p(B(0,R))} \leq |B(0,R)|^{\frac{1}{p}} \|g - g_\rho\|_{L^\infty(\mathbb{R}^n)} \leq \varepsilon.$$

Combining this with (B.0.6), we have shown

$$\|f_\rho - f\|_{L^p(\Omega)} \leq (2^{\frac{n}{p}+1} + 1)\varepsilon,$$

which holds for any $\rho \in [1, \rho_0)$. We can conclude. □