

Finite Element Exterior Calculus with Applications to the Numerical Solution of the Green–Naghdi Equations

by

Adam Morgan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Applied Mathematics

Waterloo, Ontario, Canada, 2018

© Adam Morgan 2018

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The study of finite element methods for the numerical solution of differential equations is one of the gems of modern mathematics, boasting rigorous analytical foundations as well as unambiguously useful scientific applications. Over the past twenty years, several researchers in scientific computing have realized that concepts from homological algebra and differential topology play a vital role in the theory of finite element methods. Finite element exterior calculus is a theoretical framework created to clarify some of the relationships between finite elements, algebra, geometry, and topology. The goal of this thesis is to provide an introduction to the theory of finite element exterior calculus, and to illustrate some applications of this theory to the design of mixed finite element methods for problems in geophysical fluid dynamics.

The presentation is divided into two parts. Part 1 is intended to serve as a self-contained introduction to finite element exterior calculus, with particular emphasis on its topological aspects. Starting from the basics of calculus on manifolds, I go on to describe Sobolev spaces of differential forms and the general theory of Hilbert complexes. Then, I explain how the notion of cohomology connects Hilbert complexes to topology. From there, I discuss the construction of finite element spaces and the proof that special choices of finite element spaces can be used to ensure that the cohomological properties of a particular problem are preserved during discretization. In Part 2, finite element exterior calculus is applied to derive mixed finite element methods for the Green–Naghdi equations (GN). The GN extend the more well-known shallow water equations to the regime of non-infinitesimal aspect ratio, thus allowing for some non-hydrostatic effects. I prove that, using the mixed formulation of the linearized GN, approximations of balanced flows remain steady. Additionally, one of the finite element methods presented for the fully nonlinear GN provably conserves mass, vorticity, and energy at the semi-discrete level. Several computational test cases are presented to assess the practical performance of the numerical methods, including a collision between solitary waves, the motion of solitary waves over variable bottom topography, and the breakdown of an unstable balanced state.

Acknowledgements

First, I would like to thank my supervisors, Francis Poulin and Benoit Charbonneau. In my time working with them, they have both displayed incredible patience, as evidenced by them actually reading this *kaiju* of a thesis. My project involved dragging both of them outside of the domains of their usual research, but they always showed enthusiasm and support. Francis' advice on coding has been instrumental to the success of my numerical routines, and I cannot adequately express how much I appreciate Benoit's amazing editorial skills. I would also like to thank the other members of my degree committee, Kevin Lamb and Sander Rhebergen, for helpful comments and discussions.

The study of finite element methods has formed a substantial part of my intellectual life for the past two years, and I would like to thank everyone who has helped shape my understanding of this discipline. First, Colin Cotter and Jemma Shipton have been enormously helpful in my quest to learn Firedrake. On the theoretical side of finite elements, I have learned a great deal during discussions with Keegan Kirk, whose devotion to and passion for numerical analysis is very inspiring. Thanks are also due to Sander Rhebergen (again) and Tamas Horvath, who were always happy to help me whenever I barged in on them.

I have had the privilege to work closely with many talented graduate students during my time at Waterloo, including Leon Avery (who shared with me the pains of learning to work with finite element software), Supranta Sarma Boruah (who shared with me the pains of learning canonical perturbation theory), Fabian Germ (who directed me to Cauchy's very interesting life story), Parham Hamidi (with whom I had many fruitful discussions on algebraic geometry), Anthony McCormick (my spirit-brother, the first person I met who cared about infinite-dimensional manifolds as much as I do), Jeff Samuelson ("to thine own self be true"), and Ben Storer (the Python wizard). I also would like to thank all of my friends back in Edmonton for moral support, especially Mitchell Taylor, who was always willing to listen to my complaining.

I owe a great debt to many of the wonderful teachers I have had over the past six years. First, I would like to thank Vincent Bouchard and Vakhtang Putkaradze for cultivating my initial interest in geometry and physics. I am also grateful to Gordon Swaters, from whom I learned to properly think like a theoretical physicist. I owe Manish Patnaik for sharpening my mathematical maturity and writing skills. At Waterloo, I was very fortunate to learn from Spiro Karigiannis, who taught me to love all of the gritty details of differential geometry, and Matt Satriano, who has the honour of being the person who got me to care about ring theory. Lastly, I cannot thank Xinwei Yu enough for baptizing his Math 217/317 class by fire; he is a continuing inspiration who taught me some of the most valuable lessons about mathematics and life.

Finally, I must extend gratitude to my family for their encouragement and patience. Most importantly, I would like to thank my partner Kristen Cote, without whom I would certainly be living in a dumpster

behind a 7–11. Her input on both the written and visual components of this thesis has been immeasurably helpful.

Table of Contents

Author's Declaration	ii
Abstract	iii
Acknowledgements	iv
List of Figures	x
1 Introduction	1
1.1 Prelude: The Shape of Numerical PDE Theory	1
1.2 Goals of this Thesis	2
1.3 Outline of Chapters	3
1.4 Notation	4
I An Introduction to Finite Element Exterior Calculus	6
2 Calculus on Smooth Manifolds: A User's Manual	7
2.1 Manifolds, Vectors, and Tensors	8
2.2 Basics of Differential Forms	16
2.3 Orientability and Hodge Duality	20
2.4 The Exterior Derivative	22
2.5 Integration of Differential Forms	24
3 Hilbert Spaces of Differential Forms	27
3.1 Lipschitz Manifolds	28
3.2 The Spaces $L^2\Lambda^k$	30
3.3 Weak Derivatives and the Spaces $H\Lambda^k$	35
3.4 The Trace Theorem	38

4	Hilbert Complexes	39
4.1	Some Facts from Functional Analysis	40
4.2	Hilbert Complexes: Basic Definitions	44
4.3	Cohomology	47
4.4	Harmonic Forms and the Hodge–Helmholtz Decomposition	48
4.5	The Compactness Property	52
5	Triangulations and Topology	56
5.1	Simplices and Simplicial Complexes	57
5.2	Homology of Simplicial Complexes	62
5.3	de Rham’s Theorem: Bridging Analysis and Topology	70
6	Generalities on Finite Element Methods	73
6.1	Introduction to Galerkin Methods and Finite Element Methods	74
6.2	Finite Elements	77
6.3	Local–to–Global Maps and Finite Element Spaces	81
6.4	Characterization of d–Conforming Finite Element Spaces	84
7	Polynomial Differential Forms on Simplices	87
7.1	Binomial Coefficient Identities	87
7.2	$\mathcal{P}_r\Lambda^k$	89
7.3	The Koszul Operator and $\mathcal{P}_r^-\Lambda^k$	90
8	Construction of Finite Element Spaces	99
8.1	The Lagrange Finite Element Space $\mathcal{P}_r\Lambda^0(\mathcal{T}_h)$	100
8.2	Vanishing Lemmas	105
8.3	The Trimmed Finite Element Space $\mathcal{P}_r^-\Lambda^k(\mathcal{T}_h)$	107
8.4	The Regular Finite Element Space $\mathcal{P}_r\Lambda^k(\mathcal{T}_h)$	115
9	Cohomology of Finite Element Spaces	120
9.1	Approximating Hilbert Complexes	121
9.2	Projections onto Finite Element de Rham Complexes	128
9.3	de Rham Theory à la Whitney	131
9.4	Two Big Theorems	138
9.5	The Periodic Table of the Finite Elements	139

II Applications to the Numerical Solution of the Green–Naghdi Equations	144
10 The Green–Naghdi Equations	145
10.1 Derivation of the Green–Naghdi Equations	146
10.2 Adding the Effects of Rotation	152
10.2.1 Topographic Terms	155
10.2.2 Non–dimensionalization	155
10.3 Linearized RGN	157
10.3.1 Geostrophic Balance	157
10.3.2 Linear Gravity Waves	158
11 Two Finite Element Methods for the (R)GN	161
11.1 Why is FEEC Good for Geophysical Fluid Dynamics?	162
11.2 General Framework for Mixed Formulations of (R)GN	163
11.3 Upwind Formulation of (R)GN	166
11.3.1 Weak Form of Momentum, Pseudovelocity, and Vorticity Equations	167
11.3.2 Introduction to Upwinding	167
11.3.3 Upwind Discretization of the Mass Equation	169
11.4 $H(\text{div})$ –flux Method	170
11.4.1 Weak Form	170
11.4.2 Conservation properties for the $H(\text{div})$ –flux formulation	172
11.5 Time Discretization	175
11.6 Survey of Alternative Numerical Approaches to GN and RGN	176
12 Numerical Results	178
12.1 Geostrophic Balance	179
12.2 Solitary Waves	180
12.3 Collision of Two Solitary Waves	184
12.4 Flow Over Topography	189
12.4.1 Still–Water Solutions	190
12.4.2 Linear Ramps	191
12.4.3 Gaussian Hill	194
12.5 Wave–Vortex Interaction	195
12.6 Breakdown of an Unstable Balanced State	202
12.7 Higher–Order Shape Functions: Preliminary Results	205
13 Conclusions and Future Directions	207
13.1 Summary of Results and Conclusions	207
13.2 Limitations of the Mixed Methods and Suggestions for Further Work	208

References	210
Appendix A Supplementary Firedrake Demos	219

List of Figures

5.11	Visual representation of some low-dimensional simplices. Interior points are coloured orange, and boundary points are coloured black.	59
5.12	A 2-dimensional simplicial complex.	61
5.13	A “tiling” that is not a simplicial complex. Connecting the two white vertices with an edge would turn the figure into a simplicial complex.	61
5.21	$\partial T_{\text{ref},2}$ with the induced right-hand orientation.	64
5.22	Coboundaries of the vertices of $T_{\text{ref},2}$; the captions follow the notation of Example 5.2.17.	70
6.21	DOFs for the Lagrange finite element in Example 6.2.3. Solid dots represent the points p such that eval_p is a DOF.	79
6.22	The Raviart–Thomas finite element $\text{RT}(1)$ in Example 6.2.4. The arrows represent the DOFs (evaluation of normal components at base point of the arrow).	80
6.31	Pictorial representation of global DOFs on the unit interval $[0, 1]$ with respect to the triangulation $[0, \frac{1}{2}] \cup [\frac{1}{2}, 1]$. Solid dots mark the points associated to the global DOFs.	82
6.41	The normal–tangential coordinate system defined with respect to the facet F	86
8.11	DOFs for some $\text{CG}(r)$ finite elements on intervals, corresponding to $\mathcal{P}_r \Lambda^0([0, 1])$	104
8.12	DOFs for some $\text{CG}(r)$ finite elements on triangles, corresponding to $\mathcal{P}_r \Lambda^0(T_{\text{ref},2})$	105
8.31	The degree 2 Raviart–Thomas finite element $\text{RT}(2)$ over the reference triangle, corresponding to $\star \mathcal{P}_2^- \Lambda^1(T_{\text{ref},2})$. Arrows represent edge moments of normal components against degree 1 polynomial functions, and circles represent interior moments against 1-forms with constant coefficients.	111
8.32	The degree 1 Nédélec face element of the first kind on the reference tetrahedron, $\text{N1}^f(1)$, corresponding to $\mathcal{P}_1^- \Lambda^2(T_{\text{ref},3})$	112
8.33	The degree 1 Nédélec edge element of the first kind on the reference triangle, $\text{N1}^e(1)$, corresponding to $\mathcal{P}_1^- \Lambda^1(T_{\text{ref},2})$	113
8.34	The degree 1 Nédélec edge element of the first kind on the reference tetrahedron, $\text{N1}^e(1)$, corresponding to $\mathcal{P}_1^- \Lambda^1(T_{\text{ref},3})$	114
8.35	DOFs for some $\text{DG}(r-1)$ finite elements, corresponding to $\mathcal{P}_r^- \Lambda^2(T_{\text{ref},2})$	115
8.41	The degree 1 Brezzi–Douglas–Marini finite element $\text{BDM}(1)$ over the reference triangle, corresponding to $\star \mathcal{P}_1 \Lambda^1(T_{\text{ref},2})$	118

8.42	The degree 1 Nédélec edge element of the second kind on the reference triangle, $N2^e(1)$, corresponding to $\mathcal{P}_1\Lambda^1(T_{\text{ref},2})$	118
9.11	Figure 2.4 in [5], showing two numerical solutions to (9.1.2) obtained with different FEMs. The solution u on the right is orthogonal to \mathfrak{H}_h^1 , while the solution on the left is of the form $u + q$ for some $q \in \mathfrak{H}^1$ with relatively large norm.	123
9.51	The simplest cohomology-preserving choice of finite element spaces built with the trimmed family, corresponding to the Whitney complex.	140
9.52	The simplest cohomology-preserving choice of finite element spaces built with the regular family.	140
9.53	The first two columns of the Periodic Table of the Finite Elements. Cohomology-preserving families of finite element spaces are arranged horizontally in the first column and diagonally in the second column.	141
9.54	The third and fourth columns of the Periodic Table of the Finite Elements.	142
10.11	Definition of the depth h . The solid curve represents the free surface of the fluid, and the dashed line illustrates the physical meaning of h as the distance between the flat bottom ($z = 0$ here) and the free surface.	147
10.21	Definition of h, H , and η . The green curve represents the seafloor and the blue curve represents the free surface of the fluid.	155
12.11	L^2 error in approximation of the depth for the geostrophic balance test. The $H(\text{div})$ -flux method has been used.	180
12.12	Relative error in energy for the geostrophic balance test. The $H(\text{div})$ -flux method has been used.	181
12.21	Hovmöller plot of the evolution of a solitary wave predicted by the $H(\text{div})$ -flux method, with x in metres and t in seconds. The straight red lines, representing the trajectories of the wave's peak in the xt -plane, indicate that the wave seems to maintain constant speed. . . .	182
12.22	Comparison of the two approximate solutions for depth with the exact solution h_{wave} at the end of the simulation.	183
12.23	Plot of the relative error in energy for both FEMs when 5000 elements are used. Time is measured in seconds. Note how each graph is scaled differently.	183
12.24	Relative error in the energy for solitary wave propagation. To obtain this graph, the $H(\text{div})$ -flux method with 500 elements was used.	185
12.31	Hovmöller plot of the solitary wave collision as simulated by the $H(\text{div})$ -flux method.	187
12.32	Relative error in the energy for the collision problem. The $H(\text{div})$ -flux method has been used.	187
12.33	Overtaking collision of two solitary waves at eight different times, simulated by the $H(\text{div})$ -flux method.	188
12.41	Initial conditions for the linear ramp test with $s = -\frac{4h_0}{5L}$	191
12.42	Hovmöller plot of a solitary wave traveling over a linearly increasing ramp. The $H(\text{div})$ -flux method has been used.	192

12.43	Relative error in energy for a solitary wave traveling over a linearly increasing ramp. The $H(\text{div})$ -flux method has been used.	193
12.44	Initial conditions for the linear ramp test with $s = \frac{4h_0}{5L}$	193
12.45	Hovmöller plot of a solitary wave traveling over a linearly decreasing ramp. The $H(\text{div})$ -flux method has been used.	194
12.46	Relative error in energy for a solitary wave traveling over a linearly decreasing ramp. The $H(\text{div})$ -flux method has been used.	195
12.47	Initial conditions for the Gaussian hill test.	196
12.48	Hovmöller plot of a solitary wave traveling over a Gaussian hill. The $H(\text{div})$ -flux method has been used.	197
12.49	Relative error in energy for a solitary wave traveling over a Gaussian hill. The $H(\text{div})$ -flux method has been used.	197
12.51	Plot of the initial (a) potential vorticity, (b) depth, and (c) y -component of velocity for the wave-vortex interaction test when the parameter values are $A_V = 0.1$, $S_V = 0.2$, $x_V = 50$, and $C = 1$	198
12.52	Hovmöller plot of (a) ζ and (b) the divergence of velocity with $\gamma = 0$ when $A_V = 0.1$ and $C = 1$. The $H(\text{div})$ -flux method has been used.	199
12.53	Hovmöller plot of the evolution of (a) ζ and (b) the divergence of velocity with $\gamma = 1$ when $A_V = 0.1$ and $C = 1$. The $H(\text{div})$ -flux method has been used.	200
12.54	Evolution of the wave-vortex interaction parameter $a(t)$ for (a) $\gamma = 0$ and (b) $\gamma = 1$ when $A_V = 0.1$. In both cases, the solid blue curve displays the evolution when $C = -1$, and the dotted red curve represents $C = 1$. The $H(\text{div})$ -flux method has been used.	200
12.55	Evolution of the wave-vortex interaction parameter $a(t)$ for $A_V = 0.5$ (a), $A_V = 1$ (b), and $A_V = 2$ (c) when $\gamma = 1$. The solid blue curve represents the case $C = -1$, and the dotted red curve represents $C = 1$. The $H(\text{div})$ -flux method has been used.	201
12.56	The same quantities plotted in Figure 12.55, except with $\gamma = 0$	201
12.61	Filled contour plots of the layer depth h for the instability test at eight different times. The upwind method has been used.	204
12.62	Relative error in energy for the instability test. The upwind method has been used.	205
12.71	Changes in $\delta H_{\text{GN}} \doteq \log_{10} \left(\max \frac{H_{\text{GN}}(t) - H_{\text{GN}}(0)}{H_{\text{GN}}(0)} \right)$ with respect to shape function degree r for simulations of solitary wave propagation by the upwind method. A mesh with 510 elements was used, and the Courant number was 0.8.	206

Chapter 1

Introduction

1.1 Prelude: The Shape of Numerical PDE Theory

The development of quality numerical methods for the solution of partial differential equations (PDEs) is one of the most interesting problems at the interface between mathematics and science. Often, we can predict the performance of a numerical method analytically by studying some particular property of the method in the limit of “infinite precision”. When I say that the approximation is approaching the limit of infinite precision, I mean that the parameters of the discretization are approaching whatever limit corresponds to the continuous problem. For example, if we approximate a derivative with respect to x by a difference quotient taken over a small interval Δx , infinite precision corresponds to the limit $\Delta x \rightarrow 0$. The question of what constitutes a “quality” method can vary depending on the particular problem at hand. However, there are three broad characteristics that are considered universally desirable:

- 1) **Stability**: the discrete formulation of the PDE is well-posed, and the approximate solution depends continuously on the parameters of the exact problem (rather than the discretization parameters);
- 2) **Convergence**: the approximate solution tends to the exact solution as we make the approximation infinitely precise;
- 3) **Consistency**: the approximation of the differential operator tends to the exact differential operator as we make the approximation infinitely precise.

There are more precise definitions of the above properties tailored for particular problems, but these precise definitions are always similar in spirit to the above general outlines. Such definitions usually take the form of a particular limit or inequality being satisfied. Properly defining and verifying the above three characteristics is often considered a top priority when numerical analysts begin investigating a new method.

Remark 1.1.1. *Our use of the term “stability” agrees with that most common in numerical PDE theory (see for instance [10, §3.4]).*

Now, as the name “numerical analysis” implies, most theoretical tools designed for checking stability,

convergence, and consistency are based on mathematical analysis. For example, von Neumann’s method for assessing the stability of finite difference methods is based on the theory of Fourier series [47, §16.5], and much of the study of finite element methods for linear elliptic PDEs revolves around the theory of bilinear forms on Hilbert spaces [82]. Of course, since the study of PDEs is one of the jewels of analysis, we can only expect that the study of discretizing PDEs depends critically on analysis as well. However, PDE theory also has intimate connections to algebra and topology. For instance, Hodge theory [77] demonstrates that the number of linearly independent solutions to the vector Laplace equation on a two–dimensional domain Ω is equal to the number of holes in Ω .

Although some of the most important relations between PDEs and topology have been understood for over 80 years, the significance of topological methods in numerical PDE theory has only become widely appreciated since the late 1990s. One of the most important theoretical tools created to harmonize numerical analysis with algebraic and topological techniques in PDE theory is known as **finite element exterior calculus (FEEC)**. FEEC is focused specifically on studying the mathematical structure of finite element numerical methods. Such methods involve looking for an approximate solution to a given PDE in a finite–dimensional **finite element space**, usually consisting of functions that are piecewise smooth with respect to some prescribed triangulation of the spatial domain. Some of the highlights of FEEC include

- 1) unifying many well–known finite element spaces by demonstrating how they arise as special cases of more general constructions based on differential forms;
- 2) clarifying how to choose finite element spaces in order to ensure that the algebraic and topological properties of a particular PDE are preserved at the discrete level;
- 3) using the preservation of algebraic and topological structure to provide proofs of stability and convergence.

In a nutshell, then, FEEC is about crafting good finite element methods by paying attention to the underlying algebraic and topological structures of the PDE we want to solve.

The history of FEEC up to 2010 is nicely summarized in [4, 5], and any attempt I make to present my own version of the subject’s history would just involve regurgitating the discussion in these papers. I mention, however, that the name “finite element exterior calculus” first appeared in 2006 in the work of Arnold, Falk, and Winther [4], though the importance of mimicking algebraic and topological properties of PDEs in finite element discretizations apparently originated in the early 1980s in Kotiuga’s PhD thesis. Additionally, I remark that some of the most impressive results of FEEC depend on re–interpreting certain ideas from mid–20th–century piecewise–linear topology through the lens of finite element analysis: as Christiansen demonstrated in 2005 [18], Whitney’s 1957 book on integration theory [89] contains many of the ingredients for understanding highlight 2 above.

1.2 Goals of this Thesis

Part 1 of this thesis is intended to serve as a rigorous introduction to FEEC, with an emphasis on providing careful proofs of the foundational algebraic and topological results in the theory. I am mainly interested in discussing consistency, rather than convergence and stability, in the present work. This is largely because one must master the algebraic and topological machinery required to understand how FEEC helps improve

consistency before one can understand FEEC-based proofs of stability and convergence. In terms of the above “highlights” of FEEC, then, I focus on points 1 and 2 instead of point 3.

Many the central papers in FEEC [4, 5] assume that the reader a strong background in functional analysis and its manifestations in PDE theory and finite element methods. Additionally, much of the algebraic and topological motivation underlying the developments of these papers may seem foreign to numerical analysts, many of whom have never heard of foundational topological ideas like cohomology before. In order to make FEEC more inviting and transparent, I have tried to make Part 1 as self-contained and pedagogical as possible. I assume that the reader knows only vector calculus, general topology, some functional analysis, and a bit of numerical analysis (a basic acquaintance with finite difference methods more than suffices). In place of describing all of the major successes of FEEC, I am more concerned with providing the reader with sufficient background in algebra (in particular, the theory of chain and cochain complexes) and topology in order for them to appreciate the motivation for introducing the ideas of FEEC. I have also tried to include helpful recommendations for further reading throughout, as I feel that many of the references on geometry, algebra, and topology given in some of the primary sources on FEEC may not be particularly illuminating for the newcomer. Hopefully, the reader will, after working through Part 1, be able to at least understand the jargon of contemporary research papers in this thriving discipline.

Remark 1.2.1. *An alternative, and very accessible, introductory resource on FEEC is Tiee’s splendid PhD thesis [88]. In some ways, Part 1 is complementary to Tiee’s thesis: he concentrates more on how FEEC provides information on convergence and stability estimates (highlight 3 above), while I emphasize the topological flavour of FEEC more thoroughly. If the reader is setting out to properly master the foundations of FEEC, I highly recommend they read Part 1 alongside Tiee’s thesis.*

Once I have finished taking the reader through the theory of FEEC, I use Part 2 to describe one of its concrete application in geophysical fluid dynamics. Specifically, Part 2 is devoted to developing FEEC-based discretizations and numerical simulations of the **Green–Naghdi equations (GN)** describing the vertically averaged motion of an ideal fluid. The GN are capable of describing small-scale phenomena such as solitary wave propagation and, when the effects of planetary rotation are added, they can also describe features of large-scale oceanic flows such as quasi-static vortices and jets. After introducing the GN and some of their most important properties, I present two FEEC-inspired finite element methods for their numerical solution. I exhibit numerical test cases for both of these methods to assess their performance, paying particular attention to whether or not theoretical predictions about the methods are correct. By the end of Part 2, the reader should understand why FEEC is particularly well-suited for describing large-scale oceanographic phenomena and should be able to explain the positive and negative aspects of the two finite element methods presented.

1.3 Outline of Chapters

In Chapter 2, I provide a brief guide to the calculus of differential forms. In Chapter 3 I develop a notion of L^2 and Sobolev spaces of differential forms for use throughout Part 1. In Chapter 4, I define and analyze **Hilbert complexes**, objects that provide a very useful setting for describing some algebraic aspects of PDE theory (and FEEC in particular). One of the major focuses of this chapter is understanding the **cohomology** of a Hilbert complex. In Chapter 5, I provide an overview of some of the most important ideas from algebraic topology that are relevant to FEEC. This chapter culminates in the presentation of de Rham’s Theorem on the duality between calculus on manifolds and algebraic topology. In particular,

we see that de Rham’s Theorem connects Hilbert complex cohomology to important topological data. In Chapter 6, I introduce finite element methods rigorously. In Chapter 7, I describe some simple algebraic and combinatorial properties of two families of differential forms with polynomial coefficients, in order to facilitate the construction of finite element spaces of differential forms later on. In Chapter 8, I take the reader through the FEEC construction of many well–studied finite element spaces, based on the theoretical groundwork laid in the previous chapter. In Chapter 9, I prove that certain choices of finite element spaces “fit together” to give rise to numerical approximations preserving the algebraic and topological properties of the exact problem.

In Chapter 10, I derive the GN (and their rotating analogue) using tools from Hamiltonian fluid dynamics. I also describe some physical properties of the GN that are particularly important for applications to large–scale oceanographic problems. In Chapter 11, I motivate the use of FEEC in geophysical fluid dynamics and present two finite element methods for GN built using the ideas of FEEC. I then prove that one of these numerical methods maintains some of the conservation laws associated to the GN (specifically, conservation of mass, vorticity, and energy), at least when only the spatial derivatives are discretized and the temporal derivatives remain exact. In Chapter 12, I present several numerical test cases for both methods described in Chapter 11, paying special attention to conservation of energy at the fully discrete level. In Chapter 13, I conclude by summarizing the major findings of the numerical tests, outlining the shortcomings of the two finite element methods, and suggesting specific avenues for future research.

1.4 Notation

Before beginning, I introduce some notation used throughout the thesis.

- Given a set X , its power set is denoted by 2^X and its cardinality is denoted by $|X|$.
- Given a subset A of a topological space X , its closure in X is denoted by \overline{A} and its interior is denoted by A^0 .
- $\mathbb{R}_{\geq 0}$ denotes the set of all real numbers greater than or equal to zero.
- If V is a real vector space (possibly with infinite dimension), V^* denotes its continuous dual space.
- If V and W are real vector spaces, I denote their direct sum by $V \oplus W$ (for a reminder on the definition of a direct sum, see [51, pp. 638–640]).
- Let V be a vector space, with $\{V_\alpha\}_{\alpha \in I}$ a family of subspaces of V indexed by some finite set I . The vector space $\sum_{\alpha \in I} V_\alpha$ is defined to be the span of $\cup_{\alpha \in I} V_\alpha$. I remark that $\sum_{\alpha \in I} V_\alpha = \bigoplus_{\alpha \in I} V_\alpha$ if and only if the pairwise intersections of the V_α ’s are all equal to 0.
- Let $F: U \rightarrow V$ be a smooth map between open subsets of Euclidean spaces. I denote the Jacobian of F by DF .
- δ_{ij} always denotes the Kronecker delta.
- I use $\det[a_{ij}]$ to denote the determinant of the matrix whose (i, j) –entry is equal to a_{ij} .
- The standard basis vectors for \mathbb{R}^3 in the x , y , and z directions are denoted by $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$, respectively.

- If $\mathbf{u}_1 \in \mathbb{R}^3$ and $\mathbf{u}_2 \in \mathbb{R}^2$, I use the notation “ $\mathbf{u}^1 \times \mathbf{u}^2$ ” as shorthand for “ $\mathbf{u}^1 \times (\mathbf{u}^2 + 0\hat{\mathbf{z}})$ ”. A similar convention will also be in place when $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^2$. That is, I always implicitly use placeholders of 0 in the z -component when computing any cross products involving 2-vectors.
- Let f be a smooth scalar function on \mathbb{R}^2 . Then, I define $\nabla^\perp f \doteq -\frac{\partial f}{\partial y}\hat{\mathbf{x}} + \frac{\partial f}{\partial x}\hat{\mathbf{y}} = \hat{\mathbf{z}} \times \nabla f$.

Part I

An Introduction to Finite Element Exterior Calculus

Chapter 2

Calculus on Smooth Manifolds: A User's Manual

As one might expect, understanding finite element exterior calculus depends largely on having at least a working understanding of exterior calculus. To keep Part 1 as self-contained as possible, in this chapter I provide a rapid-fire overview of what exterior calculus is and how to actually perform concrete computations with it. Our main goal is to provide a sketch of how exterior calculus summarizes and generalizes vector calculus in \mathbb{R}^n by making use of **differential forms** (roughly, generalized vector fields) on **smooth manifolds** (generalized versions of smooth curves and surfaces). By no means is this chapter intended to be a substitute for the reader working through a differential geometry book themselves: as the title suggests, my focus is on providing just enough knowledge of how exterior calculus works in order to understand its relevance to numerical analysis. Anyone looking to learn the material properly should consult the textbooks referenced throughout this chapter.

In Section 2.1, I introduce smooth manifolds before defining **vector fields** and **tensor fields** on manifolds. In Section 2.2, I define **differential forms** and discuss some basic algebraic operations we can do with them. In Section 2.3, I introduce an intrinsic notion of **orientability** for smooth manifolds, generalizing the concept of outward unit normals encountered in vector calculus. I then state that orientability may be used to define an operator called the **Hodge star**, allowing us to switch between different types of differential forms. In Section 2.4, I describe how to differentiate differential forms in a way that neatly summarizes and extends the main operations of vector calculus (specifically, the operators ∇ , $\nabla \times$, and $\nabla \cdot$). In Section 2.5, I outline how to integrate differential forms and state the general Stokes' Theorem, another result generalizing familiar objects from vector calculus to the setting of smooth manifolds.

There are many superb textbooks on differential geometry and exterior calculus, and before beginning I would like to note some of my personal favourites. Owing to the blitzkrieg presentation style of this chapter, I often defer the reader to these books so that they may obtain a deeper understanding of the material (sometimes, I also defer to them in order for the reader to obtain a shallow understanding). Flanders' little book [35] is a solid exterior calculus reference for the applied mathematician. Do Carmo's book on differential forms [28] also provides several instructive exercises and examples. I consider both John Lee's encyclopedic book on smooth manifolds [51] and Nakahara's book on topological and geometric methods in mathematical physics [62] to be references of paramount importance, though the sheer size of

both these volumes can make them seem a bit intimidating.

2.1 Manifolds, Vectors, and Tensors

In this section, I define smooth manifolds and describe how calculus can be “moved” from \mathbb{R}^n to a given smooth manifold M by discussing what it means for a scalar, vector, or tensor on M to be smooth. To do this, of course, we must also learn how to interpret vectors and tensors on manifolds. Most of the discussion in this section is a tightly-compressed version of several chapters of Lee’s comprehensive textbook [51].

In earlier times, manifolds were defined to be subsets of \mathbb{R}^n that could be described locally as sets of points satisfying some constraint equations and inequalities expressed in terms of differentiable functions (see for example Poincaré’s definition [69, §1]). From the viewpoint of modern mathematics, however, a more intrinsic definition based on topology is preferred (if the reader is unfamiliar with general topology, I recommend they skim Appendix A of [51] for a primer on the basic definitions). For the purposes of this exposition, the reader is encouraged to keep in mind the intuitive picture of what a manifold “ought to be” and not worry too much about the picky details.

Definition 2.1.1. *Let M be a Hausdorff topological space whose topology has a countable basis. Fix some natural number n and let I denote some indexing set.*

1) A **smooth atlas** on M is a family $\mathcal{A} = \{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$ of open sets $U_\alpha \subseteq M$ and homeomorphisms $\varphi_\alpha: U_\alpha \rightarrow V_\alpha$, with $V_\alpha \subseteq \mathbb{R}^n$ an open set, satisfying the following:

- for all $p \in M$, there exists $\alpha \in I$ such that $p \in U_\alpha$ (that is, the open sets U_α cover M);
- for any $\alpha, \beta \in I$, if $U_\alpha \cap U_\beta$ is nonempty, then the **transition map**

$$\varphi_\beta \circ \varphi_\alpha^{-1}: \varphi_\alpha(U_\alpha \cap U_\beta) \rightarrow \varphi_\beta(U_\alpha \cap U_\beta)$$

is a diffeomorphism.

The pairs $(U_\alpha, \varphi_\alpha)$ are called **(smooth) charts**. We sometimes call the open sets U_α **chart domains** or **coordinate patches** on M .

2) A **smooth structure** on M is a smooth atlas \mathcal{A} that is not strictly contained inside any other smooth atlas.

3) Let \mathcal{A} be a smooth structure on M . The pair (M, \mathcal{A}) is called a **smooth manifold**, and the number n is called its **dimension**.

In practice, I almost always omit reference to the specified smooth structure when discussing a smooth manifold. That is, when I say “ M is a smooth manifold”, the reader should understand that I have tacitly put a particular smooth structure on the space M . By applying the Rank–Nullity Theorem to the Jacobian matrix of any transition map, it is easy to show that the dimension of a smooth manifold is well-defined on each connected component of M . Finally, I remark that Hausdorffness and the countable basis condition can often be taken for granted in applications, so I do not focus much on these aspects of manifolds here.

We now present some basic examples of smooth manifolds. Owing to the complicated nature of the manifold definition, providing every detail is too time-consuming. For more interesting examples and more explicit details, see [51, pp. 17–24].

Example 2.1.2. Clearly, \mathbb{R}^n is a smooth n -dimensional manifold: the only chart we need is $(\mathbb{R}^n, x \mapsto x)$.

Example 2.1.3. Any open subset U of a smooth manifold M can be given the structure of a smooth manifold simply by taking the intersection of U with all of the chart domains. The dimension of U is the same as the dimension of M .

Example 2.1.4. The n -sphere S^n , defined by

$$S^n \doteq \{x \in \mathbb{R}^{n+1} \mid \|x\| = 1\},$$

is a smooth manifold of dimension n .

Example 2.1.5. The solid sphere

$$D^{n+1} \doteq \{x \in \mathbb{R}^{n+1} \mid \|x\| \leq 1\},$$

is *not* a smooth $(n + 1)$ -dimensional manifold. We may see this intuitively: inside a little neighbourhood on the interior of the solid sphere, we have $n + 1$ possible directions we can move while staying in the neighbourhood, while if our little neighbourhood is on the surface of the sphere, we can only move in n possible directions. So, there is a “jump” in dimensionality that prevents D^{n+1} from being a smooth manifold by Definition 2.1.1.

The last example may have upset the reader. Since we want a manifold to be a generalization of the vanishing set of some smooth constraints, we ought to have some way of talking about D^{n+1} as a manifold-like object. Such thinking motivates the definition of a **manifold with boundary**. A manifold with boundary is a topological space satisfying all of the conditions outlined in Definition 2.1.1, except that we allow some points on M to have open neighbourhoods that are homeomorphic to an open subset of a **half-space**

$$\mathbb{H}^n \doteq \{(x^1, \dots, x^n) \in \mathbb{R}^n \mid x^n \geq 0\}.$$

To understand how we can put smooth structures on manifolds with boundary, recall the following definition from multivariable calculus:

Definition 2.1.6. Let $S \subseteq \mathbb{R}^n$ and suppose that $F: S \rightarrow \mathbb{R}^m$. We say that F is **smooth** if, for all $p \in S$, there exists an open subset U of \mathbb{R}^n containing p and a smooth function $\tilde{F}: U \rightarrow \mathbb{R}^m$ such that

$$\tilde{F}|_{S \cap U} = F|_{S \cap U}.$$

So, saying that the transition functions on a manifold with boundary are smooth is perfectly fine, even though such transition functions may be defined on non-open domains.

Let M be an n -dimensional manifold with boundary. If a point $p \in M$ has a neighbourhood homeomorphic to an open subset of \mathbb{R}^n , we say that p is an **interior point**, else p is called a **boundary point**. Borrowing notation from multivariable calculus, we denote the set of all boundary points by ∂M . One can show that ∂M has the structure of an $(\dim M - 1)$ -dimensional smooth manifold [51, Prop. 1.38]. With this new jargon in mind, we clearly see that the solid sphere is an $(n + 1)$ -dimensional manifold with boundary. Further, $\partial D^{n+1} = S^n$.

We can also go further and define **manifolds with k -corners**, spaces that are locally homeomorphic to open subsets of

$$\{(x^1, \dots, x^n) \in \mathbb{R}^n \mid x^{n-k}, \dots, x^n \geq 0\}.$$

So, a manifold with 0-corners is just a manifold with boundary, and a manifold with 1-corners has coordinate patches that look like the corner of a solid triangle. Naturally, triangles themselves are canonical examples of manifolds with 1-corners. We usually just say “manifold with corners” when the number k is not important. In the sequel, we often have to deal with calculus problems on triangles, tetrahedra, and their higher-dimensional analogues, so our introduction of manifolds with corners is not just an exercise in stretching out abstractions. For more on manifolds with corners, see [51, pp. 415–417].

Definition 2.1.7. *Let M and N be smooth manifolds. A map $F: M \rightarrow N$ is said to be **smooth** if, for all charts (U, φ) on M and (V, ψ) on N such that $F(U) \subseteq V$, the map*

$$\hat{F} \doteq \psi \circ F \circ \varphi^{-1}: \varphi(U) \rightarrow \psi(F(U))$$

*is smooth. The map \hat{F} is called the **coordinate representation** of F . If $F: M \rightarrow N$ is a smooth bijection with smooth inverse, we say that F is a **diffeomorphism**.*

The set of all smooth, real-valued functions on M is denoted by $C^\infty \Lambda^0(M)$. This collection is also referred to as the set of **scalar fields** on M . Combining the definition of smooth maps above with Definition 2.1.6, we know what it means for a map defined on a general $S \subseteq M$ to be smooth.

Definition 2.1.8. *A function $f: M \rightarrow \mathbb{R}$ is said to be **compactly supported** if*

$$\text{supp } f \doteq \overline{\{p \in M \mid f(p) \neq 0\}}$$

is a compact subset of M .

We use the symbol $C_c^\infty \Lambda^0(M)$ to denote the set of all scalar fields on M that are compactly supported in the interior of M .

Now that we have defined smooth manifolds and smooth maps, we can begin our goal of building a generalized vector calculus in earnest. Of course, our first thought is immediately to find a suitable generalization of a vector field on \mathbb{R}^n to a smooth manifold. We usually picture vector fields in \mathbb{R}^n as smooth assignments of points $p \in \mathbb{R}^n$ to little arrows protruding from p . More formally, a smooth vector field is a smooth map $\mathbb{R}^n \rightarrow \mathbb{R}^n$, where we view the \mathbb{R}^n in the target as being “translated” so that the origin coincides with p . When working on an arbitrary smooth manifold M that is not contained in some ambient Euclidean space, we can’t directly extend such a notion of vector field to M ; Definition 2.1.1 doesn’t explicitly tell us that M comes with a free copy of \mathbb{R}^n attached to every point $p \in M$. So, we have to try to figure out a canonical way to attach a vector space to every point of M .

Definition 2.1.9. *Suppose that M is a smooth manifold of dimension n .*

*1) Let $p \in M$. The **tangent space** to M at p is the n -dimensional real vector space of all linear maps $\mathbf{u}_p: C^\infty \Lambda^0(M) \rightarrow \mathbb{R}$ satisfying*

$$\mathbf{u}_p(fg) = \mathbf{u}_p(f) g(p) + f(p) \mathbf{u}_p(g). \tag{2.1.1}$$

*We denote the tangent space at p by $T_p M$. Elements of the tangent space are called **tangent vectors** to M at p .*

2) The **tangent bundle** of M is the $2n$ -dimensional smooth manifold TM given by

$$TM \doteq \{(p, \mathbf{u}_p) \mid p \in M, \mathbf{u}_p \in T_p M\}. \quad (2.1.2)$$

Now, the reader should have noticed that I was a bit sloppy in the above definition, in that I called TM a manifold without specifying a smooth structure. One can always build a smooth structure on TM using M 's smooth structure by “linearizing” the chart maps on M in an appropriate sense. The details of the argument are simple [51, Prop. 3.18], but for the sake of time I omit them.

Although we describe the elements of TM with ordered pairs, I stress that the tangent bundle is not the Cartesian product of M with a real vector space. In particular, if $q \neq p$, the pair (q, \mathbf{u}_p) does not a priori make sense, hence TM is not guaranteed to be a Cartesian product. However, TM is *locally* a Cartesian product, in the sense that every $p \in M$ is contained in a chart domain U diffeomorphic to an open set $V \subseteq \mathbb{R}^n$ such that $U \times \pi^{-1}(p)$ is diffeomorphic to $V \times \mathbb{R}^n$.

At this point, we have succeeded in attaching a copy of \mathbb{R}^n to every point in M using the tangent space, but the reader may be wondering where the motivation for this notion of tangency comes from. Essentially, the definition of the tangent space is inspired by the one-to-one correspondence between vectors $\mathbf{u} \in \mathbb{R}^n$ and directional derivatives $\mathbf{u} \cdot \nabla$ (this follows from Taylor’s Theorem [51, pp. 53–54]). In fact, we pretty much copy over the notation for directional derivatives when working on manifolds. Studying tangency using derivatives allows us to view $T_p M$ as a linear approximation of M near p , in analogy to how the derivative of a map is its best approximation by a linear transformation. Very loosely, then, we can think of tangent vectors as “infinitesimal displacements” on M .

Suppose that (U, φ) is a coordinate chart around $p \in M$, and denote the local coordinates associated to this chart by (x^1, \dots, x^n) . Define a set of tangent vectors at p by

$$\left. \frac{\partial}{\partial x^i} \right|_p : f \mapsto \frac{\partial f}{\partial x^i}(p), \quad i = 1, \dots, n.$$

Using the correspondence $\mathbf{u} \mapsto \mathbf{u} \cdot \nabla$ in \mathbb{R}^n , we can show that

$$\left\{ \left. \frac{\partial}{\partial x^i} \right|_p \right\}_{i=1}^n$$

is a basis for $T_p M$, called the **coordinate basis**. That is, for every tangent vector $\mathbf{u}_p \in T_p M$, there exist n real numbers u^i such that

$$\mathbf{u}_p = u^i \left. \frac{\partial}{\partial x^i} \right|_p.$$

We sometimes require a way of associating tangent vectors on one manifold to tangent vectors on another manifold.

Definition 2.1.10. Let M and N be smooth manifolds and let $F: M \rightarrow N$ be a smooth map. The **pushforward** of F at $p \in M$ is the linear map $F_{*,p}: T_p M \rightarrow T_{F(p)} M$ given by

$$(F_{*,p} \mathbf{u}_p) f \doteq \mathbf{u}_p (F \circ f) \quad \forall f \in C^\infty \Lambda^0(M), \quad \mathbf{u}_p \in T_p M. \quad (2.1.3)$$

Finally, we come to defining vector fields. Since we now have a notion of “little arrows” lying tangent to M (precisely, elements of T_pM), figuring out what the generalized version of a vector field should look like is obvious.

Definition 2.1.11. Let M be a smooth manifold and let $\pi: TM \rightarrow M$ be defined by $\pi(p, \mathbf{u}_p) = p$. Let $\mathbf{u}: M \rightarrow TM$.

1) \mathbf{u} is called a **vector field** if $\pi \circ \mathbf{u}$ is the identity map on M .

2) Let \mathbf{u} be a vector field. For any chart domain U with associated local coordinates (x^1, \dots, x^n) , we define n functions $u^i: U \rightarrow \mathbb{R}$, called the **components of \mathbf{u}** , by

$$\mathbf{u}(p) = u^i(p) \frac{\partial}{\partial x^i} \Big|_p.$$

We say that \mathbf{u} is **smooth** if, in each chart domain, every component of \mathbf{u} is smooth.

3) The real vector space of all smooth vector fields on M is denoted by $\mathfrak{X}(M)$.

Let U be any chart domain on M . The coordinate basis clearly defines a family of smooth vector fields on U by way of

$$\frac{\partial}{\partial x^i} : p \mapsto \frac{\partial}{\partial x^i} \Big|_p.$$

Suppose that \mathbf{u} is a smooth vector field. When we change local coordinates $(x^1, \dots, x^n) \rightarrow (y^1, \dots, y^n)$, the Chain Rule tells us that the expression of $\mathbf{u}|_U$ in the coordinate basis changes according to

$$u^i \frac{\partial}{\partial x^i} = u^i \frac{\partial y^j}{\partial x^i} \frac{\partial}{\partial y^j}. \quad (2.1.4)$$

We now turn to constructing objects on M that are “dual” to smooth vector fields.

Definition 2.1.12. The **cotangent bundle** of M is the smooth manifold T^*M defined by

$$T^*M \doteq \{(p, \alpha_p) \mid \alpha_p \in T_p^*M\}.$$

I do not explicitly describe the smooth structure on T^*M for the sake of time, but the construction of this smooth structure is routine (following along the same lines as the construction of the tangent bundle).

Let $\{dx^i|_p\}_{i=1}^n$ denote the basis of T_p^*M dual to the coordinate basis of T_pM . That is, each dx^i is uniquely determined by the conditions

$$dx^i|_p \left(\frac{\partial}{\partial x^j} \Big|_p \right) = \delta_{ij}.$$

Using this easy basis for T_p^*M , we define one of the main building blocks of exterior calculus.

Definition 2.1.13. Let M be a smooth manifold, let $\pi: T^*M \rightarrow M$ be defined by $\pi(p, \alpha_p) = p$, and let $\alpha: M \rightarrow T^*M$.

1) α is called a **1-form** if $\pi \circ \alpha$ is the identity map on M .

2) Suppose that α is a 1-form. For any chart domain U with associated local coordinates (x^1, \dots, x^n) , we define n functions $\alpha_i: U \rightarrow \mathbb{R}$, called the **components of α** , by

$$\alpha(p) = \alpha_i(p) dx^i|_p.$$

We say that α is a **smooth differential 1-form** if, in every chart domain on M , all components of α are smooth.

3) The real vector space of all smooth differential 1-forms on M is denoted by $C^\infty \Lambda^1(M)$.

We often do not even bother saying “differential” when referring to differential 1-forms. Naturally, in each coordinate patch the maps $dx^i: p \mapsto dx^i|_p$ define a family of smooth 1-forms on U .

We may ask how the components of a 1-form are modified when we change local coordinates on U from (x^1, \dots, x^n) to (y^1, \dots, y^n) . Consider the notational setup

$$\alpha|_U = \alpha_i dx^i = \tilde{\alpha}_j dy^j.$$

To solve for the $\tilde{\alpha}_j$, we use the dual basis condition and (2.1.4) to see that

$$\begin{aligned} \tilde{\alpha}_k &= \tilde{\alpha}_j dy^j \left(\frac{\partial}{\partial y^k} \right) \\ &= \alpha_i dx^i \left(\frac{\partial}{\partial y^k} \right) \\ &= \alpha_i dx^i \left(\frac{\partial x^j}{\partial y^k} \frac{\partial}{\partial x^j} \right) \\ &= \alpha_i \frac{\partial x^j}{\partial y^k} dx^i \left(\frac{\partial}{\partial x^j} \right) \\ &= \alpha_i \frac{\partial x^i}{\partial y^k}. \end{aligned}$$

We conclude that

$$\tilde{\alpha}_k = \alpha_i \frac{\partial x^i}{\partial y^k}.$$

Notice how the change of variables formula for 1-forms resembles the change of variables rule for integrals,

$$\int_{x(a)}^{x(b)} f(x) dx = \int_a^b f(x(y)) \frac{dx}{dy} dy.$$

This similarity provides some belated motivation for the notation dx^i .

Since a 1-form defines a linear functional on the tangent space at each $p \in M$, we can “feed” a 1-form α a vector field \mathbf{u} to produce a smooth function on M :

$$\alpha(\mathbf{u}): p \mapsto \alpha_p(\mathbf{u}_p).$$

Using duality, we can equivalently view $\alpha(\mathbf{u})$ as “feeding” α to \mathbf{u} . With this picture in mind, we have that vector fields eat 1-forms to produce smooth scalars, and 1-forms eat vector fields to produce smooth scalars (note that $\alpha(\mathbf{u}) = \mathbf{u}(\alpha)$ for all vector fields \mathbf{u} and 1-forms α). This way of thinking about vectors and 1-forms helps motivate the more general concept of a **tensor field**. Before introducing tensors formally, however, we review an important definition from linear algebra.

Definition 2.1.14. *Let E and F be finite-dimensional real vector spaces, with dual spaces denoted by E^* and F^* respectively. The **tensor product** of E and F , denoted $E \otimes F$, is the real vector space of all bilinear maps $E^* \times F^* \rightarrow \mathbb{R}$.*

Since all finite-dimensional real vector spaces E satisfy $E^{**} = E$, it is not very difficult to show that, given a basis $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ of E and a basis $\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$ of F , the set of all bilinear maps on $E^* \times F^*$ of the form

$$\mathbf{e}_i \otimes \mathbf{f}_j: (\alpha, \beta) \mapsto \mathbf{e}_i(\alpha)\mathbf{f}_j(\beta)$$

constitutes a basis of $E \otimes F$.

Definition 2.1.15. *Let $p \in M$.*

1) $\mathfrak{T}_{\ell,p}^k \doteq \underbrace{(\mathbb{T}_p^*M \otimes \dots \otimes \mathbb{T}_p^*M)}_{k \text{ times}} \otimes \underbrace{(\mathbb{T}_pM \otimes \dots \otimes \mathbb{T}_pM)}_{\ell \text{ times}}$ is called the space of (k, ℓ) -**tensors** at p .

2) The **bundle of (k, ℓ) -tensors** on M is the smooth manifold \mathfrak{T}_ℓ^k defined by

$$\mathfrak{T}_\ell^k \doteq \{(p, \tau_p) \mid \tau_p \in \mathfrak{T}_{\ell,p}^k\}.$$

I omit the precise definition of the smooth structure on \mathfrak{T}_ℓ^k , as I likewise ignored the construction of the smooth structures on the tangent and cotangent bundles.

Given a local coordinate system (x^1, \dots, x^n) for a chart domain U around $p \in M$, we can represent $\tau_p \in \mathfrak{T}_{\ell,p}^k$ in terms of the real numbers

$$(\tau_p)^{j_1 \dots j_\ell}_{i_1 \dots i_k} = \tau_p \left(\frac{\partial}{\partial x^{i_1}}, \dots, \frac{\partial}{\partial x^{i_k}}, dx^{j_1}, \dots, dx^{j_\ell} \right).$$

The numbers $(\tau_p)^{j_1 \dots j_\ell}_{i_1 \dots i_k}$ are called the **components** of τ_p in the coordinate basis. Once we know all of the components of τ_p , we may write

$$\tau_p = (\tau_p)^{j_1 \dots j_\ell}_{i_1 \dots i_k} dx^{i_1} \otimes \dots \otimes dx^{i_k} \otimes \frac{\partial}{\partial x^{j_1}} \otimes \dots \otimes \frac{\partial}{\partial x^{j_\ell}}.$$

At this point, the definition of a **smooth (k, ℓ) -tensor field** is obvious:

Definition 2.1.16. *Let $\pi: \mathfrak{T}_\ell^k \rightarrow M$ be defined by $\pi(p, \tau_p) = p$. Any map $\tau: M \rightarrow \mathfrak{T}_\ell^k$ such that $\pi \circ \tau$ is the identity on M is called a (k, ℓ) -**tensor field**. If, for every chart domain U , we also have that each real-valued component function*

$$\tau_{i_1 \dots i_k}^{j_1 \dots j_\ell}: U \rightarrow \mathfrak{T}_\ell^k$$

defined by

$$\tau_{i_1 \dots i_k}^{j_1 \dots j_\ell} \cdot p \mapsto (\tau_p)_{i_1 \dots i_k}^{j_1 \dots j_\ell}$$

is smooth, then we say that τ is a **smooth** (k, ℓ) -**tensor field**.

Unless otherwise stated, all tensors in this chapter are assumed to be smooth. By convention, a $(0, 0)$ -tensor is a smooth scalar field. A $(1, 0)$ -tensor is a smooth vector field and a $(0, 1)$ -tensor is a smooth 1-form. In general, a (k, ℓ) -tensor field on M can be “fed” k vector fields and ℓ 1-forms to produce a smooth scalar field. Additionally, by calculations similar to those seen earlier in this section, we can easily write down a change of variables formula for the components of a general tensor.

As we defined compactly supported scalar fields, so can we define compactly supported tensors. Namely, a tensor field is compactly supported if and only if the closure of the complement of its vanishing set is compact. We also use the subscript c to indicate compact support in the tensorial case. For example, $\mathfrak{X}_c(M)$ denotes the set of compactly supported smooth vector fields, and $\mathcal{C}_c^\infty \Lambda^1(M)$ denotes the set of compactly supported smooth 1-forms.

Before concluding this section, we introduce some more jargon useful for handling tensors and describe a useful breed of tensors encountered frequently in geometry and its applications. If the (k, ℓ) -tensor τ satisfies

$$\tau_{i_1 \dots i_a i_{a'} \dots i_k}^{j_1 \dots j_\ell} = \tau_{i_1 \dots i_{a'} i_a \dots i_k}^{j_1 \dots j_\ell}$$

then we say that τ is **symmetric** in the indices i_a and $i_{a'}$. Conversely, if

$$\tau_{i_1 \dots i_a i_{a'} \dots i_k}^{j_1 \dots j_\ell} = -\tau_{i_1 \dots i_{a'} i_a \dots i_k}^{j_1 \dots j_\ell}$$

then we say that τ is **antisymmetric** in the indices i_a and $i_{a'}$. The same nomenclature applies to indices upstairs and non-adjacent indices as well.

A **Riemannian metric** is a symmetric $(0, 2)$ -tensor g such that, for all $p \in M$, the matrix of g_p is positive-definite. The easiest example of a Riemannian metric is the **Euclidean metric** on \mathbb{R}^n , defined by

$$g \doteq \sum_{i=1}^n dx^i \otimes dx^i.$$

We see that, for all $p \in \mathbb{R}^n$, $g_p(\mathbf{u}_p, \mathbf{u}_p) = \|\mathbf{u}_p\|$. More generally,

$$g_p(\mathbf{u}_p, \mathbf{v}_p) = \mathbf{u}_p \cdot \mathbf{v}_p.$$

So, the Euclidean metric is really just the usual dot product in disguise. With the example of the Euclidean metric in mind, we see that Riemannian metrics provide us with a way of generalizing the notions of length and angle to arbitrary smooth manifolds. A manifold M equipped with a Riemannian metric g is called a **Riemannian manifold**. We denote a Riemannian manifold using the pair (M, g) . On any (M, g) we have an alternative notion of duality between vector fields and 1-forms as follows: given $\mathbf{u} \in \mathfrak{X}(M)$ we define its **flat**, denoted $\mathbf{u}^\flat \in \mathcal{C}^\infty \Lambda^1(M)$, by

$$\mathbf{u}^\flat(\mathbf{v}) = g(\mathbf{u}, \mathbf{v}).$$

Since g_p is positive-definite for all p , the map $T_p M \rightarrow T_p^* M$ defined by

$$\mathbf{u}_p \mapsto \mathbf{u}_p^\flat$$

is an isomorphism. Therefore, the flat map is invertible. We call the inverse of the flat map the **sharp map** and denotes its action on $\alpha \in \mathcal{C}^\infty \Lambda^1(M)$ by α^\sharp . Collectively, the flat map and sharp map make up the **musical isomorphism**. In \mathbb{R}^n with the Euclidean metric, the musical isomorphism on each tangent space is effectively just the identity map: if $\mathbf{u}_p = u^i \frac{\partial}{\partial x^i} \Big|_p$ then $\mathbf{u}_p^\flat = u^i dx^i \Big|_p$.

2.2 Basics of Differential Forms

Now, we come to defining general differential forms and describing some elementary operations on them. My approach often borrows from do Carmo's brilliant textbook on differential forms [28], to which I defer for more detailed proofs.

Definition 2.2.1.

1) $\text{Alt}^k(\mathbb{T}_p M) \doteq \{\omega_p \in \mathfrak{T}_0^k \mid \omega_p \text{ is antisymmetric in every pair of indices}\}.$

2) $\text{Alt}^k(M) \doteq \{(p, \omega_p) \mid \omega_p \in \text{Alt}^k(\mathbb{T}_p M)\}.$

3) A smooth tensor field $\omega : M \rightarrow \text{Alt}^k(M)$ is called a **(smooth) differential k -form**. The number k is called the **degree** of the form. The real vector space of all differential k -forms on M is denoted by $\mathcal{C}^\infty \Lambda^k(M)$.

As the reader is no doubt unsurprised to hear, I do not intend to specify the smooth structure for $\text{Alt}^k(M)$ in this thesis. Note that I denote the set of all compact supported smooth k -forms by $\mathcal{C}_c^\infty \Lambda^k(M)$.

Some remarks on the definition of differential forms are in order. First, we usually omit saying “smooth” and “differential” in this chapter, preferring to just say k -form. Note that, since $\mathfrak{T}_0^1 = \mathbb{T}^*M$, our use of the term “1-form” is consistent. By convention, we interpret scalar fields as 0-forms. Many authors called differential forms exterior differential forms, hence the name “exterior calculus”.

We now provide several examples of differential forms, most of which are important enough to warrant their own definitions.

Definition 2.2.2. Let M and N be smooth manifolds and let $F : M \rightarrow N$ be smooth.

1) Suppose $f \in \mathcal{C}^\infty \Lambda^0(N)$. The **pullback** of f by F is the smooth function $F^*f \in \mathcal{C}^\infty \Lambda^0(M)$ given by $F^*f \doteq f \circ F$.

2) Let $\omega \in \mathcal{C}^\infty \Lambda^k(N)$ with $k \in \{1, \dots, \dim N\}$. The **pullback** of ω by F is the k -form $F^*\omega \in \mathcal{C}^\infty \Lambda^k(M)$ defined by the following: for any $\mathbf{u}_i \in \mathfrak{X}(M)$ with $i = 1, \dots, k$ and any $p \in M$, we have

$$(F^*\omega)_p(\mathbf{u}_1, \dots, \mathbf{u}_k) \doteq \omega_{F(p)}(F_{*,p}\mathbf{u}_1, \dots, F_{*,p}\mathbf{u}_k). \quad (2.2.1)$$

Notice how the pullback is defined in the only way that it can possibly make sense: ω must be fed k vector fields on N , so any attempt to define a “version” of ω on M must include a way of taking vector fields on M to vector fields on N , and this is precisely accomplished by the pushforward. In the case where N is a sufficiently nice manifold (possibly with corners) contained in M with inclusion map denoted by $\iota : N \rightarrow M$, we call the pullback by ι the **trace onto N** and use the notation $\text{tr}_N \doteq \iota^*$.

Definition 2.2.3. If $\mathbf{u} \in \mathfrak{X}(M)$ and $\omega \in C^\infty \Lambda^k(M)$, then we define $\mathbf{u} \lrcorner \omega \in C^\infty \Lambda^{k-1}(M)$ by

$$\mathbf{u} \lrcorner \omega(\mathbf{u}_1, \dots, \mathbf{u}_{k-1}) = \omega(\mathbf{u}, \mathbf{u}_1, \dots, \mathbf{u}_{k-1}) \quad \forall \mathbf{u}_i \in \mathfrak{X}(M).$$

We call $\mathbf{u} \lrcorner \omega$ the *contraction of ω with \mathbf{u}* .

Definition 2.2.4. Given $\alpha, \beta \in C^\infty \Lambda^1(M)$, we define $\alpha \wedge \beta \in C^\infty \Lambda^2(M)$ by

$$\alpha \wedge \beta(\mathbf{u}, \mathbf{v}) \doteq \det \begin{bmatrix} \alpha(\mathbf{u}) & \alpha(\mathbf{v}) \\ \beta(\mathbf{u}) & \beta(\mathbf{v}) \end{bmatrix} = \alpha(\mathbf{u})\beta(\mathbf{v}) - \alpha(\mathbf{v})\beta(\mathbf{u}) \quad \forall \mathbf{u}, \mathbf{v} \in \mathfrak{X}(M). \quad (2.2.2)$$

We call $\alpha \wedge \beta$ the *wedge product of α and β* .

By construction, $\alpha \wedge \beta$ is bilinear and antisymmetric, so it is indeed a 2-form. Notice that $\beta \wedge \alpha = -\alpha \wedge \beta$, and in particular $\alpha \wedge \alpha = 0$. Additionally, the wedge product has the distributive property:

$$\alpha \wedge (\beta + \gamma) = \alpha \wedge \beta + \alpha \wedge \gamma.$$

Example 2.2.5. To give a more concrete example, consider the 1-forms on \mathbb{R}^3 defined by

$$\alpha = y^2 dx + zy dz \quad \text{and} \quad \beta = x^2 y^3 dy + dz.$$

Then, using the distributive property and antisymmetry of \wedge ,

$$\begin{aligned} \alpha \wedge \beta &= x^2 y^5 dx \wedge dy + y^2 dx \wedge dz + x^2 y^4 z dz \wedge dy + zy dz \wedge dz \\ &= x^2 y^5 dx \wedge dy + y^2 dx \wedge dz + x^2 y^4 z dz \wedge dy. \end{aligned}$$

The wedge product generalizes naturally to accommodate finitely many factors. Given k 1-forms $\alpha^1, \dots, \alpha^k$, we define their wedge product to be the smooth k -form

$$\alpha^1 \wedge \dots \wedge \alpha^k(\mathbf{u}_1, \dots, \mathbf{u}_k) \doteq \det \begin{bmatrix} \alpha^1(\mathbf{u}_1) & \alpha^1(\mathbf{u}_2) & \dots & \alpha^1(\mathbf{u}_k) \\ \alpha^2(\mathbf{u}_1) & \alpha^2(\mathbf{u}_2) & \dots & \alpha^2(\mathbf{u}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^k(\mathbf{u}_1) & \alpha^k(\mathbf{u}_2) & \dots & \alpha^k(\mathbf{u}_k) \end{bmatrix} \quad \forall \mathbf{u}_1, \dots, \mathbf{u}_k \in \mathfrak{X}(M). \quad (2.2.3)$$

We would now like to define the wedge product of a k -form with an ℓ -form. In order to do this, we first show that the wedge product gives us a convenient local coordinate expression for k -forms. We begin by making a helpful definition that appears in different contexts throughout Part 1.

Definition 2.2.6. $\Sigma(k, n)$ denotes the set of all strictly increasing maps $\sigma: \{1, \dots, k\} \rightarrow \{1, \dots, n\}$

Lemma 2.2.7. $|\Sigma(k, n)| = \binom{n}{k}$. □

Proposition 2.2.8. [28, p. 3] $\{dx^{\sigma(1)}|_p \wedge \dots \wedge dx^{\sigma(k)}|_p\}_{\sigma \in \Sigma(k, n)}$ is a basis of $\text{Alt}^k(\mathbb{T}_p M)$.

Corollary 2.2.9. $\dim \text{Alt}^k(\mathbb{T}_p M) = \binom{n}{k}$.

□

Of course, since $\{dx^{\sigma(1)}|_p \wedge \cdots \wedge dx^{\sigma(k)}|_p\}_{\sigma \in \Sigma(k,n)}$ is a basis of $\text{Alt}^k(\mathbb{T}_p M)$, we can define a family of k -forms in a chart domain around p exactly as we saw for 1-forms:

$$dx^{\sigma(1)} \wedge \cdots \wedge dx^{\sigma(k)} : p \mapsto dx^{\sigma(1)}|_p \wedge \cdots \wedge dx^{\sigma(k)}|_p.$$

Every k -form on M can thus locally be written as a linear combination of the $dx^{\sigma(1)} \wedge \cdots \wedge dx^{\sigma(k)}$ with smooth functions as coefficients. At this point, defining the wedge product for differential forms of arbitrary degree just amounts to extension by linearity.

Definition 2.2.10. *Let $\omega \in \mathcal{C}^\infty \Lambda^k$ and $\eta \in \mathcal{C}^\infty \Lambda^\ell$ have the local coordinate expressions*

$$\omega = \sum_{\sigma \in \Sigma(k,n)} \omega_\sigma dx^{\sigma(1)} \wedge \cdots \wedge dx^{\sigma(k)} \quad \text{and} \quad \eta = \sum_{\sigma' \in \Sigma(\ell,n)} \eta_{\sigma'} dx^{\sigma'(1)} \wedge \cdots \wedge dx^{\sigma'(\ell)}.$$

Then, the **wedge product** of ω and η is the $(k + \ell)$ -form defined in local coordinates by

$$\omega \wedge \eta \doteq \sum_{\substack{\sigma \in \Sigma(k,n) \\ \sigma' \in \Sigma(\ell,n)}} \omega_\sigma \eta_{\sigma'} dx^{\sigma(1)} \wedge \cdots \wedge dx^{\sigma(k)} \wedge dx^{\sigma'(1)} \wedge \cdots \wedge dx^{\sigma'(\ell)}.$$

Before moving on, we introduce some compressed notation. If $\sigma \in \Sigma(k, n)$, then

$$dx^\sigma \doteq dx^{\sigma(1)} \wedge \cdots \wedge dx^{\sigma(k)}.$$

The change of variables rule for a general k -form is, by direct calculation,

$$\sum_{\sigma} \omega_{\sigma} dx^{\sigma} = \sum_{\sigma} \omega_{\sigma} \frac{\partial x^{\sigma(1)}}{\partial y^{j_1}} \cdots \frac{\partial x^{\sigma(k)}}{\partial y^{j_k}} dy^{j_1} \wedge \cdots \wedge dy^{j_k}. \quad (2.2.4)$$

Using this change of variables rule, we see that our definition of the wedge product is clearly independent of a particular choice of local coordinates.

In light of the above discussion, we see that n is the largest degree a form on M is allowed to have; any higher-degree form would have a local coordinate expression involving wedge products with repeated factors of some dx^i . Accordingly, we also refer to n -forms as **top forms**. Using (2.2.4) and letting S_n denote the group of permutations of $(1, \dots, n)$, the change of variables rule for a top form $\omega = \hat{\omega} dx^1 \wedge \cdots \wedge dx^n$ is

$$\begin{aligned} \hat{\omega} dx^1 \wedge \cdots \wedge dx^n &= \hat{\omega} \frac{\partial x^1}{\partial y^{j_1}} \cdots \frac{\partial x^n}{\partial y^{j_n}} dy^{j_1} \wedge \cdots \wedge dy^{j_n} \\ &= \hat{\omega} \sum_{s \in S_n} \text{sgn}(s) \frac{\partial x^1}{\partial y^{s(1)}} \cdots \frac{\partial x^n}{\partial y^{s(n)}} dy^1 \wedge \cdots \wedge dy^n \\ &= \hat{\omega} \det \left[\frac{\partial x^i}{\partial y^j} \right] dy^1 \wedge \cdots \wedge dy^n, \end{aligned}$$

where the Leibniz formula for determinants [51, p. 629] has been used to go from the second line to the third line. If the determinant is positive, then the formula above resembles the multivariable calculus change of variables rule for integrals,

$$\int_{x(\Omega)} f(x) dx^1 \cdots dx^n = \int_{\Omega} f(x(y)) \left| \det \left[\frac{\partial x^i}{\partial y^j} \right] \right| dy^1 \cdots dy^n.$$

The above correspondence provides motivation for using differential forms, rather than vector fields, as the main objects of interest in calculus (Harley Flanders put it best: “[differential forms] are the things which occur under integral signs” [35, p. 1]). We describe how to integrate differential forms in Section 2.5.

We now discuss some finer properties of the wedge product. First, we go through an interesting example.

Example 2.2.11. Consider \mathbb{R}^4 with coordinates (t, x, y, z) . Let

$$\omega = dt \wedge dx + dy \wedge dz \in \mathcal{C}^\infty \Lambda^2.$$

Then, we have that

$$\begin{aligned} \omega \wedge \omega &= dt \wedge dx \wedge dy \wedge dz + dy \wedge dz \wedge dt \wedge dx \\ &= dt \wedge dx \wedge dy \wedge dz + dt \wedge dy \wedge dz \wedge dx \\ &= 2 dt \wedge dx \wedge dy \wedge dz. \end{aligned}$$

The above example illustrates an important difference between the wedge product of general forms versus the wedge product of two 1-forms. This difference is encapsulated in the result below.

Lemma 2.2.12. *Let $\omega \in \mathcal{C}^\infty \Lambda^k$ and $\eta \in \mathcal{C}^\infty \Lambda^\ell$. Then,*

$$\omega \wedge \eta = (-1)^{k\ell} \eta \wedge \omega. \quad (2.2.5)$$

Proof. Pick any $\sigma \in \Sigma(k, n)$ and any $\tau \in \Sigma(\ell, n)$. We show that the claim holds for the special case

$$\omega = dx^{\sigma(1)} \wedge \cdots \wedge dx^{\sigma(k)} \quad \text{and} \quad \eta = dx^{\tau(1)} \wedge \cdots \wedge dx^{\tau(\ell)},$$

whence the full result follows through extension by linearity. Calculating the wedge product directly, we have

$$\begin{aligned} \omega \wedge \eta &= dx^{\sigma(1)} \wedge \cdots \wedge dx^{\sigma(k)} \wedge dx^{\tau(1)} \wedge \cdots \wedge dx^{\tau(\ell)} \\ &= -dx^{\sigma(1)} \wedge \cdots \wedge dx^{\sigma(k-1)} \wedge dx^{\tau(1)} \wedge dx^{\sigma(k)} \wedge dx^{\tau(2)} \cdots \wedge dx^{\tau(\ell)} \\ &= (-1)^k dx^{\tau(1)} \wedge dx^{\sigma(1)} \wedge \cdots \wedge dx^{\sigma(k)} \wedge dx^{\tau(2)} \wedge \cdots \wedge dx^{\tau(\ell)} \\ &= (-1)^{k\ell} dx^{\tau(1)} \wedge \cdots \wedge dx^{\tau(\ell)} \wedge dx^{\sigma(1)} \wedge \cdots \wedge dx^{\sigma(k)} \\ &= (-1)^{k\ell} \eta \wedge \omega. \end{aligned}$$

□

Finally, we remark that the wedge product respects pullbacks:

Proposition 2.2.13. *Let $F: N \rightarrow M$ be a smooth map between smooth manifolds, and let ω and η be differential forms on M . Then,*

$$F^*(\omega \wedge \eta) = (F^*\omega) \wedge (F^*\eta).$$

Proof. The proof just involves unraveling definitions, see [28, pp. 6–8] for details. □

2.3 Orientability and Hodge Duality

In this section, we introduce the notion of orientability on general smooth manifolds and use this notion to define a canonical bijection between k -forms and $(n-k)$ -forms, called **Hodge duality**. Our discussion of Hodge duality is very “bare-bones”, and I recommend the reader consult [35, §2.5–2.7] for a complete presentation.

Intuitively, the concept of orientability of a manifold M should be related to defining a consistent global notion of “up and down” on M . The reader is likely familiar with the normal vector field on a surface $M \subseteq \mathbb{R}^3$, namely a map $M \rightarrow \mathbb{R}^3$ whose image is everywhere orthogonal to $T_p M$ (with respect to Euclidean metric in \mathbb{R}^3). In courses on vector calculus, the existence of such an outward normal is taken as the definition of orientability. This definition has the disadvantage of depending on some particular “immersion” of M into a bigger Euclidean space. Since I focus on intrinsic properties in this chapter, I define orientability purely in terms of M ’s topology and smooth structure.

Definition 2.3.1. *A smooth atlas $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$ for M is said to be **oriented** if, for all α, β with $U_\alpha \cap U_\beta$ nonempty, we have*

$$\det \left[D \left(\varphi_\alpha \circ \varphi_\beta^{-1} \right) \right] > 0.$$

*If M admits an oriented smooth atlas, then we say that M is **orientable**.*

If M is a smooth surface in \mathbb{R}^3 , one can show that M is orientable in the sense of Definition 2.3.1 if and only if a smoothly varying unit normal to the surface can be defined unambiguously on all overlapping coordinate patches. Accordingly, our definition of orientability is motivated by the more familiar, intuitive notion of orientability encountered in vector calculus (though admittedly this is not completely obvious).

Let us say that oriented atlases \mathcal{A} and \mathcal{B} are **equivalent** if $\mathcal{A} \cup \mathcal{B}$ is an oriented atlas. If M is connected and $\dim M > 0$ then there are, in fact, only two oriented atlases up to equivalence. We can use one class to represent a notion of “up”, so the other class naturally represents a notion of “down”. Accordingly, we say that an **orientation** on M is a choice of either equivalence class of oriented atlases. We say that an orientable manifold M is **oriented** once we have specified an orientation.

Example 2.3.2. Let $\text{Id}: \mathbb{R}^n \rightarrow \mathbb{R}^n$ denote the identity map. Consider the atlas $\{(\mathbb{R}^n, \text{Id})\}$ for \mathbb{R}^n . Now, $D(\text{Id}) = \text{Id}$ so $\{(\mathbb{R}^n, \text{Id})\}$ is clearly oriented. Its equivalence class corresponds to the usual right-hand orientation on \mathbb{R}^n .

Remark 2.3.3. *As it turns out, one needs to demand that a manifold M is orientable in order to define the integral of a differential form over M . This is because differential forms and the geometric notion of volume measurement are intimately related; we do not have time to go through the details here, but see [51, Chapter 15] for details.*

Now, we use orientation to define a canonical correspondence between different spaces of differential forms on M . This correspondence is called **Hodge duality**. To the beginner, Hodge duality may seem like a useless coincidence of no practical utility. However, by the end of this chapter, we see that Hodge duality provides a convenient way of generalizing familiar differential operators from vector calculus to the setting of smooth manifolds. In later chapters, we also see that Hodge duality is instrumental in defining function spaces useful in studying PDEs and their numerical solutions. For the sake of time, we cannot present a complete discussion of the next result, but the reader is encouraged to see [35, §2.7] for an insightful and complete presentation.

Theorem 2.3.4. (*Hodge Duality Theorem*) Suppose that (M, g) is oriented and Riemannian. For every $p \in M$ there exists a canonical isomorphism $\star: \text{Alt}^k(\mathbb{T}_p M) \rightarrow \text{Alt}^{n-k}(\mathbb{T}_p M)$, called the **Hodge star operator**, uniquely defined by the metric and orientation. Further, the inverse of \star is given by

$$\star \star \omega_p = (-1)^{k(n-k)} \omega_p. \quad (2.3.1)$$

Clearly, the Hodge star gives rise to an isomorphism $\star: \mathcal{C}^\infty \Lambda^k \rightarrow \mathcal{C}^\infty \Lambda^{n-k}$ defined by $(\star \omega)_p \doteq \star(\omega_p)$.

I now provide some hands-on examples illustrating how to compute with the Hodge star. One can always express the Hodge star concretely in terms of the dx^σ 's [62, pp. 290–291], but the general formula is a bit daunting. So, we only stick to discussing concrete expressions for the Hodge star in simple cases. We focus on \mathbb{R}^n equipped with the Euclidean metric g (we always use the right-hand orientation without explicitly saying it). Before stating the formula for the Hodge star on (\mathbb{R}^n, g) , we go through some new notation.

Definition 2.3.5. Given $\sigma \in \Sigma(k, n)$, $\sigma^* \in \Sigma(n-k, n)$ is defined to be the unique increasing map $\{1, \dots, n-k\} \rightarrow \{1, \dots, n\}$ such that

$$(\sigma(1), \dots, \sigma(k), \sigma^*(1), \dots, \sigma^*(n-k))$$

is a permutation of $(1, 2, \dots, n)$.

Example 2.3.6. If $n = 4$, $k = 2$, and $\sigma(1) = 2$, $\sigma(2) = 4$, then $\sigma^*(1) = 1$ and $\sigma^*(2) = 3$.

We let $\text{sgn}(\sigma, \sigma^*)$ denote the sign of the permutation of $(1, \dots, n)$ obtained by concatenating σ and σ^* . In Example 2.3.6, for instance, we would have $\text{sgn}(\sigma, \sigma^*) = \text{sgn}(2, 4, 1, 3) = -1$.

Proposition 2.3.7. If \mathbb{R}^n is endowed with the Euclidean metric, then the action of the Hodge star on any k -form $\omega = \sum_\sigma \omega_\sigma dx^\sigma$ is given by

$$\star \omega = \sum_\sigma (-1)^{\text{sgn}(\sigma, \sigma^*)} \omega_\sigma dx^{\sigma^*}. \quad (2.3.2)$$

Example 2.3.8. Equip \mathbb{R}^2 with the Euclidean metric and denote coordinates by (x, y) . Then, using (2.3.2), we know that

$$\star dx = dy \quad \text{and} \quad \star dy = -dx.$$

Then, we can picture the Hodge star acting on 1-forms in \mathbb{R}^2 as a 90° counterclockwise rotation.

Example 2.3.9. Equip \mathbb{R}^3 with the Euclidean metric and denote coordinates by (x, y, z) . Then, using (2.3.2) and the antisymmetry of \wedge , we have that

$$\begin{aligned} \star dx &= (-1)^{\text{sgn}(1,2,3)} dy \wedge dz = dy \wedge dz, \\ \star dy &= (-1)^{\text{sgn}(2,1,3)} dx \wedge dz = dz \wedge dx, \quad \text{and} \\ \star dz &= (-1)^{\text{sgn}(3,1,2)} dx \wedge dy = dx \wedge dy. \end{aligned}$$

By (2.3.1), $\star: \mathcal{C}^\infty \Lambda^1(\mathbb{R}^3) \rightarrow \mathcal{C}^\infty \Lambda^2(\mathbb{R}^3)$ is its own inverse and we can easily compute the action of the Hodge star on all 2-forms on \mathbb{R}^3 . An easy way to calculate the Hodge star for 1- and 2-forms on \mathbb{R}^3 is to remember that, if (i, j, k) is a cyclic permutation of $(1, 2, 3)$, then

$$\star dx^i = dx^j \wedge dx^k.$$

2.4 The Exterior Derivative

Finally, we can finish studying algebra and move on to studying calculus. In this section, we explain a canonical way of differentiating k -forms. By the end of this section, we see that this notion of differentiation encapsulates the most fundamental differential operators in vector calculus.

Definition 2.4.1. Let $\omega \in \mathcal{C}^\infty \Lambda^k(M)$ be expressed in local coordinates by

$$\omega = \sum_{\sigma} \omega_{\sigma} dx^{\sigma}.$$

We define the **exterior derivative of ω** , denoted by $d\omega \in \mathcal{C}^\infty \Lambda^{k+1}(M)$, to be

$$d\omega = \sum_{\sigma \in \Sigma(k,n)} \sum_{j=1}^n \frac{\partial \omega_{\sigma}}{\partial x^j} dx^j \wedge dx^{\sigma}.$$

We have neglected the Einstein convention in the above definition just to prevent confusion as to where each index lives. One can check that the definition of $d\omega$ is independent of coordinates.

Example 2.4.2. Let $f = \sin(x) + xy^2 \in \mathcal{C}^\infty \Lambda^0(\mathbb{R}^2)$. Then,

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy = (-\cos(x) + y^2) dx + (2xy) dy.$$

Observe that the components of df are the same as the components of the gradient of f . That is, $df = (\nabla f)^b$ here. We can also compute the second exterior derivative of f :

$$ddf = \frac{\partial}{\partial x}(-\cos(x) + y^2) dx \wedge dx + \frac{\partial}{\partial y}(-\cos(x) + y^2) dy \wedge dx + \frac{\partial}{\partial x}(2xy) dx \wedge dy + \frac{\partial}{\partial y}(2xy) dy \wedge dy.$$

Now, $dx \wedge dx = 0$ and $dy \wedge dy = 0$, so the above becomes

$$ddf = 2y dy \wedge dx + 2y dx \wedge dy.$$

Since the wedge product is antisymmetric on 1-forms, $dy \wedge dx = -dx \wedge dy$. Therefore, $ddf = 0$.

Example 2.4.3. Let $\omega = z dx + y dy + x dz \in \mathcal{C}^\infty \Lambda^1(\mathbb{R}^3)$. Then,

$$d\omega = dz \wedge dx + 0 + dx \wedge dz = 0.$$

Consider the image of ω under the musical isomorphism,

$$\omega^\sharp = z \frac{\partial}{\partial x} + y \frac{\partial}{\partial y} + x \frac{\partial}{\partial z}.$$

We may directly compute that $\nabla \times \omega^\sharp = 0$.

Example 2.4.4. Let

$$\omega = xy dy \wedge dz + x^{32} dz \wedge dx + (z^2 + \operatorname{sech}^2(y)) dx \wedge dy \in \mathcal{C}^\infty \Lambda^2(\mathbb{R}^3).$$

Then,

$$d\omega = y \, dx \wedge dy \wedge dz + 2z \, dz \wedge dx \wedge dy = (y + 2z) \, dx \wedge dy \wedge dz.$$

Since $d\omega$ is a top form, its exterior derivative is trivially 0. Observe that the coefficient of $d\omega$ is precisely equal to the divergence of the vector field

$$\omega^\sharp = xy \frac{\partial}{\partial x} + 32x^{32} \frac{\partial}{\partial y} + (z^2 + \operatorname{sech}^2(y)) \frac{\partial}{\partial z}$$

(we justify the use of the \sharp symbol in the next paragraph).

Since the gradient, curl, and divergence have shown up as special cases of the exterior derivative (up to a shift in notation), we might guess that d condenses all of the main operations of vector calculus into one easy symbol. This is indeed the case, but before we state this result formally we articulate the relationship between vectors and forms in Euclidean spaces more clearly. Now, in every \mathbb{R}^n , the musical isomorphism gives us a correspondence $\mathcal{C}^\infty \Lambda^1(\mathbb{R}^n) \simeq \mathfrak{X}(\mathbb{R}^n)$ and, further, this isomorphism has the simple form

$$\alpha_i \, dx^i \mapsto \alpha_i \frac{\partial}{\partial x^i}.$$

So, for every n , we can canonically use 1-forms as “proxies” for vector fields and vice versa. Accordingly, we may differentiate vector fields using the exterior derivative by passing to the corresponding proxy form. In \mathbb{R}^3 , the situation is a bit more interesting: Hodge duality gives another canonical correspondence $\mathcal{C}^\infty \Lambda^1(\mathbb{R}^3) \simeq \mathcal{C}^\infty \Lambda^2(\mathbb{R}^3)$, so we actually have *two* types of differential forms that can serve as vector proxies. This means that, when we apply d to a 1-form in \mathbb{R}^3 , we are mapping a vector field to another vector field, at least up to canonical isomorphism. In the remainder of this thesis, we use the aforementioned canonical correspondences very frequently. The discussion above is summarized in the next two results, the proofs of which are routine exercises following straight from the definitions.

Proposition 2.4.5. *In \mathbb{R}^2 we have that, up to proxies,*

$$\begin{aligned} d|_{\mathcal{C}^\infty \Lambda^0} &= \nabla, \\ d|_{\mathcal{C}^\infty \Lambda^1} &= \nabla \times, \\ \star d|_{\mathcal{C}^\infty \Lambda^0} &= \nabla^\perp, \text{ and} \\ \star d \star |_{\mathcal{C}^\infty \Lambda^1} &= \nabla \cdot. \end{aligned}$$

Proposition 2.4.6. *In \mathbb{R}^3 we have that, up to proxies,*

$$\begin{aligned} d|_{\mathcal{C}^\infty \Lambda^0} &= \nabla, \\ d|_{\mathcal{C}^\infty \Lambda^1} &= \nabla \times, \text{ and} \\ \star d|_{\mathcal{C}^\infty \Lambda^2} &= \nabla \cdot. \end{aligned}$$

In light of the above proposition, we identify $d\star|_{\mathcal{C}^\infty \Lambda^1(\mathbb{R}^2)}$ and $d|_{\mathcal{C}^\infty \Lambda^2(\mathbb{R}^3)}$ with the divergence operator in the sequel, usually without comment. All in all, we have seen that the exterior derivative tidily summarizes the main differential operators of vector calculus. Therefore, d is clearly instrumental in generalizing the ideas of vector calculus to differential forms on arbitrary manifolds.

Having motivated why studying d is important, we state a result summarizing this operator’s most useful properties.

Proposition 2.4.7. [35, §3.2, 3.3]

1) $d \circ d = 0$,

2) If ω is a k -form and η is an ℓ -form, then

$$d(\omega \wedge \eta) = d\omega \wedge \eta + (-1)^k \omega \wedge d\eta.$$

3) If $F: M \rightarrow N$ is smooth and ω is any differential form, then

$$F^*d\omega = dF^*\omega.$$

Note that the property $d \circ d = 0$ generalizes the familiar vector calculus identities

$$\begin{aligned} \nabla \times \nabla &= 0 \text{ and} \\ \nabla \cdot \nabla \times &= 0 \end{aligned}$$

to arbitrary manifolds.

2.5 Integration of Differential Forms

In this section, we conclude our discussion of calculus on manifolds by explaining how to integrate differential forms. Further, we see how integration is related to exterior differentiation by Stokes' Theorem. As discussed briefly in Section 2.2, we expect differential forms to play very nicely with the notion of integration, since the change of variables rule for an n -form is essentially identical to the change of variables rule for integrals over domains in \mathbb{R}^n . With this insight in mind, a central throughline we want to emphasize is that k -forms can be integrated over k -dimensional manifolds. For example, 1-forms can be integrated over curves (to compute circulations) and 2-forms over surfaces (to compute fluxes). This idea has been used (at least implicitly) since at least 1869: Maxwell himself knew that there were fundamental differences between vector fields associated to line integrals and vector fields associated to surface integrals [57]. In modern language, we can use our knowledge of vector proxies to say that Maxwell realized that 1-forms and 2-forms were not quite the same objects. So, our definition of integration is well-motivated by classical physical reasoning.

First, we introduce how to integrate smooth n -forms on \mathbb{R}^n . This is nothing crazy at all: if $\omega = \hat{\omega} dx^1 \wedge \cdots \wedge dx^n$, then its integral over a domain Ω is defined to be

$$\int_{\Omega} \omega \doteq \int_{\Omega} \hat{\omega} dx^1 \cdots dx^n,$$

provided the integral on the right-hand side makes sense. That is, we just peel off the coefficient of ω and evaluate its Riemann integral in the usual sense. With this new paradigm for integration in mind, we can easily define integration of n -forms on chart domains. Let M be an oriented n -dimensional manifold. If $\omega \in \mathcal{C}^\infty \Lambda^n(M)$ and $(U_\alpha, \varphi_\alpha)$ is an oriented chart on M , then

$$\int_{U_\alpha} \omega \doteq \int_{\varphi_\alpha(U_\alpha)} (\varphi_\alpha^{-1})^* \omega. \tag{2.5.1}$$

At this point, we have only to extend the above definition to the entirety of M . To accomplish this, we need to somehow sew together the integrals from each coordinate patch. Fortunately, the following result makes the transition from local to global integrals easy.

Theorem 2.5.1. [90, Thm. 2.13] *Let M be a smooth manifold and let $\{U_\alpha\}_{\alpha \in I}$ be an open cover of Ω . Then, there exists a family of compactly supported smooth functions $\psi_\alpha: M \rightarrow \mathbb{R}$ such that*

- 1) $0 \leq \psi_\alpha \leq 1 \quad \forall \alpha$,
- 2) for all α , $\text{supp } \psi_\alpha \subseteq U_\alpha$,
- 3) each $p \in M$ has a neighbourhood that intersects only finitely many of the $\text{supp } \psi_\alpha$, and
- 4) $\sum_\alpha \psi_\alpha = 1$.

The collection $\{\psi_\alpha\}_{\alpha \in I}$ is called a **partition of unity on M subordinate to the cover $\{U_\alpha\}_{\alpha \in I}$** .

We can picture each element of a partition of unity as a little bump function; as one bump falls, another bump rises to make sure the sum of all the functions is still 1 everywhere.

Definition 2.5.2. *Suppose that M is oriented and let $\omega \in \mathcal{C}^\infty \Lambda^n(M)$. Cover M with charts $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$ and choose a partition of unity $\{\psi_\alpha\}_{\alpha \in I}$ subordinate to this cover. We define*

$$\int_M \omega \doteq \sum_{\alpha \in I} \int_{U_\alpha} \psi_\alpha \omega. \quad (2.5.2)$$

Proposition 2.5.3. [51, Prop. 16.5] *The notion of integral defined above does not depend on the open cover or partition of unity.*

Note that, if M is a 0-dimensional manifold (a collection of isolated points), then we interpret the integral of a scalar over M simply as a sum of point evaluations.

If N is a sufficiently nice k -dimensional manifold contained in M , then we can define the integral of $\omega \in \mathcal{C}^\infty \Lambda^k(M)$ over N through pulling ω back by the inclusion $\iota: N \rightarrow M$. Using this construction we can, for example, make sense of the integral of an element of $\mathcal{C}^\infty \Lambda^1(\mathbb{R}^3)$ over a smooth curve in \mathbb{R}^3 .

Perhaps the most important property of the integral defined above is its satisfaction of a change of variables rule. After our discussion in Section 2.5, we ought to have that this is the case. One can use the local coordinate expression for the pullback to show that the result below agrees with the usual multivariable calculus change of variables rule when working on domains in \mathbb{R}^n .

Proposition 2.5.4. (Change of Variables Rule [51, Prop. 16.6]) *Let $F: N \rightarrow M$ be a diffeomorphism whose local coordinate representation has positive Jacobian determinant everywhere. Then,*

$$\int_M \omega = \int_N F^* \omega.$$

Now, we state the general Stokes' Theorem, one of the most important results for our work in the sequel.

Theorem 2.5.5. (Stokes' Theorem [51, Thm. 16.11, 16.25]) *Let M be an oriented smooth manifold with boundary (and possibly with corners) and let $\omega \in \mathcal{C}^\infty \Lambda^{n-1}(M)$. Then,*

$$\int_M d\omega = \int_{\partial M} \text{tr}_{\partial M} \omega. \quad (2.5.3)$$

If M does not have boundary, the right-hand side of the Stokes' Theorem expression is simply 0. Using Propositions 2.4.5 and 2.4.6, we see that Stokes' Theorem includes Green's Theorem, the Divergence Theorem, and the "usual" Stokes' Theorem as special cases.

Combining Stokes' Theorem with the product rule (Proposition 2.4.7), we obtain the exterior calculus version of integration by parts:

Proposition 2.5.6. *If $\omega \in \mathcal{C}^\infty \Lambda^k(M)$ and $\eta \in \mathcal{C}^\infty \Lambda^{n-k-1}(M)$, then*

$$\int_M d\omega \wedge \eta = \int_{\partial M} \text{tr}_{\partial M} (\omega \wedge \eta) + (-1)^{k+1} \int_M \omega \wedge d\eta. \quad (2.5.4)$$

□

Before finishing this section, we introduce some simple ideas that return when we study function spaces on manifolds in Chapter 3.

Definition 2.5.7. *If ω, η are k -forms, then we define their L^2 inner product as*

$$\langle \omega, \eta \rangle_{L^2 \Lambda^k} \doteq \int_M \omega \wedge \star \eta. \quad (2.5.5)$$

In Chapter 3, we extend the L^2 inner product to forms with non-smooth components. Having defined inner products, we can talk now about adjoints in the setting of exterior calculus. The proof of the next result essentially just boils down to juggling around symbols and re-phrasing integration by parts.

Definition 2.5.8. *Let M be an oriented Riemannian n -manifold, possibly with boundary. The exterior coderivative $\delta: \mathcal{C}^\infty \Lambda^k(M) \rightarrow \mathcal{C}^\infty \Lambda^{k-1}(M)$ is defined by*

$$\delta \doteq (-1)^{nk+n+1} \star d \star. \quad (2.5.6)$$

Proposition 2.5.9. *For all $\omega \in \mathcal{C}^\infty \Lambda^k$ and $\eta \in \mathcal{C}^\infty \Lambda^{k-1}$, δ satisfies*

$$\langle \omega, d\eta \rangle_{L^2 \Lambda^k} = \int_{\partial M} \text{tr}_{\partial M} (\eta \wedge \star \omega) + \langle \delta \omega, \eta \rangle_{L^2 \Lambda^k}. \quad (2.5.7)$$

In particular, if η vanishes on ∂M , then

$$\langle \omega, d\eta \rangle_{L^2 \Lambda^k} = \langle \delta \omega, \eta \rangle_{L^2 \Lambda^k}.$$

The above result tells us that, at least formally, δ is the adjoint of d (provided we work with forms that are compactly supported inside M). We return to this observation in much greater detail in Chapter 4.

Chapter 3

Hilbert Spaces of Differential Forms

In this section, we introduce some important Hilbert spaces used in numerical PDE theory. Such spaces consist of forms constrained to satisfy some integrability conditions and, occasionally, other calculus-related regularity requirements. To build our function spaces, we need to extend the calculus of differential forms described in Chapter 2 to non-differentiable forms on non-smooth spaces. Accordingly, we describe in this chapter how to modify the notions of pullback, differentiation, and integration to work in more general circumstances. Sometimes, especially when working on non-smooth spaces, we even need to tweak the definition of a differential form itself.

The discussion in this chapter is intended to fill a gap that I feel is present in the literature; careful proofs of many results in this section, or even satisfactory definitions, are difficult to come by in numerical analysis papers. Owing to the nature of the material, the presentation style is necessarily a bit technical and dry. Accordingly, I might advise the reader to just skim the main definitions and results, returning to read the proofs and comments in this chapter as need be while they go through the rest of Part 1.

In Section 3.1, we review some basics on **Lipschitz maps** and define **Lipschitz manifolds**, a special type of non-smooth topological space. In Section 3.2, we discuss non-smooth (or “rough”) differential forms on open subsets of \mathbb{R}^n , in turn allowing us to define L^2 spaces of forms on both smooth and Lipschitz manifolds. In Section 3.3, I introduce **Sobolev spaces**, the key objects of interest in finite element analysis. These spaces consist of rough forms that are **weakly differentiable**, thus bringing differential calculus into the world of function spaces. In Section 3.4, I introduce the **trace operator**, a map allowing us to sensibly discuss boundary values for forms living in our Sobolev spaces.

We occasionally have to use some of the vernacular of measure theory in this chapter, so to make the presentation as clear as possible we briefly review the relevant definitions before proceeding. Readers looking for more details should consult [11, Chapter 4]. Recall that the Borel σ -algebra on a topological space Ω , denoted $\beta_\Omega \subseteq 2^\Omega$, is the σ -algebra generated by the open subsets of Ω . If $S \in \beta_{\mathbb{R}^n}$, we let $\lambda: \beta_S \rightarrow \mathbb{R} \cup \{\infty\}$ denote the Lebesgue measure on S . We say that $f: S \rightarrow \mathbb{R}$ is **measurable** if $f^{-1}(A) \in \beta_S$ for all $A \in \beta_{\mathbb{R}}$. Measurable functions are also called **measurable 0-forms**. The real vector space of measurable functions on S is denoted by $\Lambda^0(S)$. When we say that a property holds **almost everywhere** on S , we mean that the property holds on a set of the form $S - A$ where $A \in \beta_S$ satisfies $\lambda(A) = 0$. We abbreviate “almost everywhere” using the symbol λ -ae.

3.1 Lipschitz Manifolds

In this section, we describe how to extend some of the ideas from Chapter 2 into the realm of **Lipschitz manifolds**, essentially the least smooth domains we encounter in the sequel. The reader looking for more details on Lipschitz manifolds should consult [90, Chapter I], though be warned that, in this book, Lipschitz manifolds are called $C^{0,1}$ -manifolds.

We begin by recalling the definition of **Lipschitz maps** on normed spaces.

Definition 3.1.1. Let X and Y be real normed vector spaces with norms denoted by $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively. A map $F: X \rightarrow Y$ is said to be **Lipschitz** if there exists $C > 0$ such that, for all $x_1, x_2 \in X$, we have

$$\|F(x_1) - F(x_2)\|_Y \leq C \|x_1 - x_2\|_X.$$

Lipschitz maps are somewhere between continuous and continuously differentiable, as the next result illustrates:

Proposition 3.1.2.

1) All Lipschitz maps are continuous.

2) Let $\Sigma \subseteq \mathbb{R}^n$ be compact. If $F: \Sigma \rightarrow \mathbb{R}^m$ is continuously differentiable, then it is Lipschitz. □

The next two examples show, however, that the Lipschitz property is slightly weaker than continuous differentiability.

Example 3.1.3. Using the triangle inequality, we immediately see that the absolute value function $F(x) = |x|$ is Lipschitz on \mathbb{R} with $C = 1$. However, F is not differentiable in any neighbourhood of $x = 0$.

Example 3.1.4. Consider the map $F: \mathbb{R} \rightarrow \mathbb{R}$ given by

$$F = \begin{cases} x & \text{if } x < 0 \text{ and} \\ 2x & \text{otherwise.} \end{cases}$$

F is continuously differentiable on $\{x < 0\}$ and $\{x > 0\}$, but it is not continuously differentiable on any neighbourhood of the origin. However, we may verify directly that F is Lipschitz on \mathbb{R} with $C = 2$: clearly, we only have to check that the Lipschitz property holds when $x_1 < 0$ and $x_2 \geq 0$, and in this case we have

$$|F(x_1) - F(x_2)| = |x_1 - 2x_2| \leq |x_1 - x_2| + |x_2| \leq |x_1 - x_2| + |x_2 - x_1| = 2|x_1 - x_2|$$

and the claim holds.

With the above two examples in mind, we clearly see that any continuous piecewise-linear map is Lipschitz with C given by the maximum of the absolute values of the slopes on each piece.

The discussion above is (in a way) summarized by Rademacher's Theorem, which we state presently. We use the version of the theorem presented in [85, Thm. 1.1].

Theorem 3.1.5. (Rademacher) Let $U \subseteq \mathbb{R}^n$ be open, and denote Cartesian coordinates on U by x^1, \dots, x^n . Let $F: U \rightarrow \mathbb{R}$ be Lipschitz. Then, for almost every $p \in U$, the partial derivatives $\frac{\partial F}{\partial x^i}(p)$ exist. Further, the functions $\frac{\partial F}{\partial x^i}$ are all measurable and bounded.

We use Rademacher’s Theorem frequently in this chapter to help transplant our developments from Chapter 2 into a more general setting. I remark that, if $F: U \cap \mathbb{H}^n \rightarrow \mathbb{R}$ is Lipschitz and U is open, then Rademacher’s Theorem tells us that F is still differentiable almost everywhere: since $\{x^n = 0\} \cap (U \cap \mathbb{H}^n)$ has measure zero, the derivatives on the boundary points don’t have to be considered at all.

As we used smooth maps to define smooth manifolds, we can define “rough” manifolds using Lipschitz maps.

Definition 3.1.6.

1) Let X and Y be normed real vector spaces and let $F: X \rightarrow Y$ be a Lipschitz map. We say that F is a **Lipschitz homeomorphism** if F is invertible and its inverse is also a Lipschitz map.

2) A topological space is said to be a **Lipschitz manifold** if it satisfies Definition 2.1.1, except the transition functions are Lipschitz homeomorphisms instead of diffeomorphisms.

Similarly, we can define Lipschitz manifolds with boundary (and even Lipschitz manifolds with corners) precisely as we did in Chapter 2, just replacing the word “diffeomorphism” with “Lipschitz homeomorphism”. Of course, since plenty of corners can be modeled as graphs of Lipschitz maps, the study of Lipschitz manifolds with corners is a little bit vestigial. Likewise, a function on a Lipschitz manifold Ω is called **Lipschitz** if its coordinate expression in terms of a Lipschitz chart on Ω is a Lipschitz function; this is just the analogue of a smooth function on a smooth manifold in the Lipschitz setting.

There is, however, an important distinction between analysis on smooth manifolds and analysis on Lipschitz manifolds. Namely, we cannot canonically define the tangent bundle of a Lipschitz manifold, as the charts are not smooth enough to push forward into bundle charts. Accordingly, the study of differential forms on Lipschitz manifolds requires a bit of care, and we cannot always adapt tools from the smooth setting to describe objects in the Lipschitz setting. We return to defining forms on Lipschitz manifolds again in the next section.

The following theorem is just the “Lipschitz version” of the analogous result in the smooth case.

Theorem 3.1.7. [90, Thm. 2.13] *Let Ω be a Lipschitz manifold and let $\{U_\alpha\}_{\alpha \in I}$ be an open cover of Ω . Then, there exists a family of compactly supported Lipschitz functions $\psi_\alpha: \Omega \rightarrow \mathbb{R}$ such that*

- 1) $0 \leq \psi_\alpha \leq 1 \quad \forall \alpha$,
- 2) for all α , $\text{supp } \psi_\alpha \subseteq U_\alpha$,
- 3) each $p \in \Omega$ has a neighbourhood that intersects only finitely many of the $\text{supp } \psi_\alpha$, and
- 4) $\sum_\alpha \psi_\alpha = 1$.

The collection $\{\psi_\alpha\}_{\alpha \in I}$ is called a **partition of unity on Ω subordinate to the cover $\{U_\alpha\}_{\alpha \in I}$** .

Partitions of unity are critically important technical tools in differential geometry, used in practice to take locally defined analytical data and “spread it out” over the entire manifold. In the next section, for example, we use partitions of unity to help define function spaces on Lipschitz manifolds based on how similar function spaces are defined in local coordinates.

We now define a particular object appearing frequently throughout Part 1.

Definition 3.1.8. Let $\Omega \subseteq \mathbb{R}^n$ be open, bounded, and connected. We say that Ω is a **Lipschitz domain** if $\partial\Omega$ is a Lipschitz manifold of dimension $n - 1$.

In other words, a Lipschitz domain is a bounded, open, connected subset of \mathbb{R}^n such that every $x \in \partial\Omega$ has an open neighbourhood $U \subseteq \overline{\Omega}$ that is Lipschitz-homeomorphic to an open subset of the half-space \mathbb{H}^n . If Ω is a Lipschitz domain, then $\overline{\Omega}$ is a compact Lipschitz manifold with boundary such that the manifold boundary of $\overline{\Omega}$ coincides with the boundary of Ω as a subset of \mathbb{R}^n . The condition that the boundary be a Lipschitz $(n - 1)$ -manifold is necessary to ensure that we do not admit sets like $(0, 1) \times (0, 1) - \{(0, 0)\} \subseteq \mathbb{R}^2$ as Lipschitz domains.

Example 3.1.9. If $\partial\Omega$ is a piecewise-linear curve then Ω is a Lipschitz domain. For example, the unit square $(0, 1) \times (0, 1) \subseteq \mathbb{R}^2$ is a Lipschitz domain.

Of course, Lipschitz domains are themselves smooth manifolds, hence functions and forms of any order of differentiability may be defined on them. We must be careful, however, when applying theorems from PDE theory to forms defined on Lipschitz domains, as such theorems often assume a smooth boundary.

3.2 The Spaces $L^2\Lambda^k$

In this section we describe how to build interesting function spaces on both open subsets of \mathbb{R}^n and Lipschitz manifolds. Our presentation follows that of [75] and [85]. Throughout this section, let Ω be a Lipschitz manifold and let $U \subseteq \mathbb{R}^n$ be open. Recall that a **Hilbert space** is an inner product space X such that every Cauchy sequence in X achieves a limit in X .

We begin by defining a Hilbert space that many readers are probably familiar with.

Definition 3.2.1.

1) We say $f \in \Lambda^0(U)$ is **square-integrable** if we have that

$$\|f\|_{L^2} \doteq \left(\int_U |f|^2 \, d\lambda \right)^{\frac{1}{2}} < \infty. \quad (3.2.1)$$

We call $\|f\|_{L^2}$ the **2-norm** of f .

2) $L^2\Lambda^0(U) \doteq \{f \in \Lambda^0(U) \mid \|f\|_{L^2} < \infty\} / \text{equivalence } \lambda\text{-ae}$.

The quotient by “equivalence λ -ae” means that we consider functions that are identically equal λ -ae to be part of the same equivalence class. Notice that $L^2\Lambda^0(U)$ is a real vector space with respect to pointwise addition and scalar multiplication of equivalence classes. Sometimes, when the open set U is obvious, we simply write $L^2\Lambda^0$ or even L^2 to mean $L^2\Lambda^0(U)$.

Proposition 3.2.2. $\|\cdot\|_{L^2} : L^2\Lambda^0(U) \rightarrow \mathbb{R}$ defines a norm. This norm arises from an inner product given by

$$\langle f, g \rangle_{L^2} \doteq \int_{\Omega} fg \, d\lambda. \quad (3.2.2)$$

Further, $L^2\Lambda^0(U)$ is a Hilbert space.

□

We now want to define “ L^2 -spaces” of differential k -forms on U . This turns out to be quite painless if we use some simple constructions inspired by the theory in Chapter 2. First, we make a measure-theoretic definition.

Definition 3.2.3. Let $\pi: \text{Alt}^k(U) \rightarrow U$ denote the bundle projection. A map $\omega: U \rightarrow \text{Alt}^k(U)$ is called a *measurable k -form* if

- ω is a *section* of $\text{Alt}^k(U)$; that is, $\pi \circ \omega = \text{Id}_U$.
- $\omega^{-1}(A) \in \beta_U$ for all $A \in \beta_{\text{Alt}^k(U)}$.

The set of measurable k -forms is denoted by $\Lambda^k(U)$.

Note that this definition agrees with the definition of a measurable function in the case $k = 0$. We often say “rough k -forms” to mean “measurable k -forms”, in order to emphasize that these objects are not necessarily smooth (though all smooth forms are, trivially, measurable). Recall that the set of smooth k -forms on U is denoted by $\mathcal{C}^\infty \Lambda^k(U)$.

Remark 3.2.4. Making the transition from smooth forms to rough forms is essential for giving us the power to study PDEs on manifolds: without the freedom to work with spaces of differential forms admitting discontinuous or otherwise ill-behaved elements, we are not able to use the full power of functional analysis and modern PDE theory to help us solve problems.

Since the Hodge star makes sense on rough forms (locally, rough k -forms are linear combinations of the forms $dx^{i_1} \wedge \cdots \wedge dx^{i_k}$ with coefficients in $\Lambda^0(U)$), defining the integral of $\omega \in \Lambda^k(U)$ over U is straightforward:

$$\int_U \omega \doteq \int_U \star \omega \, d\lambda.$$

The definition of $L^2 \Lambda^0$ extends naturally to the setting of measurable k -forms using the “inner product”

$$\langle \omega, \eta \rangle = \int_U \omega \wedge \star \eta.$$

Definition 3.2.5.

1) We say $\omega \in \Lambda^k(U)$ is *square-integrable* if

$$\|\omega\|_{L^2 \Lambda^k} \doteq \left(\int_U \langle \omega, \omega \rangle \right)^{\frac{1}{2}} < \infty. \quad (3.2.3)$$

$\|\omega\|_{L^2 \Lambda^k}$ is called the *2-norm* of ω .

2) $L^2 \Lambda^k(U) \doteq \{ \omega \in \Lambda^k(U) \mid \|\omega\|_{L^2 \Lambda^k} < \infty \}$ / equivalence λ -ae.

We use the abbreviation $L^2 \Lambda^k \doteq L^2 \Lambda^k(U)$ when U is clear from context. We now show that $L^2 \Lambda^k$ may be endowed with the structure of a Hilbert space. The proof of the next theorem is straightforward, using only the formula for the Hodge star with respect to the Euclidean metric in \mathbb{R}^n and the completeness of $L^2 \Lambda^0$.

Theorem 3.2.6. $L^2\Lambda^k$ is a Hilbert space with respect to the inner product

$$\langle \omega, \eta \rangle_{L^2\Lambda^k} \doteq \int_U \langle \omega, \eta \rangle. \quad (3.2.4)$$

□

Remark 3.2.7. The construction of $L^2\Lambda^k$ extends to general smooth manifolds in a straightforward fashion. We use smooth charts to pull back the Lebesgue measure in each coordinate patch (see [56] for precise details), allowing us to define Lebesgue integrals on the manifold. We may then use $\langle \cdot, \cdot \rangle$ to define the required Hilbert space structure.

In practice, we often study a form $\omega \in L^2\Lambda^k(U)$ by considering smooth approximations of ω . The next result justifies such a strategy.

Theorem 3.2.8. Let $U \subseteq \mathbb{R}^n$ be open. The set of all smooth, compactly supported k -forms on U , denoted by $C_c^\infty\Lambda^k(U)$, is a dense subspace of $L^2\Lambda^k(U)$.

Proof. (Sketch) The proof for $k = 0$ is contained in [11, Corollary 4.23]. To obtain the result for general k , we just work component-wise using the result of the scalar case. □

In light of the above result, many authors define $L^2\Lambda^k(U)$ as the completion of $C_c^\infty\Lambda^k(U)$ with respect to the L^2 -norm (see for example [3]). This approach is advantageous in that it allows us to get to proving results about our function spaces a bit faster. However, since we deal with discontinuous functions and forms very frequently in numerical analysis, I feel that using the completion approach to L^2 spaces is aesthetically unsatisfying, hence why we develop spaces of rough forms via measure-theoretic first principles here.

For most applications, we are interested in studying L^2 spaces of forms on a Lipschitz domain Ω . The theory built so far in this section ensures that such spaces can be defined sensibly. In order to study the “boundary values” of a rough form defined on Ω , however, we must make sense of k -forms on $\partial\Omega$. Now, in general $\partial\Omega$ is a Lipschitz manifold. Recalling that Lipschitz manifolds do not have a tangent bundle, we see that the usual notion of a differential form on $\partial\Omega$ does not make sense. In order to coherently talk about boundary values of forms on Lipschitz domains, we need to re-define differential forms from the ground up. We therefore spend the rest of this section discussing how to build L^2 -spaces on Lipschitz manifolds, after the approach of [85]. Fortunately, this new definition of a differential form is very intuitive, and most of the technical issues are easily taken rectified using basic ideas from geometry or appeals to Rademacher’s Theorem.

Before defining any new objects, we need a helpful little lemma essentially telling us that the spaces $L^2\Lambda^k$ are not affected by Lipschitz coordinate changes.

Lemma 3.2.9. Let U, V be open subsets of \mathbb{R}^n and let $F: V \rightarrow U$ be Lipschitz. Suppose that $\omega \in L^2\Lambda^k(U)$. Write

$$\sum \omega_{i_1 \dots i_k} dx^{i_1} \wedge \dots \wedge dx^{i_k}$$

with each $\omega_{i_1 \dots i_k} \in L^2\Lambda^0(U)$. Then, the k -form

$$F^*\omega \doteq \sum (\omega_{i_1 \dots i_k} \circ F) \left(\frac{\partial F^{i_1}}{\partial y^j} dy^j \right) \wedge \dots \wedge \left(\frac{\partial F^{i_k}}{\partial y^j} dy^j \right) \quad (3.2.5)$$

is an element of $L^2\Lambda^k(V)$. We call $F^*\omega$ the **pullback** of ω by F .

Proof. By Rademacher's Theorem, all of the partial derivatives of F are measurable and bounded almost everywhere. Then, the proof is trivial. \square

Since Rademacher's Theorem also holds for open subsets of \mathbb{H}^n , we know that pullbacks by Lipschitz maps $F: V \cap \mathbb{H}^n \rightarrow U \cap \mathbb{H}^n$ also make sense. With the above notion of Lipschitz pullbacks, defining L^2 spaces on Lipschitz manifolds is not very difficult, as we see below.

Definition 3.2.10. Let Ω be a Lipschitz manifold (possibly with boundary) with Lipschitz atlas $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$. We define $V_\alpha \doteq \varphi_\alpha(U_\alpha)$. Choose a family of k -forms $\omega_\alpha \in L^2\Lambda^k(V_\alpha)$ indexed by $\alpha \in I$ such that, for all $\beta \in I$ with $U_\alpha \cap U_\beta$ nonempty,

$$\omega_\alpha = \left(\varphi_\alpha \circ \varphi_\beta^{-1}\right)^* \omega_\beta.$$

Then, the set

$$\omega = \{\omega_\alpha\}_{\alpha \in I}$$

is called an L^2 k -form on Ω . The real vector space of all L^2 k -forms on Ω is denoted by $L^2\Lambda^k(\Omega)$.

Note that the pullback in the definition of each ω_α makes sense in light of Lemma 3.2.9: by definition of a Lipschitz manifold, the transition maps are all Lipschitz homeomorphisms.

Essentially, L^2 k -forms on Ω may be thought of as compatible assignments of chart domains U_α to sums of the form

$$\sum \omega_{\alpha, i_1 \dots i_k} dx^{i_1} \wedge \dots \wedge dx^{i_k}$$

where the x^i are local coordinates in U_α and $\omega_{\alpha, i_1 \dots i_k} \in L^2\Lambda^0(V_\alpha)$. Our use of the sum notation is "formal" in the sense that we don't have a natural notion of addition or scalar multiplication for the ω_α 's; the notation is just there so that we can pretend we know how to perform these algebraic operations. The formal sum of two ω_α 's doesn't have to "equal" anything, like the sum of two real numbers, it's just an expression that uses the addition symbol. We have defined k -forms such that if Ω is smooth enough then ω (thought of as an assignment of charts to formal sums) can be identified with the pullback of some k -form by a smooth coordinate chart. The sum and wedge product of k -forms on a Lipschitz manifold are defined in the obvious way.

The definition of orientability of smooth manifolds carries over directly to the Lipschitz case:

Definition 3.2.11. Let Ω be a Lipschitz manifold with Lipschitz atlas $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$. We say that Ω is **orientable** if, for all α, β and almost every $p \in \varphi_\alpha(U_\alpha \cap U_\beta)$, we have

$$\det \left[D \left(\varphi_\alpha \circ \varphi_\beta^{-1} \right) (p) \right] > 0.$$

Of course, Rademacher's Theorem is being used implicitly in the above definition to ensure that the Jacobian makes sense.

Now that we have a notion of orientability, we can start talking about integration on Lipschitz manifolds. If $\omega \in L^2\Lambda^n(\Omega)$, then we define its integral over a chart domain U_α by

$$\int_{U_\alpha} \omega \doteq \int_{V_\alpha} \omega_\alpha.$$

We may then use a partition of unity $\{\psi_\alpha\}_{\alpha \in I}$ subordinate to the atlas $\{(U_\alpha, \varphi_\alpha)\}_{\alpha \in I}$ to define integrals over the whole of Ω :

$$\int_\Omega \omega \doteq \sum_\alpha \int_{V_\alpha} (\psi_\alpha \circ \varphi_\alpha^{-1}) \omega_\alpha.$$

By definition of ω , the integral defined above satisfies the change of variables formula.

We can define Riemannian metrics on Lipschitz manifolds through a similar construction: a Riemannian metric g on Ω is an assignment of each chart domain U_α to a smooth Riemannian metric g_α on V_α in a fashion compatible with Lipschitz coordinate changes. We can then define the Hodge star of an L^2 k -form on Ω by working chart-wise. Accordingly, we can turn $L^2\Lambda^k(\Omega)$ into an inner product space by defining

$$\langle \omega, \eta \rangle_{L^2\Lambda^k} \doteq \int_\Omega \omega \wedge \star \eta. \quad (3.2.6)$$

Theorem 3.2.12. *If Ω can be covered by finitely many chart domains, then the inner product (3.2.6) makes $L^2\Lambda^k(\Omega)$ into a Hilbert space.*

Proof. Let the finite atlas on Ω be denoted by $\{(U_i, \varphi_i)\}_{i=1}^N$, and let the partition of unity subordinate to this cover be denoted by $\{\psi_i\}_{i=1}^N$. Suppose without loss of generality that $\text{supp } \psi_i \subseteq U_i$. Next, let $\{\omega_n\}_{n=1}^\infty$ be any sequence in $L^2\Lambda^k(\Omega)$ that is Cauchy with respect to the norm induced by (3.2.6). Denote the value of ω_n on U_i by $\omega_{n,i} \in L^2\Lambda^k(V_i)$. Then, $\{\omega_{n,i}\}_{n=1}^\infty$ is Cauchy in $L^2\Lambda^k(V_i)$. Since $L^2\Lambda^k(V_i)$ is complete, there exists $\omega_i \in L^2\Lambda^k(V_i)$ such that $\omega_{n,i} \rightarrow \omega_i$. Now, in the notation of Definition 5.2, we define $\omega \in L^2\Lambda^k(\Omega)$ by

$$\omega = \{\omega_i\}_{i=1}^N.$$

We claim that $\omega_n \rightarrow \omega$ in $L^2\Lambda^k(\Omega)$. Since $\psi_i \leq 1$, we have that

$$\|\omega_n - \omega\|_{L^2\Lambda^k(\Omega)}^2 \leq \sum_{i=1}^N \|\omega_{n,i} - \omega_i\|_{L^2\Lambda^k(V_i)}^2.$$

Now, pick any $\epsilon > 0$. We may choose n sufficiently large so that, for all i ,

$$\|\omega_{n,i} - \omega_i\|_{L^2\Lambda^k(V_i)}^2 < \frac{\epsilon^2}{N}.$$

Therefore, for n large enough,

$$\|\omega_n - \omega\|_{L^2\Lambda^k(\Omega)} < \epsilon.$$

Since the Cauchy sequence $\{\omega_n\}_{n=1}^\infty$ was chosen arbitrarily, the proof is complete. \square

Corollary 3.2.13. *If Ω is a compact Lipschitz manifold (possibly with boundary), then $L^2\Lambda^k(\Omega)$ is a Hilbert space. In particular, if U is a Lipschitz domain then $L^2\Lambda^k(\bar{U})$ and $L^2\Lambda^k(\partial U)$ are Hilbert spaces.*

\square

With the previous corollary, we have seen that L^2 spaces of k -forms on the boundary of a Lipschitz domain make sense. Further, we have seen that the ‘‘correct notion’’ of an L^2 k -form on a Lipschitz manifold is what we might intuitively expect it to be, namely just a coordinate-independent formal sum of L^2 k -forms defined on each coordinate patch.

3.3 Weak Derivatives and the Spaces $H\Lambda^k$

In this section the “weak” notion of exterior differentiation for rough k -forms is introduced. We in turn see how the construction of Sobolev spaces of functions with domains in \mathbb{R}^n may be extended to differential forms on general manifolds. There are, of course, alternative approaches to such generalizations, but the one presented here is the standard approach used in finite element analysis [4, 5]. Throughout this section $\Omega \subseteq \mathbb{R}^n$ denotes a Lipschitz domain and *not a general Lipschitz manifold*.

Definition 3.3.1. We say that $\omega \in \Lambda^k$ is **weakly differentiable** if there exists $\xi \in \Lambda^{k+1}$ such that for all $\eta \in \mathcal{C}_c^\infty \Lambda^{k+1}$ we have

$$\langle \omega, \delta\eta \rangle_{L^2 \Lambda^k} = \langle \xi, \eta \rangle_{L^2 \Lambda^{k+1}}. \quad (3.3.1)$$

The rough $(k+1)$ -form ξ is called the **weak exterior derivative** of ω , denoted by $d\omega \doteq \xi$.

An elementary argument shows that if a k -form is weakly differentiable, then its weak derivative is unique almost everywhere. Further, all smooth k -forms are weakly differentiable with their weak derivatives corresponding to their usual exterior derivatives.

We now go through two examples demonstrating how to compute weak derivatives (actually, it may be more appropriate to say that the forthcoming examples show how to validate guesses of weak derivatives).

Example 3.3.2. Consider the function $f(x) = e^{-|x|}$ defined on the interval $(-1, 1)$. This function has a sharp “peak” at $x = 0$, so we know immediately that df does not exist in the usual sense. However, f does have a weak derivative. Define $\text{sgn}: (-1, 1) \rightarrow \{-1, 1\}$ by

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x \geq 0, \quad \text{and} \\ -1 & \text{if } x < 0. \end{cases}$$

We show that the weak derivative of f is given by

$$\xi(x) \, dx = -\text{sgn}(x) \, e^{-|x|} \, dx.$$

For any $\eta(x) \, dx \in \mathcal{C}_c^\infty \Lambda^0(-1, 1)$, we have

$$\begin{aligned} \langle \xi \, dx, \eta \, dx \rangle_{L^2 \Lambda^1} &= \int_{-1}^1 -\text{sgn}(x) \, e^{-|x|} \, \eta(x) \, dx \\ &= \int_{-1}^0 e^x \, \eta(x) \, dx - \int_0^1 e^{-x} \, \eta(x) \, dx. \end{aligned}$$

Integrating by parts and using the boundary condition on η , we have

$$\begin{aligned} \int_{-1}^1 -\text{sgn}(x) \, e^{-|x|} \, \eta(x) \, dx &= \eta(0) - \int_{-1}^0 e^x \, \eta'(x) \, dx - \eta(0) - \int_0^1 e^{-x} \, \eta'(x) \, dx \\ &= - \int_{-1}^1 e^{-|x|} \, \eta'(x) \, dx \\ &= \langle e^{-|x|}, \delta(\eta \, dx) \rangle_{L^2}. \end{aligned}$$

We have used the coderivative formula (2.5.6) to obtain the last equality. We conclude that $df(x) = \xi(x) dx$ in the weak sense. Now, in one dimension, we may use the Hodge star to canonically identify smooth 0-forms with smooth 1-forms, hence by the work above f is a weak solution to the ODE

$$|f(x)| = |f'(x)|. \quad (3.3.2)$$

(3.3.2) describes traveling wave solutions to the Camassa–Holm equation [15], a nonlinear PDE arising as an asymptotic limit of the Green–Naghdi model we study in Part 2.

Example 3.3.3. Let $S = (-1, 1) \times (-1, 1) \subseteq \mathbb{R}^2$. Consider the rough 1-form

$$\omega = \begin{cases} y dx + (1 + x^2) dy & \text{if } x > 0; \\ (1 + x^2) dy & \text{if } x \leq 0. \end{cases}$$

Notice that the x component of ω has a jump discontinuity along the y -axis. However, we show in this example that ω is still weakly differentiable with weak derivative given by

$$\xi = \begin{cases} (2x - 1) dx \wedge dy & \text{if } x > 0; \\ (2x) dx \wedge dy & \text{if } x \leq 0. \end{cases}$$

Indeed, pick any $\eta \in \mathcal{C}_c^\infty \Lambda^2(S)$ and write

$$\eta = \hat{\eta} dx \wedge dy.$$

Then, using the coderivative formula (2.5.6) and (2.3.1), we have

$$\star \delta \eta = - \star \star d \star \eta = - \star \star d \hat{\eta} = d \hat{\eta}.$$

Therefore,

$$\langle \omega, \delta \eta \rangle_{L^2 \Lambda^1} = - \int_{-1}^1 dy \int_{-1}^0 dx [(1 + x^2) \partial_x \hat{\eta}] + \int_{-1}^1 dy \int_0^1 dx [y \partial_y \hat{\eta} - (1 + x^2) \partial_x \hat{\eta}].$$

Upon integrating by parts and using the boundary conditions on η as in the previous example, we have that

$$\begin{aligned} \langle \omega, \delta \eta \rangle_{L^2 \Lambda^1} &= \int_{-1}^1 dy \int_{-1}^0 dx [2x \hat{\eta}] + \int_{-1}^1 dy \int_0^1 dx [(2x - 1) \hat{\eta}] \\ &= \int_S \hat{\eta} \xi \\ &= \langle \xi, \eta \rangle_{L^2 \Lambda^2}. \end{aligned}$$

Hence, $\xi = d\omega$ in the weak sense. We conclude that weak differentiability does not imply global continuity.

Both of the above examples illustrate an important property of weak differentiation: if $\omega \in \Lambda^k(\Omega)$ is weakly differentiable and piecewise smooth with respect to a “nice” partition $\Omega = \cup_i \Omega_i$, then

$$(d\omega)|_{\Omega_i} = d(\omega|_{\Omega_i}).$$

Now, we define the main family of Hilbert spaces used through this thesis. Such spaces allow us to apply powerful methods from functional analysis to help solve problems in PDE theory.

Definition 3.3.4. $H\Lambda^k \doteq \{\omega \in L^2\Lambda^k \mid d\omega \in L^2\Lambda^{k+1}\}$ is called the **Sobolev space** of k -forms on Ω .

One could define different Sobolev spaces by switching out the “2” in $L^2\Lambda^k$ for some $p \geq 1$. We could also use another notion of a derivative on Ω , say the covariant derivative induced by an affine connection, to define Sobolev spaces (this choice is by far the more popular one in geometric analysis). However, such constructions are not required for the present purposes, and so there is no harm in simply calling $H\Lambda^k$ the Sobolev space of k -forms on Ω .

Remark 3.3.5. Notice that $H\Lambda^n = L^2\Lambda^n$ trivially. Then, the Sobolev space of top forms is full of “bad” discontinuous elements. We eventually see how this seemingly burdensome fact actually helps us immensely in numerical PDEs, in that it indicates how the geometric perspective naturally accommodates the famous discontinuous Galerkin methods (defined later on in Part 1).

Theorem 3.3.6.

1) The following defines an inner product on $H\Lambda^k$:

$$\langle \omega, \eta \rangle_{H\Lambda^k} = \langle \omega, \eta \rangle_{L^2\Lambda^k} + \langle d\omega, d\eta \rangle_{L^2\Lambda^{k+1}}. \quad (3.3.3)$$

The norm associated to $\langle \cdot, \cdot \rangle_{H\Lambda^k}$ is called the **Sobolev norm** on k -forms.

2) With respect to the inner product defined above, $H\Lambda^k$ is a Hilbert space.

Proof. The proof that $\langle \cdot, \cdot \rangle_{H\Lambda^k}$ is indeed an inner product is straightforward, and we do not go through the details here. Showing that $H\Lambda^k$ is Hilbert is not much more difficult, and we essentially copy the usual proof for Sobolev spaces of scalar functions (see, for example, [11, Proposition 8.1]) over to the geometric notation.

Since $H\Lambda^k$ is an inner product space, the only thing remaining to show is that it is complete. Let $\{\omega_n\}_{n=1}^\infty$ be a Cauchy sequence in $H\Lambda^k$. Then, in particular, it must be true that $\{\omega_n\}_{n=1}^\infty$ is Cauchy in $L^2\Lambda^k$ and $\{d\omega_n\}_{n=1}^\infty$ is Cauchy in $L^2\Lambda^{k+1}$. Since $L^2\Lambda^\ell$ is complete for all $\ell = 0, \dots, n$ we know that there exist $\omega \in L^2\Lambda^k$ and $\xi \in L^2\Lambda^{k+1}$ such that $\omega_n \rightarrow \omega$ and $d\omega_n \rightarrow \xi$.

We now show that $\xi = d\omega$. To do this, pick any $\eta \in \mathcal{C}_c^\infty\Lambda^{k+1}$. Then,

$$\begin{aligned} \langle \omega, d\eta \rangle_{L^2\Lambda^k} &= \int_{\Omega} \langle \lim_{n \rightarrow \infty} \omega_n, d\eta \rangle \\ &= \lim_{n \rightarrow \infty} \int_{\Omega} \langle \omega_n, d\eta \rangle \\ &= \lim_{n \rightarrow \infty} \int_{\Omega} \langle d\omega_n, \eta \rangle \\ &= \int_{\Omega} \langle \xi, \eta \rangle \\ &= \langle \xi, \eta \rangle_{L^2\Lambda^{k+1}}. \end{aligned}$$

□

Since the exterior derivative is nilpotent, we know that $dH\Lambda^k \subseteq H\Lambda^{k+1}$.

3.4 The Trace Theorem

Many physically and mathematically important problems in PDE theory involve solving a differential equation for a continuously differentiable function u defined on a prescribed domain Ω subject to the constraint $u|_{\partial\Omega} = \tilde{u}$. Of course, since the solution u has enough regularity to be differentiated, it must be continuous. Therefore, its restriction to a subset of the domain is unambiguously defined. However, when studying PDEs with the machinery of weak solutions and Sobolev spaces, prescribing boundary values no longer make sense: $\partial\Omega$ has Lebesgue measure zero, hence a weak solution (an *equivalence class* of functions) can take arbitrary values on the boundary. This problem is rectified by proving **Trace Theorems**, as they are called in the PDE theory literature. Our discussion of traces follows a small portion of the very comprehensive presentation in Tiee’s PhD thesis [88, §1.7].

Loosely, Trace Theorems tell us that the restriction map taking continuous functions on Ω to continuous functions on $\partial\Omega$ can be extended to a bounded linear map taking Sobolev forms on Ω to L^2 forms on $\partial\Omega$ (sometimes, the extended restriction map actually takes values in a space of “distributions” containing the L^2 forms). Of course, there are many different Trace Theorems suited for different Sobolev spaces and domain regularity conditions. We only need one such theorem for our purposes, allowing us to define boundary values for elements of $H\Lambda^k$. For the sake of time, we do not prove this result (or even state it with absolute precision); a complete discussion would take us too far adrift into the ocean of Fourier analysis and distribution theory, and would likely be just a rehash of Tiee’s excellent presentation anyway.

Theorem 3.4.1. (*Trace Theorem*) *Let $\Omega \subseteq \mathbb{R}^n$ be a Lipschitz domain and let $\iota: \partial\Omega \rightarrow \Omega$ denote the inclusion map. There exists a Banach space $H^{-\frac{1}{2}}\Lambda^k(\partial\Omega) \supseteq L^2\Lambda^k(\partial\Omega)$ and a bounded linear operator*

$$\text{tr}_{\partial\Omega}: H\Lambda^k(\Omega) \rightarrow H^{-\frac{1}{2}}\Lambda^k(\partial\Omega)$$

such that, if $\omega \in C^\infty\Lambda^k(\overline{\Omega})$, then

$$\text{tr}_{\partial\Omega}\omega = \iota^*\omega.$$

For our purposes, forming $\text{tr}_{\partial\Omega}\omega$ amounts to prescribing the boundary values of $\omega \in H\Lambda^k(\Omega)$. More discussion on this theorem in the case of differential forms of arbitrary degree may be found in [4, pp. 19–20].

We can use the trace operator to define subspaces of $H\Lambda^k$ determined by boundary values.

Definition 3.4.2. $\mathring{H}\Lambda^k(\Omega) \doteq \{\omega \in H\Lambda^k(\Omega) \mid \text{tr}_{\partial\Omega}\omega = 0\}$.

The subspace defined above appears when we introduce how to **weakly formulate** PDEs in Section 6.1. For this application, we need a little lemma that we prove now.

Lemma 3.4.3. $C_c^\infty\Lambda^k(\Omega)$ is dense in $\mathring{H}\Lambda^k(\Omega)$.

Proof. This is essentially the “Sobolev version” of Theorem 3.2.8, and we likewise prove it using appeals to the scalar case. By [11, Thm. 8.12] the claim holds for $k = 0$, and we obtain the general result by working component-wise. \square

Chapter 4

Hilbert Complexes

In this chapter we introduce special sequences called **Hilbert complexes**. The study of Hilbert complexes provides a convenient framework for proving various important results in PDE theory. Among these results is the **Hodge–Helmholtz decomposition**, which generalizes the Helmholtz decomposition from vector calculus to the setting of the Sobolev spaces $H\Lambda^k$. Additionally, in Chapter 9 we see that Hilbert complexes provide novel techniques for analyzing the quality of numerical methods. Indeed, we focus on studying Hilbert complexes primarily because such complexes allow us to cleanly articulate some of the most vital properties of a PDE that should be preserved during discretization.

The general theory of Hilbert complexes was introduced in 1992 by Brüning and Lesch [14], where the authors’ intent was largely to abstract the theory of elliptic complexes on smooth manifolds. Hilbert complexes first appeared explicitly as tools in numerical PDE theory in the 2010 work of Arnold, Falk and Winther [5], though the importance of Hodge–Helmholtz decomposition in numerical analysis had been investigated a few years earlier in, for example [4, 18]. The utility of Hodge–Helmholtz decompositions in pure geometry has been understood for over eighty years; see Schwarz’ monograph [77] for a historical discussion along this vein and a detailed account of the mathematical theory.

The name “Hodge–Helmholtz decomposition” is not standard; indeed, in both the finite element literature and the geometry literature it is simply called “Hodge decomposition”. I chose to add the attribution to Helmholtz in this exposition because I anticipate that many readers are coming to this work from an applied mathematics background. While applied mathematicians are certainly familiar with Helmholtz decompositions in vector calculus, they may have never heard of the Hodge decomposition of differential forms on manifolds. By adding “Helmholtz” to the name, I stress that this decomposition is a natural generalization of a familiar technique. I suppose I also chose this naming convention because there’s a bit of Vladimir Arnol’d in me, angry that Hodge often gets all the credit for something that Helmholtz *really* discovered (albeit in the most trivial case). That little bit of Arnol’d also makes me want to call the Hodge–Helmholtz decomposition the “Hodge–Helmholtz–Morrey–Friedrichs–Kodaira–de Rham–Weyl decomposition” at least once (inspired by the joke from [7, p. 192] and the historical discussion of decomposition theorems in [77, pp. 3–4]), so now that I’ve gotten that out of the way we can get back to being serious.

In Section 4.1, we review some facts from the theory of Hilbert spaces and linear operators. In Section 4.2, we actually define Hilbert complexes and discuss transformations between them. In Section 4.3, we introduce an object called the **cohomology** of a Hilbert complex, which is related to an important problem

from vector calculus. In Section 4.4, we show how to apply the Hodge–Helmholtz decomposition to a wide class of Hilbert complexes. We also connect cohomology to a generalized version of Laplace’s equation. In Section 4.5, we verify that the “best parts” of the abstract theory, including the Hodge–Helmholtz decomposition, can be applied to a Hilbert complex derived from the spaces $L^2\Lambda^k$ and $H\Lambda^k$. Such a verification is accomplished by checking that the complex in question satisfies a special **compactness property**.

4.1 Some Facts from Functional Analysis

In this section we briefly review some relevant definitions and results from functional analysis, especially the theory of densely–defined closed operators on Hilbert spaces. This analytical material is covered beautifully in classic books such as Reed and Simon [74] and Yosida [91]. The forthcoming exposition includes a detailed guide of where readers may find more thorough discussion, as I must forgo the presentation of many proofs for the sake of brevity. Throughout this section, W denotes a Hilbert space with inner product $\langle \cdot, \cdot \rangle: W \times W \rightarrow \mathbb{R}$ unless stated otherwise. All vector spaces we deal with are assumed to be real.

First, we recall some facts about orthogonality.

Definition 4.1.1. *Let $S \subseteq W$. Then,*

$$S^\perp \doteq \{w \in W \mid \langle w, s \rangle = 0 \ \forall s \in S\}. \tag{4.1.1}$$

S^\perp is called the **orthogonal complement** of S in W .

Since inner products are positive definite, $S \cap S^\perp$ is either 0 or the empty set.

Proposition 4.1.2. *For any $S \subseteq W$, S^\perp is a closed subspace.*

Proof. Showing that S^\perp is a subspace is trivial, and S^\perp is closed since the inner product is continuous in each argument. □

Proposition 4.1.2 provides a handy algebraic characterization of closedness for subspaces: if we can show that some subspace of W arises as an orthogonal complement, then it is automatically closed.

The next theorem is probably the most frequently cited result in Hilbert complex theory, to say nothing of its applications in other areas of analysis and differential equations.

Theorem 4.1.3. (*Projection Theorem*) *Let W be a Hilbert space with $V \subseteq W$ a closed subspace. Then, $W = V \oplus V^\perp$. Further, if we write $w \in W$ as $w = v + v^\perp$ with $v \in V$ and $v^\perp \in V^\perp$, then v is the best approximation to w in V and v^\perp is the best approximation to w in V^\perp :*

$$\begin{aligned} \|w - v\| &= \inf_{\tilde{v} \in V} \|w - \tilde{v}\| \text{ and} \\ \|w - v^\perp\| &= \inf_{\tilde{v} \in V^\perp} \|w - \tilde{v}\|. \end{aligned}$$

Proof. See [74, p. 42]. □

Corollary 4.1.4. *If S is a subspace of W then $\overline{S} = S^{\perp\perp}$.*

Proof. Clearly, $S \subseteq S^{\perp\perp}$. Taking the closure of both sides and using Proposition 4.1.2, we get $\overline{S} \subseteq S^{\perp\perp}$. To show the other inclusion, pick any $\sigma \in S^{\perp\perp}$. By continuity of the inner product, the orthogonal complement of S is equal to the orthogonal complement of \overline{S} . We use the Projection Theorem to find $\overline{s} \in \overline{S}$ and $s^\perp \in S^\perp$ such that

$$\sigma = \overline{s} + s^\perp.$$

Taking the inner product of both sides with s^\perp , we see that $\langle s^\perp, s^\perp \rangle = 0$ hence $s^\perp = 0$. □

In the case where W is a space of functions, the Projection Theorem says that every function in W can be interpolated in a given closed subspace V by a unique function that is *guaranteed* to minimize the approximation error.

Now, we discuss linear operators.

Definition 4.1.5.

1) A **linear operator** T between Hilbert spaces W and V , denoted $T: W \rightarrow V$, is a linear map from a subspace $\text{dom } T \subseteq W$, called the **domain of T** , to V .

2) If $\overline{\text{dom } T} = W$, then we say that T is **densely-defined**.

3) A linear operator is said to be **closed** if the set

$$\Gamma(T) \doteq \{(w, Tw) \in W \times V \mid w \in \text{dom } T\} \tag{4.1.2}$$

is closed in $W \times V$. We call $\Gamma(T)$ the **graph of T** .

4) If $\text{dom } T = W$ and there exists $C > 0$ such that $\|Tw\| \leq C \|w\|$, then we say that T is **bounded**.

Remark 4.1.6. The notation $T: W \rightarrow V$ is, unfortunately, a little confusing, as T isn't really a map from the set W to the set V . However, such notation is standard in analysis and PDE theory.

We use the notation $\text{range } T$ ($\ker T$) to denote the range (kernel) of T as a linear map $\text{dom } T \rightarrow V$. Often, we say that a linear operator is **unbounded** if we want to emphasize that it is not bounded. Showing that boundedness is equivalent to continuity is not difficult, even in much more abstract situations [91, pp. 42–43]. Bounded operators also tend to be easier to impose nice structures on (we discuss this a bit in subsequent paragraphs), but unfortunately many operators of interest in mathematical physics are unbounded.

The next theorem shows how boundedness and closedness of an operator are related to one another.

Theorem 4.1.7. (*Closed Graph Theorem*) Let $T: W \rightarrow V$ be a linear operator between Hilbert spaces such that $\text{dom } T = W$. Then, T is bounded if and only if it is closed.

Proof. See [74, p. 83]. □

The following result is similar the Closed Graph Theorem, but just distinct enough to avoid being a corollary.

Proposition 4.1.8. *If a linear operator $T: W \rightarrow V$ is closed then $\text{dom } T$ becomes a Hilbert space when endowed with the **graph inner product***

$$\langle w_1, w_2 \rangle_{\text{dom } T} \doteq \langle w_1, w_2 \rangle_W + \langle Tw_1, Tw_2 \rangle_V.$$

Further, $T: \text{dom } T \rightarrow V$ is bounded. □

Taken together, the Closed Graph Theorem and Proposition 4.1.8 give meaning to the heuristic saying that closed operators are “almost continuous”.

Proposition 4.1.9. *The weak exterior derivative $d: L^2\Lambda^k \rightarrow L^2\Lambda^{k+1}$ is densely-defined and closed.*

Proof. First, $\text{dom } d|_{L^2\Lambda^k} = H\Lambda^k \supseteq C_c^\infty\Lambda^k$. Combining this insight with Theorem 3.2.8, we know that d is always densely-defined. To complete the proof, the reader can easily verify that d is closed directly from the definitions [88, Example 1.11.2]. □

Naturally, we are often concerned with determining if an operator is invertible. Luckily, there is a simple criterion that makes dealing with inverses of bounded operators easy.

Theorem 4.1.10. (*Inverse Mapping Theorem*) *A bounded, bijective linear operator must necessarily have a bounded inverse.*

Proof. See [74, p. 83] or [91, p. 77]. □

One of the most important constructions in operator theory is that of the **adjoint**. Loosely, if $T: W \rightarrow V$ is a differential operator, then the adjoint of T is the differential operator $T^*: V \rightarrow W$ that we obtain by “integrating T by parts”. Indeed, when we apply Hilbert complex machinery to study Sobolev spaces of differential forms, we see how integration by parts formulas familiar from differential geometry appear naturally when studying the adjoint of the exterior derivative. We now define the adjoint precisely.

Definition 4.1.11. *Suppose that $T: W \rightarrow V$ is a densely-defined linear operator between Hilbert spaces. Denote the inner product on W by $\langle \cdot, \cdot \rangle_W$ and the inner product on V by $\langle \cdot, \cdot \rangle_V$. The **adjoint** of T , denoted by T^* , is the unbounded linear operator $T^*: V \rightarrow W$ with*

$$\text{dom } T^* \doteq \{v \in V \mid \exists w \in W \text{ such that } \langle w, \tilde{w} \rangle_W = \langle v, T\tilde{w} \rangle_V \ \forall \tilde{w} \in \text{dom } T\}$$

and $T^*v \doteq w$ in the above notation.

Lemma 4.1.12. *The adjoint is indeed a function, and it is unique.* □

The next example is important enough to merit its own proposition.

Proposition 4.1.13. *Let $\Omega \subseteq \mathbb{R}^n$ be a Lipschitz domain. The weak exterior derivative $d: L^2\Lambda^k \rightarrow L^2\Lambda^{k+1}$ has as its adjoint the **weak exterior coderivative** given by*

$$\delta = (-1)^{nk+n+1} \star d\star, \tag{4.1.3}$$

defined on the domain

$$\mathring{H}^*\Lambda^k(\Omega) \doteq \{\omega \in L^2\Lambda^k(\Omega) \mid \delta\omega \in L^2\Lambda^{k-1}(\Omega), \ \text{tr}_{\partial\Omega} \star \omega = 0\}.$$

Proof. The proof of this result requires a little more distribution theory that I am prepared to present here, so I must direct the reader to [5, Theorem 4.1] and [88, Theorem 1.7.5, pp. 69–70] for a complete discussion. \square

The following theorem, when combined with Proposition 4.1.9, implies that the adjoint of the exterior derivative behaves nicely. So, the content of Proposition 4.1.13 is just that the adjoint is precisely what we would naïvely guess it to be given some experience studying smooth differential forms.

Theorem 4.1.14. *If T is a densely-defined closed linear operator between Hilbert spaces, then T^* is also a densely-defined closed linear operator. Further, $T^{**} = T$.*

Proof. The claim is proved in [74, pp. 252–253] for the special case $W = V$, and extending their arguments to full generality is straightforward. \square

Studying adjoints allows us to say a great deal about subspaces associated to a given operator, as the following results demonstrate.

Proposition 4.1.15. *Let T be a densely-defined closed operator. Then, $\ker T^* = (\text{range } T)^\perp$.*

Proof. The inclusion “ \subseteq ” follows from the definition of T^* . The other direction is similar, though we must also use that T is densely-defined. \square

Corollary 4.1.16. *Let T be a densely-defined closed operator. Then, $\ker T$ is closed.*

Proof. Using Theorem 4.1.14 and Proposition 4.1.15, we have that $\ker T = (\text{range } T^*)^\perp$ hence the kernel of T is an orthogonal complement. Applying Proposition 4.1.2 completes the argument. \square

Theorem 4.1.17. *(Closed Range Theorem) Let $T: W \rightarrow V$ be a closed, densely-defined operator. Then, $\text{range } T$ is closed in V if and only if $\text{range } T^*$ is closed in W .*

Proof. See [91, pp. 205–206]. \square

We conclude this review section by introducing some harmless abstraction that leads to high-quality notation. Given densely-defined closed operators $S: U \rightarrow V$ and $T: V \rightarrow W$ there is no a priori way to form the function composition $S \circ T$, since we cannot guarantee that $\text{range } T \subseteq \text{dom } S$. However, if S and T are bounded then their composition can indeed be defined. We therefore have a family of “objects”, namely Hilbert spaces, and a family of composable “morphisms”, namely bounded linear operators. We refer to the communal collection of objects and morphisms by the name **Hilb**. This notation allows us to replace the sentence “ W is a Hilbert space” with the symbols $W \in \text{Obj}(\mathbf{Hilb})$ (meaning that W is an object in **Hilb**). Similarly, instead of saying that “ T is a bounded linear operator between Hilbert spaces” we can say that $T \in \text{Mor}(\mathbf{Hilb})$ (meaning that T is a morphism in **Hilb**). Alternatively, if we want to emphasize the domain and range of $T: W \rightarrow \widetilde{W}$, we can say that $T \in \text{Hom}(W, \widetilde{W})$; Hom is short for “homomorphism”.

The previous paragraph implies that **Hilb** is a **category**, a collection of objects and morphisms together with a rule for associatively composing morphisms with compatible domains and ranges. Often, the objects in a category are sets with some algebraic or topological structure, the morphisms are maps “respecting” the structure of the objects, and the composition rule is just regular composition of maps. Categories

following this general template include **Hilb**, **Top** (where the objects are topological spaces and the morphisms are continuous maps), and **Vect** (where the objects are real vector spaces and the morphisms are linear maps).

The general theory of categories provides mathematicians with a suitable language for articulating fundamental relationships between mathematical and logical constructs, especially in algebraic topology. For the purposes of our exposition, however, we only use category theory as a source of convenient jargon and notation. Indeed, the new notation introduced in the previous paragraph (Obj, Mor, and Hom) comes from category theory; the meaning of this notation in the general case should be clear. More profound results and more sophisticated applications of categories can be found in [36, 61, 81].

Remark 4.1.18. *Without too much trouble, we can form categories where the objects are sets and the morphisms are so-called **partial functions** that are not defined everywhere [81, Example 1.2.3]. Essentially, the composition of partial functions cannot be usual function composition, but something a little more subtle. Accordingly, one could form a category where the objects are Hilbert spaces and the morphisms are unbounded operators, provided that one is careful about defining the composition rule. We ignore such constructions in the sequel, largely for the sake of saving time and avoiding distractions (we want the theory to work for us, not the other way around).*

4.2 Hilbert Complexes: Basic Definitions

Having completed a review of the required topics from analysis, we are finally ready to define Hilbert complexes. Developing basic definitions is the sole province of this section, with deeper constructions and theorems all relegated to the remaining three sections.

Definition 4.2.1. *Let W^1, W^2, \dots be Hilbert spaces and let $d^k: W^k \rightarrow W^{k+1}$ be densely-defined closed linear operators satisfying*

$$1) \text{ range } d^k \subseteq \text{dom } d^{k+1} \text{ and}$$

$$2) d^{k+1}d^k = 0$$

for all k . Then, we say that the data $\{(W^k, d^k)\}_k$ define a **Hilbert complex**. The operators d^k are called the **differentials** of the complex. If all the differentials are bounded, the Hilbert complex is said to be **bounded**.

Hilbert complexes are best represented by drawing a diagram:

$$\dots \xrightarrow{d^{k-2}} W^{k-1} \xrightarrow{d^{k-1}} W^k \xrightarrow{d^k} W^{k+1} \xrightarrow{d^{k+1}} \dots \quad (4.2.1)$$

In addition to viewing a Hilbert complex as a family of pairs $\{(W^k, d^k)\}_k$ or a diagram like (4.2.1), we can also view a Hilbert complex as a single pair (W, d) where

$$W \doteq \bigoplus_k W^k$$

and $d: W \rightarrow W$ is the densely-defined closed linear map given by

$$d|_{W^k} \doteq d^k.$$

Of course, $d \circ d = 0$. W is an example of a **graded vector space**. “Graded” just means that W decomposes as a direct sum of vector spaces indexed by integers. Since W is graded and each term in the direct sum is a Hilbert space, W is called a **graded Hilbert space**. Any graded Hilbert space W is itself a Hilbert space with inner product given by the sum of the inner products on each W^k .

Further, we know that $d: W^k \rightarrow W^{k+1}$. In the graded vector space jargon, this means that d is a **degree +1** linear operator: an operator on a graded Hilbert space W has degree ℓ if it maps W^k to $W^{k+\ell}$. Heuristically, a graded vector space W can be thought of like a ladder where each rung represents one of the W^k . Applying a degree +1 operator moves us up one rung, applying a degree -32 operator moves us down 32 rungs, and applying a degree 0 operator keeps us on the same rung we’re already on.

Definition 4.2.2. A *subcomplex* of a Hilbert complex (W, d) is a Hilbert complex (V, d) with

$$1) V^k \subseteq W^k \text{ and}$$

$$2) d^k V^k \subseteq V^{k+1}$$

for all k .

We note that the differentials of the subcomplex must necessarily be the differentials of the supercomplex. Given a Hilbert complex (W, d) , we can always extract a subcomplex (V, d) by taking $V^k \doteq \text{dom } d^k$ for all k . The Hilbert complex so obtained is called the **domain complex** of (W, d) . By Proposition 4.1.8, domain complexes are always bounded.

Example 4.2.3. Let $\Omega \subseteq \mathbb{R}^n$ be a Lipschitz domain. The **unbounded L^2 de Rham complex**,

$$0 \rightarrow L^2 \Lambda^0(\Omega) \xrightarrow{d} L^2 \Lambda^1(\Omega) \xrightarrow{d} \dots \xrightarrow{d} L^2 \Lambda^n(\Omega) \rightarrow 0, \quad (4.2.2)$$

is of great interest in PDE theory and geometry. Sometimes we do not bother writing down the L^2 when referring to this sequence. We remark that the map $0 \rightarrow L^2 \Lambda^0(\Omega)$ is simply the inclusion of the zero vector (the only linear map with domain 0). In Proposition 4.1.9 we showed that d is densely-defined and closed, hence the unbounded L^2 de Rham complex indeed satisfies Definition 4.2.1.

The domain complex of (4.2.2) is called the **L^2 de Rham complex**, represented by

$$0 \rightarrow H \Lambda^0(\Omega) \xrightarrow{d} H \Lambda^1(\Omega) \xrightarrow{d} \dots \xrightarrow{d} L^2 \Lambda^n(\Omega) \rightarrow 0. \quad (4.2.3)$$

Much of the analysis we undertake in the sequel concerns the L^2 de Rham complex rather than its unbounded counterpart. However, we still study the unbounded complex for the purpose of obtaining good estimates: bounding the L^2 norm of a function is often more useful than bounding its Sobolev norm.

Definition 4.2.4. A *morphism of bounded Hilbert complexes* $\Phi: (V, d) \rightarrow (\tilde{V}, \tilde{d})$ is a set of bounded linear maps $\Phi^k: V^k \rightarrow \tilde{V}^k$ such that $\tilde{d}^k \Phi^k = \Phi^{k+1} d^k$ for all k . We indicate that Φ is a morphism by writing $\Phi \in \text{Hom}((V, d), (\tilde{V}, \tilde{d}))$.

A set of bounded linear maps $\Phi^k: V^k \rightarrow \tilde{V}^k$ defines a morphism of bounded Hilbert complexes if and only if the diagram

$$\begin{array}{ccc} V^k & \xrightarrow{d^k} & V^{k+1} \\ \Phi^k \downarrow & & \downarrow \Phi^{k+1} \\ \tilde{V}^k & \xrightarrow{\tilde{d}^k} & \tilde{V}^{k+1} \end{array} \quad (4.2.4)$$

commutes for all k . From the graded vector space point of view, Φ is a degree 0 bounded operator $V \rightarrow \tilde{V}$ such that $\tilde{d}\Phi = \Phi d$. We may define a category **HilbC** where the objects are bounded Hilbert complexes, the morphisms are defined by way of Definition 4.2.4, and the composition of morphisms is usual function composition.

Remark 4.2.5. *In the previous section, we saw that there was no canonical way to define a category where the objects are Hilbert spaces, the morphisms are densely-defined operators, and the composition is usual function composition. Similarly, defining a category whose objects are arbitrary Hilbert complexes and whose morphisms satisfy Definition 4.2.4 can only be accomplished by the sweat of our brows: if (W, d) and (\tilde{W}, \tilde{d}) are arbitrary Hilbert complexes and $\Phi: W \rightarrow \tilde{W}$ is a degree 0 operator satisfying $\tilde{d}\Phi = \Phi d$ then range $\Phi^k \subseteq \text{dom } \tilde{d}^k$. So, even before we worry about how to define composition, we have to restrict our class of morphisms even beyond the specifications of Definition 4.2.4.*

At the end of the avalanche of abstract definitions that was this section, the reader may be worried that they have descended into the mire of what Arnol'd would call “unnecessary scholastic pseudoscience” [8, p. VI]. The developments above, however, have very concrete and useful manifestations in numerical analysis. Specifically, a differential equation can be interpreted as a relationship between elements of different rungs of the de Rham complex. For instance, let the function $f \in \mathcal{C}^1 \Lambda^0(\mathbb{R})$ be defined by the simple ODE

$$\star df = f.$$

This equation relates the 0-form f to its exterior derivative, a 1-form. From a Hilbert complex point of view, therefore, a numerical approximation of a PDE can be phrased as a relationship between elements of different rungs of a finite-dimensional subcomplex

$$\dots \rightarrow \Lambda_h^{k-1} \xrightarrow{d} \Lambda_h^k \xrightarrow{d} \Lambda_h^{k+1} \xrightarrow{d} \dots$$

of the de Rham complex. That is, we want to find an approximate solution to a PDE that lives inside an appropriate Λ_h^k . Heuristically, we can formalize the passage from a PDE to its discretization by building a morphism π_h projecting the de Rham complex down to the appropriate subcomplex (so that each map $\pi_h^k: H\Lambda^k \rightarrow \Lambda_h^k$ is a projection onto the subspace Λ_h^k). Such a morphism is illustrated in the commuting diagram

$$\begin{array}{ccc} H\Lambda^k & \xrightarrow{d} & H\Lambda^{k+1} \\ \pi_h^k \downarrow & & \downarrow \pi_h^{k+1} \\ \Lambda_h^k & \xrightarrow{d} & \Lambda_h^{k+1} \end{array} \quad (4.2.5)$$

The advantage of discretizing a PDE using a morphism is that, since morphisms must commute with differentials, it does not matter whether we differentiate first or discretize first. For example, if a 2-form $\mathbf{B} \in H\Lambda^2(\mathbb{R}^3)$ satisfies $d\mathbf{B} = 0$ (using vector proxies, this means that $\nabla \cdot \mathbf{B} = 0$), then our approximation $\pi_h \mathbf{B}$ to \mathbf{B} in Λ_h^2 satisfies $d\pi_h \mathbf{B} = 0$ as well. Physically, such a \mathbf{B} could represent a magnetic field, in turn implying that an approximation of \mathbf{B} built with the Hilbert complex structure in mind is guaranteed to still satisfy Gauss' Law. So, by paying attention to certain mathematical properties of a problem, we can develop numerical methods that provably respect certain physical properties of that problem.

4.3 Cohomology

In this section we introduce an important graded vector space associated to a given Hilbert complex called the **cohomology**. In applications to the L^2 de Rham complex, studying cohomology tells us whether the PDE $d\eta = \omega$ has a solution $\eta \in H\Lambda^{k-1}$ for any given $\omega \in \ker d|_{H\Lambda^k}$. This PDE is of paramount importance in vector calculus: once we switch to using vector proxies, being able to find a solution for arbitrary ω amounts to proving that all curl-free vector fields are gradients of a scalar field. Of course, all students of advanced calculus learn that every curl-free field on a simply connected domain in \mathbb{R}^3 must be a gradient, though this is far from being universally true for vector fields on arbitrary manifolds. We return to the relationship between the cohomology of Hilbert complexes and topology many times throughout Part 1.

Definition 4.3.1. *Consider a Hilbert complex*

$$\dots \xrightarrow{d^{k-2}} W^{k-1} \xrightarrow{d^{k-1}} W^k \xrightarrow{d^k} W^{k+1} \xrightarrow{d^{k+1}} \dots \quad (4.3.1)$$

1) The k^{th} **cohomology space** of the complex is

$$H^k \doteq \ker d^k / \text{range } d^{k-1}. \quad (4.3.2)$$

2) The complex is said to be **exact** if $H^k = 0$ for all k .

3) The **cohomology** of a Hilbert complex (W, d) is the graded vector space

$$H^\bullet \doteq \bigoplus_k H^k. \quad (4.3.3)$$

Returning to the question of solving $d\eta = \omega$, we see that a solution $\eta \in H\Lambda^{k-1}$ exists (for all $\omega \in \ker d|_{H\Lambda^k}$) if and only if $H^k = 0$. As the dimension of H^k gets bigger, we can find more and more linearly independent $\omega \in \ker d \subseteq H\Lambda^k$ for which the PDE has no solutions. We study the relationship between PDEs and cohomology more in the next section.

Example 4.3.2. Let $\Omega \subseteq \mathbb{R}^3$ be a simply connected Lipschitz domain. The **extended** L^2 de Rham complex

$$0 \rightarrow \mathbb{R} \rightarrow H\Lambda^0(\Omega) \xrightarrow{d} H\Lambda^1(\Omega) \xrightarrow{d} H\Lambda^2(\Omega) \xrightarrow{d} H\Lambda^3(\Omega) \rightarrow 0$$

is exact (the proof of this fact is trivial using a bit of algebraic topology). Intuitively, we might expect that at least the third and fourth cohomologies vanish because all curl-free smooth vector fields on Ω are gradients (of course, the elements of the Hilbert spaces here are not necessarily smooth, so this intuitive justification is by no means a rigorous proof).

Note that we must squeeze the \mathbb{R} in between the first two terms of the usual L^2 de Rham complex to guarantee exactness. If we had not done this we would have $H^1 = \mathbb{R}$ because the only scalar fields on a simply connected domain with vanishing derivative are constants, and the image of 0 in $H\Lambda^0$ is just 0. We often abuse terminology and say that the L^2 de Rham complex over a simply connected domain is exact, but the reader should be aware we really mean that it is only exact once we pass to the extended complex.

We might naturally expect that Hilbert complexes related by a morphism have cohomologies that are also, in some sense, related. Indeed, morphisms in **HilbC** always give rise to maps between cohomology spaces.

Definition 4.3.3. Suppose that $(V, d), (\tilde{V}, \tilde{d}) \in \text{Obj}(\mathbf{HilbC})$ with cohomologies denoted by H^\bullet and \tilde{H}^\bullet , respectively. Let $\Phi \in \text{Hom}((V, d), (\tilde{V}, \tilde{d}))$. Then, define a map $H\Phi: H^\bullet \rightarrow \tilde{H}^\bullet$ by

$$(H\Phi)^k([v]) = [\Phi^k(v)] \quad \forall v \in V^k. \quad (4.3.4)$$

We call $H\Phi$ the *induced map on cohomology*.

Commutativity of the diagram in (4.2.4) ensures that the induced map on cohomology is well-defined. Additionally, the map induced by the identity morphism on (V, d) is the identity map on H^\bullet . Eventually, the construction above is applied to show that the discretized versions of the L^2 de Rham complex discussed in the previous section somehow “mimic” cohomology as long as some very modest estimates are satisfied.

Proposition 4.3.4. Let $\Phi, \Psi \in \text{Mor}(\mathbf{HilbC})$. If the composition $\Psi\Phi$ is defined, then $H(\Psi\Phi) = H\Psi H\Phi$. □

Let $\mathbf{Vect}^{\mathbb{Z}}$ denote the category where the objects are graded vector spaces, the morphisms are linear maps, and the composition is usual function composition. We have seen that the basic objects of cohomology theory arise from a sort of “map of categories” $H: \mathbf{HilbC} \rightarrow \mathbf{Vect}^{\mathbb{Z}}$ taking a complex (V, d) to its cohomology H^\bullet and a morphism Φ to the induced map $H\Phi$. Formally, H is called a **functor**. More specifically, Proposition 4.3.4 tells us that H is a **covariant** functor, so called because it distributes through composition of morphisms.

We can connect seemingly disjoint disciplines by building a functor between the main categories of interest in each discipline. For example, the cohomology functor $H: \mathbf{HilbC} \rightarrow \mathbf{Vect}^{\mathbb{Z}}$ allows us answer the *analytic* question “does $d\eta = \omega$ have a solution $\eta \in H\Lambda^{k-1}$ for all $\omega \in \ker d|_{H\Lambda^k}$?” by answering the *algebraic* question “is the vector space H^k trivial?” Provided that computing the dimension of H^k is easy (in Chapter 5, we show that such calculations are indeed not difficult, at least for many physical applications), we have therefore simplified the problem in our original category \mathbf{HilbC} by moving over to the category $\mathbf{Vect}^{\mathbb{Z}}$. This example illustrates how thinking in terms of categories and functors helps us problem-solve: we take a hard problem in a category \mathcal{A} and try to reduce it to a simple problem in another category \mathcal{B} by constructing a functor $F: \mathcal{A} \rightarrow \mathcal{B}$. The systematic functorial simplification of problems in the category **Top** is the main goal of algebraic topology, introduced in the next chapter.

4.4 Harmonic Forms and the Hodge–Helmholtz Decomposition

In this section we concretely connect Hilbert complexes to vector calculus by proving that an analogue of the Helmholtz decomposition naturally emerges in the abstract setting of Hilbert complexes. We also discuss the relationship between this generalized decomposition and a generalized version of Laplace’s equation. Before diving into substantial material, however, we must introduce a definition that narrows down a nice class of Hilbert complexes for further analysis.

Definition 4.4.1. A Hilbert complex (W, d) is said to be **closed** if $\text{range } d^{k-1}$ is closed in W^k for all k .

Throughout this section we only study closed Hilbert complexes. The reason we prefer to deal with closed complexes is twofold: it generally leads to tidier theorems, and it is also usually achievable in practice. In the next section, we prove that the L^2 de Rham complex is closed, though for the time being we just assume this to be true.

We motivate the next few definitions by way of an interesting example. Consider the L^2 de Rham complex with functions defined over a Lipschitz domain in \mathbb{R}^n . In the previous section, we saw how cohomology tells us whether the PDE $d\eta = \omega$ can always be solved for η given $\omega \in \ker d$. There is, in fact, another “PDE-centric” approach to investigating L^2 de Rham cohomology. To see this, start by using the Projection Theorem to write

$$\ker d = \text{range } d \oplus \left(\ker d \cap (\text{range } d)^\perp \right) \subseteq H\Lambda^k.$$

In light of the above decomposition, if $H^k \neq 0$ then there is some $\omega \in \ker d \cap (\text{range } d)^\perp$. Therefore, for every $\xi \in H\Lambda^{k-1}$ we have

$$0 = \langle \omega, d\xi \rangle = \langle \delta\omega, \xi \rangle$$

hence $\delta\omega = 0$. An appeal to Proposition 4.1.15 shows that $\delta\omega = 0$ implies $\omega \in (\text{range } d)^\perp$. We therefore see that finding if a complex has non-trivial cohomology is equivalent to finding non-trivial solutions to the system

$$d\omega = 0, \tag{4.4.1a}$$

$$\delta\omega = 0. \tag{4.4.1b}$$

On passing to vector proxies, solving (4.4.1) amounts to finding a nonzero (and square-integrable) vector field that is both curl-free and divergence-free. Recall from advanced calculus that no such smooth nonzero vector fields exist on simply connected domains in \mathbb{R}^3 (provided we prescribe vanishing normal component boundary conditions, as warranted by the restrictions on the domain of the coderivative).

We would now like to generalize the maxim “solve (4.4.1) to find out about cohomology” to arbitrary (closed) Hilbert complexes. To this end, we make a useful definition.

Definition 4.4.2. *The space of **harmonic k -forms** is*

$$\mathfrak{H}^k \doteq \ker d^k \cap (\text{range } d^{k-1})^\perp. \tag{4.4.2}$$

Our use of the word “harmonic” is explained over the next few pages. Applying Proposition 4.1.15, we see that $(\text{range } d^{k-1})^\perp = \ker (d^{k-1})^*$. Therefore, elements of \mathfrak{H}^k are precisely the solutions of

$$d^k w = 0, \tag{4.4.3a}$$

$$(d^{k-1})^* w = 0, \tag{4.4.3b}$$

which is clearly the generalization of (4.4.1).

Interestingly, finding all harmonic forms tells us *everything* about cohomology.

Proposition 4.4.3. (*Abstract Hodge Theorem*) *For closed Hilbert complexes, $\mathfrak{H}^k \simeq H^k$.*

Proof. By the Projection Theorem,

$$\ker d^k = \text{range } d^{k-1} \oplus \left(\ker d^k \cap (\text{range } d^{k-1})^\perp \right) = \text{range } d^{k-1} \oplus \mathfrak{H}^k. \tag{4.4.4}$$

Now, the Second Isomorphism Theorem says that if A and B are vector spaces then $(A \oplus B)/A$ is isomorphic to B [31, p. 349]. Applying this isomorphism theorem to (4.4.4) completes the proof. \square

Example 4.4.4. By Example 4.3.2 and Proposition 4.4.3, $\mathfrak{H}^k = 0 \forall k$ for the (extended) L^2 de Rham complex on a simply connected domain in \mathbb{R}^3 .

Now, we switch gears and discuss the Hodge–Helmholtz decomposition of Hilbert complexes. First, recall the classical Helmholtz decomposition theorem: given any sufficiently regular square-integrable vector field \mathbf{u} on \mathbb{R}^3 , there exists a scalar field ϕ and a vector field \mathbf{A} such that

$$\mathbf{u} = \nabla\phi + \nabla \times \mathbf{A}.$$

While ϕ and \mathbf{A} are certainly not unique, the vectors $\mathbf{u}_1 = \nabla\phi$ and $\mathbf{u}_2 = \nabla \times \mathbf{A}$ are. Recall that $\nabla \times \mathbf{u}_1 = 0$ and $\nabla \cdot \mathbf{u}_2 = 0$ so, qualitatively, Helmholtz decomposition implies that every nice vector field can be split into a piece that does not locally “swirl around” and a piece that does not “spread out”. This insight has applications in oceanography, for example: large-scale steady flows are characterized by having a (vertically averaged) velocity field with zero divergence, so taking the divergence of whatever velocity field we’re working with isolates perturbations to the steady solution. The Helmholtz decomposition is also of paramount importance in electromagnetic theory.

The utility of Helmholtz decomposition in mathematical physics naturally leads to the question of whether or not a similar result holds for vector fields, or more properly differential forms, defined on arbitrary domains. This question is more than a geometer’s whim, as non-trivial topologies appear frequently in physical applications. For example, plasma physicists investigating tokamak fusion technology must study magnetohydrodynamic flows within a torus. Additionally, an oceanographer interested in studying flows around islands would be working with vector fields defined over a domain with “holes”. The question of how to extend Helmholtz decomposition into the realm of arbitrary manifolds was an important problem in geometric analysis, drawing the attention of de Rham, Hodge, Weyl, Kodaira, Friedrichs, and Morrey (again, I refer the reader to the historical discussion in [77, pp. 4–5]). Their work indicates that we must in general add a *third* term to the Helmholtz decomposition, representing the presence of “holes” in the domain.

In the abstract Hilbert complex setting, the aforementioned third term in the Hodge–Helmholtz decomposition instead represents the non-triviality of the cohomology spaces. However, in the next chapter we discuss de Rham’s Theorem, which articulates how the cohomology of the L^2 de Rham complex incorporates topological information about the domain. With de Rham’s Theorem in mind, there is no distinction between considering the third term as a contribution from topological non-triviality or cohomological non-triviality.

Having motivated the Hodge–Helmholtz decomposition physically and hinted at its attractive connection to topology, we finally come to stating and proving it.

Theorem 4.4.5. (*Hodge–Helmholtz Decomposition*) *If (W, d) is a closed Hilbert complex then each W^k admits an orthogonal decomposition*

$$W^k = \text{range } d^{k-1} \oplus \mathfrak{H}^k \oplus \text{range } (d^k)^*. \tag{4.4.5}$$

Proof. By the Projection Theorem,

$$W^k = \ker d^k \oplus (\ker d^k)^\perp.$$

Then, apply the orthogonal decomposition of $\ker d^k$ used in the proof of Proposition 4.4.3 to obtain

$$W^k = \text{range } d^{k-1} \oplus \mathfrak{H}^k \oplus (\ker d^k)^\perp.$$

Combining Corollary 4.1.4, Theorem 4.1.14, and Proposition 4.1.15 (note that Corollary 4.1.4 is applicable only because (W, d) is closed), we see

$$(\ker d^k)^\perp = \overline{\text{range } (d^k)^*}.$$

Applying Theorem 4.1.17 completes the proof. \square

Remark 4.4.6. *Proving Hodge–Helmholtz decomposition just amounts to applying the Projection Theorem twice; we have a very powerful, far-reaching result whose proof is a trivial corollary of a foundational theorem from Hilbert space theory. Showing that this general result actually applies to Hilbert complexes of practical interest, however, can be difficult, as we see in the next section.*

A quick check with vector proxies shows that the Hodge–Helmholtz decomposition indeed reduces to the regular Helmholtz decomposition when the domain is \mathbb{R}^3 , since there are no nonzero harmonic forms in this case.

Now, we explain our use of the name “harmonic” in referring to \mathfrak{H}^k . Our goal is to reduce the system (4.4.3) to a single equation written in terms of the **abstract Hodge Laplacian**,

$$\Delta^k = d^{k-1}(d^{k-1})^* + (d^k)^*d^k. \quad (4.4.6)$$

The abstract Hodge Laplacian is an unbounded operator $W^k \rightarrow W^k$ with domain

$$\text{dom } \Delta^k = \{u \in \text{dom } d^k \cap \text{dom } (d^{k-1})^* \mid (d^{k-1})^*u \in \text{dom } d^{k-1}, d^k u \in \text{dom } (d^k)^*\}.$$

In the case of the L^2 de Rham complex, we refer to Δ^k as the **Hodge Laplacian**. For the L^2 de Rham complex on \mathbb{R}^n , the Hodge Laplacian on scalar functions reduces to the familiar Laplacian operator up to a factor of -1 . We can easily verify this claim explicitly in the case $n = 2$ (the general case just requires a little extra time). Letting (x, y) denote Cartesian coordinates on \mathbb{R}^2 , we have

$$\begin{aligned} \Delta^0 f &= \delta d f \\ &= -(\star d)^2 f \\ &= -\star d \star (\partial_x f \, dx + \partial_y f \, dy) \\ &= -\star d(-\partial_y f \, dx + \partial_x f \, dy) \\ &= -(\partial_x^2 + \partial_y^2) f. \end{aligned}$$

So, in a natural special case, the abstract Hodge Laplacian reduces to minus one times the most famous second-order differential operator in mathematics.

Remark 4.4.7. *Since the domain of the coderivative δ only includes k -forms ω such that*

$$\text{tr}_{\partial\Omega} \star \omega = 0 \quad \text{and} \quad \text{tr}_{\partial\Omega} \star d\omega = 0,$$

finding the kernel of Δ^k is a problem that includes boundary conditions. Therefore, determining $\ker \Delta^0$ reduces to solving the scalar Laplace equation with a homogeneous Neumann boundary condition (hence all solutions are constant on each connected component of Ω). Similarly, finding $\ker \Delta^n$ amounts to solving the scalar Laplace equation with a homogeneous Dirichlet boundary condition.

We now show that $\ker \Delta^k = \mathfrak{H}^k$, using a trick from [60, Prop. 2.14, p. 120]. Considering the above special case, such a result explains why we call elements of \mathfrak{H}^k “harmonic”. Clearly, any $u \in W^k$ satisfying (4.4.3) belongs to $\ker \Delta^k$. Now, suppose that $\Delta^k u = 0$. Then, because δ and d are adjoints of one another,

$$0 = \langle \Delta u, u \rangle = \langle \delta u, \delta u \rangle + \langle du, du \rangle.$$

Since each term on the right-hand side is positive and the inner product is positive definite, we must have that $\delta u = 0$ and $du = 0$. We summarize our work with a proposition:

Proposition 4.4.8. $\ker \Delta^k = \mathfrak{H}^k$.

□

In review, we have explored two equivalent ways of understanding Hilbert complex cohomology: solving the system (4.4.3) and finding all linearly independent solutions of $\Delta^k u = 0$. When working with scalar functions on subsets of \mathbb{R}^n , these approaches amount to solving Laplace’s equation. Additionally, we have seen that all “rungs” of a closed Hilbert complex admit a Hodge–Helmholtz decomposition. Therefore, we can apply techniques from vector calculus to study L^2 and Sobolev spaces of differential forms. Later on, we construct discretizations of the de Rham complex that “preserve” the Hodge–Helmholtz decomposition, allowing us to maintain the use of this powerful analytical tool when studying computer simulations.

4.5 The Compactness Property

In the previous section, we proved many nice results for closed Hilbert complexes, chief among them the Hodge–Helmholtz decomposition. Determining closure in practice may be accomplished by way of the **compactness property**. The purpose of this section is to introduce the compactness property and investigate some of its consequences. Once we have defined the compactness property in the abstract Hilbert complex setting, we present a theorem of R. Picard [68] guaranteeing that the L^2 de Rham complex indeed has this property. Following this, we may show that the unbounded L^2 de Rham complex is closed. Throughout this section let (W, d) denote an arbitrary Hilbert complex.

Definition 4.5.1. *Let X and Y be Hilbert spaces. An operator $T: X \rightarrow Y$ is said to be **compact** if T maps bounded subsets of X to subsets of Y with compact closure.*

Some authors refer to sets with compact closure as **relatively compact** or **precompact**.

Definition 4.5.2. *Let (W, d) be a Hilbert complex and define a subspace $A \subseteq W^k$ by*

$$A \doteq \text{dom } d^k \cap \text{dom } (d^{k-1})^*.$$

Let $\langle \cdot, \cdot \rangle$ denote the inner product on W^k . We give A the structure of a Hilbert space by equipping it with the inner product

$$\langle u, v \rangle_A \doteq 2\langle u, v \rangle + \langle d^k u, d^k v \rangle + \langle (d^{k-1})^* u, (d^{k-1})^* v \rangle.$$

*We say that (W, d) has the **compactness property** if, for all k , the inclusion $\iota: A \hookrightarrow W^k$ is a compact operator.*

For future reference, we explicitly write out the norm on A induced by the inner product. If $\| \cdot \|$ is the norm on W^k , then

$$\|u\|_A^2 \doteq 2\|u\|^2 + \|d^k u\|^2 + \|(d^{k-1})^* u\|^2.$$

Remark 4.5.3. *In the above definition, we stated that A had the structure of a Hilbert space, but we did not prove its completeness. The proof is not difficult, however. If a sequence $(u_n)_n$ in A is Cauchy with respect to $\|\cdot\|_A$ then it is also Cauchy with respect to $\|\cdot\|_{\text{dom } d^k}$ and $\|\cdot\|_{\text{dom } (d^{k-1})^*}$. Since $\text{dom } d^k$ and $\text{dom } (d^{k-1})^*$ are complete, $(u_n)_n$ converges to a limit in each of these spaces. Further, both of these limits must be equal (since the graph norms on the domains bound the W^k norm). This proves that $(u_n)_n$ achieves a limit in A , hence A with the inner product $\langle \cdot, \cdot \rangle_A$ is Hilbert.*

Remark 4.5.4. *In [5] the authors require the subspace A to be dense in order for the compactness property to hold, but for our purposes density is a superfluous requirement.*

The compactness property immediately implies that the spaces \mathfrak{H}^k are especially well-behaved.

Proposition 4.5.5. *If a Hilbert complex has the compactness property, then $\dim \mathfrak{H}^k < \infty$ for all k .*

Proof. Clearly, $\mathfrak{H}^k \subseteq A$. Hence, we can view \mathfrak{H}^k as a normed space with two different norms, $\|\cdot\|$ and $\|\cdot\|_A$. If $q \in \mathfrak{H}^k$ then

$$\|q\|_A = \sqrt{2} \|q\|. \quad (4.5.1)$$

By (4.5.1) we know that

$$\iota|_{\mathfrak{H}^k} \left(\overline{B_0(\sqrt{2})} \right) = \overline{B_0(1)} \subseteq \mathfrak{H}^k,$$

where the balls are taken in the appropriate norms. Since (W, d) has the compactness property, $\overline{B_0(1)}$ is compact in $\|\cdot\|$. A corollary of the Riesz Lemma [91, §3.2, pp. 84–85] tells us that the unit ball in a normed space is compact if and only if the space is finite-dimensional. Therefore, $\dim \mathfrak{H}^k < \infty$. \square

Now, we switch out of the abstract Hilbert complex setting and focus on the L^2 de Rham complex for the remainder of this section. We begin by presenting a key result of R. Picard. For a theorem of this difficulty, I would usually outsource the proof to the primary source, but the basic thrust of Picard’s argument is so intuitive that I feel we must at least go over the big ideas.

Theorem 4.5.6. *(R. Picard [68]) Let $\Omega \subseteq \mathbb{R}^n$ be a Lipschitz domain. Then, the unbounded L^2 de Rham complex*

$$0 \rightarrow L^2\Lambda^0(\Omega) \xrightarrow{d} L^2\Lambda^1(\Omega) \xrightarrow{d} \dots$$

has the compactness property: for all k , $H\Lambda^k(\Omega) \cap \mathring{H}^\Lambda^k(\Omega)$ is a compactly included subspace of $L^2\Lambda^k(\Omega)$.*

Proof. (Sketch) Before we begin, notice that certain techniques from geometric analysis are not applicable in this situation. In particular, the forms in $H\Lambda^k$ are not smooth enough for Gaffney’s Inequality to hold [77, §2.1, pp. 59–66], and the domain Ω is not smooth enough for the Kondrachov–Rellich Theorem to hold [11, Thm. 9.16, p. 285]. In particular, even though Lipschitz domains are smooth manifolds (because they are open subsets of \mathbb{R}^n), their boundaries lack the regularity required to apply the Kondrachov–Rellich Theorem. There are extensions of the Kondrachov–Rellich Theorem to “rougher” domains [90, §I.7], but unfortunately even such generalizations are not strong enough to prove the claim.

The main strategy of Picard’s proof is to regularize the domain so that the aforementioned classical methods are usable. We start by reducing to the case where Ω is open, relatively compact, and Lipschitz-homeomorphic to the unit ball $B \doteq B_0(1) \subseteq \mathbb{R}^n$. By a simple partition of unity argument, we can always

reduce to this simple case through shrinking the chart domains on Ω if need be. Denote the Lipschitz homeomorphism mapping Ω to the open ball by $\varphi: B \rightarrow \Omega$. We want to show that the pullback

$$\varphi^*: L^2\Lambda^k(\Omega) \rightarrow L^2\Lambda^k(B) \quad (4.5.2)$$

(interpreted in the sense of Lemma 3.2.9) is an isomorphism. Now, φ^* is clearly a linear bijection, so by the Inverse Mapping Theorem (Theorem 4.1.10) we need only show that φ^* is bounded. Let x^1, \dots, x^n denote restricted Cartesian coordinates on B . Then, using multi-index notation, we have

$$\varphi^*\omega = (\omega_I \circ \varphi) \frac{\partial \varphi^I}{\partial x^J} dx^J. \quad (4.5.3)$$

Since the partial derivatives are bounded by Rademacher's Theorem (Theorem 3.1.5), we immediately have that φ^* is a bounded operator and therefore an isomorphism. Further, since pullbacks commute with the exterior derivative we have that φ^* restricts to an isomorphism

$$\varphi^*: H\Lambda^k(\Omega) \rightarrow H\Lambda^k(B). \quad (4.5.4)$$

Unfortunately, φ^* does not quite restrict to a homeomorphism

$$\varphi^*: H\Lambda^k(\Omega) \cap \mathring{H}^*\Lambda^k(\Omega) \rightarrow H\Lambda^k(B) \cap \mathring{H}^*\Lambda^k(B),$$

but it does get "close enough". In particular, there exists a symmetric, bounded, positive-definite linear operator $\epsilon: L^2\Lambda^k(B) \rightarrow L^2\Lambda^k(B)$ such that

$$\varphi^*: H\Lambda^k(\Omega) \cap \mathring{H}^*\Lambda^k(\Omega) \rightarrow H\Lambda^k(B) \cap \epsilon^{-1}\left(\mathring{H}^*\Lambda^k(B)\right).$$

is a homeomorphism.

Now, we can use classical methods (Gaffney, Kondrachov–Rellich) to establish compactness of the embedding

$$H\Lambda^k(B) \cap \mathring{H}^*\Lambda^k(B) \hookrightarrow L^2\Lambda^k(B).$$

Using [68, Lemma 2], however, we know that

$$H\Lambda^k(B) \cap \epsilon^{-1}\left(\mathring{H}^*\Lambda^k(B)\right) \hookrightarrow L^2\Lambda^k(B)$$

is also compact. Since φ^* is bounded, the proof is complete. \square

Theorem 4.5.7. (*Strong Poincaré Inequality*, [4, Thm. 2.2, p. 23]) *There exists a constant c such that, for all $\omega \in H\Lambda^k(\Omega) \cap \mathring{H}^*\Lambda^k(\Omega) \cap (\mathfrak{H}^k)^\perp$,*

$$\|\omega\| \leq c(\|d\omega\| + \|\delta\omega\|).$$

Proof. (Sketch) The proof presented in [4] is hard to improve on, and we only review the main steps of the argument for the sake of completeness. If the estimate in the claim does not hold, we can find a sequence $(\omega_n)_n$ in $H\Lambda^k(\Omega) \cap \mathring{H}^*\Lambda^k(\Omega) \cap (\mathfrak{H}^k)^\perp$, normalized so that all ω_n lie on the unit sphere, such that $(d\omega_n)_n$ and $(\delta\omega_n)_n$ both tend to 0 as $n \rightarrow \infty$. By compactness of the embedding in Theorem 4.5.6 we know $(\omega_n)_n$ has a subsequence that converges to a (necessarily harmonic) limit ω in the L^2 norm. However, $(\mathfrak{H}^k)^\perp$ is closed hence $\omega \in (\mathfrak{H}^k)^\perp$ as well. Therefore, $\omega = 0$, contradicting the assumption that all ω_n live on the unit sphere. \square

Corollary 4.5.8. (*Weak Poincaré Inequality*) For all $\omega \in (\ker d)^\perp \cap H\Lambda^k(\Omega)$,

$$\|\omega\| \leq c \|d\omega\|.$$

Proof. Immediately, we know that $\omega \in (\mathfrak{H}^k)^\perp$ because $\mathfrak{H}^k \subseteq \ker d$. Observe now that $(\ker d)^\perp = \overline{\text{range } \delta}$ by Proposition 4.1.15 and Corollary 4.1.4. Since $\delta \circ \delta = 0$ and $\ker \delta$ is closed, $\overline{\text{range } \delta} \subseteq \ker \delta$. We have shown that $(\ker d)^\perp \subseteq \ker \delta$. Therefore, $\omega \in \text{dom } \delta = \mathring{H}^* \Lambda^k(\Omega)$. All in all we know that

$$\omega \in H\Lambda^k(\Omega) \cap \mathring{H}^* \Lambda^k(\Omega) \cap (\mathfrak{H}^k)^\perp,$$

so the Strong Poincaré Inequality may be applied to finish the proof. \square

Corollary 4.5.9. [4, Thm. 2.3, p. 23] *The unbounded L^2 de Rham complex is closed.*

Proof. Choose any $\xi \in L^2 \Lambda^k$ such that there exists $(\omega_n)_n \in H\Lambda^{k-1}$ with $d\omega_n \rightarrow \xi$ in the L^2 norm; that is, $\xi \in \overline{\text{range } d}$. Without loss of generality we may choose $\omega_n \in (\ker d)^\perp$ if ξ is nonzero. Applying the Weak Poincaré Inequality, we see that for all $n, m \in \mathbb{N}$

$$\|\omega_n - \omega_m\| \leq c \|d\omega_n - d\omega_m\|.$$

Since $(d\omega_n)_n$ converges and $L^2 \Lambda^{k-1}$ is complete, the above estimate tells us there exists $\omega \in L^2 \Lambda^{k-1}$ such that $\omega_n \rightarrow \omega$ in the L^2 norm. Next, pick any $\eta \in C_c^\infty \Lambda^k$. We have

$$\langle \omega_n, \delta\eta \rangle = \langle d\omega_n, \eta \rangle$$

by definition of the weak derivative. Taking the limit as n goes to infinity,

$$\langle \omega, \delta\eta \rangle = \langle \xi, \eta \rangle$$

hence $d\omega = \xi$. Therefore, $\text{range } d$ is closed. \square

So, all of the nice results from the previous section (Hodge–Helmholtz decomposition, Hodge Theorem) apply to the unbounded L^2 de Rham complex, and therefore to the L^2 de Rham complex as well. Additionally, we know that the spaces of harmonic forms for the L^2 de Rham complex are finite–dimensional, through successive application of Theorem 4.5.6 and Proposition 4.5.5. The latter result allows us to prove some very interesting corollaries when we discuss the approximation of Hilbert complexes in later chapters. Specifically, we prove in Chapter 9 that building a suitable discretizing projection morphism (see the discussion at the end of Section 4.2) yields an isomorphism between the cohomology of the de Rham complex and its finite–dimensional “approximation”. A crucial step in this proof involves using the finite–dimensionality of the spaces \mathfrak{H}^k .

Remark 4.5.10. *All results on the L^2 de Rham complex presented in this section require the spatial domain Ω to be bounded. While boundedness of the domain may be taken for granted in numerical analysis, unbounded domains are a common sight in other applications (scattering theory, for instance).*

Chapter 5

Triangulations and Topology

In this chapter, we study some basic aspects of algebraic topology, especially the relation between topology and differential calculus. The purpose in spending a chapter on algebraic topology, when our ultimate goal is understanding the numerical treatment of PDEs, is to provide motivation for why one would even think of the ideas underlying FEEC in the first place. That is, algebraic topology is intimately related to PDEs by way of de Rham’s Theorem (a result hinted at in Chapter 4). We may try, therefore, to obtain information about certain PDEs, such as the Laplace equation on k -forms, by answering topological questions. Since there are such deep connections between topology and PDEs, we expect that developing numerical methods somehow reflecting those connections is well-advised.

Algebraic topology is about understanding a topological space X by associating to it some algebraic data $F(X)$. This data could be, for example, an object in **Vect**, the category of real vector spaces. The algebraic data should be simpler to work with than the topological space itself so that, by proving some simple result about $F(X)$ (for example, that its dimension as a real vector space is nonzero), we can “pull back” and prove a more difficult result about X (for example, that not every closed path in X can be shrunk down to a point). We also try to make sure that $F(X)$ is an **invariant** of X ; specifically, if Y is homeomorphic to X then $F(X)$ should be “isomorphic” to $F(Y)$. If $F(X)$ is a vector space, for example, then in categorical language the above construction amounts to building a functor $F: \mathbf{Top} \rightarrow \mathbf{Vect}$. So, although algebraic topology is famous for its abstractness and technical difficulty, its central goal is very easy to comprehend.

Including a proper discussion of the techniques of algebraic topology, even to the level required to just barely prove the theorems required here, would double the length of Part 1. Accordingly, the purpose of this chapter is to introduce the subject gently and, often, qualitatively. I focus on providing plenty of simple, hands-on examples to illustrate the basic intuition underlying the abstract constructions that dominate algebraic topology. Additionally, I emphasize certain aspects of the historical development of algebraic topology, as I believe that seeing how mathematical concepts chronologically built off of one another helps one understand why modern perspectives are the way they are. Another reason to periodically mention the history of topology in our presentation is that the interaction between topology and numerical PDE theory is very close in spirit to the way early topologists, including Poincaré and de Rham, thought about their subject. Most of the historical citations presented here are taken from Dieudonné’s comprehensive history [27]. I also frequently use material from Stillwell’s annotated translation of Poincaré’s topological work [69].

To set the stage for our investigation, I would like to spend some time outlining the general history of the algebraic topology we study in this chapter. Prior to the 1890s, algebraic topology consisted mostly of a few apparently disparate methods for solving certain geometric problems in 2 or 3 dimensions [69, pp. vii–viii]. Much attention was paid to numerical invariants of shapes (the number of holes in a surface, for example), rather than algebraic invariants. Around 1892, Poincaré began a systematic general investigation into the properties of higher–dimensional spaces without recourse to any sort of metric. In his 1895 paper *Analysis situs*, Poincaré outlined how to associate algebraic objects, specifically groups and (implicitly) free modules, to shapes of arbitrary dimension. *Analysis situs* introduced many of the major ideas of modern topology, especially **homotopy** and **homology**, two different ways of talking about “holes” in shapes; in this thesis, we concern ourselves exclusively with homology, defined precisely in Section 5.2. However, Poincaré’s intuitionistic approach to topology lacked any rigor in the modern sense, and some of his definitions were known to be inadequate even in his lifetime.

In the decades after *Analysis situs*, most research in algebraic topology was concentrated on putting Poincaré’s ideas on a firmer logical footing. Homology theory, for example, was revolutionized in the 1920s by mathematicians like Brouwer and Vietoris [55]. Using Emmy Noether’s new abstract algebra, topologists could formally present the significance of Poincaré’s implicit use of free modules in developing homology. This new paradigm facilitated the development of more general homology theories, allowing for algebraic techniques to be used in the study of a wider variety of spaces. In 1931, de Rham proved the theorem bearing his name as part of his doctoral thesis [24], thus interweaving algebraic topology and differential calculus. With this, we are essentially finished with our overview of relevant episodes in the history of pure topology; the only topological ideas we use in the sequel that did not fall out of numerical PDE theory are related to Whitney’s proof of de Rham’s Theorem [89], discussed in Section 9.3.

Any book covering rudimentary algebraic topology may be consulted for further details on the material in this chapter. Readers completely unfamiliar with algebraic topology may benefit from consulting the books of Ghrist [36] or Zomorodian [92] while studying the discussion here. Both of these little books emphasize visual intuition and practical utility over generality and rigour, making them ideal entry points into algebraic topology for those without a background in pure mathematics. [36] is written in a very casual, conversational tone, though the breadth of material covered in this book may be a bit intimidating. [92] is gentler, introducing most of the requisite abstract algebra as it is needed and focusing on less material than Ghrist’s book. For sources more in line with traditional perspectives on algebraic topology, see Munkres’ textbook [61] or the relevant chapters of Nakahara’s tract on mathematical physics [62].

In Section 5.1, I introduce simplices and simplicial complexes, focusing on special simplicial complexes that arise from triangulating domains in Euclidean spaces. In Section 5.2, I introduce the theory of homology for simplicial complexes and discuss its intuitive interpretation in terms of characterizing the number of “holes” in a complex. I also introduce abstract chain and cochain complexes before discussing **simplicial cohomology**. In Section 5.3, we discuss de Rham’s Theorem on the duality between algebraic topology and differential calculus.

5.1 Simplices and Simplicial Complexes

In this section we discuss some of the main building blocks of algebraic topology, namely simplices and simplicial complexes. A simplex is really just a generalization of a triangle or tetrahedron to arbitrary dimensions, and a simplicial complex is a topological space obtained by sticking several simplices together according to some simple rules. One can get reasonably far in algebraic topology by working with “approx-

imations” of general topological spaces by simplicial complexes, and this is essentially the perspective we take throughout Part 1. We begin by introducing some basic facts about simplices, including barycentric coordinates, before moving on to defining simplicial complexes and triangulations of domains in \mathbb{R}^n . We conclude by discussing a technical condition usually imposed on triangulations appearing in the numerical analysis of PDEs.

Definition 5.1.1. Let $\{p_0, \dots, p_k\}$ be a collection of $k + 1$ points in \mathbb{R}^n such that the vectors $\{p_j - p_0\}_{j=1}^k$ are linearly independent. The k -**simplex with vertices** p_0, \dots, p_k is the convex hull of $\{p_0, \dots, p_k\}$, denoted by $[p_0 \dots p_k]$. That is,

$$x \in [p_0 \dots p_k] \iff \exists \lambda_0, \dots, \lambda_k \in \mathbb{R}_{\geq 0} \text{ with } \sum_{i=0}^k \lambda_i = 1 \text{ such that } x = \sum_{i=0}^k \lambda_i p_i. \quad (5.1.1)$$

The nonnegative real numbers λ_i are called the **barycentric coordinates** of x . The number k is called the **dimension** of the simplex.

Note that k -simplices have $k + 1$ vertices.

Example 5.1.2. We define the **standard n -simplex** in \mathbb{R}^n to be

$$T_{\text{ref},n} = [(0, 0, \dots, 0) (1, 0, \dots, 0) (0, 1, 0, \dots, 0) \cdots (0, 0, \dots, 1)]. \quad (5.1.2)$$

Then, $T_{\text{ref},1} = [(0) (1)]$ is the unit interval, $T_{\text{ref},2} = [(0, 0) (1, 0) (0, 1)]$ is a right isosceles triangle, and $T_{\text{ref},3} = [(0, 0, 0) (1, 0, 0) (0, 1, 0) (0, 0, 1)]$ is a triangular pyramid. See Figure 5.11 for an illustration. In numerical analysis, $T_{\text{ref},n}$ is also called the **reference element** for reasons explained in Chapter 6.

The barycentric coordinates on an n -simplex in \mathbb{R}^n give rise to maps $\lambda_i: T \rightarrow [0, 1]$ taking a point $x \in T$ to its i^{th} barycentric coordinate. If we view the points of T as being represented by Cartesian coordinates (x_1, \dots, x_n) , the maps λ_i collectively describe an affine change of coordinates $x_i \mapsto \lambda_i$ for $i = 1, \dots, n$. Of course, the convexity constraint $\sum_{i=0}^n \lambda_i = 1$ means that once we find $\lambda_1, \dots, \lambda_n$, we automatically know λ_0 . That is, the barycentric coordinate system really just depends on n numbers. This becomes important when we start dealing with polynomials on simplices in the sequel: a polynomial in the barycentric coordinates is a polynomial in n independent variables, not $n + 1$ independent variables.

Definition 5.1.3. Let $T = [p_0 \dots p_k]$. A **subsimplex** or **face** of T is any d -simplex $[p_{i_0}, \dots, p_{i_d}]$ with $0 \leq d \leq k$ and $\{i_0, \dots, i_d\} \subseteq \{0, \dots, k\}$. The set of all subsimplices of T is denoted by $\Delta(T)$. The set of all d -subsimplices of T is denoted by $\Delta_d(T)$. A **facet** of a k -simplex T is an element of $\Delta_{k-1}(T)$.

Lemma 5.1.4.

$$|\Delta_d(T)| = \binom{k+1}{d+1}. \quad \square$$

Proposition 5.1.5. Let T be a k -simplex and let F be a facet of T . Then, there is some $j \in \{0, \dots, k\}$ such that $F = V(\lambda_j) \doteq \{x \in T \mid \lambda_j(x) = 0\}$.

Proof. Since F is an $(k - 1)$ -subsimplex of T , it is generated by d of T 's vertices. Suppose the vertex we do not use is p_j for some $j \in \{0, \dots, k\}$. Then, by definition, $\lambda_j(x) = 0$ for all $x \in F$. \square

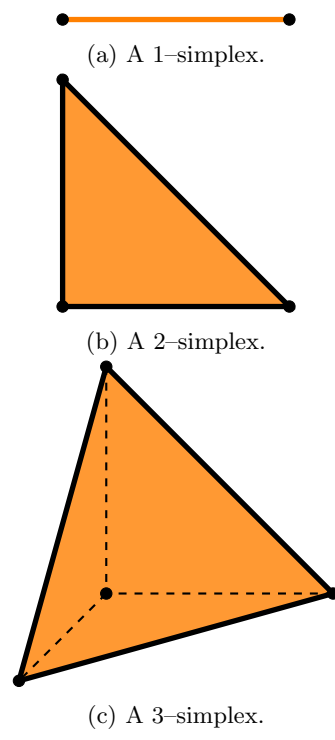


Figure 5.11: Visual representation of some low-dimensional simplices. Interior points are coloured orange, and boundary points are coloured black.

We remark that the boundary of a k -simplex T , denoted ∂T , satisfies

$$\partial T = \bigcup_{F \in \Delta_{k-1}(T)} F.$$

Notice that the boundary of a simplex is not a simplex. For example, the boundary of the 1-simplex $[0, 1]$ is $\{0, 1\}$: this is a union of two simplices, but not a simplex itself.

We may construct more complicated shapes using simplices.

Definition 5.1.6. A *simplicial complex* \mathcal{T} is a finite collection of simplices such that

- for all $T \in \Delta_n(\mathcal{T})$ and all $f \in \Delta(T)$, $f \in \mathcal{T}$ and
- any two $T, T' \in \mathcal{T}$ are either disjoint or $T \cap T' \in \Delta(T) \cap \Delta(T')$.

The number

$$\dim \mathcal{T} \doteq \max_{T \in \mathcal{T}} \dim T$$

is called the **dimension** of the simplicial complex.

We say $f \in \Delta_d(\mathcal{T})$ to indicate that f is a d -simplex contained in the complex \mathcal{T} , directly importing the notation established above. Most authors allow simplicial complexes to contain an arbitrary number of simplices, rather than just a finite number as we have done. However, finite complexes are sufficient for applications in numerical analysis.

We always view our simplicial complexes \mathcal{T} as living within some ambient Euclidean space \mathbb{R}^n . We use $|\mathcal{T}|$ to denote the subset of \mathbb{R}^n consisting of all the points that live in at least one simplex in \mathcal{T} (remember, \mathcal{T} is a set of simplices, not a point set). We call $|\mathcal{T}|$ the **polyhedron** of \mathcal{T} . Of course, $|\mathcal{T}|$ carries the subspace topology inherited from \mathbb{R}^n . In this sense, we are able to view simplicial complexes as topological spaces.

Example 5.1.7. Of course, every simplex T gives rise to a simplicial complex $\mathcal{T} = \Delta(T)$.

Example 5.1.8. Figure 5.12 depicts an example of a simplicial complex, and Figure 5.13 depicts a non-example. Note the importance of the intersection property in Definition 5.1.6 in determining whether or not a “tiling” of some polygon constitutes a simplicial complex.

For various reasons, we often have to deal with all of the simplices in \mathcal{T} that are “close to” some fixed $T \in \mathcal{T}$. The following definition helps make this precise:

Definition 5.1.9. Let \mathcal{T} be a simplicial complex and let $T \in \mathcal{T}$. The **star of T in \mathcal{T}** is defined by

$$\text{Star}(T) \doteq \{f \in \mathcal{T} \mid T \in \Delta(f)\} \subseteq \mathcal{T}.$$

Notice that $\text{Star}(T)$ can also be interpreted as a point set, consisting of all those points for which there exists $T' \in \mathcal{T}$ with $p \in T'$ and $T \in \Delta(T')$. We use both interpretations of $\text{Star}(T)$ in the sequel; whether we are referring to the star as a point set or not is almost always unambiguous from context.

The primarily simplicial complexes of interest in numerical analysis arise from partitioning domains within Euclidean spaces.

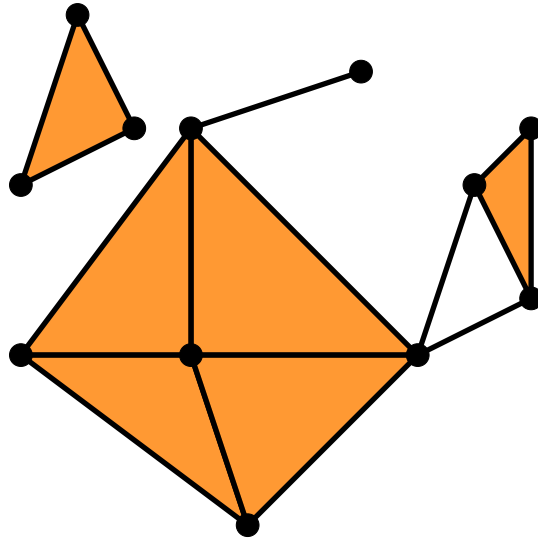


Figure 5.12: A 2-dimensional simplicial complex.

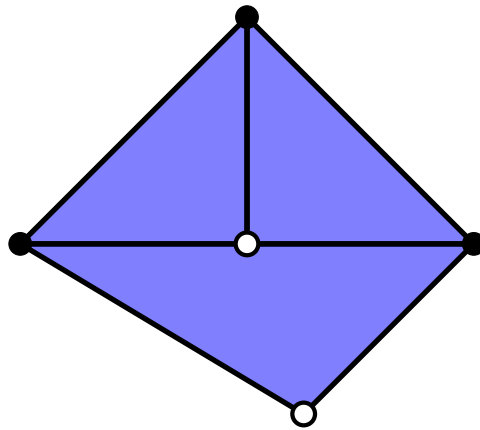


Figure 5.13: A "tiling" that is not a simplicial complex. Connecting the two white vertices with an edge would turn the figure into a simplicial complex.

Definition 5.1.10. A Lipschitz domain $\Omega \subseteq \mathbb{R}^n$ is said to be **polyhedral** if there exist finitely many n -simplices T_i such that

- $\bar{\Omega} = \cup_i T_i$ and
- for all $i \neq j$, $T_i \cap T_j$ is either the empty set or some $f \in \Delta(T_i) \cap \Delta(T_j)$.

We then say that the n -dimensional simplicial complex $\mathcal{T}_h \doteq \cup_i \Delta(T_i)$ is a **triangulation** of Ω .

Sometimes, we also refer to a triangulation as a **mesh**. The use of a subscript h is a convention arising from numerical analysis; more details are discussed in the next paragraph. However, usage of this subscript extends beyond explicitly indicating the size of simplices: when the reader sees a subscript h appearing in the sequel, they should infer that it indicates passage to the discretized setting.

Now, we discuss the interpretation of the subscript h in numerical analysis, and introduce a common technical requirement imposed on meshes in practice. h is usually chosen to be the **mesh size parameter**, defined by

$$h \doteq \max_{T \in \mathcal{T}_h} \text{diam}(T).$$

So, the smaller h is, the smaller all of the simplices partitioning Ω must be. In applications, we are usually using triangulations to approximate solutions to PDEs for functions or forms defined over the domain Ω , with our approximate solution living in some space of forms with components that are piecewise smooth with respect to the mesh. We intuitively expect that our approximation should resemble the exact solution more closely as $h \rightarrow 0$. In light of the above discussion, we usually define a family of meshes over Ω indexed by the mesh size parameter h and then examine the quality of approximation as $h \rightarrow 0$. However, we need to impose an auxiliary requirement in order to guarantee that the triangulation stays well-behaved as we make the simplices smaller and smaller. The following definition of such “good behaviour” is adapted from [19, p. 5].

Definition 5.1.11. Let $\Omega \subseteq \mathbb{R}^n$ be a polyhedral domain equipped with a family of triangulations $\{\mathcal{T}_h\}_h$ indexed by the real number h . We say that $\{\mathcal{T}_h\}_h$ is **shape-regular** if there exists $C > 0$, independent of h , such that

$$(\text{diam}(T))^n \leq C \text{vol}(T) \quad \forall T \in \Delta_n(\mathcal{T}_h).$$

Heuristically, a family of triangulations is shape-regular if none of the component simplices “stretch out” too much as $h \rightarrow 0$.

5.2 Homology of Simplicial Complexes

In this section we introduce a graded vector space associated to a simplicial complex, called the **homology**. As cohomology allows us to study Hilbert complexes algebraically, homology allows us to study the topology of a simplicial complex algebraically. Before introducing homology, we need to define **orientations** on simplices and spaces of **chains** in simplicial complexes. Following these basic definitions, we discuss homology more precisely and go through some practical examples. We then discuss the relationship between the homology of simplicial complexes and the cohomology of Hilbert complexes in a bit more detail, and

then conclude by explaining how homology can be “dualized” to establish a notion of cohomology for simplicial complexes.

We note that, in this section, we develop homology theory “with real coefficients”. Such a choice makes the connection between algebraic topology and Hilbert complexes very clear. In introductory texts, however, homology is usually introduced “with integer coefficients” (see for example [92]). The reader looking to compare the discussion here to that presented in the aforementioned references should not become confused or worry much about this disparity: we can essentially switch out \mathbb{R} for \mathbb{Z} to transition between the two homology theories without gaining or losing any substantial information (at least in the context of applications to numerical analysis).

We begin by describing how to equip simplices with orientations. We remark that any ordering $<$ of a finite set $\{p_0, \dots, p_k\}$ is associated to a permutation of $\{p_0, \dots, p_k\}$ according to the rule

$$p_{i_0} < p_{i_1} < \dots < p_{i_k} \leftrightarrow \begin{pmatrix} p_0 & \dots & p_k \\ p_{i_0} & \dots & p_{i_k} \end{pmatrix}.$$

Recall that $\begin{pmatrix} p_0 & \dots & p_k \\ p_{i_0} & \dots & p_{i_k} \end{pmatrix}$ denotes the permutation mapping $p_j \mapsto p_{i_j}$

Definition 5.2.1. *Let T be a simplex. Two orderings of Δ_0 with associated permutations s_1 and s_2 are said to be **equivalent** if there exists an even permutation s such that $s_2 = s \circ s_1$. An **orientation on T** is an equivalence class of orderings of $\Delta_0(T)$.*

If $\dim T > 0$ there are precisely two orientations on T : one corresponding to even permutations of the identity, and the other corresponding to odd permutations of the identity. To indicate that we are switching to the opposite orientation on an oriented simplex T , we write $-T$.

Example 5.2.2. The orderings

$$(0, 0) < (1, 0) < (0, 1) \quad \text{and} \quad (1, 0) < (0, 1) < (0, 0)$$

represent the same orientation on $T_{\text{ref},2}$. This orientation coincides with the right-handed orientation that $T_{\text{ref},2}$ inherits from \mathbb{R}^2 : if you curl the fingers of your right hand in the “increasing” direction, your thumb points upward in both cases. Conversely, the ordering $(0, 0) < (0, 1) < (1, 0)$ coincides with the induced left-hand orientation.

Now, we introduce the notion of **chains** in simplicial complexes. Given some real numbers c_i and some $T_i \in \Delta_k(\mathcal{T})$, we can define an expression c by

$$c \doteq \sum c_i T_i.$$

The set of all such expressions c is called the set of k -**chains in \mathcal{T}** , denoted by $C_k(\mathcal{T})$. Of course, the sum in the definition of a k -chain is purely formal (see the discussion below Definition for an outline of what it means for a sum to be “formal”).

Defining addition and scalar multiplication simplex-wise, we endow $C_k(\mathcal{T})$ with the structure of a vector space. Note how our notation for chains naturally includes the notation for reversing orientation (multiplication of a simplex by -1). A chain of the form $c = T_i$ is said to be **elementary**. Clearly, the set of all elementary k -chains is a basis of $C_k(\mathcal{T})$. In the jargon of abstract algebra, we have that $C_k(\mathcal{T})$ is the **free real vector space generated by $\Delta_k(\mathcal{T})$** .

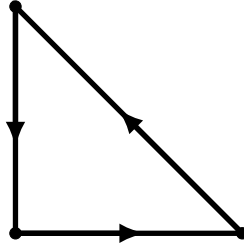


Figure 5.21: $\partial T_{\text{ref},2}$ with the induced right-hand orientation.

Remark 5.2.3. We can also define chains as real-valued functions c on the set oriented k -simplices of \mathcal{T} such that $c(-T) = -c(T)$. This perspective is taken in [61]. However, we prefer to stress the formal sum interpretation of chains, as it allows us to move a bit faster.

Now, we turn to the construction of a special type of chains carrying concrete geometric meaning. First, we go through an easy example in order to motivate the general case.

Example 5.2.4. Let $T_{\text{ref},2} = [(0,0) (1,0) (0,1)]$ be the standard triangle in \mathbb{R}^2 , equipped with the right-handed orientation. We know that $\partial T_{\text{ref},2}$ (as a subset of \mathbb{R}^2) is the union of 1-simplices of T

$$\partial T_{\text{ref},2} = [(0,0) (1,0)] \cup [(1,0) (0,1)] \cup [(0,1) (0,0)].$$

$\partial T_{\text{ref},2}$ then corresponds to a 1-chain on $\Delta(T)$,

$$\partial T_{\text{ref},2} = [(0,0) (1,0)] + [(1,0) (0,1)] + [(0,1) (0,0)] \in C_1(\Delta(T)).$$

This expression is clearly well-defined irrespective of the chosen representative of the right-handed orientation.

Let $p_0 = (0,0)$, $p_1 = (1,0)$, and $p_2 = (0,1)$. For $i = 0, 1, 2$, we let $[p_0 \dots \widehat{p}_i \dots p_2]$ denote the facet of $T_{\text{ref},2}$ obtained by omitting the i^{th} vertex; the hat means “take this vertex out”. For example,

$$[(0,0) (1,0)] = [p_0 p_1 \widehat{p}_2] \quad \text{and} \quad [(0,1) (0,0)] = -[p_0 \widehat{p}_1 p_2],$$

where we have picked up a negative sign in the second expression because of the orientation. With this new notation in mind, we may write the 1-chain corresponding to the boundary as

$$\partial T_{\text{ref},2} = \sum_{i=0}^2 (-1)^i [p_0 \dots \widehat{p}_i \dots p_2]. \tag{5.2.1}$$

This is just a weighted sum over the facets of $T_{\text{ref},2}$. See Figure 5.21 for an illustration of the 1-chain $\partial T_{\text{ref},2}$ (the arrows point in the direction of increase with respect to the equivalent orders associated to the right-hand orientation).

More generally, given an oriented n -simplex $T = [p_0 \dots p_n]$ we can define an $(n-1)$ -chain ∂T by replacing the union symbol in the expression for the boundary ∂T with an addition symbol. We must be careful to ensure that the facets are ordered with respect to the induced orientation from T (an ordering on a set

induces an ordering on all subsets), as we have done in the specific case of the standard triangle. The hat notation carries over nicely to this general situation and we have by direct comparison with (5.2.1) that

$$\partial T = \sum_{i=0}^n (-1)^i [p_0 \dots \widehat{p}_i \dots p_n]. \quad (5.2.2)$$

Remark 5.2.5. We recall that each facet of T is identified with the vanishing set of one barycentric coordinate function. The hat notation reflect this: $[p_0 \dots \widehat{p}_i \dots p_n]$ is the vanishing set of the barycentric coordinate associated to the vertex p_i . The operation of taking out a vertex of T amounts to setting the barycentric coordinate corresponding to that vertex to zero.

Since $C_k(\mathcal{T})$ is a vector space, the above construction allows us to define a **boundary operator** on arbitrary k -chains $\partial: C_k(\mathcal{T}) \rightarrow C_{k-1}(\mathcal{T})$ through extension by linearity. The boundary of a general k -chain

$$c = \sum c_i T_i$$

is therefore given by

$$\partial c = \sum c_i \partial T_i.$$

We sometimes use the notation $\partial_k \doteq \partial|_{C_k(\mathcal{T})}$ to emphasize some special property of the boundaries of k -chains. Taken together, the spaces of chains and the boundary operators give us a sequence

$$\cdots \leftarrow C_{k-1} \xleftarrow{\partial_k} C_k \xleftarrow{\partial_{k+1}} C_{k+1} \leftarrow \cdots .$$

The resemblance between the diagram above and the diagrams we saw when studying Hilbert complexes in Chapter 4 is noteworthy to say the least, but we postpone such discussions until the end of this section.

Example 5.2.6. The boundary of the standard triangle $T_{\text{ref},2}$ is not a simplex, but it is a 1-chain. The developments of the preceding paragraph imply that we can define $\partial \partial T_{\text{ref},2} \in C_0(T_{\text{ref},2})$. Using (5.2.1) and the definition of ∂ , we have that

$$\begin{aligned} \partial \partial T_{\text{ref},2} &= \partial [(0,0) (1,0)] + \partial [(1,0) (0,1)] + \partial [(0,1) (0,0)] \\ &= [(0,0)] - [(1,0)] + [(1,0)] - [(0,1)] + [(0,1)] - [(0,0)] \\ &= 0. \end{aligned}$$

So, the boundary of the boundary is $0 \in C_0(T_{\text{ref},2})$. This corresponds with a fact familiar from multivariable calculus: the boundary of a boundary of a domain in \mathbb{R}^n is empty.

The last example reflects an important and intuitive nilpotency rule.

Proposition 5.2.7. [92, Theorem 4.3] $\partial \circ \partial = 0$.

Naturally, one can ask under what conditions all chains in the kernel of ∂ arise as boundaries of higher-dimensional chains. Towards solving this problem, we make some helpful definitions.

Definition 5.2.8.

1) $Z_k \doteq \ker \partial_k$ is called the set of k -cycles.

2) $B_k \doteq \text{range } \partial_{k+1}$ is called the set of k -boundaries.

3) The k^{th} homology space is the quotient

$$H_k \doteq Z_k/B_k.$$

4) The (total) homology of \mathcal{T} is the graded real vector space

$$H_\bullet \doteq \bigoplus_{k=0}^{\dim \mathcal{T}} H_k.$$

Of course, the quotient in the definition of H_k makes sense because $B_k \subseteq Z_k$ by Proposition 5.2.7. Notice that all the indices k in the above definition are downstairs.

Remark 5.2.9. The reason for the name k -boundaries is obvious, and the name k -cycles is chosen because they correspond to cyclic paths (that is, complete circuits) when dealing with two-dimensional complexes (see the examples below).

With these definitions in mind, we can re-phrase the question above Definition 5.2.8 as “under what conditions is $Z_k = B_k$?” or equivalently “under what conditions is $H_k = 0$?”. Before we provide a precise answer, we study some simple examples illustrating how the question is connected to topology. By the same token, these examples tell us how topological information is encoded in the spaces H_k .

Example 5.2.10. Consider the same notational setup as Example 5.2.4. We compute the three potentially nonzero homology spaces H_2 , H_1 , and H_0 of $T_{\text{ref},2}$. H_2 is very easy to compute: $\partial T_{\text{ref},2}$ is nonzero, hence the only way a 2-chain can be in Z_2 is if that chain is just 0. Then, the top homology space is trivial. Moreover, if we let f_i denote the edge of $T_{\text{ref},2}$ opposite the vertex p_i ($i = 0, 1, 2$), the aforementioned example yields that $B_1 = \text{span}(f_1 + f_2 + f_3)$.

Now, suppose that $c \in Z_1$. Write c in terms of the edges f_i as

$$c = \sum c_i f_i.$$

Then, since c is a 1-cycle, the computations of Examples 5.2.4 and 5.2.6 give us

$$0 = (c_2 - c_0) p_0 + (c_0 - c_1) p_1 + (c_1 - c_2) p_2.$$

This can only be true if $c_0 = c_1 = c_2$. Accordingly, $Z_1 = \text{span}(f_1 + f_2 + f_3)$; that is, any 1-cycle must be a multiple of $f_1 + f_2 + f_3 = \partial T_{\text{ref},2}$ (this example helps substantiate use of the name “cycle”, after the discussion in Remark 5.2.9). Therefore, $H_1 = 0$.

Finally, we come to the zeroth homology space. Since $\partial_0 = 0$ identically, $Z_0 = C_0$. Then, Z_0 has a canonical basis consisting of all the vertices p_i . We conclude that $Z_0 \simeq \mathbb{R}^3$. Now, direct calculation (similar to Example 5.2.6) shows that any 0-boundary c must have the form $c = \sum c_i p_i$ with $\sum c_i = 0$. Then, $B_0 \simeq \mathbb{R}^2$. Using the Second Isomorphism Theorem for vector spaces, we conclude that $H_0 \simeq \mathbb{R}$. In summary, the only non-trivial homology space for the standard 2-simplex is H_0 , which is just a copy of \mathbb{R} .

Example 5.2.11. Now, let $T = \partial T_{\text{ref},2}$. The same calculations we used in Example 5.2.10 tell us

$$H_0 \simeq \mathbb{R} \quad \text{and} \quad H_1 \simeq \mathbb{R}.$$

The zeroth homology space is the same, but when we take out the interior of the simplex the first homology space becomes non-trivial. This leads us to ask what differences between $T_{\text{ref},2}$ and $\partial T_{\text{ref},2}$ can lead to this change in homology. Aside from the obvious (their dimensions are not the same), the two simplicial complexes have different topological features. Namely, $T_{\text{ref},2}$ is **contractible** (the space can be “shrunk down” to a point) while its boundary is not. Such topological disparity is indeed the cause of the difference in homology spaces.

With these examples in mind, we are ready to discuss the general relationship between topology and homology spaces, though we must outsource the proofs to the references. The next theorem gives an immediate, albeit weak, connection between homology and topology.

Theorem 5.2.12. [61, Cor. 18.2] *Homeomorphic simplicial complexes have isomorphic homologies.*

So, if the homologies of two different complexes disagree, then the complexes cannot be homeomorphic. That is, examining homology gives us a necessary condition for topological equivalence, and this condition may be understood using purely algebraic objects. Homotopy theory shows that there is a much weaker sufficient condition for two complexes to have isomorphic homologies [61, Chapter 2].

The next result shows that the zeroth homology is always very easy to compute.

Theorem 5.2.13. [61, Thm. 7.2] *The dimension of H_0 is equal to the number of connected components of the simplicial complex \mathcal{T} .*

The other homology spaces may also be interpreted topologically (after the discussion in Example 5.2.11) in the vein of the following heuristic:

- 1) $\dim H_1 =$ number of 1-dimensional holes in \mathcal{T} ,
- 2) $\dim H_2 =$ number of 2-dimensional holes in \mathcal{T} ,

and so on for the higher homology spaces. In the above, we understand the “dimension” of a hole to be the largest number n such that a homeomorphic copy of the n -sphere fits inside the hole. For example, the boundary of a triangle has a 1-dimensional hole because we can squeeze a circle of sufficiently small radius into the region where the interior of the triangle used to be. Even though we can fit a single point (0-sphere) in the interior too, the circle is the sphere of largest dimension fitting inside the hole, hence the hole is 1-dimensional. In particular, the homology spaces of a contractible complex must all be trivial, except for the zeroth homology.

Notice how the dimensions of the homology spaces, rather than the particular presentation of the spaces themselves, are the primary source of topological data arising from homology theory. Consequently, these dimensions have their own special names: they are called the **Betti numbers** of \mathcal{T} . Betti numbers were among the first great discoveries in topology, pre-dating both Poincaré’s *Analysis situs* and Emmy Noether’s formal introduction of homology spaces by several decades. To the reader who is wondering why we bother studying homology if the Betti numbers contain all topological information, I recommend a historical article written by Mac Lane [55]. Mac Lane’s essay provides an interesting perspective on the

utility and novelty of Noether’s abstract–algebraic viewpoint in topology and, indeed, in mathematics as a whole.

Now, we discuss the similarities between the diagram

$$\cdots \leftarrow C_{k-1} \xleftarrow{\partial_k} C_k \xleftarrow{\partial_{k+1}} C_{k+1} \leftarrow \cdots$$

and the Hilbert complexes encountered in Chapter 4, which could be written as

$$\cdots \rightarrow W^{k-1} \xrightarrow{d^{k-1}} W^k \xrightarrow{d^k} W_{k+1} \rightarrow \cdots .$$

If the Hilbert complex is bounded then, from a very abstract perspective, the two sequences above are almost identical: each term in the sequence is an object in some category (**Vect** and **Hilb**, respectively), and all of the maps taking one term to the other are morphisms in that category (linear maps, bounded linear operators). Further, these morphisms reduce to 0 when applied one after the other, giving rise to two similar objects: homology for chain complexes and cohomology for Hilbert complexes. Therefore, from a sufficiently abstract viewpoint, the only substantial difference between the sequence of chain spaces and a bounded Hilbert complex is the direction that the arrows point.

In the twentieth century, mathematicians noticed patterns like the one discussed above. Sequences that had the same general properties as the sequence of chain spaces (every term in the sequence is an object in some category, there are nilpotent morphisms between each term) were appearing in many different problems, primarily in topology and geometry. Accordingly, such sequences earned a special name: if the arrows point in the direction of decreasing k , we say that the sequence is a **chain complex**. Otherwise, we say that the sequence is a **cochain complex**. The discipline of **homological algebra** (see for instance [61]) was conceived to examine the properties of chain and cochain complexes in full generality.

Homological algebra grew up hand–in–hand with category theory, sharing with it the dream of understanding topological spaces by studying in isolation certain invariant algebraic objects associated to them. So, even though these subjects are often perceived by outsiders as dungeons of self–indulgent abstract fussiness, they maintain their roots in very natural topological ideas. We do not use too much homological algebra or category theory in this thesis, beyond treating them as a convenient source of unifying jargon and notation. Indeed, for applications in the numerical analysis of PDEs, it is often preferable to study topology from the classical perspective of Poincaré, rather than from the perspective of Eilenberg, Mac Lane, and their contemporaries. However, techniques from category theory, homological algebra, and other famously abstract disciplines have recently shown themselves to be useful in solving real–world problems, especially in data science and computing. For an enlightening discussion on the growing importance of such techniques in applied mathematics, computer science, and statistics, I recommend Ghrist’s essay [37].

We conclude this section by discussing a theory “dual” to simplicial homology. Denote the dual space to $C_k(\mathcal{T})$ by $C^k(\mathcal{T})$. The functionals in $C^k(\mathcal{T})$ are called k –**cochains**. Since elementary k –chains form a basis of $C_k(\mathcal{T})$, we immediately have a dual basis of $C^k(\mathcal{T})$. Elements of this basis are called **elementary k –cochains**, and they are in one–to–one correspondence with elements of $\Delta_k(\mathcal{T})$. As with elementary chains, we can identify elementary cochains with faces of \mathcal{T} . For example, depending on context the symbol $[p_0 \dots p_k]$ can represent a k –simplex, or the same k –simplex viewed as a chain or cochain.

Denote the adjoint of the boundary operator ∂_{k+1} by

$$\partial_k^* : C^k(\mathcal{T}) \rightarrow C^{k+1}(\mathcal{T}) .$$

We often call ∂_k^* the **coboundary operator**. Since we do not assume a particular inner product structure on the chain spaces $C_k(\mathcal{T})$, the adjoint of ∂_{k+1} is defined by the following condition: for all $\alpha \in C^k(\mathcal{T})$ and $c \in C_k(\mathcal{T})$,

$$\partial_k^* \alpha(c) = \alpha(\partial_{k+1} c).$$

The cochain spaces and coboundary operators together form a cochain complex

$$\dots \rightarrow C^{k-1} \xrightarrow{\partial_{k-1}^*} C^k \xrightarrow{\partial_k^*} C^{k+1} \rightarrow \dots,$$

since $\partial_k^* \partial_{k-1}^* = 0$ using the nilpotency of ∂ . We can define cohomology for the above cochain complex as we did for Hilbert complexes earlier on.

Definition 5.2.14.

- 1) $Z^k \doteq \ker \partial_k^*$ is called the space of *k-cocycles*.
- 2) $B^k \doteq \text{range } \partial_{k-1}^*$ is called the space of *k-coboundaries*.
- 3) The k^{th} *simplicial cohomology space* is the quotient

$$H^k \doteq Z^k / B^k.$$

- 4) The *(total) simplicial cohomology* of \mathcal{T} is

$$H^\bullet = \bigoplus_{k=0}^{\dim \mathcal{T}} H^k.$$

Lemma 5.2.15. $H^k \simeq (H_k)^*$.

Proof. One checks routinely that the map $H^k \rightarrow (H_k)^*$ defined by

$$[\alpha] \mapsto ([c] \mapsto \alpha(c))$$

gives us the required isomorphism. □

Recall that all of the topological information we need from a simplicial complex's homology is encoded in the dimensions of the spaces H_k . Since

$$\dim H_k = \dim H^k,$$

we know that simplicial cohomology gives us the same topological information as homology.

In order to get anywhere with simplicial cohomology, we need a concrete formula for computing with the coboundary operator. This formula uses stars, introduced in Definition 5.1.9.

Proposition 5.2.16. Denote the vertices of the simplicial complex \mathcal{T} by p_j . Let $c = [p_0 \dots p_k] \in C^k(\mathcal{T})$. Then,

$$\partial_k^* c = \sum_{\substack{j \neq 0, \dots, k \\ p_j \in \text{Star}(c)}} [p_j p_0 \dots p_k]. \tag{5.2.3}$$

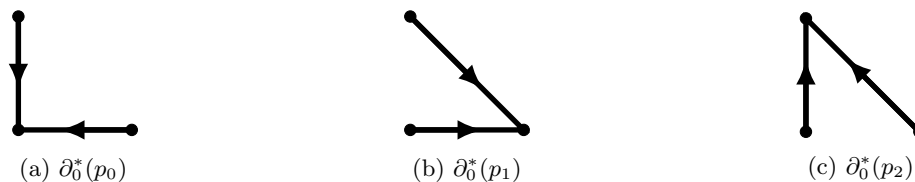


Figure 5.22: Coboundaries of the vertices of $T_{\text{ref},2}$; the captions follow the notation of Example 5.2.17.

Proof. This formula may be found in [61, p. 253] or [89, p. 362]. The proof amounts to a direct verification. \square

Explicitly saying that we sum over $j \neq 0, \dots, k$ is a bit redundant, but it helps us maintain cohesion when computing coboundaries.

Example 5.2.17. In this example, we compute the action of the coboundary operator on the vertices of $T_{\text{ref},2}$ (that is, the elementary 0-cochains on $T_{\text{ref},2}$). We use the same notation introduced in Example 5.2.4. By (5.2.3), we see that

$$\partial_0^*(p_0) = [p_1 p_0] + [p_2 p_0],$$

$$\partial_0^*(p_1) = [p_0 p_1] + [p_2 p_1], \text{ and}$$

$$\partial_0^*(p_2) = [p_0 p_2] + [p_1 p_2].$$

All three of these coboundaries correspond to sharp corners, illustrated in Figure 5.22. This example demonstrates that coboundaries and cocycles look like “picket fences” [61, p. 254] while boundaries and cycles look like closed circuits.

In summary, we used this section to introduce an algebraic invariant of a simplicial complex, the homology H_\bullet . The k^{th} homology space can be interpreted intuitively as measuring the extent to which closed circuits in a simplicial complex fail to arise as boundaries of higher-dimensional subsimplices. We then drew some comparisons between the homology of simplicial complexes and the cohomology of Hilbert complexes, and discussed how the study of homology served as a springboard for the development of category theory and homological algebra. Finally, we “dualized” homology to develop a notion of cohomology for simplicial complexes. Trivially, simplicial cohomology contains the same topological data carried by homology. In the next section, we see that simplicial cohomology and Hilbert complex cohomology are essentially the same object up to isomorphism. Such connections between topology and analysis are very useful in the theoretical and numerical treatment of PDEs, and provide a focal point for our examination of the consistency of finite element methods in Chapter 9.

5.3 de Rham’s Theorem: Bridging Analysis and Topology

In this section we examine (mostly qualitatively) de Rham’s Theorem, a deep result exhibiting the correspondence between algebraic topology and calculus on manifolds. The presentation borrows heavily from

John Lee’s introduction to de Rham’s Theorem [51, Chapter 18], and readers wishing to fill in some of the technical details missing here ought to consult Lee’s book.

First, we discuss cochain complexes appearing in calculus, rather than algebraic topology. Recall from Chapter 4 that the L^2 de Rham complex on a polyhedral domain $\Omega \subseteq \mathbb{R}^n$ is defined by

$$0 \rightarrow H\Lambda^0(\Omega) \xrightarrow{d} H\Lambda^1(\Omega) \xrightarrow{d} \dots \xrightarrow{d} L^2\Lambda^n(\Omega) \rightarrow 0, \quad (5.3.1)$$

where we remember that the exterior derivative d is understood in the weak sense. We can define a “smooth” version of the above Hilbert complex, called the **smooth de Rham complex**, using smooth k -forms and the exterior derivative (not just the *weak* exterior derivative). The smooth de Rham complex is written as

$$0 \rightarrow \mathcal{C}^\infty\Lambda^0(\Omega) \xrightarrow{d} \mathcal{C}^\infty\Lambda^1(\Omega) \xrightarrow{d} \dots \xrightarrow{d} \mathcal{C}^\infty\Lambda^n(\Omega) \rightarrow 0. \quad (5.3.2)$$

The smooth de Rham complex (5.3.2) is a cochain complex, but not a Hilbert complex. We denote the cohomology of (5.3.1) by $H_{L^2, dR}^k$ and the cohomology of (5.3.2) by H_{dR}^k . Near the end of this section, we cite results indicating that these two cohomologies are equivalent, but we have no a priori reason to expect that this is the case.

As discussed in Sections 4.3 and 4.4, studying the cohomology of the L^2 de Rham complex reveals whether or not the PDE $\omega = d\eta$ has a weak solution for any given $\omega \in \ker d|_{H\Lambda^k}$. Clearly, then, the cohomology of the smooth de Rham complex tells us about whether the same PDE has *smooth, strong* solutions for any given $\omega \in \ker d|_{\mathcal{C}^\infty\Lambda^k}$.

Now, since $\bar{\Omega}$ is equal to the polyhedron of some triangulation \mathcal{T}_h , we can associate simplicial cohomology spaces to $\bar{\Omega}$. Of course, this triangulation is not unique: in the 2D case, for instance, we can chop each triangle in \mathcal{T}_h into four triangular pieces to obtain another triangulation whose polyhedron is $\bar{\Omega}$. As it turns out, the cohomology spaces associated to $\bar{\Omega}$ are independent of the particular triangulation we use. To prove this rigorously, we would need to develop the theory of **singular cohomology** [51, Chapter 18], but for our purposes such details are unimportant. The only things we need to know are that one can canonically talk about the cohomology of $\bar{\Omega}$ in a topologically invariant way, and that this cohomology can be computed by picking any particular triangulation. Actually, we can also associate “singular cohomology” spaces to Ω , and these spaces are isomorphic to the simplicial cohomology spaces of $\bar{\Omega}$ defined above. All in all, talking about the cohomology spaces H^k of Ω is absolutely no problem.

Using the notion of integration over manifolds with corners, we know the integral of a k -form over a k -simplex makes sense. Extending by linearity, we can immediately develop a notion for the integral of a k -form over a k -chain. The integral provides a natural pairing between k -forms and k -chains: if $c \in C_k(\bar{\Omega})$ (this notation is sloppy, but remember that the actual triangulation doesn’t matter) and $\omega \in \mathcal{C}^\infty\Lambda^k(\bar{\Omega})$, then

$$(c, \omega) \mapsto \int_c \omega \in \mathbb{R}.$$

By way of the natural pairing of chains and forms, we have a linear map from the k^{th} cohomology space of

$$0 \rightarrow \mathcal{C}^\infty\Lambda^0(\bar{\Omega}) \xrightarrow{d} \mathcal{C}^\infty\Lambda^1(\bar{\Omega}) \xrightarrow{d} \dots \xrightarrow{d} \mathcal{C}^\infty\Lambda^n(\bar{\Omega}) \rightarrow 0 \quad (5.3.3)$$

to $(H_k)^*$, defined by

$$[\omega] \mapsto \left([c] \mapsto \int_c \omega \right). \quad (5.3.4)$$

Combining the above with Lemma 5.2.15 yields a homomorphism from the k^{th} cohomology of (5.3.3) to H^k , referred to as the **de Rham homomorphism**. By Stokes’ Theorem, the de Rham homomorphism is well-defined (see [51, p. 483] for the details). Using singular cohomology theory, we can play a similar game to get another de Rham homomorphism mapping H_{dR}^k (the k^{th} cohomology space of (5.3.2)) to H^k ; explaining the details would require us to define integrals of k -forms on Ω over “singular” k -chains, which would take us too far afield. The end product of this line of investigation is the following result:

Theorem 5.3.1. (*de Rham’s Theorem* [51, Thm. 18.14]) *The de Rham homomorphism $H_{\text{dR}}^k \rightarrow H^k$ is an isomorphism.*

We now know that the smooth de Rham complex and the simplicial cochain complex contain exactly the same information. In PDE theory, however, we often prefer to work in the setting of weakly differentiable, rather than smooth, functions and forms. Hence, most of our attention in the sequel is focused on the L^2 de Rham complex rather than the smooth one. We would like to see if the $H_{L^2\text{dR}}^k$ encodes the same topological data as H_{dR}^k . Since

$$C^\infty \Lambda^k(\Omega) \subseteq H\Lambda^k(\Omega),$$

we know that the inclusion map induces a canonical injection $H_{\text{dR}}^k \rightarrow H_{L^2\text{dR}}^k$.

Proposition 5.3.2. *The injection induced by inclusion is an isomorphism $H_{\text{dR}}^k \simeq H_{L^2\text{dR}}^k$.*

The above proposition has been proven at least twice in the geometry literature (on more general manifolds), first by Cheeger in 1980 [17, p. 94] and second by Brüning and Lesch in 1992 [14, pp. 109–110]. Cheeger’s proof uses a smoothing argument, very much in the spirit of analytical PDE theory, while the proof of Brüning and Lesch follows from a more general result in Hilbert complex theory.

Combining de Rham’s Theorem with Proposition 5.3.2, we have that

$$H_{L^2\text{dR}}^k \simeq H^k$$

for Lipschitz domains. Then, we have finally constructed the bridge between Hilbert complexes and topology promised in Chapter 4. In light of Proposition 5.3.2, we often do not bother making the distinction between L^2 and smooth de Rham cohomologies, and just refer to both spaces as *the* de Rham cohomology.

In summary, we have seen that the L^2 de Rham complex on a polyhedral domain Ω is intimately related to the topology of Ω . By studying de Rham’s Theorem (and its extension to the L^2 setting by Cheeger and Brüning and Lesch), we have delivered on a promise made in Chapter 4. Namely, we now see why understanding the “third term” in the Hodge–Helmholtz decomposition as a cohomological object is equivalent to understanding it as a topological object. Combining this insight with the Hodge Theorem, we see that the Hodge Laplacian has a kernel whose dimension is entirely determined by the topology of our spatial domain. All in all, we certainly believe by now that there are profound connections between PDEs, Hilbert complexes, and topology. In Chapter 9, we illustrate how to solve PDEs numerically using methods that properly reflect these connections. Before we can get to this, however, we must spend three chapters developing some more basic theory for the particular family of numerical methods we intend to study.

Chapter 6

Generalities on Finite Element Methods

Now that we have learned about the algebra and topology underlying PDE theory, we are ready to start studying the discretization and numerical solution of differential equations. In this section, we develop some theory regarding **finite element methods (FEMs)** for numerically solving PDEs. FEMs are a member of a general family of discretization schemes called **Galerkin methods**. Galerkin methods involve re-formulating a given PDE “weakly” on some function space V , then looking for a solution within a finite-dimensional space V_h , often a subspace of V . When using an FEM, the space V_h consists of functions that are piecewise smooth with respect to a triangulation of the domain; the subscript “ h ” universally indicates passage to the discrete setting. The remainder of Part 1 of this thesis is devoted to explicitly constructing usable spaces V_h and proving that the resulting discretizations “preserve” the cohomology of the non-discretized problem. FEMs have been around for over half a century, the first one appearing in a primitive form in a 1943 paper of Courant [23], but only in the last two decades has their connection to Hilbert complexes and topology become widely recognized. We discuss structure-preserving FEMs and their history more in later sections; the present section is primarily concerned with the basic aspects of the theory behind FEMs, without mention of their algebraic and topological properties.

FEMs are perhaps the most versatile of the many numerical methods crafted to study PDEs, largely due to their strengths in handling complicated domains and different boundary conditions. Finite difference numerical methods, however, may become a little unwieldy in non-trivial spatial geometries (see the discussion at the end of [47, Ch. 8]). Even the powerful spectral methods lose some formal accuracy in the absence of periodic boundary conditions [47, Ch. 10] (granted, there are ways of adapting spectral methods to more general boundary conditions using Chebyshev polynomials). Another important group of numerical PDE-solving techniques, the finite volume methods, have been adapted into the service of special FEMs called **discontinuous Galerkin methods (DGMs)**. The study of finite volume methods without consideration of finite elements, however, remains an active research area. For a discussion on the relationship between finite volume methods and DGMs, see [29, §1.1], though be cautioned that they view FEMs as special DGMs, not the other way around as we do; our nomenclature here follows that typical of finite element exterior calculus. Additionally, the practical implementation of FEMs has become much easier in the past decade with the introduction of software packages like FEniCS [2] and Firedrake [72]. These packages automate a great deal of the more tedious aspects of coding up FEMs. For example,

instead of having to manually input a basis for the space V_h , we can just select an appropriate V_h from a catalogue of common choices; the spaces we construct in Chapter 8 are included in this catalogue.

The literature on FEMs is incredibly expansive, running the gamut from textbooks written for engineering undergraduates to advanced monographs written for professional researchers in PDE theory. There are, however, some reference texts that merit special mention. Of course, much of FEM theory is unified by the finite element exterior calculus we develop in this thesis, but even with such modern geometric formalism in hand one should still consult the standard texts to develop an appropriate view of the “big picture”. Those looking to immediately start problem-solving with FEMs should consult the (freely available) introductory FEniCS book [54]. This text includes enough theory to understand FEM codes in the FEniCS environment, and more than enough computer examples to satisfy the curious reader. On the more theoretical side of FEMs, one can hardly do better than the book of Ciarlet [20], a classic of numerical analysis. The little text by Strang and Fix [82] is also excellent, though some sections are a bit outdated. Another standard, albeit much younger, book on FEM theory is that of Boffi, Brezzi, and Fortin [10]. For a modern textbook addressing FEM theory as well as common numerical linear algebra issues encountered in implementing FEM codes, see the book of Elman, Silvester, and Wathen [33]. The reader interested in seeing FEMs presented alongside other numerical methods should consult Iserles’ textbook [47]. DGMs are a very important component of contemporary research in scientific computing, and the text of Dolejší and Feistauer [29] provides a modern survey of the subject. As previously mentioned, DGMs are intimately connected with finite volume methods, so understanding finite volume methods is critical to developing good DGMs. LeVeque’s book [52] provides a very accessible general reference on finite volume methods.

In Section 6.1, I introduce Galerkin methods by way of a detailed example before briefly outlining FEMs. In Section 6.2, we define abstract **finite elements**. In Section 6.3, we use finite elements to construct special finite-dimensional subspaces $\Lambda_h^k \subseteq L^2\Lambda^k$. In Section 6.4, we characterize those Λ_h^k that are contained inside $H\Lambda^k$.

6.1 Introduction to Galerkin Methods and Finite Element Methods

We illustrate the use of Galerkin methods by going through a simple example step-by-step. After this example is done, we explain in broad strokes how FEMs arise as special Galerkin methods.

Let Ω denote a polygonal domain in \mathbb{R}^2 , let $\Delta: C^\infty\Lambda^0(\Omega) \rightarrow C^\infty\Lambda^0(\Omega)$ denote the Laplace operator $\Delta = \delta d = -[\partial_x^2 + \partial_y^2]$, and let $f \in C^\infty\Lambda^0(\Omega)$. Consider solving Poisson’s equation

$$\Delta u = f \tag{6.1.1}$$

subject to the Dirichlet boundary condition

$$\text{tr}_{\partial\Omega} u = 0. \tag{6.1.2}$$

To construct a numerical solution to the problem above, we would like to approximate u in terms of some family of simple functions. For instance, triangulating Ω and choosing our approximating functions to be continuous and piecewise smooth with respect to this triangulation seems like a good idea (indeed, we come back to this idea in a few pages). If we were to study such piecewise smooth functions, however, our

approximate solution would not be \mathcal{C}^2 , or even \mathcal{C}^1 , unless we do some extra work (specifically, matching derivatives of our approximate solution at interfaces). We could also consider using approximating functions that are linear combinations of eigenfunctions of Δ , but of course such an approach is not practical if these eigenfunctions cannot be analytically computed. All in all, we see that we cannot necessarily preserve differentiability by passing to simple approximating functions.

Accordingly, we switch to studying (6.1.1) through a **weak form**. This effectively amounts to using integration by parts in order to lessen the regularity requirements on u . To begin deriving the weak form of the Poisson equation, pick any **test function** $v \in \mathcal{C}_c^\infty \Lambda^0(\Omega)$ and take the L^2 inner product of (6.1.1) with v to obtain

$$\langle \Delta u, v \rangle_{L^2} = \langle f, v \rangle_{L^2}.$$

Next, use integration by parts on the left-hand side to obtain

$$\langle du, dv \rangle_{L^2 \Lambda^1} = \langle f, v \rangle_{L^2}.$$

To solve the problem above, all we need is $u \in \mathring{H}\Lambda^0$ (if we are willing to understand the derivatives in the weak sense). Accordingly, we can recast our boundary value problem thus: determine $u \in \mathring{H}\Lambda^0(\Omega)$ such that

$$\langle du, dv \rangle_{L^2 \Lambda^1} = \langle f, v \rangle_{L^2} \quad \forall v \in \mathcal{C}_c^\infty \Lambda^0. \quad (6.1.3)$$

As it turns out, we can go even further with “weakening” the problem. Since $\mathcal{C}_c^\infty \Lambda^\ell$ is dense in $\mathring{H}\Lambda^\ell$ by Lemma 3.4.3, solving (6.1.3) is equivalent to solving

$$\langle du, dv \rangle_{L^2 \Lambda^1} = \langle f, v \rangle_{L^2} \quad \forall v \in \mathring{H}\Lambda^0. \quad (6.1.4)$$

Note how the test functions v satisfy the homogeneous Dirichlet boundary condition. If our solution u was supposed to satisfy *inhomogeneous* Dirichlet conditions, v would still have to vanish on the boundary. However, if we instead had Neumann boundary conditions on u , we would not require any constraints on the boundary values of v : if we did prescribe the boundary values of the test functions when dealing with Neumann conditions, then the boundary term in the weak form would reduce to a constant.

In numerical PDE theory, any boundary condition that results in one having to find weak solutions by restricting to a proper subset of $H\Lambda^k$ is said to be **essential**. Since the Dirichlet boundary condition for our formulation of the Poisson equation yield constraints on the space where we look for a solution u (that is, we demand that $u \in \mathring{H}\Lambda^0$, not just $H\Lambda^0$), this boundary condition is essential. If Neumann boundary conditions were imposed instead, we would not have enough constraints to expect a solution in any set smaller than $H\Lambda^0$. However, we would have boundary integrals appearing in the weak form (6.1.4); in light of the discussion in the above paragraph, test functions do not vanish on the boundary for Neumann conditions. Any boundary condition that leaves the solution space unconstrained but rather manifests itself in the weak form is said to be **natural**. Figuring out which boundary conditions are natural and which ones are essential is a critical step in deriving a correct Galerkin method. Note that Dirichlet conditions are not always essential, and Neumann conditions are not always natural; see [54, §2.2] for a canonical example.

Having completed our digression on boundary conditions, we return to studying the weak Poisson equation (6.1.4). We begin by modifying the notation. Let $A: \mathring{H}\Lambda^0 \times \mathring{H}\Lambda^0 \rightarrow \mathbb{R}$ denote the bilinear form $(u, v) \mapsto \langle du, dv \rangle_{L^2 \Lambda^1}$ and $F: \mathring{H}\Lambda^0 \rightarrow \mathbb{R}$ denote the functional $v \mapsto \langle f, v \rangle_{L^2}$. Then, (6.1.4) becomes

$$A(u, v) = F(v) \quad \forall v \in \mathring{H}\Lambda^0. \quad (6.1.5)$$

Now, we get to the heart of the Galerkin method. We choose a finite-dimensional subspace $\Lambda_h^0 \subseteq \mathring{H}\Lambda^0$ consisting of some simple functions and restrict the problem (6.1.5) to this subspace. That is, we solve the following problem for an approximate solution $u_h \in \Lambda_h^0$:

$$A(u_h, v) = F(v) \quad \forall v \in \Lambda_h^0. \quad (6.1.6)$$

This is now a problem in finite-dimensional linear algebra and, consequently, it can be phrased in terms of vectors and matrices. Pick a basis $\{\phi_j\}_j$ of Λ_h^0 . Then, form the **stiffness matrix** \mathbf{A} with entries

$$A_{ij} = A(\phi_i, \phi_j)$$

and the **load vector** \mathbf{F} with entries

$$F_i = F(\phi_i).$$

If we write

$$u_h = \sum_j u_j \phi_j$$

and define a vector \mathbf{u}_h whose entries are

$$(\mathbf{u}_h)_j = u_j,$$

then finding an approximate solution to our original problem just amounts to finding the solution to

$$\mathbf{A}\mathbf{u}_h = \mathbf{F}, \quad (6.1.7)$$

since \mathbf{A} is symmetric. From here, one can pick their favourite numerical method for solving linear systems to get \mathbf{u}_h (the conjugate gradient method is a good tool for solving this particular problem [33, Ch. 2]). This completes our description of the simplest Galerkin method for the scalar Poisson equation. We summarize the crucial steps of deriving a Galerkin method for a generic boundary value problem:

- 1) recast the governing PDE in a weak form, using integration by parts to “peel off” derivatives onto the test functions (thus weakening the regularity requirements on the approximate solution), and decide which boundary conditions are essential and which ones are natural;
- 2) restate the weak form of the PDE as a relationship between bilinear and linear (or, more generally, nonlinear) forms on an appropriate $H\Lambda^k$;
- 3) choose a finite-dimensional space Λ_h^k (often a subspace of $H\Lambda^k$) and solve the weak form on Λ_h^k using a numerical method for systems of algebraic equations (in the nonlinear case, a Newton-type root-finding method is often useful).

Remark 6.1.1. *Dirichlet boundary conditions appear in a Galerkin discretization as prescribed entries in the approximate solution vector \mathbf{u}_h .*

Galerkin methods provide an intuitive strategy for discretization, provided that we have a good choice of Λ_h^k . We are then left with the task of constructing such spaces in practice. This is the domain of the finite element method (FEM). In broad strokes, the FEM tells us how to construct Λ_h^k using a triangulation of the domain Ω and a set of rules for uniquely determining a form in $L^2\Lambda^k$ that is piecewise smooth with respect to this triangulation. These “rules” tell us how to reconstruct an element of Λ_h^k bit-by-bit from local data associated to each simplex in the triangulation.

However, we often want to make sure that the function space we obtain is actually contained in $H\Lambda^k$. This property is called **d-conformity**, with the d reminding us of the importance of weak exterior differentiability. In more customary numerical analysis jargon, d-conformity is sometimes called **$H(\text{div})$ -conformity** or **$H(\text{curl})$ -conformity** depending on the form degree k ; this nomenclature arises because the exterior derivative reduces to the divergence or curl in familiar situations. As a serious health warning, most finite element researchers use the term “conformity” in practice to mean that the subspace Λ_h^k consists of forms whose coefficients all have L^2 weak derivatives. Such FEMs are widely regarded as “vanilla”: good for basic problems, but not especially versatile. The reader should be aware that our development of d-conforming FEMs using the spaces $H\Lambda^k$ includes mixed methods and discontinuous Galerkin methods (these terms are defined later, and the unfamiliar reader should not panic). These are d-conforming by our definition, but not “conforming” by the more common definition.

We must also prove that the subspace Λ_h^k has satisfactory approximation properties. For instance, the error in approximation of $\omega \in H\Lambda^k$ by $\omega_h \in \Lambda_h^k$ should tend to zero as we make our triangulation of Ω more and more refined. The satisfaction of such approximation conditions is briefly considered in the concrete cases discussed in later chapters. There are also algebraic properties related to Hilbert complex structure that Λ_h^k ought to satisfy, and these are also discussed in later chapters.

Remark 6.1.2. *If the approximate solution and test functions live in different spaces, some authors call the scheme a **Petrov-Galerkin method** (see for example [88]). So, when reading numerical analysis papers, one may come across FEMs that the authors explicitly say are not Galerkin methods. For the purposes of this thesis, I call any method that includes solving a weak formulation on a finite-dimensional space a Galerkin method, regardless of where the test functions live. The reader experienced in numerical analysis may find this convention sloppy, but for the present purposes it doesn't cause too much trouble.*

To summarize, in this section I introduced the Galerkin method for the discretization and numerical solution of PDEs. This involves reducing to a weak form and solving some algebraic equations on a finite-dimensional space Λ_h^k . Then, we qualitatively introduced the FEM, which helps us choose a Λ_h^k using a triangulation of the domain.

6.2 Finite Elements

In this section we introduce the definition of a finite element, the basic building block for FEMs. Finite elements provide a sort of standard rule for locally constructing piecewise smooth functions and forms. The procedure for stitching these local data together to obtain functions defined over the whole domain is the province of the next section. The formal definition of finite elements that we present here was popularized in Ciarlet's 1978 textbook [20].

Definition 6.2.1. *A **finite element** consists of the following data:*

- 1) *a simply connected, compact set $T \subseteq \mathbb{R}^n$ with Lipschitz boundary and nonempty interior;*
- 2) *a finite-dimensional subspace $V(T) \subseteq C^\infty \Lambda^k(T)$, called the **shape space**;*
- 3) *a set of linear functionals $\ell_i^T: V(T) \rightarrow \mathbb{R}$, called the **degrees of freedom (DOFs)**, forming a basis of the dual space $V^*(T)$.*

We refer to a finite element by way of the triplet $(T, V(T), \{\ell_i^T\}_i)$

In the numerical analysis literature, T is sometimes synecdochically called a “finite element”, or even just an “element”. For the purposes of most applications, T is either a simplex or hypercube. From here on, we only study the case of T as a simplex. For the purposes of this exposition, the shape spaces always consist of differential forms with polynomial coefficients (we often refer to such forms succinctly as “polynomial differential forms”).

We often refer to the requirement that $\{\ell_i^T\}_i$ be a basis by saying that the DOFs must be **unisolvent**. Unisolvence is equivalent to showing that, given arbitrary scalars $v_1, \dots, v_{\dim V(T)}$, there exists a unique $v \in V(T)$ such that $\ell_i^T(v) = v_i$ for all i . Hence, every $v \in V(T)$ can be uniquely reconstructed using just the numbers $\ell_i^T(v)$. Often, we introduce DOFs by defining their span, rather than the basis vectors ℓ_i^T themselves. Once we prove that the span is the whole of $V^*(T)$, we can select a basis useful for practical computations.

The next result gives us a useful test for determining whether or not a set of functionals on $V(T)$ is actually unisolvent.

Lemma 6.2.2. *Let W be a finite-dimensional real vector space, and let $A \subseteq W^*$ be a subspace. Define a subspace U of W by*

$$U \doteq \{w \in W \mid \alpha(w) = 0 \ \forall \alpha \in A\} = \bigcap_{\alpha \in A} \ker \alpha. \quad (6.2.1)$$

The subspace U is equal to 0 if and only if $A = W^$.*

Proof. The implication “ \Leftarrow ” is obvious. To show the other direction, first let $\dim W = N$ and $\dim A = M$. Pick a basis $\{\mathbf{e}_j\}_{j=1}^N$ of W and a corresponding dual basis $\{\mathbf{e}^j\}_{j=1}^N$ of W^* . That is,

$$\mathbf{e}^j(\mathbf{e}_k) = \delta_{jk}.$$

Finally, pick a basis $\{\alpha^i\}_{i=1}^M$ of A . For all i , therefore, we may write $\alpha^i = \alpha_j^i \mathbf{e}^j$ for some $\alpha_j^i \in \mathbb{R}$. Now, consider the linear operator $L: W \rightarrow \mathbb{R}^M$ whose $M \times N$ matrix has an i, j entry given by $L_j^i = \alpha_j^i$. We show that $\ker L = U$.

If $w = w^j \mathbf{e}_j \in U$ then for all $i = 1, \dots, M$ we have that the i^{th} component of $L(w)$ in the standard basis on \mathbb{R}^M is given by

$$(L(w))_i = \alpha_j^i w^j = \alpha_i(w).$$

So, $L(w) = 0$ if and only if $\alpha_i(w) = 0$ for all i . Since $\{\alpha_i\}_{i=1}^M$ is a basis for A , this tells us that $\ker L = U$.

By the Rank–Nullity Theorem (see for example [10, Cor. 3.1.2]), $N = \dim U + \text{rank } L$ and by hypothesis $\dim U = 0$. We conclude that $N = \text{rank } L$. Since $\{\alpha_i\}_{i=1}^M$ is a basis, we know that $\text{rank } L = M$, thus $M = N$. \square

We conclude this section by providing some examples of finite elements. Chapters 7 and 8 are devoted to constructing more general finite elements.

Example 6.2.3. Let $T = [0, 1]$, let $V(T)$ be the space of degree one polynomials on T , and let eval_p denote the map taking smooth functions on T to their value at p . Then, we define linear functionals on $V(T)$ by

$$\ell_1^T = \text{eval}_0 \quad \text{and} \quad \ell_2^T = \text{eval}_1.$$



Figure 6.21: DOFs for the Lagrange finite element in Example 6.2.3. Solid dots represent the points p such that eval_p is a DOF.

The ℓ_i^T are indeed unisolvent: they are clearly linearly independent, and Lemma 6.2.2 tells us that they span $V(T)^*$. Hence, the triplet $(T, V(T), \{\ell_1^T, \ell_2^T\})$ defines a finite element, referred to as the **degree 1 Lagrange element** or, in compressed form, CG(1) (the CG stands for “continuous Galerkin”). See Figure 6.21 for an illustration of the DOFs for CG(1). The generalization of the CG(1) element to triangles (rather than just intervals) appeared in Courant’s 1943 paper [23], making it the first finite element ever studied. In Courant’s work, however, the term “finite element” is not used.

Example 6.2.4. In this example, we define a finite element where the shape space consists of 1–forms rather than scalars. Actually, we define the shape space as a set of vector fields and implicitly use the canonical correspondence between vectors and 1–forms in \mathbb{R}^2 to maintain consistency with Definition 6.2.1. Let $T = T_{\text{ref},2} = [(0, 0) (1, 0) (0, 1)]$ and let

$$V(T) \doteq \{\mathbf{u} \in \mathfrak{X}(T) \mid \mathbf{u} = (b_1 + ax, b_2 + ay), \quad a, b_1, b_2 \in \mathbb{R}\}.$$

So, $V(T)$ is a 3–dimensional vector space. Denote the outward normal on the boundary of T by \mathbf{n} , then define linear functionals on $V(T)$ by

$$\ell_1^T(\mathbf{u}) = \text{eval}_{(0, \frac{1}{2})}(\mathbf{u} \cdot \mathbf{n}), \quad \ell_2^T(\mathbf{u}) = \text{eval}_{(\frac{1}{2}, 0)}(\mathbf{u} \cdot \mathbf{n}), \quad \text{and} \quad \ell_3^T(\mathbf{u}) = \text{eval}_{(\frac{1}{2}, \frac{1}{2})}(\mathbf{u} \cdot \mathbf{n}).$$

We claim that the ℓ_i^T are unisolvent. If the ℓ_i^T were not linearly independent then we could find constants $c_1, c_2, c_3 \in \mathbb{R}$ such that, for all $a, b_1, b_2 \in \mathbb{R}$,

$$(c_1, c_2, c_3) \cdot (b_1, b_2, b_1 + b_2 + a) = 0.$$

So, (c_1, c_2, c_3) is orthogonal to every vector in \mathbb{R}^3 . Therefore, $c_i = 0$ for all i .

Next, we show that the ℓ_i^T span the whole dual space. If all of the ℓ_i^T vanish on

$$\mathbf{u} = (b_1 + ax, b_2 + ay)$$

then in particular we have that

$$\begin{aligned} \ell_1^T(\mathbf{u}) = 0 &\Rightarrow b_1 = 0 \text{ and} \\ \ell_2^T(\mathbf{u}) = 0 &\Rightarrow b_2 = 0, \end{aligned}$$

hence

$$\ell_3^T(\mathbf{u}) = 0 \Rightarrow a = 0.$$

An application of Lemma 6.2.2 completes the argument. Then, we have indeed defined a finite element, called the **lowest Raviart–Thomas element** and denoted by RT(1). In other work, the lowest Raviart–Thomas element is denoted by RT(0). RT(1) was introduced by Raviart and Thomas in a 1975 conference paper [73] (the proceedings of the conference became more widely available after being published as lecture

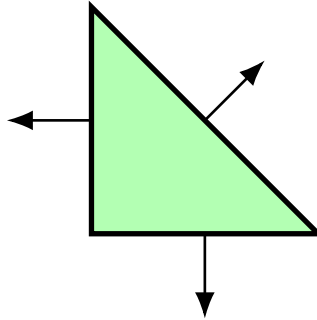


Figure 6.22: The Raviart–Thomas finite element $RT(1)$ in Example 6.2.4. The arrows represent the DOFs (evaluation of normal components at base point of the arrow).

notes in 1977). Raviart and Thomas were actually able to build a finite element like $RT(1)$ for arbitrary shape polynomial degree, and we provide a modern perspective on their construction in Chapter 8.

Figure 6.22 gives a visual representation of the DOFs for $RT(1)$. We could also define DOFs by way of taking point evaluations of *tangential* components, in which case we would be constructing the **lowest Nédélec edge element of the first kind**, denoted $N1^e(1)$. We discuss the many different finite elements bearing Nédélec’s name in the next chapter.

We introduce many more finite elements in later chapters and develop techniques for proving unisolvence in an elegant, unified manner. We also extend the Lagrange and Raviart–Thomas elements studied above to arbitrary polynomial degree. I would like to point out, however, that we cannot always choose DOFs to be simple point evaluation functionals. In full generality the DOFs for most finite elements are defined by taking a k -form to its weighted integral over some subsimplex $f \in \Delta(T)$. In this case we say that the DOF is **associated** to the face f .

Remark 6.2.5. *The reader may have noticed that the finite elements studied above were defined with respect to reference simplices. In practice, of course, we would like to consider simplices of varying sizes and spatial positions. However, it is enough to only define a finite element on the reference n -simplex in order to get finite elements for every n -simplex in a triangulation; as long as we’re dealing with polyhedral domains, each n -simplex T in our triangulation may be mapped to $T_{\text{ref},n}$ by way of an invertible affine transformation, so all computations on T can be pulled back to $T_{\text{ref},n}$. For more information on the relationship between affine transformation and finite elements (specifically, the finite elements we define in Chapter 8), see [4, §3.4].*

In studying the above examples the reader may have guessed that there are many finite elements, and keeping track of them all may be a steep task. FEEC shows how plenty of the most common finite elements are derived from two main families of differential forms with polynomial coefficients. Keeping track of the different finite elements and their relationships with one another is therefore facilitated by adopting a geometric perspective. Further, FEEC paved the way for the creation of the Periodic Table of the Finite Elements [6], which concisely summarizes many common finite elements on a single convenient chart. The Periodic Table also contains some subtly encoded algebraic information, discussed in further detail in Chapter 9.

6.3 Local-to-Global Maps and Finite Element Spaces

Suppose that Ω is a polyhedral domain equipped with a triangulation \mathcal{T}_h . Further, suppose that for all $T \in \Delta_n(\mathcal{T}_h)$ we have a finite element $(T, V(T), \{\ell_i^T\})$. In this section, we turn to constructing finite-dimensional subspaces $\Lambda_h^k \subseteq L^2\Lambda^k(\bar{\Omega})$ by sewing together the finite elements associated to each component simplex T of the triangulation. Our strategy for building Λ_h^k begins by selecting those $\omega \in L^2\Lambda^k(\bar{\Omega})$ satisfying $\omega|_T \in V(T) \forall T \in \Delta_n(\mathcal{T}_h)$. We call the subspace of all such forms $V(\mathcal{T}_h)$. From there, we narrow ourselves down to dealing with $\omega \in V(\mathcal{T}_h)$ satisfying a certain compatibility condition with respect to “equivalent” DOFs. Most of our work in this section involves defining this compatibility condition rigorously.

Remark 6.3.1. *Our conventions for visually representing the DOFs are taken directly from [54, Ch.3]; the DOF diagrams presented here are essentially identical to those found in this book.*

Remark 6.3.2. *The forms in $V(\mathcal{T}_h)$ are, a priori, multiply-valued on inter-element boundaries. Recall, however, that strictly speaking these forms are actually equivalence classes, so this apparent multi-valuedness is no cause for alarm. If we wanted to be more careful, we could say that $V(\mathcal{T}_h)$ consists of those forms in $L^2\Lambda^k(\bar{\Omega})$ such that, for all $T \in \Delta_n(\mathcal{T}_h)$, there exists $\xi \in V(T)$ and a representative $\tilde{\omega}$ of ω such that $\xi = \tilde{\omega}|_T$.*

Recall that, by prescribing values for every DOF on the simplex T , we determine a unique $v \in V(T)$. We would now like to introduce the notion of **global DOFs** from which we can analogously reconstruct certain k -forms in $V(\mathcal{T}_h)$. To distinguish the DOFs of each finite element T from these global DOFs, we call the ℓ_i^T **local DOFs**. Our approach to defining global DOFs mostly follows that of [54, Ch.2], extending the special case discussed explicitly in [20, Ch.2].

Definition 6.3.3. *Let $N_T = \dim V(T)$ and*

$$N = \sum_{T \in \Delta_n(\mathcal{T}_h)} N_T.$$

1) *A **local-to-global map** is a map $\iota_T: \{1, \dots, N_T\} \rightarrow \{1, \dots, N\}$.*

2) *Pick a local-to-global map ι_T for $T \in \Delta_n(\mathcal{T}_h)$, and choose $i \in \{1, \dots, N_T\}$. We then have a multi-valued linear functional $\ell_{\iota_T(i)}$ on $V(\mathcal{T}_h)$ defined by*

$$\ell_{\iota_T(i)}(\omega) = \ell_i^T(\omega|_T).$$

*Any such $\ell_{\iota_T(i)}$ is called a **global DOF**.*

Let ℓ_i^T and $\ell_j^{T'}$ be local DOFs associated to simplices T and T' , respectively. Suppose that $\iota_T(i) = \iota_{T'}(j)$. Then, a global DOF $\ell_{\iota_T(i)}$ is single-valued at $\omega \in V(\mathcal{T}_h)$ if and only if

$$\ell_i^T(\omega|_T) = \ell_j^{T'}(\omega|_{T'}).$$

That is, multi-valuedness only arises when local DOFs corresponding to the same global DOF disagree at ω . Restricting ourselves only to dealing with $\omega \in V(\mathcal{T}_h)$ on which the global DOFs are single-valued seems like a good choice for specifying Λ_h^k , but we make some more remarks on Definition 6.3.3 before going further.



Figure 6.31: Pictorial representation of global DOFs on the unit interval $[0, 1]$ with respect to the triangulation $[0, \frac{1}{2}] \cup [\frac{1}{2}, 1]$. Solid dots mark the points associated to the global DOFs.

Example 6.3.4. Local-to-global maps are not a deep construction. Indeed, they amount to nothing more than just re-labeling local DOFs to account for the finite elements over the whole of \mathcal{T}_h . For instance, suppose that we partition $[0, 1]$ into the two subintervals $T = [0, \frac{1}{2}]$ and $T' = [\frac{1}{2}, 1]$, define shape spaces to be degree one polynomials on T or T' , and choose local DOFs to be endpoint evaluation after Example 6.2.3. That is, we define a degree 1 Lagrange finite element on each subinterval. Using the notation

$$\begin{aligned} \ell_1^T &= \text{eval}_0, & \ell_2^T &= \text{eval}_{\frac{1}{2}} & \text{and} \\ \ell_1^{T'} &= \text{eval}_{\frac{1}{2}}, & \ell_2^{T'} &= \text{eval}_1, \end{aligned}$$

we define local-to-global maps by

$$\begin{aligned} \iota_T(1) &= 1, & \iota_T(2) &= 2 & \text{and} \\ \iota_{T'}(1) &= 2, & \iota_{T'}(2) &= 3. \end{aligned}$$

See Figure 6.31. This choice ensures that we do not count evaluation at $\frac{1}{2}$ twice as a global DOF, which is an intuitive requirement: we shouldn't count evaluating at the same point, integrating over the same edge, et cetera as different global DOFs. The restriction to those forms in $V(\mathcal{T}_h)$ on which local DOFs corresponding to the same global DOF agree, therefore, is a very natural choice.

Now, the motivation for the precise definition of a finite element space should be clear.

Definition 6.3.5. Let $\Omega \subseteq \mathbb{R}^n$ be a polyhedral domain with given triangulation \mathcal{T}_h . Suppose that for each $T \in \Delta_n(\mathcal{T}_h)$ we have a finite element $(T, V(T), \{\ell_i^T\}_i)$ and a local-to-global map ι_T . We define the **finite element space** associated to this data to be the finite-dimensional subspace $\Lambda_h^k \subseteq L^2\Lambda^k(\bar{\Omega})$ consisting of those $\omega \in L^2\Lambda^k(\bar{\Omega})$ satisfying

- 1) for all $T \in \Delta_n(\mathcal{T}_h)$, $\omega|_T \in V(T)$ and
- 2) all global DOFs are single-valued on ω .

The finite element space, therefore, is simply the dual of the span of all global DOFs. Imposing the constraint of single-valuedness means that number of total global DOFs, and therefore the dimension of Λ_h^k , should be less than N (in the notation of Definition 6.3.3). In Example 6.3.4, for instance, we have that $N = 4$ but the total number of global DOFs is 3 (we remark that there does not need to be a global DOF for every $i \in \{1, \dots, N\}$). Further, this example illustrates how the specific DOFs and the constraint of single-valuedness encode information about the regularity of forms inside Λ_h^k : if a piecewise linear scalar function has $\text{eval}_{\frac{1}{2}}$ as a single-valued global DOF, then that function must automatically be continuous across the whole of $[0, 1]$. Though the functions in the finite element space corresponding to this example are continuous, we have no guarantee that a general finite element space is contained in $C\Lambda^k(\Omega)$ or $H\Lambda^k(\Omega)$.

Recall that we said a local DOF ℓ_i^T was associated to $f \in \Delta(T)$ if ℓ_i^T could be represented as a weighted integral over f . Similarly, we can say that a global DOF ℓ_i is **associated to** $f \in \mathcal{T}_h$ if

- 1) there exists $T \in \Delta_n(\mathcal{T}_h)$ such that $f \in \Delta(T)$, and
- 2) there exists a local DOF ℓ_j^T associated to f such that T 's local-to-global map takes ℓ_j^T to ℓ_i .

One may worry that, if there are distinct $T, T' \in \mathcal{T}_h$ such that $f \in \Delta(T) \cap \Delta(T')$, then the above definition is a little ambiguous. In particular, we may think that a DOF ℓ_i associated to f gives different values on $\omega \in \Lambda_h^k$ depending on whether we compute $\ell_i(\omega)$ using $\omega|_T$ or $\omega|_{T'}$ (see Definition 6.3.3). However, since global DOFS are single-valued on the finite element space, there is actually no ambiguity at all. Every DOF we consider in this thesis is associated to a particular $f \in \mathcal{T}_h$.

By construction of the global DOFs $\{\ell_j\}_j$, Λ_h^k admits a canonical **nodal basis** $\{\phi_j\}_j$, defined to be the dual basis with respect to the DOFs. That is, the ϕ_j are all uniquely determined by

$$\ell_i(\phi_j) = \delta_{ij}.$$

Accordingly, all $\omega \in \Lambda_h^k$ can be written as

$$\omega = \sum_j \ell_j(\omega) \phi_j.$$

The following property of nodal bases makes them especially desirable for computational purposes.

Proposition 6.3.6. *Suppose that the global DOF ℓ_j is associated to $f \in \mathcal{T}_h$. Let ϕ_j denote the corresponding nodal basis vector. For all $T \in \Delta_n(\mathcal{T}_h)$ such that $f \notin \Delta(T)$, $\phi_j|_T = 0$.*

Proof. Pick any admissible T and let its local-to-global map be denoted by ι_T . Then, none of the local DOFs ℓ_i^T are mapped to the global DOF ℓ_j . By definition of the dual basis and global DOFs,

$$0 = \ell_{\iota_T(i)}(\phi_j) = \ell_i^T(\phi_j|_T).$$

Since the local DOFs are unisolvent, $\phi_j|_T = 0$. □

In light of the above Proposition, we say that the nodal basis is **local**: each element is supported only near the subsimplex associated to the corresponding DOF. Locality of the nodal basis is computationally significant because it leads to sparse stiffness matrices (see Section 6.1). Many clever algorithms for solving sparse linear systems exist (see [47, Ch. 11] for some elementary examples) and are readily available through software packages such as PETSc [9] (which is directly accessible in FEniCS and Firedrake).

Before concluding, I would like to comment on how some of the constructions developed in this section appear (or don't appear) in practical computer implementations. When using FEniCS or Firedrake, the user specifies the parameters of the triangulation and the finite elements to be used (chosen from a catalogue using a simple keyword). The local-to-global maps are “under the hood”; they do not need to be specified by the user. Further, these maps are always chosen so that local DOFs corresponding to the same operation on the same subsimplex are mapped to the same global DOF, as we discussed in the example above Definition 6.3.5.

In practice, trying to plot solutions to a PDE using just a list of global DOFs requires a measure of tedious work, as there is no canonical labeling scheme for the local-to-global maps in either of the software packages mentioned. Additionally, because DOFs are averages just as often as they are point evaluations, merely assembling the DOFs doesn't give us enough information for plotting. Accordingly, to plot numerical solutions we often just create a Python array whose entries are approximations of the solution's value at each vertex of each simplex in the triangulation.

6.4 Characterization of d-Conforming Finite Element Spaces

As discussed in the previous section, we have absolutely no a priori guarantee that a finite element space Λ_h^k gives rise to a “good” (in a sense to be defined more precisely in Chapter 9) Galerkin scheme. In full generality, we cannot even be sure of the d-conformity property $\Lambda_h^k \subseteq H\Lambda^k(\Omega)$. Accordingly, we would like to find some conditions to help us establish $\Lambda_h^k \subseteq H\Lambda^k(\Omega)$. Towards this goal, we have a nice characterization of piecewise smooth k -forms that are in $H\Lambda^k(\Omega)$.

Theorem 6.4.1. *Let \mathcal{T}_h be a triangulation of $\Omega \subseteq \mathbb{R}^n$ and denote the elements of $\Delta_n(\mathcal{T}_h)$ by T_i . Suppose $\omega \in L^2\Lambda^k(\overline{\Omega})$ restricts to a smooth k -form on each component n -simplex T_i . We have that $\omega \in H\Lambda^k(\Omega)$ if and only if for all i, j and all $F \in \Delta_{n-1}(T_i) \cap \Delta_{n-1}(T_j)$, we have $\text{tr}_F(\omega|_{T_i}) = \text{tr}_F(\omega|_{T_j})$ (that is, the facet traces of ω are **single-valued**).*

Proof. Let us first restrict ourselves to the case of a triangulation $T_1 \cup T_2$, so that the T_i 's share only a single common face F . Choose an arbitrary smooth $(k+1)$ -form η that is compactly supported in Ω . Then, integrating by parts (valid on each T_i because $\omega|_{T_i}$ is smooth), we have that

$$\begin{aligned} \int_{T_i} d(\omega|_{T_i}) \wedge \star\eta &= \int_{T_i} \omega|_{T_i} \wedge \star\delta\eta + \int_{\partial T_i} \text{tr}_{\partial T_i}(\omega|_{T_i} \wedge \star\eta) \\ &= \int_{T_i} \omega|_{T_i} \wedge \star\delta\eta + \int_F \text{tr}_F(\omega|_{T_i}) \wedge \text{tr}_F(\star\eta). \end{aligned}$$

We have used that the support of η is contained inside Ω to go from the first line to the second one. Now, without loss of generality we suppose that F is oriented according to the induced orientation on ∂T_1 (and therefore opposite to the induced orientation on ∂T_2). Adding up the above expressions for $i = 1$ and $i = 2$ gives us that

$$\int_{T_1} d(\omega|_{T_1}) \wedge \star\eta + \int_{T_2} d(\omega|_{T_2}) \wedge \star\eta = \langle \omega, \delta\eta \rangle_{L^2\Lambda^k} + \int_F \text{tr}_F(\omega|_{T_1} - \omega|_{T_2}) \wedge \text{tr}_F \star\eta. \quad (6.4.1)$$

Suppose $\text{tr}_F(\omega|_{T_1}) = \text{tr}_F(\omega|_{T_2})$. Then, (6.4.1) reduces to

$$\int_{T_1} d(\omega|_{T_1}) \wedge \star\eta + \int_{T_2} d(\omega|_{T_2}) \wedge \star\eta = \langle \omega, \delta\eta \rangle_{L^2\Lambda^k}. \quad (6.4.2)$$

Let $\xi \in \Lambda^{k+1}(\Omega)$ be defined by $\xi|_{T_i} = d(\omega|_{T_i})$ for $i = 1, 2$. Then, by (6.4.2)

$$\langle \xi, \eta \rangle_{L^2\Lambda^{k+1}} = \langle \omega, \delta\eta \rangle_{L^2\Lambda^k}.$$

Since η was arbitrary, we conclude that $\xi = d\omega$. Further, the quantity

$$\langle d\omega, d\omega \rangle_{L^2\Lambda^{k+1}} = \sum_i \int_{T_i} d(\omega|_{T_i}) \wedge \star d(\omega|_{T_i})$$

is finite because it is the sum of finitely many finite numbers (each $d(\omega|_{T_i})$ is a smooth form on the compact set T_i). So, $d\omega \in L^2\Lambda^{k+1}(\Omega)$ and therefore $\omega \in H\Lambda^k(\Omega)$.

Conversely, if $\omega \in H\Lambda^k$ then in particular ω is weakly differentiable. Further, since ω is piecewise smooth it must be true that $(d\omega)|_{T_i} = d(\omega|_{T_i})$. Therefore, we can combine the two integrals on the

left-hand side of (6.4.1). Accordingly, for all smooth $(k + 1)$ -forms η compactly supported inside Ω , we have that

$$\langle d\omega, \eta \rangle_{L^2\Lambda^{k+1}} = \langle \omega, \delta\eta \rangle_{L^2\Lambda^k} + \int_F \operatorname{tr}_F (\omega|_{T_1} - \omega|_{T_2}) \wedge \operatorname{tr}_F \star \eta. \quad (6.4.3)$$

However, ω is weakly differentiable, so we also have

$$\langle d\omega, \eta \rangle_{L^2\Lambda^{k+1}} = \langle \omega, \delta\eta \rangle_{L^2\Lambda^k}. \quad (6.4.4)$$

Using (6.4.3) and (6.4.4) together, we obtain

$$\int_F \operatorname{tr}_F (\omega|_{T_1} - \omega|_{T_2}) \wedge \operatorname{tr}_F \star \eta = 0.$$

Since the smooth form η is arbitrary, this tells us that $\operatorname{tr}_F (\omega|_{T_1}) = \operatorname{tr}_F (\omega|_{T_2})$. The proof of the special case is now complete.

Now, we consider a general triangulation $\cup_i T_i$. For any $F \in \Delta_{n-1}(\cup_i T_i)$ we know by definition of a triangulation that either $F \subseteq \partial\Omega$ or $F \cap \partial\Omega$ has measure zero; that is, every facet is either a boundary or interior facet. Additionally, we know that for every interior facet F there exists a unique pair of simplices, T_+ and T_- , such that $F \in \Delta_{n-1}(T_+) \cap \Delta_{n-1}(T_-)$. Without loss of generality we assume that F is oriented compatibly with the orientation induced by T_+ , similar to the simplifying assumption we made in the special case.

Choose $\eta \in \mathcal{C}_c^\infty \Lambda^k(\Omega)$. Integrating by parts on each T_i and using the boundary conditions on η , we have that

$$\sum_i \int_{T_i} d(\omega|_{T_i}) \wedge \star \eta = \sum_i \int_{T_i} \omega \wedge \star \delta \eta + \sum_{\substack{\text{Interior} \\ \text{Facets}}} \int_F \operatorname{tr}_F (\omega|_{T_+} - \omega|_{T_-}) \wedge \operatorname{tr}_F \star \eta. \quad (6.4.5)$$

If the facet traces of ω are single-valued then we can just repeat the steps of the special case to see that $\omega \in H\Lambda^k(\Omega)$. On the other hand, if $\omega \in H\Lambda^k(\Omega)$ we may pick $\eta \in \mathcal{C}_c^\infty \Lambda^k(T_i \cup T_j) \subseteq \mathcal{C}_c^\infty \Lambda^k(\Omega)$ so that (6.4.5) reduces to (6.4.1) upon re-labeling. Then, using the arguments of the special case again, we know the facet traces are single-valued. The general proof is now complete. \square

Corollary 6.4.2. *All $\omega \in \Lambda_h^k$ have single-valued traces on facets if and only if $\Lambda_h^k \subseteq H\Lambda^k(\Omega)$.* \square

A result of Crouzeix and Raviart, discussed in [20, Remark 2.3.7, pp. 91–92] and [54, Ch. 3], shows that there exist finite element spaces of scalar functions where the facet traces are not single-valued.

At first glance, the condition for d-conformity established in Corollary 6.4.2 seems excessively restrictive. However, one must recognize that the condition of a k -form having single-valued traces on facets becomes weaker and weaker as k increases. Certainly, any scalar function with single-valued facet traces must be continuous across the whole domain. Conversely, if $\omega \in H\Lambda^n(\Omega) = L^2\Lambda^n(\Omega)$ then $\operatorname{tr}_F \omega = 0$ for all facets F , so all top forms trivially have single-valued facet traces. Accordingly, single-valued facet traces on k -forms starts as a very rigid restriction when $k = 0$, and eventually becomes meaningless when $k = n$.

For $0 < k < n$, the situation is somewhere in between. For example, if ω is a piecewise smooth 1-form on a triangulated polygonal domain in \mathbb{R}^2 , ω has single-valued traces on edges if and only if the *tangential* component of ω across each element boundary is continuous. To see this we can use normal-tangential

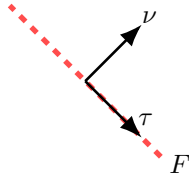


Figure 6.41: The normal–tangential coordinate system defined with respect to the facet F .

coordinates (ν, τ) , defined with respect to an edge F . That is, ν is the displacement perpendicular to F and τ is the displacement parallel to F (see Figure 6.41). In (ν, τ) coordinates,

$$\omega = \omega_{\perp} d\nu + \omega_{\parallel} d\tau.$$

Since $d\nu = 0$ on F , we conclude that

$$\text{tr}_F \omega = \omega_{\parallel} d\tau.$$

So, ω_{\perp} can have a jump discontinuity across F , but ω_{\parallel} must be continuous. Therefore, using differential forms provides us with a nice way of including all kinds of finite element spaces with funny inter–element continuity conditions without leaving the Sobolev spaces we defined earlier.

Chapter 7

Polynomial Differential Forms on Simplices

In the last chapter, we introduced FEMs in a general setting. We only constructed a handful of real examples of low-order finite elements, and our proofs that these examples actually satisfy Definition 6.2.1 relied on the DOFs being point evaluation functionals. In applications, we often need more complicated DOFs and higher-order shape functions. Therefore, we should seek a more unified approach to constructing finite elements; namely, an approach where unisolvence can be proven for arbitrary shape function polynomial degree. The development of such a unified picture of FEM theory is undertaken in the next chapter. In the present chapter, we focus on building a foundational knowledge of two families of polynomial differential forms defined over simplices. Upon mastering the theory of such polynomial forms, we will be in a much better position to confidently and succinctly define shape spaces and DOFs. The spaces we study in this section are discussed extensively in [4], though the presentation style borrows more from the review paper [3].

In Section 7.1, we prove some easy combinatorial identities that facilitate dimension-counting arguments used throughout this chapter and the next one. In Section 7.2, we introduce the simplest family of polynomial k -forms, denoted $\mathcal{P}_r \Lambda^k$ and referred to as the **regular** polynomial forms. In Section 7.3, we introduce the **Koszul operator** and use it to define another family of polynomial k -forms, denoted by $\mathcal{P}_r^- \Lambda^k$ and referred to as the **trimmed** polynomial forms.

7.1 Binomial Coefficient Identities

In this section, we present some simple results that are used frequently in the sequel. The statements proven are so short and simple that any attempt at discussing them would be mere filler. Accordingly, I have written this section in a spartan style.

Lemma 7.1.1.

$$\binom{a}{b} = \binom{a-1}{b-1} + \binom{a-1}{b}. \quad (7.1.1)$$

□

Lemma 7.1.2.

$$\sum_{j=0}^a \binom{a+1}{j+1} \binom{b-1}{j} = \binom{a+b}{b}. \quad (7.1.2)$$

Proof. By the Binomial Theorem, we know that the coefficient of x^a in $(1+x)^{a+b}$ is $\binom{a+b}{a}$. But, consider that

$$(1+x)^{a+b} = (1+x)^{a+1}(1+x)^{b-1} = \sum_{i=0}^{a+1} \sum_{j=0}^{b-1} \binom{a+1}{i} \binom{b-1}{j} x^{i+j}.$$

The above expression indicates that the coefficient of x^a in $(1+x)^{a+b}$ is also given by

$$\sum_{j=0}^a \binom{a+1}{a-j} \binom{b-1}{j} = \sum_{j=0}^a \binom{a+1}{j+1} \binom{b-1}{j}.$$

□

Lemma 7.1.3.

$$\binom{a}{b} \binom{b}{c} = \binom{a}{c} \binom{a-c}{b-c}. \quad (7.1.3)$$

Proof.

$$\begin{aligned} \binom{a}{b} \binom{b}{c} &= \frac{a!}{(a-b)!c!(b-c)!} \\ &= \frac{a!(a-c)!}{(a-b)!c!(b-c)!(a-c)!} \\ &= \binom{a}{c} \binom{a-c}{b-c}. \end{aligned}$$

□

Lemma 7.1.4. For $a \geq b$,

$$\sum_{j=0}^{a-b} \binom{a}{b+j} \binom{c}{j} = \binom{a+c}{a-b}. \quad (7.1.4)$$

Proof. We proceed as in Lemma 7.1.2. The coefficient of x^{a-b} in the polynomial $(1+x)^{a+c}$ is $\binom{a+c}{a-b}$, by the Binomial Theorem. However,

$$(1+x)^{a+c} = (1+x)^a(1+x)^c = \sum_{i=0}^a \sum_{j=0}^c \binom{a}{i} \binom{c}{j} x^{i+j},$$

hence the coefficient of x^{a-b} is also equal to

$$\sum_{j=0}^{a-b} \binom{a}{a-b-j} \binom{c}{j} = \sum_{j=0}^{a-b} \binom{a}{b+j} \binom{c}{j}.$$

□

7.2 $\mathcal{P}_r\Lambda^k$

In this section we study the most basic family of polynomial differential forms used to develop FEMs. Let $\mathcal{P}_r(x_1, \dots, x_n)$ denote the vector space of polynomials in n variables with real coefficients and degree less than or equal to r . Recall that a polynomial $p \in \mathcal{P}_r(x_1, \dots, x_n)$ is said to be **homogeneous (of degree r)** if, for all $\alpha \in \mathbb{R}$,

$$p(\alpha x_1, \dots, \alpha x_n) = \alpha^r p(x_1, \dots, x_n).$$

Let $\mathcal{H}_r(x_1, \dots, x_n)$ denote the real vector space of homogeneous polynomials in n variables with real coefficients and degree precisely r (together with the zero function to maintain the vector space structure). Given an n -simplex T , we use the notation $\mathcal{P}_r(T)$ and $\mathcal{H}_r(T)$ to denote the same sets of polynomials, now viewed as vector spaces of real-valued functions on T . Then,

$$\mathcal{P}_r\Lambda^0(T) = \bigoplus_{s=0}^r \mathcal{H}_s(T). \quad (7.2.1)$$

More generally, let $\mathcal{P}_r\Lambda^k(T)$ denote the subspace of $\Lambda^k(T)$ consisting of all k -forms with coefficients in $\mathcal{P}_r(T)$; the subspace $\mathcal{H}_r\Lambda^k(T)$ is defined similarly.

Proposition 7.2.1.

$$\dim \mathcal{P}_r(T) = \binom{n+r}{r}. \quad (7.2.2)$$

Proof. We start by finding $\dim \mathcal{H}_s(T)$ for any s , then use the decomposition (7.2.1) to complete the proof. Observe that computing $\dim \mathcal{H}_s(T)$ is equivalent to finding the number of distinct ways to fill $s+n-1$ slots with $n-1$ bars and s dots: a bar means we switch to the next variable in our monomial, and a dot means we multiply the expression by the variable we're currently considering. For example, the dot-bar diagrams corresponding to the monomials spanning $\mathcal{H}_2(x_1, x_2)$ are as follows:

$$\begin{aligned} \cdot \cdot \mid &\leftrightarrow x_1^2, \\ \cdot \mid \cdot &\leftrightarrow x_1 x_2, \\ \mid \cdot \cdot &\leftrightarrow x_2^2. \end{aligned}$$

Once we place the s dots in the $s+n-1$ slots, the placement of the bars is completely determined. Therefore, the number of ways we can distribute the dots and bars is just the number of ways we can choose where to put the dots. This proves that $\dim \mathcal{H}_s(T) = \binom{s+n-1}{s}$.

With (7.2.1) in mind, we have that

$$\begin{aligned} \dim \mathcal{P}_r\Lambda^0(T) &= \sum_{s=0}^r \dim \mathcal{H}_s(T) \\ &= \sum_{s=0}^r \binom{s+n-1}{s} \\ &= 1 + \binom{n}{1} + \binom{n+1}{2} + \dots + \binom{n+r-1}{r}. \end{aligned}$$

By iteratively applying Lemma 7.1.1, we see

$$\begin{aligned}
\binom{n+r}{r} &= \binom{n+r-1}{r} + \binom{n+r-1}{r-1} \\
&= \binom{n+r-1}{r} + \binom{n+r-2}{r-1} + \binom{n+r-2}{r-2} \\
&= \dots \\
&= \binom{n+r-1}{r} + \binom{n+r-2}{r-1} + \dots + \binom{n+1}{2} + \binom{n}{1} + 1.
\end{aligned}$$

The proof is now complete. \square

Corollary 7.2.2.

$$\dim \mathcal{P}_r \Lambda^k(T) = \binom{n+r}{r} \binom{n}{k}. \quad (7.2.3)$$

\square

7.3 The Koszul Operator and $\mathcal{P}_r^- \Lambda^k$

In this section we introduce a surprisingly powerful tool from homological algebra, the **Koszul operator**. This fantastic gadget allows us to prove statements about various polynomial cochain complexes with ease. Additionally, the Koszul operator reveals how to construct a second useful family of polynomial differential forms, the **trimmed** polynomial forms $\mathcal{P}_r^- \Lambda^k$. In forthcoming chapters, we see how trimmed forms are essential to finite element analysis in that they clarify how to select DOFs effectively given a choice of shape space.

Definition 7.3.1. *Let*

$$X \doteq x^i \frac{\partial}{\partial x^i} \in \mathfrak{X}(\mathbb{R}^n).$$

The Koszul operator $\kappa: \Lambda^k(\mathbb{R}^n) \rightarrow \Lambda^{k-1}(\mathbb{R}^n)$ is defined by

$$\kappa\omega \doteq X \lrcorner \omega. \quad (7.3.1)$$

The exterior derivative d generalizes the differential operators of vector calculus in \mathbb{R}^3 (∇ , $\nabla \cdot$, $\nabla \times$) to forms on arbitrary manifolds. In a much looser sense, the Koszul operator is analogously related to different notions of vector multiplication in \mathbb{R}^3 .

Proposition 7.3.2. *Under the correspondences $\text{Alt}^k(\mathbb{R}^3) \simeq \mathbb{R}^3$ for $k = 1, 2$, we have*

$$\omega \in \Lambda^2(\mathbb{R}^3) \Rightarrow \kappa\omega = X \times \omega \quad \text{and} \quad (7.3.2)$$

$$\omega \in \Lambda^1(\mathbb{R}^3) \Rightarrow \kappa\omega = X \cdot \omega. \quad (7.3.3)$$

Proof. Using Cartesian coordinates (x, y, z) on \mathbb{R}^3 , we know that $dx(X) = x$, $dy(X) = y$, and $dz(X) = z$. Now, pick $\omega \in \Lambda^k(\mathbb{R}^3)$. For $k=2$, we have

$$\begin{aligned}\kappa\omega &= X \lrcorner (a \, dy \wedge dz + b \, dz \wedge dx + c \, dx \wedge dy) \\ &= a(y \, dz - z \, dy) + b(z \, dx - x \, dz) + c(x \, dy - y \, dx) \\ &= (bz - cy) \, dx + (cx - az) \, dy + (ay - bx) \, dz \\ &= (x, y, z) \times (a, b, c) \\ &= X \times \omega.\end{aligned}$$

For $k=1$, the proof is trivial:

$$\kappa\omega = X \lrcorner (a \, dx + b \, dy + c \, dz) = ax + by + cz = (x, y, z) \cdot (a, b, c) = X \cdot \omega.$$

□

To get a better “feel” for how differential forms change under the action of the Koszul operator, we go through some concrete examples of how to calculate $\kappa\omega$ for a given ω .

Example 7.3.3. Let $\omega = xy \, dx + x^2 \, dy \in \Lambda^1(\mathbb{R}^2)$. Then,

$$\begin{aligned}\kappa\omega &= X \lrcorner (xy \, dx + x^2 \, dy) \\ &= xy(x) + x^2(y) \\ &= 2x^2y \in C^\infty(\mathbb{R}^2).\end{aligned}$$

Example 7.3.4. Let $\omega = (x^2 + y^2) \, dx \wedge dy \in \Lambda^2(\mathbb{R}^2)$. Then,

$$\begin{aligned}\kappa\omega &= X \lrcorner [(x^2 + y^2) \, dx \wedge dy] \\ &= (x^2 + y^2)(x) \, dy - (x^2 + y^2)(y) \, dx \\ &= -(x^2y + y^3) \, dx + (x^3 + xy^2) \, dy \in \Lambda^1(\mathbb{R}^2).\end{aligned}$$

Further,

$$\begin{aligned}\kappa^2(\omega) &= \kappa(\kappa\omega) \\ &= X \lrcorner [-(x^2y + y^3) \, dx + (x^3 + xy^2) \, dy] \\ &= -x^3y - xy^3 + x^3y + xy^3 \\ &= 0.\end{aligned}$$

These two examples illustrate some important general results about κ .

Lemma 7.3.5. κ maps $\mathcal{H}_r\Lambda^k$ to $\mathcal{H}_{r+1}\Lambda^{k-1}$ and $\mathcal{P}_r\Lambda^k$ to $\mathcal{P}_{r+1}\Lambda^{k-1}$.

□

Proposition 7.3.6. $\kappa^2 = 0$.

Proof. The proof is trivial: for $k = 0, 1$ the claim holds because contracting a vector field with a scalar gives the zero function, and for $k \geq 2$ the claim holds because ω is antisymmetric. \square

The next theorem describes the most important property of the Koszul operator, aside from its nilpotency.

Theorem 7.3.7. (*The Magic Formula*) *For all $\omega \in \mathcal{H}_r \Lambda^k$, we have that*

$$(\kappa d + d\kappa)\omega = (r + k)\omega. \quad (7.3.4)$$

Proof. [4, Thm. 3.1] contains two different proofs: one by direct calculation, and the other by an elegant geometric argument. We present the latter approach here, though it requires some acquaintance with Lie derivatives (see [51, pp. 227–231, pp. 372–373] for an introduction). First, recall Cartan’s Magic Formula for computing Lie derivatives (see for example [51, Thm. 14.35]): if $X \in \mathfrak{X}(\mathbb{R}^n)$ and $\omega \in \mathcal{C}^\infty \Lambda^k(\mathbb{R}^n)$ then the Lie derivative of ω along the flow of X , denoted $\mathcal{L}_X \omega$, is given by

$$\mathcal{L}_X \omega = X \lrcorner d\omega + d(X \lrcorner \omega). \quad (7.3.5)$$

If we choose $X = x^i \frac{\partial}{\partial x^i}$ then Cartan’s formula tells us that

$$\mathcal{L}_X \omega = (\kappa d + d\kappa)\omega. \quad (7.3.6)$$

So, we need only show that $\mathcal{L}_X \omega = (r + k)\omega$ when $\omega \in \mathcal{H}_r \Lambda^k$.

In order to compute the Lie derivative, however, we must find the flow of X . That is, we would like to find the smooth map $\Theta_t(p) : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$ such that, for all $p \in \mathbb{R}^n$, the initial value problem

$$\frac{d}{dt} \Theta_t(p) = X \circ \Theta_t(p), \quad (7.3.7)$$

$$\Theta_0(p) = p \quad (7.3.8)$$

is satisfied.

However, $X \circ \Theta_t(p) = \Theta_t(p)$ (viewed as an element of $T_{\Theta_t(p)} \mathbb{R}^n$). Then, (7.3.7) is trivially solved by

$$\Theta_t(p) = e^t p. \quad (7.3.9)$$

At this point, the proof proceeds by direct calculation. First, we compute the pullback of ω by $\Theta_t(p)$. For $v_1, \dots, v_k \in T_p \mathbb{R}^n$, we have

$$\begin{aligned} (\Theta_t^* \omega)_p(v_1, \dots, v_k) &= \omega_{\Theta_t(p)} \left((\Theta_t)_{*,p} v_1, \dots, (\Theta_t)_{*,p} v_k \right) \\ &= \omega_{e^t p} (e^t v_1, \dots, e^t v_k) \\ &= e^{kt} \omega_{e^t p} (v_1, \dots, v_k). \end{aligned}$$

The k -multilinearity of ω has been used in the last step.

Now, the entries of ω are homogeneous polynomials of degree r . Accordingly, we can pull out the scalar multiple in the subscript of $\omega_{e^t p}$ to obtain

$$(\Theta_t^* \omega)_p(v_1, \dots, v_k) = e^{(r+k)t} \omega_p(v_1, \dots, v_k).$$

Therefore, for all $\omega \in \mathcal{H}_r\Lambda^k$,

$$\Theta_t^*\omega = e^{(r+k)t}\omega. \quad (7.3.10)$$

Using (7.3.10) together with (7.3.6) and the definition of the Lie derivative, we have that

$$(\kappa d + d\kappa)\omega = \mathcal{L}_X\omega = \left. \frac{d}{dt} \right|_{t=0} \Theta_t^*\omega = \left. \frac{d}{dt} \right|_{t=0} e^{(r+k)t}\omega = (r+k)\omega.$$

The above calculation completes the proof. \square

Remark 7.3.8. In [4] and [5], the Magic Formula is called the **Homotopy Formula**. I chose to stick with the former nomenclature for two reasons. Firstly, many readers may be unfamiliar with the term “homotopy”; secondly, the Magic Formula makes the theory of FEMs so easy to handle that it really does seem like sorcery.

Now, we present some interesting corollaries of the Magic Formula. First, we see how this formula facilitates computation of the (co)homology of some simple (co)chain complexes.

Corollary 7.3.9. Suppose that $r \geq 1$, and consider the two complexes defined below (if $s < 0$, then $\mathcal{P}_s\Lambda^k = 0$):

1) the **polynomial de Rham complex**

$$0 \rightarrow \mathcal{P}_r\Lambda^0 \xrightarrow{d} \mathcal{P}_{r-1}\Lambda^1 \xrightarrow{d} \cdots \xrightarrow{d} \mathcal{P}_{r-n}\Lambda^n \rightarrow 0 \quad \text{and}$$

2) the **Koszul complex**

$$0 \leftarrow \mathcal{P}_r\Lambda^0 \xleftarrow{\kappa} \mathcal{P}_{r-1}\Lambda^1 \xleftarrow{\kappa} \cdots \xleftarrow{\kappa} \mathcal{P}_{r-n}\Lambda^n \leftarrow 0.$$

For both of these complexes, the bottom (co)homology is equal to \mathbb{R} , and every other (co)homology space is 0.

Proof. Since each term in both sequences decomposes as $\mathcal{P}_r\Lambda^k = \bigoplus_{s=0}^r \mathcal{H}_s\Lambda^k$ and the operators d and κ both map forms with homogeneous polynomial coefficients to forms with homogeneous polynomial coefficients, it suffices to prove the claim for the each of the following subsequences (with $0 \leq s \leq r$):

$$\mathcal{H}_s\Lambda^0 \xrightarrow{d} \mathcal{H}_{s-1}\Lambda^1 \xrightarrow{d} \cdots \xrightarrow{d} \mathcal{H}_{s-n}\Lambda^n \quad \text{and} \quad (7.3.11a)$$

$$\mathcal{H}_s\Lambda^0 \xleftarrow{\kappa} \mathcal{H}_{s-1}\Lambda^1 \xleftarrow{\kappa} \cdots \xleftarrow{\kappa} \mathcal{H}_{s-n}\Lambda^n. \quad (7.3.11b)$$

We begin with (7.3.11a). We want to show that if $d\omega = 0$ for $\omega \in \mathcal{H}_s\Lambda^k$, then there exists $\eta \in \mathcal{H}_{s+1}\Lambda^{k-1}$ such that $\omega = d\eta$. Using the Magic Formula,

$$(s+k)\omega = d\kappa\omega.$$

First, let $k = 0$. Then, if $s > 0$, the above expression implies that $\omega = 0$. Accordingly, we have that

$$\ker(d|_{\mathcal{P}_r\Lambda^0}) = \ker(d|_{\mathcal{H}_0\Lambda^0}) = \mathcal{H}_0\Lambda^0 = \mathbb{R}.$$

Therefore, the bottom cohomology is equal to \mathbb{R} .

For $k > 0$, we certainly know that $s + k > 0$. Let $\eta = \frac{1}{s+k}\kappa\omega$. By the above calculation we know that $\omega = d\eta$ and, trivially, $\eta \in \mathcal{H}_{s+1}\Lambda^{k-1}$. Therefore, every closed element of $H_s\Lambda^k$ is exact (for $k > 0$).

A perfectly symmetric argument shows the exactness of (7.3.11b) modulo the bottom homology: given $\eta \in \mathcal{H}_{s+1}\Lambda^{k-1}$ such that $\kappa\eta = 0$ we let $\omega = \frac{1}{s+k}d\eta \in \mathcal{H}_s\Lambda^k$ so by the Magic Formula $\eta = \kappa\omega$. The proof is complete. \square

In light of the above corollary, we often simply say that the polynomial de Rham and Koszul complexes are exact (the bottom (co)homology is usually not very interesting anyway, so ignoring it is often okay).

Corollary 7.3.10. *Let $r \geq 1$ and $0 \leq k \leq n$. Then, we have*

$$\mathcal{H}_r\Lambda^k = \kappa\mathcal{H}_{r-1}\Lambda^{k+1} \oplus d\mathcal{H}_{r+1}\Lambda^{k-1}. \quad (7.3.12)$$

Proof. The Magic Formula tells us that every element of $\mathcal{H}_r\Lambda^k$ may be written as the sum of a k -form in $\kappa\mathcal{H}_{r-1}\Lambda^{k+1}$ and another k -form in $d\mathcal{H}_{r+1}\Lambda^{k-1}$. Then, we have only to show that this decomposition is unique. It suffices to prove

$$\kappa\mathcal{H}_{r-1}\Lambda^{k+1} \cap d\mathcal{H}_{r+1}\Lambda^{k-1} = 0.$$

Pick $\omega \in d\mathcal{H}_{r+1}\Lambda^{k-1}$, then $d^2 = 0$ implies that $d\omega = 0$. If $\omega \in \kappa\mathcal{H}_{r-1}\Lambda^{k+1}$ as well, then $\kappa^2 = 0$ implies that $\kappa\omega = 0$. Using the Magic Formula we know that

$$(\text{nonzero constant}) \cdot \omega = 0,$$

hence $\omega = 0$. \square

Corollary 7.3.11. *If $r + k \neq 0$ then $\kappa|_{d\mathcal{P}_r\Lambda^k}$ and $d|_{\kappa\mathcal{P}_r\Lambda^k}$ are injective.*

Proof. We only prove that $\kappa|_{d\mathcal{P}_r\Lambda^k}$ is injective, as the second statement can be established using similar arguments. Consider $\omega \in \mathcal{H}_r\Lambda^k$. If $\kappa d\omega = 0$ then by the Magic Formula $(r+k)\omega = d\kappa\omega$. Taking the exterior derivative of both sides yields $d\omega = 0$ hence κ is injective on $d\mathcal{P}_r\Lambda^k$. The general result follows since $\mathcal{P}_r\Lambda^k = \bigoplus_{s=0}^r \mathcal{H}_s\Lambda^k$. \square

So, the kernel of $\kappa: \mathcal{P}_{r-1}\Lambda^{k+1} \rightarrow \mathcal{P}_r\Lambda^k$ only intersects $d\mathcal{P}_r\Lambda^k$ trivially, and likewise for $d|_{\kappa\mathcal{P}_r\Lambda^k}$.

Now, we are ready to define the alternative to $\mathcal{P}_r\Lambda^k$ mentioned at the beginning of this section.

Definition 7.3.12.

$$\mathcal{P}_r^-\Lambda^k \doteq \mathcal{P}_{r-1}\Lambda^k \bigoplus \kappa\mathcal{H}_{r-1}\Lambda^{k+1}. \quad (7.3.13)$$

$\mathcal{P}_r^-\Lambda^k$ is called the space of **trimmed** polynomial k -forms.

Clearly, $\mathcal{P}_r^-\Lambda^k \subseteq \mathcal{P}_r\Lambda^k$. The trimmed space consists of elements of $\mathcal{P}_r\Lambda^k$ where the terms of highest polynomial degree are constrained to have a very specific form. Indeed, this is why they are called trimmed forms: the dimension of $\mathcal{P}_r\Lambda^k$ is “trimmed” down by imposing a special restriction. In the next chapter we show how the spaces $\mathcal{P}_r^-\Lambda^k$ appear naturally in many well-known aspects of FEM theory. For instance, the Raviart–Thomas finite element studied in Example 6.2.4 has as its shape space $\star\mathcal{P}_1\Lambda^1(T_{\text{ref},2})$ (here, \star is the Hodge star). Before we can even get close to proving anything about finite elements, however, we need to learn more about the trimmed spaces. In particular, we must carefully calculate the dimension of the trimmed spaces in order to prove unisolvence later on.

Lemma 7.3.13.

$$\mathcal{P}_r^- \Lambda^0 = \mathcal{P}_r \Lambda^0.$$

Proof. By Corollary 7.3.10, $\kappa \mathcal{H}_{r-1} \Lambda^1 = \mathcal{H}_r \Lambda^0$. □

So, there is no distinction between the “regular” and trimmed polynomial scalar functions.

Lemma 7.3.14. *Let T be an n -simplex. Then, for $r + k \geq 1$, we have*

$$\dim(\kappa \mathcal{H}_r \Lambda^k(T)) = \binom{n+r}{n-k} \binom{r+k-1}{k-1}. \quad (7.3.14)$$

Proof. By the Rank–Nullity Theorem,

$$\dim(\kappa \mathcal{H}_r \Lambda^k) = \dim \mathcal{H}_r \Lambda^k - \dim(\ker \kappa|_{\mathcal{H}_r \Lambda^k}).$$

Since the Koszul complex is exact, $\ker \kappa|_{\mathcal{H}_r \Lambda^k} = \kappa \mathcal{H}_{r-1} \Lambda^{k+1}$. Then, using the arguments of Proposition 7.2.1,

$$\dim(\kappa \mathcal{H}_r \Lambda^k) = \binom{n+r-1}{n-1} \binom{n}{k} - \dim(\kappa \mathcal{H}_{r-1} \Lambda^{k+1}). \quad (7.3.15)$$

The above expression indicates that we can easily establish the claim using “backwards” induction after [4, Thm. 3.3]. Let $k = n$. Since $\mathcal{H}_{r-1} \Lambda^{n+1} = 0$, we have

$$\begin{aligned} \dim(\kappa \mathcal{H}_r \Lambda^n) &= \binom{n+r-1}{r} - \dim(\kappa \mathcal{H}_{r-1} \Lambda^{n+1}) \\ &= \binom{n+r-1}{n-1}. \end{aligned}$$

So, the claim holds in the case $k = n$.

Now, suppose that the claim holds for $k = n, n-1, \dots, \ell+1$. We prove that it also holds for $k = \ell$.

Using (7.3.15), we see

$$\begin{aligned}
\dim(\kappa\mathcal{H}_r\Lambda^\ell) &= \binom{n+r-1}{r} \binom{n}{\ell} - \dim(\kappa\mathcal{H}_{r-1}\Lambda^{\ell+1}) \\
&= \binom{n+r-1}{n-1} \binom{n}{\ell} - \binom{n+r-1}{n-\ell-1} \binom{r+\ell-1}{\ell} \\
&= \frac{(n+r-1)!}{\ell!} \left[\frac{n!}{r!(n-\ell)!(n-1)!} - \frac{(r+\ell-1)!}{(r-1)!(n-\ell-1)!(r+\ell)!} \right] \\
&= \frac{(n+r-1)!}{\ell!} \left[\frac{n}{r!(n-\ell)!} - \frac{1}{(r+\ell)(r-1)!(n-\ell-1)!} \right] \\
&= \frac{(n+r-1)!}{r!\ell!(n-\ell)!} \left[n - \frac{(n-\ell)r}{(r+\ell)} \right] \\
&= \frac{(n+r)!}{r!(\ell-1)!(n-\ell)!} \frac{1}{\ell(n+r)} \left[\frac{\ell(n+r)}{r+\ell} \right] \\
&= \frac{(n+r)!}{r!(\ell-1)!(n-\ell)!(r+\ell)} \\
&= \left[\binom{n+r}{n-\ell} (r+\ell-1)! \right] \left[\frac{1}{r!(\ell-1)!} \right] \\
&= \binom{n+r}{n-\ell} \binom{r+\ell-1}{\ell-1}.
\end{aligned}$$

The proof is now finished. □

Proposition 7.3.15.

$$\dim \mathcal{P}_r^- \Lambda^k(T) = \binom{n+r}{n-k} \binom{r+k-1}{k}. \quad (7.3.16)$$

Proof. Using the definition of $\mathcal{P}_r^- \Lambda^k(T)$ we immediately know

$$\dim \mathcal{P}_r^- \Lambda^k = \dim \mathcal{P}_{r-1}^- \Lambda^k + \dim \kappa\mathcal{H}_{r-1} \Lambda^{k+1}.$$

Therefore, by Corollary 7.2.2 and Lemma 7.3.14, we have

$$\begin{aligned}
\dim \mathcal{P}_r^- \Lambda^k &= \binom{n+r-1}{n} \binom{n}{k} + \binom{n+r-1}{n-k-1} \binom{r+k-1}{k} \\
&= \frac{(n+r-1)!}{k!(r-1)!(n-k)!} + \binom{n+r-1}{n-k-1} \binom{r+k-1}{k} \\
&= \left[\frac{(n+r-1)!}{(k+r-1)!(n-k)!} + \binom{n+r-1}{n-k-1} \right] \binom{r+k-1}{k} \\
&= \left[\binom{n+r-1}{n-k} + \binom{n+r-1}{n-k-1} \right] \binom{r+k-1}{k}.
\end{aligned}$$

Applying Lemma 7.1.1 completes the proof. □

I encourage the reader to take a little coffee break after working through all these messy calculations.

We conclude this section by examining the relationship between the trimmed spaces and the exterior derivative. Some of the results discussed below may seem like recreational trivialities, but all of these little lemmas and corollaries serve to eventually help us prove broad, elegant results in FEM theory.

Lemma 7.3.16. *Let $\omega \in \mathcal{P}_r^- \Lambda^k$ satisfy $d\omega = 0$. Then, $\omega \in \mathcal{P}_{r-1} \Lambda^k$.*

Proof. Since $\omega \in \mathcal{P}_r^- \Lambda^k$ we may write $\omega = \omega_1 + \kappa\omega_2$ with $\omega_1 \in \mathcal{P}_{r-1} \Lambda^k$ and $\omega_2 \in \mathcal{H}_{r-1} \Lambda^{k+1}$. Accordingly, the coefficients of $\kappa\omega_2$ either all have degree precisely r , or they are all equal to 0. If $\kappa\omega_2 \neq 0$ then Lemma 7.3.11 implies that all coefficients of $d\kappa\omega_2$ have degree precisely $r - 1$.

On the other hand, $d\omega_1$ has coefficients with degree less than or equal to $r - 2$. Therefore, the only way that $0 = d\omega_1 + d\kappa\omega_2$ can be true is if each of the two terms in the sum are identically zero. So, $\kappa\omega_2 = 0$. \square

Corollary 7.3.17.

$$\ker d|_{\mathcal{P}_r^- \Lambda^k} = \ker d|_{\mathcal{P}_{r-1} \Lambda^k}.$$

\square

Corollary 7.3.18. *Define the **generalized Whitney complex** by*

$$0 \rightarrow \mathcal{P}_r^- \Lambda^0 \xrightarrow{d} \mathcal{P}_r^- \Lambda^1 \xrightarrow{d} \dots \xrightarrow{d} \mathcal{P}_r^- \Lambda^n \rightarrow 0. \quad (7.3.17)$$

The bottom cohomology space of the generalized Whitney complex is equal to \mathbb{R} , and every other cohomology space is trivial.

Proof. First, we verify that the complex actually makes sense; that is, we must show that $d\mathcal{P}_r^- \Lambda^k \subseteq \mathcal{P}_r^- \Lambda^{k+1}$. However, this follows from just unraveling the definitions:

$$d\mathcal{P}_r^- \Lambda^k \subseteq d\mathcal{P}_r \Lambda^{k+1} \subseteq \mathcal{P}_{r-1} \Lambda^{k+1} \subseteq \mathcal{P}_r^- \Lambda^{k+1}.$$

Now, we can begin the proof in earnest. Note that the result on the bottom cohomology is obvious, so we have only to prove that the remaining cohomology spaces are trivial. Since $d^2 = 0$, we know that

$$d\mathcal{P}_r^- \Lambda^k = d \left(\mathcal{P}_r^- \Lambda^k \oplus d\mathcal{H}_{r+1} \Lambda^{k+1} \right).$$

Using Corollary 7.3.10, the above statement becomes

$$d\mathcal{P}_r^- \Lambda^k = d\mathcal{P}_r \Lambda^k.$$

Since the polynomial de Rham complex is exact, however, we also know that

$$d\mathcal{P}_r^- \Lambda^k = \ker d|_{\mathcal{P}_{r-1} \Lambda^k}.$$

Corollary 7.3.17 then yields

$$d\mathcal{P}_r^- \Lambda^k = \ker d|_{\mathcal{P}_r^- \Lambda^k},$$

hence the cohomology is trivial for $k > 0$. \square

We could also phrase the above claim as “the generalized Whitney complex is exact if we ignore the first term”. As with the Koszul and polynomial de Rham complexes, we often abuse terminology and simply say that the generalized Whitney complex is exact.

The **Whitney complex** is the generalized Whitney complex in the case $r = 1$. In his treatise on integration theory [89], Whitney (building off earlier work of de Rham and Weil) used this complex to provide a relatively simple proof of de Rham’s Theorem. In Chapter 9 we study Whitney’s perspective on de Rham theory in much more detail. Upon doing so, we see that the Whitney complex actually encodes the same information as the de Rham complex and, further, that a version of de Rham’s Theorem can be proved using only tools from FEM theory.

We have covered a lot of ground in this section, so before finishing a brief review is in order. First, we introduced the Koszul operator κ on differential forms. Subsequently, we developed some theory about how κ acts on polynomial forms. By far the most important result regarding κ is the Magic Formula, which we saw facilitated many proofs on the (co)homology of polynomial forms. After gaining a strong understanding of the Koszul operator, we defined the trimmed polynomial spaces $\mathcal{P}_r^- \Lambda^k$, counted their dimensions, and talked a little bit about their cohomological properties. Having become comfortable with polynomial forms, both regular and trimmed, we are now ready to turn to a unified, elegant, and rigorous construction of usable finite element spaces.

Chapter 8

Construction of Finite Element Spaces

In this chapter we show that the polynomial forms constructed in the previous chapter can be used to build two families of finite element spaces. After introducing these two families, we demonstrate how they may be understood as casting well-known aspects of FEM theory in a new light. Specifically, many commonly used finite element spaces arise as members of one of the two main families, so FEEC nicely unifies many of the useful tools of FEM theory. As in the previous chapter, we are essentially just re-packaging results from [3] and [4]. In particular, our developments follow [3, §2.2, §3] and [4, §5] very closely.

The insight that so many common finite elements appear naturally when working with differential forms was presented by Arnold, Falk, and Winther in 2006 [4, Tables 5.1 and 5.2]. Over the course of eight years, Arnold and Logg refined this insight into a concise way of presenting the main ideas of FEEC, the Periodic Table of the Finite Elements [6]. This table features all of the “commonly used” finite elements referenced above and also visually encodes information on the homological properties of special FEMs. Upon learning what the Periodic Table says and how to read it, one can quickly use the table to design usable, structure-preserving FEMs. We discuss the Periodic Table more thoroughly in Chapter 9, paying particular attention to its homological properties, but we introduce it here for the sake of stressing its use as a novel tool in scientific computing. We also make reference to the Periodic Table throughout this chapter, mostly using it as a source of good notation.

Normally, I would present a little historical review at this point, introducing the aforementioned “commonly used” finite elements and presenting a rough chronology of their discovery. However, I feel that such discussion is best left to when we actually see how these elements arise from the two main families. That is, I present relevant sources on each of the elements as the elements are introduced.

In Section 8.1, we define the general Lagrange finite element space (generalizing Example 6.2.3) and prove unisolvence for the associated DOFs. In Section 8.2, we develop some technical lemmas that help speed up proofs later on. In Section 8.3, we build a family of finite element spaces with shape functions based on the spaces $\mathcal{P}_r^- \Lambda^k(T)$ and discuss this family’s relation to several famous finite element spaces known before FEEC. In Section 8.4, we do the same as we did in Section 8.3 except with shape functions based on the spaces $\mathcal{P}_r \Lambda^k(T)$.

Throughout this chapter $\Omega \subseteq \mathbb{R}^n$ denotes a polyhedral domain, \mathcal{T}_h denotes a triangulation of Ω , and

$T \in \Delta_n(\mathcal{T}_h)$. λ_i denotes the i^{th} barycentric coordinate on the simplex T (see Section 5.1 for a reminder on how such coordinates are defined). Whenever we refer to spaces of polynomial forms without specifying their domain, the domain is implicitly considered to be the simplex T . For example, $\mathcal{P}_r\Lambda^k \doteq \mathcal{P}_r\Lambda^k(T)$.

8.1 The Lagrange Finite Element Space $\mathcal{P}_r\Lambda^0(\mathcal{T}_h)$

In this section we present the easiest example of a finite element space obtained within the formalism of FEEC. We show that this space consists of continuous scalar functions that are, with respect to a given triangulation of the domain, piecewise polynomials.

Definition 8.1.1. *Let \mathcal{T}_h be a triangulation of Ω . The **Lagrange finite element** consists of the following data:*

1) *The shape spaces are given by $V(T) = \mathcal{P}_r\Lambda^0(T) \ \forall T \in \mathcal{T}_h$.*

2) *Pick any $T \in \Delta_n(\mathcal{T}_h)$ and $f \in \Delta(T)$. Define a space of real-valued functionals on $V(f)$ by*

$$W(f) \doteq \left\{ u \in V(T) \mapsto \int_f \text{tr}_f u q \mid q \in \mathcal{P}_{r-\dim f-1}\Lambda^{\dim f}(f) \right\}. \quad (8.1.1)$$

Then, the span of the degrees of freedom on T is given by

$$\sum_{f \in \Delta(T)} W(f). \quad (8.1.2)$$

*Define local-to-global maps by requiring that all local DOFs representing the same weighted integral over the same face f are mapped to the same global DOF. The finite element space obtained using the above data is called the **Lagrange space**, denoted by $\mathcal{P}_r\Lambda^0(\mathcal{T}_h)$.*

Before going further, we make some remarks on the DOFs for the Lagrange finite element. We have chosen to define the span of the DOFs, rather than specifically choosing the DOFs themselves. If we prove that the span is equal to the entire dual space $(\mathcal{P}_r\Lambda^0(T))^*$, then by choosing a basis for $\mathcal{P}_{r-\dim f-1}\Lambda^{\dim f}(f)$ for every $f \in \Delta(T)$ we obtain a set of DOFs. Choosing to work with the span, rather than a particular basis, is usually a convenient choice. We continue with this style of studying DOFs in the sequel, only choosing to exhibit specific functionals in the examples.

The central goal of this section is to prove that $\mathcal{P}_r\Lambda^0(\mathcal{T}_h)$ is, in fact, a finite element space according to Definition 6.3.5. In light of the above paragraph, we need only show that the span of the DOFs is $(\mathcal{P}_r\Lambda^0(T))^*$. Our proof of unisolvence for the Lagrange elements lays the groundwork for our investigations of more complicated finite element spaces, so we take great pains to go through every detail now in order to make later work easier.

Theorem 8.1.2. *For all $T \in \Delta_n(\mathcal{T}_h)$, we have*

$$(\mathcal{P}_r\Lambda^0(T))^* = \bigoplus_{f \in \Delta(T)} W(f). \quad (8.1.3)$$

Proof. First, we use induction on $\dim T$ to show that

$$(\mathcal{P}_r \Lambda^0(T))^* = \sum_{f \in \Delta(T)} W(f), \quad (8.1.4)$$

then we apply the binomial coefficient identities from Section 7.1 to prove that the sum is actually direct. We remark that the inclusion $\sum_{f \in \Delta(T)} W(f) \subseteq (\mathcal{P}_r \Lambda^0(T))^*$ is clear.

The case $\dim T = 0$ is trivial: the only functions on T are constants, so if $u \in \mathcal{P}_r \Lambda^0(T)$ is killed by all the functionals in $W(f)$ then $u = 0$, hence by Lemma 6.2.2 we are done.

Let $n = \dim T$ and suppose that the claim holds for simplices with dimension $n - 1$. Let $u \in \mathcal{P}_r \Lambda^0(T)$ be such that

$$u \in \bigcap_{\alpha \in \sum_{f \in \Delta(T)} W(f)} \ker \alpha.$$

Pick an arbitrary facet $F \in \Delta_{n-1}(T)$. Recall that F can be written as the vanishing set of one of the barycentric coordinates. In symbols,

$$F = V(\lambda_i) \quad \text{for some } i \in \{0, \dots, n\}.$$

Because $\text{tr}_f \circ \text{tr}_F = \text{tr}_f$, we know that

$$\text{tr}_F u \in \bigcap_{\alpha \in \sum_{f \in \Delta(F)} W(f)} \ker \alpha.$$

Since (8.1.4) holds for $n - 1$ simplices and F is an $n - 1$ simplex, the above expression implies that $\text{tr}_F u = 0$.

Write u as a polynomial in the barycentric coordinates λ_i on T . Using $\text{tr}_F u(\lambda_0, \dots, \lambda_n) = 0$ and $F = V(\lambda_i)$, we know that λ_i divides the polynomial u . The facet F was chosen arbitrarily, however, so for all $i = 0, \dots, n$ we have that λ_i divides u . Therefore, there exists $p \in \mathcal{P}_{r-n-1}(T)$ such that

$$u(\lambda_0, \dots, \lambda_n) = p(\lambda_0, \dots, \lambda_n) \prod_{i=0}^n \lambda_i.$$

Since $p \in \mathcal{P}_{r-n-1}(T)$, the functional

$$u \mapsto \int_T u p \, d\text{Vol}$$

lives in $W(T)$. But, because u vanishes under the action of all such functionals,

$$\int_T p^2 \prod_{i=0}^n \lambda_i \, d\text{Vol} = 0.$$

Since $\lambda_i \geq 0$, the above implies that $p = 0$, hence $u = 0$ as well. Applying Lemma 6.2.2, we conclude that (8.1.4) holds for all n .

Next, we show that the sum in (8.1.4) is direct. First, observe that for all $f \in \Delta(T)$ we have a surjective linear map $\mathcal{P}_{r-\dim f-1}(f) \rightarrow W(f)$ taking a polynomial q to the functional $u \mapsto \int_f u q$. For all f , then, $\dim W(f) \leq \dim \mathcal{P}_{r-\dim f-1}(f)$.

Now, we do some dimension counting. First, observe that a sum over faces in $\Delta(T)$ can be replaced by summing over face dimension d , provided we keep track of how many d -dimensional faces are in T . With this in mind,

$$\sum_{f \in \Delta(T)} \dim \mathcal{P}_{r-\dim f-1}(f) = \sum_{d=0}^n |\Delta_d(T)| \dim \mathcal{P}_{r-d-1}(\lambda_{i_1}, \dots, \lambda_{i_d}).$$

Clearly, we have $|\Delta_d(T)| = \binom{n+1}{d+1}$. Then, using Proposition 7.2.1, we see that

$$\begin{aligned} \sum_{f \in \Delta(T)} \dim \mathcal{P}_{r-\dim f-1}(f) &= \sum_{d=0}^n \binom{n+1}{d+1} \binom{r-d-1+d}{r-d-1} \\ &= \sum_{d=0}^n \binom{n+1}{n-d} \binom{r-1}{d}. \end{aligned}$$

Combining the above with Lemma 7.1.1, we get

$$\sum_{f \in \Delta(T)} \dim \mathcal{P}_{r-\dim f-1}(f) = \binom{n+r}{r}.$$

Using Proposition 7.2.1 again,

$$\sum_{f \in \Delta(T)} \dim \mathcal{P}_{r-\dim f-1}(f) = \dim \mathcal{P}_r \Lambda^0(T). \quad (8.1.5)$$

If there exists f such that $\dim W(f) < \dim \mathcal{P}_{r-\dim f-1}(f)$ then by (8.1.4)

$$\begin{aligned} \dim \mathcal{P}_r(T) &= \dim \sum_{f \in \Delta(T)} W(f) \\ &\leq \sum_{f \in \Delta(T)} \dim W(f) \\ &< \sum_{f \in \Delta(T)} \dim \mathcal{P}_{r-\dim f-1}(f). \end{aligned}$$

Using (8.1.5), the above expression becomes

$$\dim \mathcal{P}_r(T) < \dim \mathcal{P}_r \Lambda^0(T),$$

which is a contradiction. Therefore, $\dim W(f) = \dim \mathcal{P}_{r-\dim f-1}(f)$ for all f . Hence, the surjection taking polynomial weight functions to their corresponding functionals is actually a vector space isomorphism: no two weight functions give rise to the same functional.

Now, we show that if $f_i \neq f_j$ then $W(f_i) \cap W(f_j) = 0$. Using (8.1.4), (8.1.5), and the isomorphism

$W(f) \simeq \mathcal{P}_{r-\dim f-1}(f)$, we have

$$\begin{aligned} \binom{n+r}{r} &= \sum_f \dim W(f) - \sum_{f_i \neq f_j} \dim [W(f_i) \cap W(f_j)] \\ &= \sum_f \dim \mathcal{P}_{r-\dim f-1}(f) - \sum_{f_i \neq f_j} \dim [W(f_i) \cap W(f_j)] \\ &= \binom{n+r}{r} - \sum_{f_i \neq f_j} \dim [W(f_i) \cap W(f_j)]. \end{aligned}$$

Therefore,

$$0 = \sum_{f_i \neq f_j} \dim [W(f_i) \cap W(f_j)].$$

However, each term in the sum is nonnegative. This implies that $\dim [W(f_i) \cap W(f_j)] = 0$ for all $f_i \neq f_j$. Therefore, $W(f_i) \cap W(f_j) = 0$ for all $f_i \neq f_j$. The proof is now (finally) complete. \square

With Theorem 8.1.2, we have finished proving that the Lagrange elements are in fact finite elements, and therefore the space $\mathcal{P}_r \Lambda^0(\mathcal{T}_h)$ is in fact a finite element space. Following our discussion in Section 6.4, we know that the Lagrange space is a subspace of $H\Lambda^0$: by definition of the DOFs and Theorem 8.1.2, the facet traces of any element of $\mathcal{P}_r \Lambda^0(\mathcal{T}_h)$ are unique, hence by Corollary 6.4.2 the Lagrange space is contained in $H\Lambda^0$. Additionally, single-valuedness of the facet traces implies that all elements of the Lagrange space are continuous across the entire domain.

Now, we go through two examples connecting the Lagrange elements we have just developed to the Lagrange elements studied in Example 6.2.3. The purpose of these examples is largely to start getting a feeling for how to switch between the FEEC and “pre-FEEC” ways of thinking about specific finite elements, as a great level of comfort with such switching is necessary for understanding modern FEM theory and implementation. In particular, we illustrate why we can choose the DOFs to be point evaluation functionals so that our construction in Definition 8.1.1 agrees with Example 6.2.3.

Example 8.1.3. In this example we study the Lagrange finite element over the reference 1-simplex $T = [0, 1]$ and show that we can choose our DOFs to be point evaluation functionals, as in Example 6.2.3. We make use of the more traditional notation $\text{CG}(r)$ to denote the Lagrange element with shape space $\mathcal{P}_r \Lambda^0(T)$ (see Example 6.2.3 for a reminder on the notation). We also use $\text{CG}(r)$ to denote the Lagrange finite element space.

We consider two simple cases: $\text{CG}(1)$ and $\text{CG}(2)$. Examining Definition 8.1.1, we have that the only nontrivial $W(f)$ ’s correspond to the vertices $p_i \in \Delta_0(T)$:

$$W(p_i) \doteq \left\{ u \in V(T) \mapsto \int_{p_i} \text{tr}_{p_i} u \, q = u(p_i) \, q \mid q \in \mathcal{P}_0 \Lambda^0(p_i) \simeq \mathbb{R} \right\}.$$

Therefore, the only DOFs for $\text{CG}(1)$ are moments against scalars over the vertices, meaning that we can just choose DOFs to be point evaluation in this case. So, the general construction of Lagrange elements reduces to the element from Example 6.2.3 with no trouble.

For $\text{CG}(2)$, things get a little more interesting. If we wanted to naïvely generalize the Lagrange element from Example 6.2.3 to degree 2 shape functions, we would just add a third point evaluation DOF, $\text{eval}_{\frac{1}{2}}$.



Figure 8.11: DOFs for some $\text{CG}(r)$ finite elements on intervals, corresponding to $\mathcal{P}_r\Lambda^0([0, 1])$.

At first glance, however, it is not clear that this DOF is naturally accounted for in Definition 8.1.1. By this definition, the third DOF for $\text{CG}(2)$ “should” be taken to be a scalar multiple of the averaging functional

$$u \mapsto \int_0^1 u \, dx.$$

However, by Theorem 8.1.2 we know that $\text{eval}_{\frac{1}{2}}$ must be in the span of the averaging functional and the vertex DOFs eval_0 and eval_1 . A quick safety calculation verifies this explicitly: for all $u \in \mathcal{P}_2\Lambda^0([0, 1])$, we have

$$\text{eval}_{\frac{1}{2}}(u) = \frac{3}{2} \int_0^1 u \, dx - \frac{1}{4}\text{eval}_0(u) - \frac{1}{4}\text{eval}_1(u).$$

The above identity is simply the statement that Simpson’s Rule for numerical integration is exact for degree 2 polynomials defined on $[0, 1]$. So, Definition 8.1.1 does still allow us to choose point evaluations for our DOFs. See Figure 8.11 for an illustration of the DOFs for the first two Lagrange elements on $[0, 1]$.

The reader may wonder why we have even bothered writing the DOFs as moments if we were just going to go back to using point evaluations anyway. The reason is because considering the Lagrange DOFs as linear combinations of moments allows us to generalize ideas from the proof of Theorem 8.1.2 to finite element subspaces of $H\Lambda^k$ for $k \geq 1$; in the rest of this chapter we see that unisolvence can be proved for all sorts of finite elements using essentially the same arguments we applied to prove Theorem 8.1.2. Additionally, by proving Theorem 8.1.2 with moments instead of point evaluation functionals, we can keep the notation clean: it is certainly easier to start with moments, prove everything we need, and then switch to point evaluations than it is to pick what points we’ll evaluate at (these points change depending on the dimension of our simplex and the shape function degree!) and then prove unisolvence. In brief, working with moments allows us to prove Theorem 8.1.2 for all n and r at once, with no added notational complexity.

Example 8.1.4. In this example we briefly discuss CG elements on the reference triangle $T = T_{\text{ref},2}$. The DOF diagrams for some low-degree CG elements are shown in Figure 8.12, with black dots still representing point evaluations as in the previous example. The argument for why all DOFs can be chosen to be point evaluations is a carbon copy of what we saw in Example 8.1.3, since all functionals on $\mathcal{P}_r\Lambda^0(T)$ can be written as moments against lower-degree polynomials by Theorem 8.1.2. Notice that there are no interior DOFs for the $\text{CG}(r)$ element on triangles for $r < 3$, so an element of the $\text{CG}(1)$ or $\text{CG}(2)$ finite element space is completely determined by its values on edges. As previously mentioned, $\text{CG}(1)$ was used by Courant in 1943 [23], making it “the original” finite element.

Many authors in the numerical analysis community define Lagrange space much differently than we have done here. In a more classical treatment of finite elements, one would likely use the umbrella term “Lagrange finite element space” to mean any (not necessarily conforming) finite element space of piecewise polynomial scalar functions with DOFs corresponding to point evaluations. Our Lagrange spaces corresponds to the special case of “continuous Lagrange spaces” using this classical terminology. FEEC accommodates discontinuous scalar functions, provided that such functions are viewed as the coefficient of a top form; we discuss discontinuous finite element spaces in detail in Section 8.3.

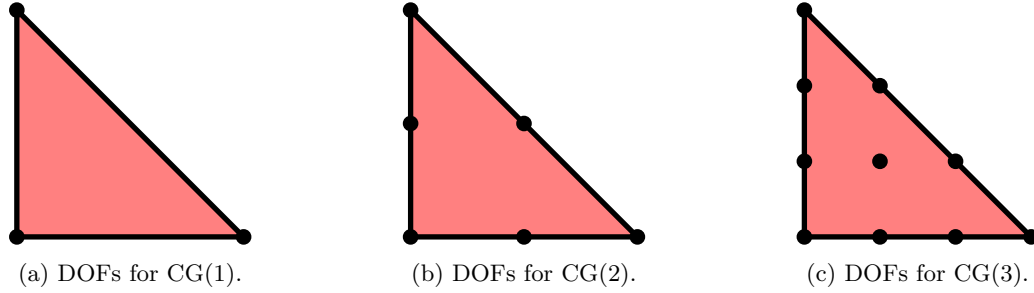


Figure 8.12: DOFs for some $\text{CG}(r)$ finite elements on triangles, corresponding to $\mathcal{P}_r\Lambda^0(T_{\text{ref},2})$.

8.2 Vanishing Lemmas

In this section we focus on establishing some little results in isolation in order to facilitate the construction of finite element spaces that we undertake in the rest of the chapter. We often use notation that last appeared in Chapter 2. For example, recall that, given natural numbers k and n , $\Sigma(k, n)$ denotes the set of all strictly increasing maps $\sigma: \{1, \dots, k\} \rightarrow \{1, \dots, n\}$. Given $\sigma \in \Sigma(k, n)$, $\sigma^* \in \Sigma(n - k, n)$ denotes the unique increasing map $\{1, \dots, n - k\} \rightarrow \{1, \dots, n\}$ such that $\{\sigma(1), \dots, \sigma(k), \sigma^*(1), \dots, \sigma^*(n - k)\}$ is a permutation of $\{1, 2, \dots, n\}$.

Remember that any point in an n -simplex T actually has $n + 1$ associated barycentric coordinates (each vertex of T defines a barycentric coordinate function), so the coordinate system on T defined by the λ_i 's is overdetermined. However, we must have that

$$\sum_{i=0}^n \lambda_i = 1,$$

so we can leave out *one* of the λ_i without losing any information. In this chapter, we choose to consistently leave out the coordinate λ_0 . Therefore, we have a coordinate system in T defined by n independent variables $\lambda_1, \dots, \lambda_n$. This coordinate system is clearly related to the Cartesian coordinate system (x^1, \dots, x^n) by way of an affine transformation (that is, a composition of a translation and a linear map), so we immediately obtain the following:

Lemma 8.2.1. *If $u \in \mathcal{C}^\infty\Lambda^k(T)$, then there exist scalars $u_\sigma \in \mathcal{C}^\infty\Lambda^0(T)$ such that u can be written as*

$$u = \sum_{\sigma \in \Sigma(k, n)} u_\sigma \, d\lambda_{\sigma(1)} \wedge d\lambda_{\sigma(2)} \wedge \cdots \wedge d\lambda_{\sigma(k)}. \quad (8.2.1)$$

□

Now, we have established enough notation to begin our proofs.

Lemma 8.2.2. *If $u \in \mathcal{P}_r\Lambda^k$ satisfies $\text{tr}_{\partial T}u = 0$ and*

$$\int_T u \wedge q = 0 \quad \forall q \in \mathcal{P}_{r-n+k}\Lambda^{n-k},$$

then $u = 0$.

Proof. By Lemma 8.2.1, we may write u as

$$u = \sum_{\sigma \in \Sigma(k, n)} u_\sigma \, d\lambda_{\sigma(1)} \wedge d\lambda_{\sigma(2)} \wedge \cdots \wedge d\lambda_{\sigma(k)}, \quad (8.2.2)$$

with $u_\sigma \in \mathcal{P}_r$. Pick any $i = 1, \dots, n$ and let F_i be the face of T defined by $F_i = V(\lambda_i)$. By hypothesis, $\text{tr}_{F_i} u = 0$. So, we know that λ_i divides u_σ provided that $i \notin \text{range } \sigma$ (if $i \in \text{range } \sigma$ then $\text{tr}_{F_i} d\lambda_{\sigma(1)} \wedge \cdots \wedge d\lambda_{\sigma(k)} = 0$, hence we get no information about u_σ). Therefore, for all $\sigma \in \Sigma(k, n)$ there exists some $p_\sigma \in \mathcal{P}_{r-n-k}$ such that

$$u_\sigma = p_\sigma \lambda_{\sigma^*(1)} \cdots \lambda_{\sigma^*(n-k)}. \quad (8.2.3)$$

Now, we would like to pick a special weight function q that we can use to establish the vanishing of u everywhere. Let

$$q = \sum_{\sigma \in \Sigma(k, n)} (-1)^{\text{sgn}(\sigma, \sigma^*)} p_\sigma \, d\lambda_{\sigma^*(1)} \wedge \cdots \wedge d\lambda_{\sigma^*(n-k)} \in \mathcal{P}_{r-n-k} \Lambda^{n-k}.$$

Then, by hypothesis, we know that

$$\begin{aligned} 0 &= \int_T u \wedge q \\ &= \int_T \left[\sum_{\sigma \in \Sigma(k, n)} p_\sigma \lambda_{\sigma^*(1)} \cdots \lambda_{\sigma^*(n-k)} \, d\lambda_{\sigma(1)} \wedge \cdots \wedge d\lambda_{\sigma(k)} \right] \wedge \left[\sum_{\sigma \in \Sigma(k, n)} (-1)^{\text{sgn}(\sigma, \sigma^*)} p_\sigma \, d\lambda_{\sigma^*(1)} \wedge \cdots \wedge d\lambda_{\sigma^*(n-k)} \right]. \end{aligned}$$

Now, all of the terms with repeated factors of $d\lambda_i$ vanish, so all of the cross terms in the above expression disappear. Further, the powers of -1 appearing in q allow us to rearrange the n -form in the integrand to obtain the volume form in barycentric coordinates. All in all, we get

$$0 = \sum_{\sigma \in \Sigma(k, n)} \int_T p_\sigma^2 \lambda_{\sigma^*(1)} \cdots \lambda_{\sigma^*(n-k)} \, d\lambda_1 \wedge \cdots \wedge d\lambda_n.$$

All λ_i are strictly positive in the interior of T , and the vertices of T may without loss of generality be taken to be in the positive orientation. Therefore, the integrand is nonnegative. By elementary analysis, $p_\sigma = 0 \quad \forall \sigma$, and we conclude that $u = 0$. \square

The next result strengthens Lemma 8.2.2, since the space of trimmed polynomials on T is a subspace of $\mathcal{P}_r \Lambda^k$.

Lemma 8.2.3. *If $u \in \mathcal{P}_r \Lambda^k$ satisfies $\text{tr}_{\partial T} u = 0$ and*

$$\int_T u \wedge q = 0 \quad \forall q \in \mathcal{P}_{r-n+k}^- \Lambda^{n-k},$$

then $u = 0$.

Proof. For any $p \in \mathcal{C}^\infty \Lambda^{n-k-1}$, integration by parts tells us that

$$\int_T du \wedge p = (-1)^{k+1} \int_T u \wedge dp. \quad (8.2.4)$$

If we specialize to $p \in \mathcal{P}_{r-n+k}\Lambda^{n-k-1}$ then in particular $dp \in \mathcal{P}_{r-n+k}^-\Lambda^{n-k}$. With this in mind, we use (8.2.4) and the hypothesis to obtain

$$\int_T du \wedge p = 0 \quad \forall p \in \mathcal{P}_{r-n+k}\Lambda^{n-k-1}. \quad (8.2.5)$$

Since traces commute with d , $\text{tr}_{\partial T}u = 0$ implies that $\text{tr}_{\partial T}du = 0$. Using (8.2.5) and Lemma 8.2.2, we get $du = 0$.

Going back to (8.2.4), we know that

$$\int_T u \wedge dp = 0 \quad \forall p \in \mathcal{C}^\infty\Lambda^{n-k-1}. \quad (8.2.6)$$

Pick any $q \in \mathcal{P}_{r-n+k}\Lambda^{n-k}$. By Corollary 7.3.10, there exists $q^- \in \mathcal{P}_{r-n+k}^-\Lambda^{n-k}$ and $p \in \mathcal{H}_{r-n+k+1}\Lambda^{n-k-1}$ such that $q = q^- + dp$. Then, we have that

$$\int_T u \wedge q = \int_T u \wedge q^- + \int_T u \wedge dp = 0.$$

Now, apply Lemma 8.2.2 again to see that $u = 0$. \square

Finally, we give an analogue of Lemma 8.2.3 applicable to trimmed polynomial forms vanishing against regular weights.

Lemma 8.2.4. *If $u \in \mathcal{P}_r^-\Lambda^k$ satisfies $\text{tr}_{\partial T}u = 0$ and*

$$\int_T u \wedge q = 0 \quad \forall q \in \mathcal{P}_{r+k-n-1}\Lambda^{n-k},$$

then $u = 0$.

Proof. The proof is just a regurgitation of the arguments used to obtain Lemma 8.2.3. Using the integration by parts formula we know that, for all $p \in \mathcal{P}_{r-n+k}\Lambda^{n-k-1}$,

$$\int_T du \wedge p = 0.$$

Since differentiation commutes with pullback, $\text{tr}_{\partial T}du = 0$. Therefore, Lemma 8.2.2 indicates that $du = 0$. Further, Lemma 7.3.16 tells us that $u \in \mathcal{P}_{r-1}\Lambda^k$. Paying close attention to the degree of the weight functions in (8.3.1), we see that we are now allowed to apply Lemma 8.2.2 to u . Therefore, $u = 0$ and the proof is complete. \square

8.3 The Trimmed Finite Element Space $\mathcal{P}_r^-\Lambda^k(\mathcal{T}_h)$

In this section we introduce a general family of finite elements with shape spaces consisting of trimmed polynomial differential forms. These finite elements in turn give rise to useful finite element spaces, which we denote by $\mathcal{P}_r^-\Lambda^k(\mathcal{T}_h)$ in analogy to our notation for Lagrange spaces. We begin by presenting the main definitions and unisolvence proofs, exactly as we did in Section 8.1. Then, we go through some quick examples illustrating how several famous finite elements, such as the Raviart–Thomas elements from Example 6.2.4, arise as “copies” of $\mathcal{P}_r^-\Lambda^k(\mathcal{T}_h)$ for special choices of k .

Definition 8.3.1. Let \mathcal{T}_h be a triangulation of Ω . The **trimmed finite element** of polynomial degree r and form degree k consists of the following data:

1) The shape spaces are given by $\mathcal{P}_r^- \Lambda^k(T) \quad \forall T \in \mathcal{T}_h$.

2) Pick any $T \in \Delta_n(\mathcal{T}_h)$ and $f \in \Delta(T)$. Define a space of real-valued functionals on $V(f)$ by

$$W(f) \doteq \left\{ u \in V(T) \mapsto \int_f \text{tr}_f u \wedge q \mid q \in \mathcal{P}_{r+k-\dim f-1} \Lambda^{\dim f-k}(f) \right\}. \quad (8.3.1)$$

Then, the span of the degrees of freedom on T is given by

$$\sum_{f \in \Delta(T)} W(f). \quad (8.3.2)$$

Local-to-global maps are defined taking any local DOFs associated to the same weighted integral over the same face to the same global DOF. The finite element space associated to the above data is called the **trimmed finite element space** of polynomial degree r and form degree k , denoted by $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$.

Theorem 8.3.2. In the notation of Definition 8.3.1,

$$\mathcal{P}_r^- \Lambda^k(T)^* = \bigoplus_{f \in \Delta(T)} W(f). \quad (8.3.3)$$

Proof. First, we use induction on $\dim T$ to prove that

$$\mathcal{P}_r^- \Lambda^k(T)^* = \sum_{f \in \Delta(T)} W(f). \quad (8.3.4)$$

If $\dim T = 0$ then the proof is trivial. Suppose now that $\dim T = n$ and that (8.3.4) holds for $(n-1)$ -simplices. If $u \in \mathcal{P}_r^- \Lambda^k(T)$ is killed by every functional in each $W(f)$ then, by the induction hypothesis, $\text{tr}_{\partial T} u = 0$. Combining the definition of $W(T)$ with Lemma 8.2.4, we know that $u = 0$. Therefore, we have proved (8.3.4).

Now, we have only to count dimensions, as we did for the Lagrange elements, to prove that the sum in (8.3.4) is actually direct. First, observe that

$$\begin{aligned} \sum_{f \in \Delta(T)} \dim \mathcal{P}_{r+k-\dim f-1} \Lambda^{\dim f-k}(f) &= \sum_{d=0}^n |\Delta_d(T)| \binom{r+k-1}{r+k-d-1} \binom{d}{d-k} \\ &= \sum_{d=k}^n \binom{n+1}{d+1} \binom{r+k-1}{d} \binom{d}{k} \\ &= \sum_{j=0}^{n-k} \binom{n+1}{j+k+1} \binom{r+k-1}{j+k} \binom{j+k}{k}. \end{aligned}$$

Using Lemma 7.1.3 with $a = r+k-1$, $b = j+k$, and $c = k$ the above becomes

$$\sum_{f \in \Delta(T)} \dim \mathcal{P}_{r+k-\dim f-1} \Lambda^{\dim f-k}(f) = \binom{r+k-1}{k} \sum_{j=0}^{n-k} \binom{n+1}{j+k+1} \binom{r-1}{j}.$$

If we now apply Lemma 7.1.4 with $a = n + 1, b = k + 1$, and $c = r - 1$, then we see that

$$\sum_{j=0}^{n-k} \binom{n+1}{j+k+1} \binom{r-1}{j} = \binom{n+r}{n-k}.$$

Putting this all together and using Proposition 7.3.15, we have

$$\sum_{f \in \Delta(T)} \dim \mathcal{P}_{r+k-\dim f-1} \Lambda^{\dim f-k}(f) = \binom{r+k-1}{k} \binom{n+r}{n-k} = \dim \mathcal{P}_r^- \Lambda^k(T).$$

From here, arguments analogous to those we applied to prove unisolvence for Lagrange elements allow us to conclude that

$$\mathcal{P}_r^- \Lambda^k(T)^* = \bigoplus_{f \in \Delta(T)} W(f).$$

□

Upon finishing the above proof we know that the $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$ are truly finite element spaces. We can actually obtain *two* distinct sets of finite element spaces from the trimmed family. In particular, we can define a set of finite element spaces $\star \mathcal{P}_r^- \Lambda^{n-k}(\mathcal{T}_h)$ using shape spaces $\star \mathcal{P}_r^- \Lambda^{n-k}(T)$ and the same DOF span that Definition 8.3.1 prescribes for $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$. As we see over the course of the next few pages, the finite element spaces $\star \mathcal{P}_r^- \Lambda^{n-k}(\mathcal{T}_h)$ and $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$ are not the same, but we consistently refer to them both under the banner of “trimmed finite element spaces” for the remainder of Part 1.

Before introducing specific examples of how the above finite elements are related to “pre-FEEC” FEM theory, let’s investigate the general inter-element continuity requirements a form in $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$ must satisfy. Now, we can uniquely reconstruct $u \in \mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$ if we know how each global DOF acts on u . As a trivial corollary, facet traces of forms in $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$ are single-valued. Applying Corollary 6.4.2, we know that $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h) \subseteq H\Lambda^k(\Omega)$. As we discussed at the end of Section 6.4, the constraint on facet traces becomes less and less restrictive as k increases: scalars are continuous, top forms are discontinuous, and all other forms in between have some coefficients that are continuous and some that are not. So, a wide variety of forms inhabit the spaces $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$, making this family well-suited for handling a wide variety of practical applications.

We now go through several examples connecting the $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$ family to some renowned finite element spaces. In studying these examples, we begin to see how FEEC provides a unified and simplified notation for FEM theory. Of course, good notation is not the only reason that FEEC is worth studying (in Chapter 9, we see that FEEC gives us better numerical approximations by paying attention to algebra and topology), but nevertheless good notation makes our lives simpler and makes the subject a little more accessible.

Example 8.3.3. By comparing DOFs in the case $k = 0$, we see that the Lagrange space $\mathcal{P}_r \Lambda^0(\mathcal{T}_h)$ is precisely equal to $\mathcal{P}_r^- \Lambda^0(\mathcal{T}_h)$. So, the trimmed finite element spaces generalize the Lagrange elements.

Example 8.3.4. In this example we show how the finite element space associated to the lowest Raviart–Thomas element RT(1) from Example 6.2.4 is related to the trimmed space $\mathcal{P}_1^- \Lambda^1(\mathcal{T}_h)$ or, more accurately, the modified trimmed space $\star \mathcal{P}_1^- \Lambda^1(\mathcal{T}_h)$. Then, we show how to obtain generalized Raviart–Thomas elements for arbitrary shape function polynomial degree, thus re-obtaining the results of [73] from a different point of view. Throughout this example, let $T = T_{\text{ref},2}$.

In two dimensions, we have that

$$\kappa(dx \wedge dy) = -y dx + x dy.$$

Then, upon switching to vector proxies,

$$\mathcal{P}_1^- \Lambda^1(T) = \{u \in \mathfrak{X}(T) \mid u(x, y) = (b_2 + ay, b_1 - ax), a, b_1, b_2 \in \mathbb{R}\}.$$

When we take the Hodge star (which just amounts to 90° counterclockwise rotation of vectors in \mathbb{R}^2) and absorb a negative sign into b_1 , we see that

$$\star \mathcal{P}_1^- \Lambda^1(T) = \{u \in \mathfrak{X}(T) \mid u(x, y) = (b_1 + ax, b_2 + ay), a, b_1, b_2 \in \mathbb{R}\}. \quad (8.3.5)$$

The right-hand side of the above expression is the shape space for RT(1). Of course, $\star \mathcal{P}_1^- \Lambda^1(T)$ is the same space as $\mathcal{P}_1^- \Lambda^1(T)$, but working under the image of the Hodge star allows us to show that the DOFs for RT(1) follow naturally from the DOFs for $\mathcal{P}_1^- \Lambda^1(T)$.

We now prove that the Raviart–Thomas DOFs arise from the DOFs prescribed for the trimmed family in the special case $k = 1, r = 1$. First, observe that the DOFs for $\mathcal{P}_1^- \Lambda^1(T)$ associated to $f \in \Delta(T)$ may be written as moments over f weighed against elements of

$$\mathcal{P}_{1-\dim f} \Lambda^{\dim f-1}(f).$$

In the notation of Definition 8.3.1, we must therefore have that $W(T) = 0$ and, for all $p_i \in \Delta_0(T)$, $W(p_i) = 0$. Accordingly, the DOFs for $\mathcal{P}_1^- \Lambda^1(T)$ can only be associated to edges. Further, the space spanned by the weight functions is $\mathcal{P}_0 \Lambda^0(T) \simeq \mathbb{R}$, so the only candidates for DOFs associated to f are nonzero multiples of the averaging functional

$$\star u \mapsto \frac{1}{\text{length}(f)} \int_f \text{tr}_f \star u.$$

Let the unit outward normal on f be denoted by \mathbf{n} . Then, for all smooth 1-forms u defined on T ,

$$\text{tr}_f \star u = \mathbf{u} \cdot \mathbf{n}$$

once we switch to vector proxies. Using (8.3.5), however, we see immediately that all edge traces of $u \in \star \mathcal{P}_1^- \Lambda^1(T)$ are just constants. When restricted to $\star \mathcal{P}_1^- \Lambda^1(T)$, therefore, the averaging functional on f is the same as the point evaluation functional associated to any point in the interior of f . So, for the sake of drawing a nice symmetric picture like Figure 6.22, we may as well pick our DOFs to be

$$\text{eval}_{(\frac{1}{2}, 0)}, \text{eval}_{(0, \frac{1}{2})}, \text{ and } \text{eval}_{(\frac{1}{2}, \frac{1}{2})}.$$

These are the DOFs for the Raviart–Thomas element in Example 6.2.4. So, making use of the Hodge star, we see that the trimmed finite element with $k = 1, r = 1$ coincides with RT(1).

Naturally, then, we may define a general Raviart–Thomas finite element RT(r) with shape spaces consisting of $\mathcal{P}_r^- \Lambda^1(T)$ and DOFs prescribed by Definition 8.3.1. That is, the span of the DOFs consists of

- 1) all edge moments against elements of $\mathcal{P}_{r-1} \Lambda^0(f)$ and

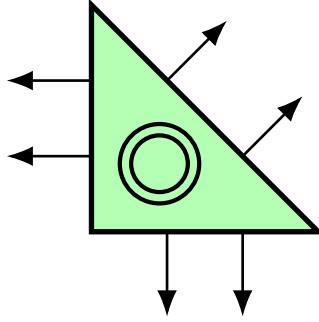


Figure 8.31: The degree 2 Raviart–Thomas finite element $\text{RT}(2)$ over the reference triangle, corresponding to $\star\mathcal{P}_2^-\Lambda^1(T_{\text{ref},2})$. Arrows represent edge moments of normal components against degree 1 polynomial functions, and circles represent interior moments against 1–forms with constant coefficients.

2) all interior moments against elements of $\mathcal{P}_{r-2}\Lambda^1(T)$.

So, $\text{RT}(r)$ has r DOFs per edge and $r(r-1)$ interior DOFs, for a total of $r(r+2)$ DOFs (we may check that this is equal to $\dim \mathcal{P}_r^-\Lambda^1(T)$). We illustrate the DOFs for $\text{RT}(2)$ in Figure 8.31: each circle indicates a distinct interior moment, while the arrows are now used to denote edge moments, rather than just point evaluations, of normal components.

Now, we discuss the finite element space $\star\mathcal{P}_r^-\Lambda^1(\mathcal{T}_h)$ obtained by sewing together the Raviart–Thomas elements. We also call this space $\text{RT}(r)$, since it is always clear from context whether we are referring to the finite element or the finite element space. There are two important properties of $\text{RT}(r)$ that bear mentioning before we finish this example. First, the requirement that all global DOFs for $\text{RT}(r)$ are single–valued implies that normal components across inter–element boundaries are continuous. Second, if we recall that $d\star$ acting on a 1–form u is the same as $\nabla\cdot$ acting on the corresponding vector proxy \mathbf{u} (upon dropping the area element $dx\wedge dy$), then conformity implies that all vector fields living in $\text{RT}(r)$ have a weak divergence that lives in $L^2(\Omega)$. In pre–FEEC nomenclature, we would then say that the Raviart–Thomas finite element spaces are $H(\text{div})$ –**conforming**, where the H indicates that the divergence is understood in the weak sense. These two properties exemplify a general relationship in FEM theory: $H(\text{div})$ –conforming finite element spaces over domains in \mathbb{R}^2 consist of vector fields whose normal components across inter–element boundaries are all continuous (this is an almost trivial corollary of Theorem 6.4.1). So, when we see a DOF diagram where the arrows are orthogonal to the facets, we should immediately realize that we are dealing with an $H(\text{div})$ –conforming approximation.

Example 8.3.5. In this example we use trimmed elements to introduce a natural generalization of the Raviart–Thomas elements to tetrahedra. Now, the most important property of the RT finite element spaces is their $H(\text{div})$ –conformity. So, any extension of RT spaces to 3D should be built with the (weak) divergence operator in mind. More precisely, the generalized RT space should consist of differential forms on which the exterior derivative reduces to the divergence operator. In \mathbb{R}^3 , however, we know that the exterior derivative on 2–forms is equal to $\nabla\cdot$ if we harmlessly drop the volume element as we did in 2D (notice how we do not have to use the Hodge star to obtain the divergence operator, as we did in 2D). Therefore, the clear choice of shape space for a “3D RT element” is $\mathcal{P}_r^-\Lambda^2(T_{\text{ref},3})$. Using Definition 8.3.1, we see that the correct DOFs are

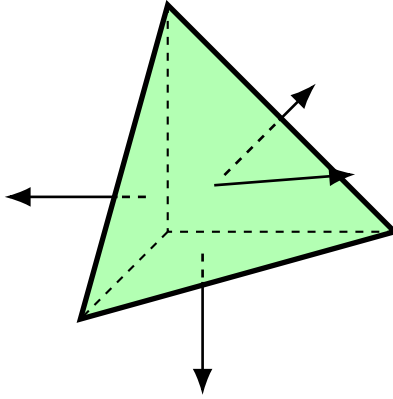


Figure 8.32: The degree 1 Nédélec face element of the first kind on the reference tetrahedron, $N1^f(1)$, corresponding to $\mathcal{P}_1^- \Lambda^2(T_{\text{ref},3})$.

- 1) moments over the facet f against elements of $\mathcal{P}_{r-1} \Lambda^0(f)$ and
- 2) interior moments against elements of $\mathcal{P}_{r-2} \Lambda^1$.

We see that the DOFs indeed resemble those for the RT elements: the degree of the weight functions for facet and interior moments is the same for both triangles and tetrahedra.

The tetrahedral finite element constructed in the previous paragraph is often referred to as the **Nédélec face element of the first kind** and denoted by $N1^f(r)$. This element was introduced by Nédélec in a seminal 1980 paper [64]. As in the previous example, we use $N1^f(r)$ to refer to both the finite element and the finite element space interchangeably. We continue to use this convention for finite element spaces throughout the thesis without further comment. Of course, the finite element space $N1^f(r)$ is $H(\text{div})$ -conforming. Figure 8.32 shows the DOF diagram for $N1^f(1)$, with the arrows aligned orthogonally to the faces of the tetrahedron. Since the $N1^f$ element really is just the “tetrahedral version” of the RT element, the DOF diagram is coloured the same way as the DOF diagrams for RT elements. Sometimes, the $N1^f(r)$ element is actually called the Raviart–Thomas element too (see for instance [54]). In this thesis, however, we try to keep our notation roughly in line with that of the Periodic Table [6].

Example 8.3.6. Let $T = T_{\text{ref},2}$. Way back in Example 6.2.4, we briefly mentioned that changing the DOFs for the lowest RT elements from normal component evaluation on edges to tangential component evaluation on edges also yielded a finite element. This finite element is called the **lowest Nédélec edge element of the first kind** on T and denoted by $N1^e(1)$ (by now, the reader has no doubt gathered that Nédélec is a busy guy). Since the edge trace of a 1-form on T is just the tangential component (relative to that edge) of the corresponding vector proxy, the DOFs for $N1^e(1)$ coincide perfectly with those of the trimmed element with shape space $\mathcal{P}_1^- \Lambda^1(T)$. That is, the finite element $N1^e(1)$ is just a special case of the trimmed elements constructed in Definition 8.3.1; we don’t even have to rotate the shape spaces using the Hodge star this time. The DOF diagram for $N1^e(1)$ is shown in Figure 8.33.

At this point, we can easily generalize the $N1^e(1)$ element to arbitrary shape function degree: $N1^e(r)$ is just another notation for the trimmed finite element with shape space $\mathcal{P}_r^- \Lambda^1(T)$ (modulo switching to vector proxies). The DOFs are

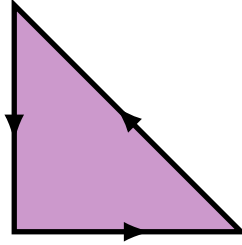


Figure 8.33: The degree 1 Nédélec edge element of the first kind on the reference triangle, $N1^e(1)$, corresponding to $\mathcal{P}_1^- \Lambda^1(T_{\text{ref},2})$.

- 1) moments of tangential components over edges against elements of $\mathcal{P}_{r-1} \Lambda^0(f)$ and
- 2) interior moments against elements of $\mathcal{P}_{r-2} \Lambda^1(T)$.

If u is a smooth 1-form on T with vector proxy \mathbf{u} and $\hat{\mathbf{z}}$ denotes the unit vector orthogonal to the plane, then

$$du = [\hat{\mathbf{z}} \cdot \nabla \times (\mathbf{u} + 0\hat{\mathbf{z}})] dx \wedge dy.$$

So, we can say that d acting on 1-forms in 2D corresponds with the curl operator. By d -conformity, then, any element of $N1^e(r)$ must have a weak curl (since that element must be the vector proxy of a 1-form with a weak derivative). In more customary FEM theory terminology, this means that the $N1^e$ spaces are $H(\text{curl})$ -conforming.

Now, as we have defined them, the $N1^e$ elements on triangles have two minor issues. First, there is a difficulty of proper attribution (discussed in [54]) in that Nédélec himself worked only with tetrahedra in [64]. Additionally, the $N1^e$ elements on triangles are not nearly as common in applications as the RT elements, and they do not appear on the Periodic Table (even though they would “fit”). We chose to include the triangular version of Nédélec’s finite elements in our work primarily for pedagogical reasons: 2D is easier to understand than 3D and, in the next example, we see that there are essentially no differences between Nédélec’s 3D $H(\text{curl})$ -conforming finite element spaces and the 2D finite element spaces presented above. So, while referring to the triangular $N1^e$ elements as “famous” might be a bit of stretch, they quickly generalize to unambiguously famous elements in 3D.

Example 8.3.7. Let $T = T_{\text{ref},3}$. This example is effectively a repetition of Example 8.3.5 with “ $H(\text{div})$ ” replaced with “ $H(\text{curl})$ ”. Our goal is to generalize the Nédélec edge elements of the first kind to tetrahedra (thus constructing the finite element spaces that Nédélec actually worked with and rectifying the issues brought up at the end of Example 8.3.6). We also refer to the generalized tetrahedral Nédélec edge elements as $N1^e$; since there ought not to be any confusion about whether a PDE involves 2 or 3 spatial dimensions, this notation does not cause trouble.

Recalling that the exterior derivative becomes the curl when acting on 1-forms in \mathbb{R}^3 , we reason that the finite element used to build the $H(\text{curl})$ -conforming generalization of $N1^e(r)$ is the trimmed element with shape space $\mathcal{P}_r^- \Lambda^1(T)$. Hence, the DOFs are

- 1) moments over the edges e weighed against elements of $\mathcal{P}_{r-1} \Lambda^0(e)$,

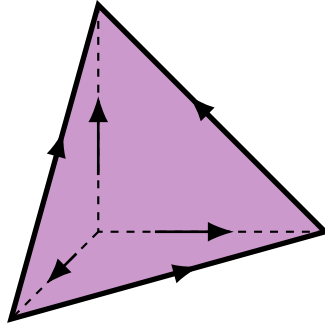


Figure 8.34: The degree 1 Nédélec edge element of the first kind on the reference tetrahedron, $N1^e(1)$, corresponding to $\mathcal{P}_1^- \Lambda^1(T_{\text{ref},3})$.

- 2) moments over the facets f weighed against elements of $\mathcal{P}_{r-2} \Lambda^1(f)$, and
- 3) interior moments against $\mathcal{P}_{r-3} \Lambda^2(T)$.

We illustrate the DOFs for $N1^e(1)$ in Figure 8.34. Notice how our comments in Example 8.3.6 on normal versus tangential arrows in DOF diagrams apply in the 3D case too: Figure 8.34 shows the DOFs for an $H(\text{curl})$ -conforming element and involves arrows tangential to the facets while Figure 8.32 shows the DOFs for an $H(\text{div})$ -conforming element and involves arrows normal to the facets. Nédélec introduced his edge elements side-by-side with his face elements in [64].

Notice how the DOFs for the lowest $H(\text{div})$ -conforming elements we constructed are all associated to facets, while the DOFs for the lowest $H(\text{curl})$ -conforming elements are all associated to edges. Intuitively, we should expect this: in vector calculus, the divergence is related to integrals over 2D regions (by the Divergence Theorem) while the curl is related to integrals over lines or curves (by the classical Stokes' Theorem). The mathematical justification for this intuitive comparison is, essentially, contained in Theorem 6.4.1.

I would like to re-iterate that, in 3D, we did not have to use the Hodge star to obtain well-known finite elements at all, while in 2D we had to use the Hodge star to switch between $N1^e$ and RT elements. This is because, in 2D, we only have one type of differential forms (specifically, 1-forms) that can serve as vector proxies. Accordingly, we have to choose between whether we want our trimmed finite element space to be $H(\text{curl})$ or $H(\text{div})$ -conforming when we work in 2D. In 3D this choice is unnecessary because we have both 1-forms (where d acts as the curl) and 2-forms (where d acts as the divergence) capable of serving as vector proxies. Unfortunately, many authors ignore the importance of the Hodge star in obtaining $H(\text{div})$ -conforming spaces in 2D, making the connection of such spaces to FEEC a bit confusing for the beginner; even the elegant Periodic Table does not emphasize that we must take the Hodge star to get $H(\text{div})$ -conformity.

Example 8.3.8. Let $T = T_{\text{ref},n}$. In this example, we study the trimmed finite elements with shape spaces $\mathcal{P}_r^- \Lambda^n(T)$. Denote these finite elements by $\text{DG}(r-1)$ (the notation is explained over the course of the next paragraph). We might initially think that such finite elements are precise copies of $\text{CG}(r)$, as n -forms are dual to scalars. However, notice that $\mathcal{P}_r^- \Lambda^n(T) = \mathcal{P}_{r-1} \Lambda^n(T)$, so there is already one discrepancy between Lagrange elements and elements made of trimmed top forms. This discrepancy explains why we denote



Figure 8.35: DOFs for some $DG(r-1)$ finite elements, corresponding to $\mathcal{P}_r^-\Lambda^2(T_{\text{ref},2})$.

the degree in $DG(r-1)$ the way we do. Additionally, elements of the CG finite element spaces are globally continuous. Conversely, the finite element space we would obtain by patching the DG elements together has no inter-element continuity restrictions, since all top forms trivially have single-valued facet traces. So, the $DG(r-1)$ finite element space is full of discontinuous functions, and this is why we denote the space the way we do: the “DG” stands for “discontinuous Galerkin”. DG spaces have been used explicitly since the early 1970s, though some of the ideas underlying the use of discontinuous approximating functions date back to Godunov’s 1959 work on finite volume methods [39]. For a comprehensive historical overview of DG spaces, see [29, §1.2].

Examining Definition 8.3.1, we see that the only DOFs for $DG(r-1)$ are internal moments against $\mathcal{P}_{r-1}(T)$. Intuitively, we might expect that this is the case: since any form in the $DG(r-1)$ space is discontinuous, we are not allowed to take its traces on lower-dimensional subsimplices (or, at least, we gain nothing from doing so since the trace vanishes) hence the only DOFs must be associated to element interiors. Figure 8.35 shows the DOF diagrams for the first two DG spaces on triangles. As with the CG diagrams, solid dots represent point evaluation.

To review, in this section we have defined the $\mathcal{P}_r^-\Lambda^k(\mathcal{T}_h)$ family of finite element spaces. Then, we showed how plenty of famous finite elements (CG, RT, $N1^e$, $N1^f$, and DG) arise as special cases of our general construction. We have seen that FEEC provides a clear and concise way of referring to many important finite elements: while one may call the element in Example 8.3.5 the Nédélec face element of the first kind or the Raviart–Thomas element, saying that the element is simply the trimmed element with shape space $\mathcal{P}_r\Lambda^2(T_{\text{ref},3})$ is unambiguous.

8.4 The Regular Finite Element Space $\mathcal{P}_r\Lambda^k(\mathcal{T}_h)$

In this section, we introduce another general family of finite element spaces, denoted by $\mathcal{P}_r\Lambda^k(\mathcal{T}_h)$. We often refer to these spaces as making up the **regular family** because, when restricted to a simplex T , their elements live in the “regular”, unmodified space of polynomial k -forms $\mathcal{P}_r\Lambda^k(T)$. Our introduction to the regular family follows the same template as our introduction of the trimmed family in the previous section: we define the relevant finite elements, then prove unisolvence of the DOFs, discuss inter-element continuity, and explain how members of the regular family correspond to more well-known finite element spaces.

Definition 8.4.1. Let \mathcal{T}_h be a triangulation of Ω . The **regular finite element** consists of the following data:

1) The shape spaces are given by $\mathcal{P}_r\Lambda^k(T) \forall T \in \mathcal{T}_h$.

2) Pick any $T \in \Delta_n(\mathcal{T}_h)$ and $f \in \Delta(T)$. Define a space of real-valued functionals on $V(f)$ by

$$W(f) \doteq \left\{ u \in V(T) \mapsto \int_f \text{tr}_f u \wedge q \mid q \in \mathcal{P}_{r+k-\dim f}^- \Lambda^{\dim f - k}(f) \right\}. \quad (8.4.1)$$

Then, the span of the degrees of freedom on T is given by

$$\sum_{f \in \Delta(T)} W(f). \quad (8.4.2)$$

Choose local-to-global maps as in Definition 8.3.1. The finite element space associated to this data is called the regular finite element space, denoted by $\mathcal{P}_r\Lambda^k(\mathcal{T}_h)$.

Now, recall that the Lagrange finite element space is precisely the trimmed finite element space of scalars. However, our notation for Lagrange elements agrees with the *regular* finite element space of scalars. Before proving unisolvence for the regular family, we demonstrate that this apparently ambiguous notation is actually not an issue: the space of scalars for both the trimmed and regular families is just the Lagrange space.

Lemma 8.4.2. We have that

$$\mathcal{P}_r^- \Lambda^0(\mathcal{T}_h) = \mathcal{P}_r \Lambda^0(\mathcal{T}_h),$$

where the space on the right-hand side is interpreted as $\mathcal{P}_r \Lambda^k(\mathcal{T}_h)$ in the case $k = 0$.

Proof. By Lemma 7.3.13 all of the shape spaces are the same, so it remains to check that the functionals in the span of the DOFs are also all the same. Observe that, if $f \in \Delta_d(T)$ for $T \in \Delta_n(\mathcal{T}_h)$, then

$$\mathcal{P}_{r-d}^- \Lambda^d(f) = \mathcal{P}_{r-d-1} \Lambda^d(f) \bigoplus \kappa \mathcal{H}_{r-d+1} \Lambda^{d+1}(f) = \mathcal{P}_{r-d-1} \Lambda^d(f) \bigoplus \kappa(0) = \mathcal{P}_{r-d-1} \Lambda^d(f).$$

So, the spaces where the polynomial weight functions of the DOFs live are identical. The proof is complete. \square

Theorem 8.4.3. In the notation of Definition 8.4.1, we have

$$\mathcal{P}_r \Lambda^k(T)^* = \bigoplus_{f \in \Delta(T)} W(f). \quad (8.4.3)$$

Proof. Using Lemma 8.2.3 and induction on n , we know immediately that $\mathcal{P}_r \Lambda^k(T)^* = \sum_{f \in \Delta(T)} W(f)$. All that remains is the dimension count. Using Proposition 7.3.15 and arguments we've already seen for Lagrange spaces, we get

$$\begin{aligned} \sum_{f \in \Delta(T)} \dim \mathcal{P}_{r-\dim f+k}^- \Lambda^{\dim f - k} &= \sum_{d=0}^n |\Delta_d(T)| \dim \mathcal{P}_{r-d+k}^- \Lambda^{d-k} \\ &= \sum_{d=0}^n \binom{n+1}{d+1} \binom{r+k}{k} \binom{r-1}{d-k}. \end{aligned}$$

If we apply Lemma 7.1.4 with $a = n + 1$, $b = k + 1$, and $c = r - 1$ then

$$\begin{aligned}
\sum_{f \in \Delta(T)} \dim \mathcal{P}_{r-\dim f+k}^- \Lambda^{\dim f-k} &= \binom{r+k}{k} \binom{n+r}{n-k} \\
&= \frac{(n+r)!}{r!k!(n-k)!} \\
&= \frac{(n+r)!n!}{n!r!k!(n-k)!} \\
&= \binom{n+r}{r} \binom{n}{k} \\
&= \dim \mathcal{P}_r \Lambda^k(T),
\end{aligned}$$

where we have applied Corollary 7.2.2 in obtaining the last equality. Earlier arguments can now be applied to complete the proof. \square

The same argument used to establish d-conformity for the trimmed family (presented after the proof of Theorem 8.3.2) yields that all of the regular finite element spaces are d-conforming:

$$\mathcal{P}_r \Lambda^k(\mathcal{T}_h) \subseteq H \Lambda^k(\Omega).$$

So, again, the $\mathcal{P}_r \Lambda^k(\mathcal{T}_h)$ spaces allow for many different inter-element continuity conditions. We can also use the Hodge star to obtain a modified version of the regular family (useful for obtaining $H(\text{div})$ -conforming spaces in 2D), as in the previous subsection.

Now, we go through a few examples that exhibit how members of the regular family showed up in FEM theory long before FEEC was formally introduced. These examples are much shorter than those presented at the end of Section 8.3 for the simple reason that the arguments used to establish the “FEEC-classical” correspondence are essentially identical regardless of whether we deal with the trimmed or regular family. We have already gone through such arguments in detail with the last bout of examples; everything essentially just boils down to connecting simple properties of differential forms to their analogues in vector calculus, and then checking that all the polynomial degrees work out.

Example 8.4.4. Clearly, the Lagrange spaces and discontinuous spaces are also in the regular family. Note, however, that $\mathcal{P}_r^- \Lambda^n(\mathcal{T}_h)$ consists of piecewise polynomials of degree at most $r - 1$, while $\mathcal{P}_r^- \Lambda^n(\mathcal{T}_h)$ consists of piecewise polynomials of degree at most r . So, there are some subtle differences between the way that the discontinuous spaces work in the different families.

Example 8.4.5. Let $T = T_{\text{ref},2}$. In this example we construct the **Brezzi–Douglas–Marini element** on T , denoted $\text{BDM}(r)$, where r is the shape form degree. $\text{BDM}(r)$ was originally introduced in a 1985 collaboration between Douglas and the husband-and-wife team of Brezzi and Marini [13]. The BDM elements patch together to give an $H(\text{div})$ -conforming finite element space, essentially making BDM the regular version of RT. $\text{BDM}(r)$ is the “rotated” regular element, with shape space $\star \mathcal{P}_r \Lambda^1(T)$ and DOFs consisting of

- 1) edge moments of normal components against $\mathcal{P}_r^- \Lambda^0(f)$ and
- 2) interior moments against $\mathcal{P}_{r-1}^- \Lambda^1(T)$.

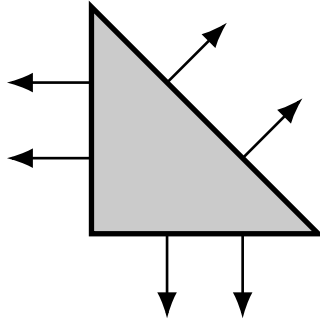


Figure 8.41: The degree 1 Brezzi–Douglas–Marini finite element BDM(1) over the reference triangle, corresponding to $\star\mathcal{P}_1\Lambda^1(T_{\text{ref},2})$.

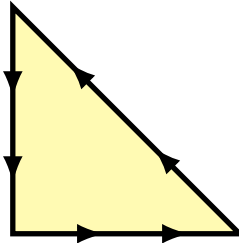


Figure 8.42: The degree 1 Nédélec edge element of the second kind on the reference triangle, $\text{N}2^e(1)$, corresponding to $\mathcal{P}_1\Lambda^1(T_{\text{ref},2})$.

The DOFS are illustrated in Figure 8.41.

On tetrahedra, the Brezzi–Douglas–Marini element generalizes to the **Nédélec face element of the second kind**, denoted $\text{N}2^f(r)$, similarly to how $\text{RT}(r)$ generalizes to $\text{N}1^f(r)$. Though $\text{N}2^f$ elements were actually introduced by Brezzi, Douglas, Durán, and Fortin in 1987 [12], we maintain the attribution to Nédélec to stay consistent with the Periodic Table (of course, the developments of [12] are very much inspired by Nédélec’s work on extending RT elements to 3D). Since we do not focus much on 3D examples in the sequel, we are content to stick with the tetrahedral elements from the trimmed family. Further discussion on $\text{N}2^f$ is outsourced to the references [10, 12, 54].

Example 8.4.6. For the sake of completeness, we introduce an $H(\text{curl})$ -conforming space built with regular elements. If we just take away the “ $-$ ” superscripts in Example 8.3.6, we can construct an $H(\text{curl})$ -conforming finite element with shape spaces consisting of regular 1-forms, $\mathcal{P}_r\Lambda^1(T)$. This element is the **Nédélec edge element of the second kind** on triangles, denoted $\text{N}2^e(r)$. The DOF diagram for $\text{N}2^e(1)$ is shown in Figure 8.42.

When contrasted with Figures 6.22 and 8.33, Figures 8.41 and 8.42 illustrate an important difference between the trimmed and regular elements. Namely, the trimmed spaces have fewer degrees of freedom. Accordingly, we expect that computing with finite elements arising from the trimmed family is faster than working with the regular spaces.

In summary, we have developed a second family of finite element spaces that can be used in conjunction with (or in place of) the trimmed spaces. We have also shown that the regular spaces, constructed in the FEEC framework, are really just familiar finite element spaces cast in a new, unified notation, precisely as we saw was the case for the trimmed spaces. Some of the well-studied finite elements that appear in the regular family include CG, BDM, $N2^e$, $N2^f$, and DG. Now that we have a veritable arsenal of finite element spaces, we can actually start analyzing whether or not they give rise to good numerical approximations. Of course, “good” is a protean term in numerical analysis, but in the next chapter we see how developing FEMs with FEEC in mind leads to simulations with desirable properties.

Chapter 9

Cohomology of Finite Element Spaces

In Chapter 8 we showed how to construct finite element spaces satisfying Definition 6.3.5. However, we did not prove that they actually have good approximation properties. In the Introduction, we described three properties that a procedure for numerically solving a PDE should have. Of course, the precise definitions of these properties depend on the particular problem under consideration, but the definitions are always along the following lines:

- 1) **Stability**: the discretized problem should be well-posed (in particular, the norm of the numerical solution should be controlled by the parameters of the exact problem, irrespective of the discretization parameters);
- 2) **Convergence**: the approximate solution should tend to the exact solution in the limit of “infinite precision” (for FEMs, this usually means the limit of mesh size tending to zero);
- 3) **Consistency**: our approximation of the differential operator appearing in the PDE should tend to the exact differential operator in the limit of “infinite precision”.

FEEC provides a unified framework for discussing all three of the above properties, though as mentioned earlier we only really focus on discussing consistency in this thesis. This focus is due to time constraints as well as pedagogy; in order to understand the analytical formalism used to prove stability and convergence in modern papers like [5], one must be comfortable with the algebraic formalism used to discuss consistency improvements.

When we start thinking about discretization as a morphism between Hilbert complexes (see the discussion below Remark 4.2.5), we start to view consistency as a homological property. Rather than saying a consistent method is one that just satisfies some particular error estimate (see, for instance, [20, p. 191]), we might also say that a consistent method should “mimic” the Hilbert complex structure of the problem we want to solve. This leads us to think about finding a natural diagnostic for determining whether or not structure is being preserved during discretization. Since cohomology is such a vital object in Hilbert complex theory (cohomology describes the kernel of the Hodge Laplacian by Theorem 4.4.3 and, therefore,

determines Hodge–Helmholtz decomposition), we might try to investigate whether or not our numerical method “has cohomology” isomorphic to the de Rham cohomology. In order to follow this line of questioning, however, we need to define a notion of cohomology associated to an FEM. We therefore have a little program for investigating the consistency of FEMs:

- 1) describe how some collections of finite element spaces can be viewed as subcomplexes of the de Rham complex, called **finite element de Rham complexes**;
- 2) construct a “discretization morphism” mapping the de Rham complex to a finite element de Rham complex, after the discussion below Remark 4.2.5;
- 3) assess conditions under which the cohomology of the finite element complex is the same as the cohomology of the full de Rham complex (for ease of reference, we call the cohomology of the subcomplex the **discrete cohomology** and the cohomology of the supercomplex the **exact cohomology**).

The present chapter is devoted to going through the above steps in detail. We eventually see that we can guarantee the existence of an isomorphism between the exact and discrete cohomologies by choosing finite element spaces carefully; that is, the consistency of an FEM is largely determined by the choice of approximating spaces. We also see that studying the consistency of FEMs through the above program reveals novel connections between the “classical” spaces studied in Chapter 8. Such connections are beautifully and concisely represented in the Periodic Table of the Finite Elements [6] (see also Figures 9.53, 9.54).

The origins of FEEC lie in many different papers and books spread over a wide variety of subjects including numerical analysis, differential topology, geometric analysis, and electric engineering. The historical development of the homological perspective on FEMs is summarized very thoroughly in [4, Introduction] and [5, §1.3] and, as mentioned in the Introduction, I do not intend to repeat their comprehensive discussion here.

In Section 9.1, we develop some abstract results on the approximation theory of Hilbert complexes. In Section 9.2, we discuss the abstract results of the previous section in the specific context of FEMs, stating the existence of a discretization operator and examining some of its properties. In Section 9.3, we study the simplest finite element de Rham complex, called the **Whitney complex**, in order to build some strong consistency results. In Section 9.4, we discuss two important theorems: first, we prove de Rham’s Theorem using only tools from FEM theory, and then we prove one of the most important results in FEEC, the Christiansen Triple–Decker Theorem [18]. The latter theorem provides useful conditions for establishing an isomorphism between discrete and exact cohomologies. In Section 9.5, we discuss some simple examples of cohomology–preserving choices of finite element spaces before describing how the Periodic Table can be used to quickly make such choices in more generality.

Throughout this chapter, Ω denotes a polyhedral domain in \mathbb{R}^n and $\{\mathcal{T}_h\}_h$ denotes a shape–regular family of triangulations of Ω (see Definition 5.1.11).

9.1 Approximating Hilbert Complexes

In this section we define **bounded cochain projections**, helping us formalize what we mean by “approximating” an abstract bounded Hilbert complex (V, d) by a subcomplex (V_h, d) . Often, the complex (V, d)

appears in applications as the domain complex of a bigger, unbounded Hilbert complex (see the discussion under Definition 4.2.2 for a reminder on the definition of a domain complex). Tacitly, we always think of the spaces V_h^k making up the complex V_h as finite-dimensional (since our ultimate goal is to have a subcomplex usable for practical computations), but we often do not need to assume finite-dimensionality to establish our theorems. We then show that, if the error in approximation of harmonic forms is less than 100%, then the cohomology of (V_h, d) is isomorphic to the cohomology of (V, d) . We are thus able to characterize when the most important algebraic invariant of (V, d) is preserved upon passing to the discrete setting. In the next section, we actually show that this abstract theory is applicable to approximations of the L^2 de Rham complex built from the finite element spaces presented in Chapter 8.

Before diving into general theory, we present an example motivating our focus on cohomology preservation a bit more specifically. Along the way, we discuss some of the subtleties associated with approximating Hodge–Helmholtz decompositions. Let (V, d) be the domain complex of a closed Hilbert complex (W, d) . Using the constructions from Section 4.4, we can formulate an abstract version of the Poisson equation thus: for a fixed nonzero $f \in W^k$, find $u \in V^k \cap \text{dom } d^{k-1} *$ such that

$$(d^k d^{k-1} * + d^{k-1} * d^k) u = \Delta^k u = f. \quad (9.1.1)$$

If u satisfies (9.1.1) then, for every $q \in \mathfrak{H}^k$, $u + q$ satisfies (9.1.1) too. An elementary argument shows that if we also demand that $u \in (\mathfrak{H}^k)^\perp$ then the solution to (9.1.1) is unique. If \mathfrak{H}^k is finite-dimensional, as it often is in applications, we can guarantee uniqueness of a solution u by imposing $\dim \mathfrak{H}^k$ constraints on u . Remembering that \mathfrak{H}^k is isomorphic to the k^{th} cohomology of (V, d) by the Hodge Theorem (Theorem 4.4.3), we can equivalently say that cohomology helps us determine uniqueness of solutions to (9.1.1).

A specific example introduced in [5, Example 2.3.3, Figure 2.4] indicates that we cannot ignore the above correspondence between well-posedness and cohomology when developing FEMs. This example involves solving for a 1-form u on the annulus

$$\Omega = \left\{ \frac{1}{4} \leq x^2 + y^2 \leq 1 \right\} \subseteq \mathbb{R}^2$$

such that the Poisson equation

$$\Delta^1 u = x \, dy \quad (9.1.2)$$

is satisfied. We impose the boundary conditions

$$\text{tr}_{\partial\Omega} \star u = 0 \quad \text{and} \quad \text{tr}_{\partial\Omega} \star du = 0,$$

corresponding to the vanishing of both

- 1) the normal component of u and
- 2) the vertical component of the curl of u

on the boundary; see Remark 4.4.7 for a reminder on why the above boundary conditions must be satisfied by the harmonic 1-forms on Ω . Vector proxies of approximate solutions of (9.1.2) obtained using two distinct FEMs are shown in Figure 9.11. [5] contains a few more specific details, but for our purposes all we have to know is that the solution on the left is calculated assuming that both its components live in CG(1), while the solution on the right is calculated using a **mixed formulation**; we discuss mixed finite

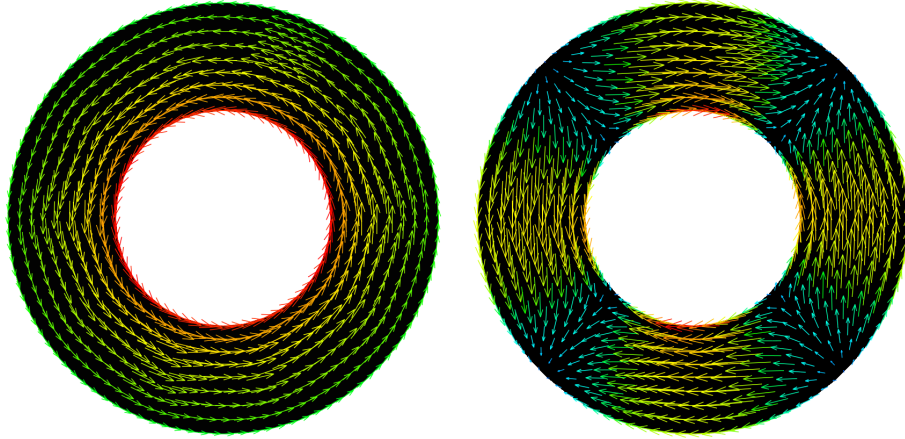


Figure 9.11: Figure 2.4 in [5], showing two numerical solutions to (9.1.2) obtained with different FEMs. The solution u on the right is orthogonal to \mathfrak{H}_h^1 , while the solution on the left is of the form $u + q$ for some $q \in \mathfrak{H}^1$ with relatively large norm.

element methods more in Section 9.5 and Part 2. Both solutions display very different behaviour, leading us to suspect that there are issues with at least one of the FEMs. As it turns out, the solution on the right is correct and unique (this is proved in [5]), while the solution on the left has some issues that warrant more thorough investigation.

We now briefly discuss why the approximate solution on the left side of Figure 9.11 is technically correct, but not at all useful. Since Ω has a single 1–dimensional hole, we know that its first cohomology space is 1–dimensional, hence by the Hodge Theorem and the de Rham Theorem $\dim \mathfrak{H}^1 = 1$. Therefore, solutions to (9.1.2) are not unique unless we add the constraint $u \in (\mathfrak{H}^1)^\perp$ to our problem. Now, in the solution shown on the right of Figure 9.11, the method requires that the discrete solution must be orthogonal to the space of (approximately) harmonic 1–forms \mathfrak{H}_h^1 (this space is discussed in more detail in a few more paragraphs). No such constraint appears in the fully continuous method shown on the left, and this is the cause of the disparity between the two approximate solutions. According to the authors of [5], the approximate solution on the left includes a contribution from \mathfrak{H}^1 that completely obfuscates the solution on the right: if the harmonic part of the approximation has a large enough norm, then the simulation looks like it is outputting a solution of $\Delta^1 u = 0$, rather than a solution of (9.1.2). We have thus seen that the unique solution to (9.1.2) cannot be obtained using *any* FEM; unless we explicitly demand discrete orthogonality in the method, we cannot be sure that our simulation “sniffs out” the unique solution.

Remark 9.1.1. *In practice, one must be able to compute all of the approximately harmonic forms in order to impose orthogonality constraints on solutions to the Poisson equation. This reduces to finding a basis for the kernel of a certain matrix, described in more detail in [5, p. 31].*

However, developing a cohomology–preserving approximation is nontrivial. Suppose that we would like to develop an approximate solution u_h to (9.1.1) using a finite–dimensional subcomplex (V_h, d) of (V, d) . In light of the discussion above, we anticipate the need to impose certain constraints on our approximate solution u_h in order to guarantee uniqueness. These constraints should intuitively be related to the discrete

version of the Hodge Laplacian on the subcomplex (V_h, d) . Specifically, we would like to try obtaining a unique approximate solution by demanding that u_h is in the orthogonal complement of the space of discrete harmonic forms,

$$\mathfrak{H}_h^k \doteq \ker d^k|_{V_h^k} \cap (dV_h^{k-1})^\perp.$$

We do not know in general whether $\dim \mathfrak{H}_h^k = \dim \mathfrak{H}^k$ unless we have better information on how well elements of (V_h, d) approximate elements of (V, d) . Of course, if we can prove that $\mathfrak{H}_h^k \simeq \mathfrak{H}^k$ (or, equivalently, that the cohomology of the subcomplex is isomorphic to the cohomology of the supercomplex), then we indeed have the right number of constraint equations.

Since (V_h, d) is a subcomplex, the image/kernel of $d^k|_{V_h^k}$ is contained in the image/kernel of d^k . However, there are no such clean inclusions for the adjoint operators d^{k*} because the adjoint for the subcomplex, denoted d_h^{k*} , is a priori distinct from d^{k*} . Why is this the case? We want the Hodge–Helmholtz decomposition of V_h^k to be orthogonal with respect to the inner product on W^k , $\langle \cdot, \cdot \rangle$. Therefore, for $u \in V_h^{k+1}$, $d_h^{k*}u$ is computed by solving the system

$$\langle d_h^{k*}u, v \rangle = \langle u, d^k v \rangle \quad \forall v \in V_h^k.$$

Since the V_h^k 's are finite-dimensional, this system trivially has a unique solution regardless of what u we pick. In particular, we can define $d_h^{k*}u$ even if $u \notin \text{dom } d^{k*} \cap V_h^{k+1}$. While the discrete adjoint d_h^{k*} and the exact adjoint d^{k*} must agree with one another on the set $\text{dom } d^{k*} \cap V_h^{k+1}$, the two operators are very different beasts. In the same vein, we have that

$$(dV_h^k)^\perp \supseteq (dV^k)^\perp$$

(orthogonal complements are taken with respect to $\langle \cdot, \cdot \rangle$). Therefore, the space of discrete harmonic k -forms \mathfrak{H}_h^k is in general a superset of $\mathfrak{H}^k \cap V_h^k$. In particular, a discrete harmonic form is not necessarily an exact harmonic form. We might then have that $\dim \mathfrak{H}_h^k \neq \dim \mathfrak{H}^k$, thus the discrete Hodge–Helmholtz decomposition

$$V_h^k = \text{range } d^k \oplus \mathfrak{H}_h^k \oplus \text{range } d_h^{k*}$$

may not be very useful; in passing to the discrete setting, we could conceivably accrue falsely harmonic forms, or even have exactly harmonic forms that are not discretely harmonic due to error associated with the approximation of d^{k*} . Therefore, we simply do not know if approximating a solution to the abstract Poisson equation by studying (V_h, d) reproduces the correct amount of uniqueness constraints.

Therefore, we need to impose more specific restrictions on the approximating subcomplex (V_h, d) in order to make sure that the spaces of exact and discrete harmonic forms are isomorphic. Fortunately, such restrictions are nicely phrased in terms of the discretization operators referenced in the introduction to this chapter and the paragraph below Remark 4.2.5.

Remark 9.1.2. *Establishing an isomorphism $\mathfrak{H}_h^k \simeq \mathfrak{H}^k$ is important for problems other than the Poisson equation. In the case of the L^2 de Rham complex, establishing such an isomorphism is vital because harmonic forms are, by the theorems of Hodge and de Rham, intimately connected to the topology of the spatial domain. Therefore, if we are trying to use an FEM that does not accurately reproduce the space \mathfrak{H}^k , we are ignoring important topological information.*

Now, we turn to developing theory that allows us to get some rough conditions for establishing an isomorphism $\mathfrak{H}_h^k \simeq \mathfrak{H}^k$. As mentioned in the opening paragraph of this section, we only deal with bounded

Hilbert complexes in this section. Recall that \mathbf{HilbC} denotes the category whose objects are bounded Hilbert complexes and whose morphisms are described by Definition 4.2.4. We now define bounded cochain projections formally.

Definition 9.1.3. *Let $(V, d) \in \text{Obj}(\mathbf{HilbC})$ with (V_h, d) a subcomplex. If $\pi_h \in \text{Hom}((V, d), (V_h, d))$ satisfies*

1) π_h^k is surjective and

2) $\pi_h^k \circ \pi_h^k = \pi_h^k$

for all k , then π_h is called a **bounded cochain projection**.

By definition of a morphism in \mathbf{HilbC} , any bounded cochain projection π_h must also satisfy the following two requirements for all k :

1) $\pi_h^k: V^k \rightarrow V_h^k$ is a bounded linear operator, and

2) we have a commuting diagram

$$\begin{array}{ccc} V^k & \xrightarrow{d} & V^{k+1} \\ \pi_h^k \downarrow & & \downarrow \pi_h^{k+1} \\ V_h^k & \xrightarrow{d} & V_h^{k+1} \end{array}$$

In terms of scientific computing, we can think of the supercomplex (V, d) as the setting for the exact problem (that is, the weak form of a given PDE posed over some suitable Sobolev space) and the subcomplex (V_h, d) as the setting for the discrete problem. Bounded cochain projections can then be understood simply as structure-preserving maps taking functions or forms considered in the exact problem to their discrete analogues. That is, we deal with $u \in V^k$ when we are studying a PDE with pen and paper, but we deal with $\pi_h^k u \in V_h^k$ when we are studying an approximate version of that PDE on a computer. From this point of view, bounded cochain projections are literally operators formalizing the passage from the exact problem to the discrete problem. In the parlance of traditional finite element theory, π_h^k might be called an **interpolation operator**.

Remark 9.1.4. *Since a bounded cochain projection π_h must be surjective, the induced map on cohomology $H\pi_h$ must also be surjective. Using the Hodge Theorem (Theorem 4.4.3), we conclude that if $\mathfrak{H}^k = 0$, then $\mathfrak{H}_h^k = 0$ too.*

Suppose that (V, d) arises as the domain complex of a Hilbert complex (W, d) . One may initially think that constructing bounded cochain projections is trivial when the spaces V_h^k are all finite-dimensional: since finite-dimensional subspaces are closed we may apply the Projection Theorem to obtain a family of bounded projection operators $\text{Proj}^k: W^k \rightarrow V_h^k$, defined by taking an element of W^k to its best approximation in V_h^k (since these operators are bounded on each W^k , they are bounded on each V^k). However, the operators Proj^k do not in general commute with the differentials d^k . For example, let $V^k = H\Lambda^k(-1, 1)$

for $k = 0, 1$ and let $V_h^0 = \mathcal{P}_1\Lambda^0(-1, 1)$ and $V_h^1 = \mathcal{P}_0\Lambda^1(-1, 1)$. The diagram

$$\begin{array}{ccc} H\Lambda^0(-1, 1) & \xrightarrow{d} & H\Lambda^1(-1, 1) \\ \text{Proj}^0 \downarrow & & \downarrow \text{Proj}^1 \\ \mathcal{P}_1\Lambda^0(-1, 1) & \xrightarrow{d} & \mathcal{P}_0\Lambda^1(-1, 1) \end{array}$$

does not commute; a straightforward computation with the orthonormal basis $\left\{\sqrt{\frac{1}{2}}, \sqrt{\frac{3}{2}}x\right\}$ of $\mathcal{P}_1\Lambda^0(-1, 1)$ shows that

$$d(\text{Proj}^0 \sin x) \neq \text{Proj}^1(d \sin x).$$

Therefore, the naïve projections onto closed subspaces are not strong enough for our purposes. We are left with an amusing paradox: we do not desire to approximate our exact solution in terms of its “best approximation”.

The above discussion indicates that π_h^k does not provide the best approximation to $u \in V^k$ in V_h^k . However, the following result shows that the error in approximation by π_h^k decreases as the error in the best approximation decreases. Some numerical analysts say that such an error estimate is **quasioptimal** [5].

Proposition 9.1.5. *If π_h is a bounded cochain projection, then for all k there exists some constant $C > 0$ depending only on π_h^k such that for any $u \in V^k$,*

$$\|u - \pi_h^k u\| \leq C \inf_{v \in V_h^k} \|u - v\|. \quad (9.1.3)$$

Proof. Since π_h^k is equal to the identity map on the subspace V_h^k , the result follows immediately from the definition of a bounded operator. \square

Next, we prove a simple, elegant theorem establishing how bounded cochain projections help us to discretely reproduce cohomology, provided a very modest estimate is satisfied.

Theorem 9.1.6. [5, Theorem 3.4] *Let $(V, d) \in \text{Obj}(\mathbf{HilbC})$ be closed with (V_h, d) a subcomplex, and let $\pi_h \in \text{Hom}((V, d), (V_h, d))$ be a bounded cochain projection. Denote the cohomology of (V, d) by \mathbf{H}^\bullet and the cohomology of (V_h, d) by \mathbf{H}_h^\bullet . Suppose that, for every nonzero $q \in \mathfrak{H}^k$, we have*

$$\frac{\|q - \pi_h^k q\|}{\|q\|} < 1. \quad (9.1.4)$$

Then, the induced map $\mathbf{H}\pi_h: \mathbf{H}^\bullet \rightarrow \mathbf{H}_h^\bullet$ is an isomorphism.

Proof. Surjectivity of the induced map follows immediately from surjectivity of π_h^k . Now, we prove that the induced map is injective. If $u \in \ker d^k$ satisfies $\mathbf{H}\pi_h^k(\bar{u}) = 0$, we want to show that $u \in d^{k-1}V^{k-1}$ (so that $\bar{u} = 0$). Since the complex is closed, we may use the Hodge–Helmholtz decomposition to find $v \in V^{k-1}$ and $q \in \mathfrak{H}^k$ such that

$$u = d^{k-1}v + q.$$

By definition of the induced map, $H\pi_h^k(\bar{u}) = 0$ means that $\pi_h^k u \in d^{k-1}V_h^{k-1}$. Since π_h^k commutes with the differentials, we know that $\pi_h^k d^{k-1}v \in d^{k-1}V_h^{k-1}$ as well, hence $\pi_h^k q \in d^{k-1}V_h^{k-1}$. However, $d^{k-1}V_h^{k-1} \subseteq d^{k-1}V^{k-1}$, and $d^{k-1}V^{k-1} \subseteq (\mathfrak{H}^k)^\perp$ by orthogonality of the Hodge–Helmholtz decomposition. Therefore,

$$\langle q, \pi_h^k q \rangle = 0.$$

If q is nonzero, then the Pythagorean Theorem gives

$$\frac{\|q - \pi_h^k q\|}{\|q\|} \geq 1.$$

The above inequality contradicts the estimate (9.1.4), hence $u = d^{k-1}v$ and the proof is complete. \square

In view of the Hodge Theorem (Theorem 4.4.3), we ought to expect that controlling the error in approximation of harmonic forms is necessary to establish that an approximating subcomplex has the same cohomology as its supercomplex. Therefore, the above result is a little more intuitive than it might initially seem. However, the weakness of the error estimate (9.1.4) is still quite a pleasant surprise, leading us to hope that we can actually develop cohomology–mimicking numerical schemes in practical applications.

We now turn to modifying Theorem 9.1.6 in order to make it more convenient for use in the next section. When numerically solving a PDE, we often want to control the error in our approximation of the exact solution by adjusting some parameter in the discretization scheme. When using an FEM, for instance, taking the limit as the mesh size parameter goes to zero should generate a sequence of approximate solutions that converges to the exact solution. Accordingly, for this last result we treat h as a real–valued parameter.

Corollary 9.1.7. *Let $(V, d) \in \text{Obj}(\mathbf{HilbC})$ be closed. Suppose that, for all k , $\dim \mathfrak{H}^k < \infty$. Additionally, suppose there exists a family of subcomplexes (V_h, d) and bounded cochain projections π_h indexed by the real number $h > 0$ such that, for every $u \in V^k$, we have*

$$\lim_{h \rightarrow 0} \|u - \pi_h^k u\| = 0. \tag{9.1.5}$$

Then, for h sufficiently small, the induced map on cohomology $H\pi_h$ is an isomorphism.

Proof. Let $m = \dim \mathfrak{H}^k$ and pick an orthonormal basis $\{e_i\}_{i=1}^m$ of \mathfrak{H}^k . We define the ℓ^1 norm of

$$q = \sum_{i=1}^m q_i e_i \in \mathfrak{H}^k$$

by

$$\|q\|_{\ell^1} \doteq \sum_{i=1}^m |q_i|.$$

Since all norms on finite–dimensional spaces are equivalent there exists $C > 0$ such that, for all $q \in \mathfrak{H}^k$,

$$\|q\|_{\ell^1} \leq C \|q\|.$$

Now, for each $i = 1, \dots, m$ let $h_i > 0$ be small enough so that

$$\|e_i - \pi_{h_i}^k e_i\| < \frac{1}{C} \|e_i\|.$$

Such an h_i can always be selected because (9.1.5) holds. Choose $h < \min h_i$.

Pick any $q \in \mathfrak{H}^k$. Since the e_i form a basis, there exist $q_1, \dots, q_m \in \mathbb{R}$ such that

$$q = \sum_{i=1}^m q_i e_i.$$

Then, we have

$$\|q - \pi_h^k q\| \leq \sum_{i=1}^m |q_i| \|e_i - \pi_{h_i}^k e_i\| < \frac{1}{C} \sum_{i=1}^m |q_i| \|e_i\| = \frac{1}{C} \sum_{i=1}^m |q_i| = \frac{1}{C} \|q\|_{\ell^1} \leq \|q\|.$$

Applying Theorem 9.1.6, the proof is complete. \square

Remark 9.1.8. Notice how Theorem 9.1.6 and Corollary 9.1.7 both rely on (V, d) being closed; without closure, we cannot use the Hodge–Helmholtz decomposition.

The above corollary says that, provided the kernel of the Hodge Laplacian is finite-dimensional and our approximation of (V, d) by (V_h, d) can be refined so that the error in approximation is arbitrarily small, the cohomologies of the subcomplex and supercomplex coincide. Given Theorem 9.1.6, this result is intuitive: if we can make the approximation error arbitrarily small by varying the parameter h , then obviously the error in approximating harmonic forms can eventually be made less than 100%. Finite-dimensionality of \mathfrak{H}^k is necessary to show that we can go from the pointwise statement (9.1.5) to the uniform bound (9.1.4).

In this section, we began by studying approximations of the Poisson equation in order to provide some concrete motivation. We then introduced bounded cochain projections, morphisms representing the passage from an exact problem to a discrete problem. Following this, we studied conditions under which a bounded, closed Hilbert complex could be approximated in a cohomology-preserving way. In particular, we saw a reasonably weak condition for establishing an isomorphism between the exact and discrete cohomologies given a sufficiently fine mesh. Now, we are ready to begin studying the approximation theory of Hilbert complexes in the context of finite elements.

9.2 Projections onto Finite Element de Rham Complexes

In this section we show how the finite element spaces constructed in Chapter 8 can be used to form subcomplexes of the L^2 de Rham complex. Then, with the developments of the previous section in mind, we have only to establish the existence of a bounded cochain projection in order to start talking about how well FEMs preserve cohomology. Fortunately, the construction of such projections is explicitly described in the FEM literature. The details of the relevant proofs, however, are very technical and beyond the scope of this thesis. Accordingly, in the discussion of bounded cochain projections associated to FEMs presented here, we only mention a few of the big ideas and state slightly incomplete versions of the relevant

results. After this loose discussion of projection operators, we apply results from the previous section to understand the homological properties of finite element approximations of the L^2 de Rham complex,

$$0 \rightarrow H\Lambda^0(\Omega) \xrightarrow{d} H\Lambda^1(\Omega) \xrightarrow{d} \dots \xrightarrow{d} L^2\Lambda^n(\Omega) \rightarrow 0,$$

written more succinctly as $(H\Lambda^\bullet, d)$.

We begin by stringing together the finite element spaces from Chapter 8 to form Hilbert complexes. Recall that, in Chapter 7, we investigated various cochain complexes associated to the regular and trimmed polynomial spaces on an n -simplex T . In particular, we studied the **polynomial de Rham complex**,

$$0 \rightarrow \mathcal{P}_r\Lambda^0(T) \xrightarrow{d} \mathcal{P}_{r-1}\Lambda^1(T) \xrightarrow{d} \dots \xrightarrow{d} \mathcal{P}_{r-n}\Lambda^n(T) \rightarrow 0,$$

and the **generalized Whitney complex**,

$$0 \rightarrow \mathcal{P}_r^-\Lambda^0(T) \xrightarrow{d} \mathcal{P}_r^-\Lambda^1(T) \xrightarrow{d} \dots \xrightarrow{d} \mathcal{P}_r^-\Lambda^n(T) \rightarrow 0.$$

Both of these complexes naturally extend to more complicated cochain complexes,

$$0 \rightarrow \mathcal{P}_r\Lambda^0(\mathcal{T}_h) \xrightarrow{d} \mathcal{P}_{r-1}\Lambda^1(\mathcal{T}_h) \xrightarrow{d} \dots \xrightarrow{d} \mathcal{P}_{r-n}\Lambda^n(\mathcal{T}_h) \rightarrow 0 \quad \text{and} \quad (9.2.1a)$$

$$0 \rightarrow \mathcal{P}_r^-\Lambda^0(\mathcal{T}_h) \xrightarrow{d} \mathcal{P}_r^-\Lambda^1(\mathcal{T}_h) \xrightarrow{d} \dots \xrightarrow{d} \mathcal{P}_r^-\Lambda^n(\mathcal{T}_h) \rightarrow 0. \quad (9.2.1b)$$

Since all the regular and trimmed finite element spaces are d -conforming, we know that both (9.2.1a) and (9.2.1b) are subcomplexes of the L^2 de Rham complex. Keeping with the nomenclature of Chapter 8, we call (9.2.1a) the **regular finite element de Rham complex** and (9.2.1b) the **trimmed finite element de Rham complex** (the special case $r = 1$, the **Whitney complex**, is of paramount importance in the next two sections). In other sources [4, 5], (9.2.1a) is called the **Sullivan–Whitney complex**, since this complex was used by Sullivan in his work on topology in the 1970s (see the historical references in [5]). As remarked in [4], we could build many more finite element subcomplexes of the de Rham complex by mixing and matching the regular and trimmed spaces. For simplicity we only stick to the two complexes defined in (9.2.1). Both of these complexes are denoted by (Λ_h^\bullet, d) interchangeably, following our notation for the L^2 de Rham complex.

We can now associate discrete cohomology spaces to FEMs. Using the techniques of the last section, we examine how well the cohomology of (Λ_h^\bullet, d) agrees with the de Rham cohomology. In order to begin comparing the exact and discrete cohomologies, however, we need to construct bounded cochain projections $\pi_h: H\Lambda^\bullet \rightarrow \Lambda_h^\bullet$.

Recall from the previous section that projections onto best approximations cannot be taken as bounded cochain projections, since they are not, in general, morphisms. So, a naïve first guess at π_h does not work. Now, in Chapters 6 and 8 we emphasized that a finite element space Λ_h^k is defined so that every $\omega \in \Lambda_h^k$ can be uniquely reconstructed from the values that all of the global DOFs take on ω . Accordingly, we might think that our best hope of obtaining a usable π_h involves somehow extending the following reconstruction procedure to elements of $H\Lambda^k$:

- 1) choose $u \in \mathcal{C}\Lambda^k(\Omega)$,
- 2) assemble the values of the global DOFs acting on u , denoted by $\ell_i(u)$, and

3) reconstruct the unique element of Λ_h^k determined by the numbers $\ell_i(u)$.

The above procedure defines a projection $I_h^k: \mathcal{C}\Lambda^k(\Omega) \rightarrow \Lambda_h^k$ commuting with the exterior derivative [4, §4.9]. Unfortunately, since the DOFs for all of our finite element spaces from Chapter 8 depend on traces onto faces, I_h^k does not canonically extend to the entirety of $H\Lambda^k$: elements of $H\Lambda^k$ are not necessarily piecewise smooth, so taking traces on faces is not necessarily well-defined. Therefore, if we want to somehow modify I_h^k to transform it into a bounded cochain projection, we might anticipate having a great deal of work ahead of us. The details of such a modification are quite technical, so we simply state the end result without proof and defer precise discussions to the references mentioned in the next paragraph.

A rigorous construction of bounded cochain projections for finite element de Rham complexes was presented in 2006 by Arnold, Falk, and Winther [4]. Their results relied on certain additional assumptions about the triangulation \mathcal{T}_h beyond shape-regularity, but in 2008 Christiansen and Winther [19] showed that these assumptions could be somewhat relaxed. The 2010 paper of Arnold, Falk, and Winther [5, §5.4, §5.5] contains a concise summary of the modified construction.

Theorem 9.2.1. [5, Theorem 5.9] *Let Λ_h^k denote either of the finite element spaces $\mathcal{P}_r\Lambda^k(\mathcal{T}_h)$ or $\mathcal{P}_r^-\Lambda^k(\mathcal{T}_h)$, with (Λ_h^\bullet, d) denoting the corresponding finite element de Rham complex. There exists a bounded projection $\pi_h^k: L^2\Lambda^k(\Omega) \rightarrow \Lambda_h^k$ such that, for all $u \in L^2\Lambda^k(\Omega)$,*

$$\lim_{h \rightarrow 0} \|u - \pi_h^k u\|_{L^2\Lambda^k} = 0. \quad (9.2.2)$$

Further, when restricted to $H\Lambda^k(\Omega)$, the projections π_h^k collectively define a morphism of bounded Hilbert complexes $\pi_h: (H\Lambda^\bullet, d) \rightarrow (\Lambda_h^\bullet, d)$.

The above theorem tells us that we have bounded cochain projections from the de Rham complex to both finite element complexes in (9.2.1) and, additionally, these projections provide good approximations to L^2 forms (due to (9.2.2)). Actually, the theorem we stated here is slightly incomplete: more precise results in the aforementioned references tell us that the smoother u is, the faster the interpolation error $\|u - \pi_h^k u\|_{L^2\Lambda^k}$ tends to 0.

We now present a simple corollary indicating how well our finite element complexes are able to replicate de Rham cohomology.

Corollary 9.2.2. *Consider the same notational setup as Theorem 9.2.1. For h sufficiently small, the cohomology of (Λ_h^\bullet, d) is isomorphic to the cohomology of $(H\Lambda^\bullet, d)$ via the induced map $H\pi_h$.*

Proof. By Theorem 9.2.1 we have the existence of a bounded cochain projection, hence the results of Section 9.1 apply. Since $(H\Lambda^\bullet, d)$ has the compactness property (Theorem 4.5.6), its associated space of harmonic k -forms is finite-dimensional by Proposition 4.5.5. By Corollary 4.5.9 $(H\Lambda^\bullet, d)$ is closed. Combining both of these facts with (9.2.2), we apply Corollary 9.1.7 to finish the proof. \square

Remark 9.2.3. *Though the proof of the above corollary was mostly just straightforward application of earlier results, there is one subtlety worth remarking on. The identity (9.2.2) is written in terms of the $L^2\Lambda^k$ norm, while the similar statement in Corollary 9.1.7 is written in terms of the $H\Lambda^k$ norm. However, if we pick any $u \in H\Lambda^k$, we know that (9.2.2) holds for both u and du . Then, since π_h is a morphism, we immediately have that*

$$\lim_{h \rightarrow 0} \|u - \pi_h^k u\|_{H\Lambda^k}^2 = 0 \quad \forall u \in H\Lambda^k. \quad (9.2.3)$$

So, Corollary 9.1.7 is indeed applicable.

In later sections, we strengthen the above result profoundly, ultimately showing that the induced map on cohomology is an isomorphism *regardless* of the mesh size parameter h .

In summary, we have indicated that some of the finite element spaces constructed in Chapter 8 fit together to form subcomplexes of the L^2 de Rham complex. We then stated the existence of bounded cochain projections $\pi_h: (H\Lambda^\bullet, d) \rightarrow (\Lambda_h^\bullet, d)$. Finally, we used the abstract tools from Section 9.1 to show that certain finite element approximations reproduce exact cohomology.

9.3 de Rham Theory à la Whitney

For the time being, we briefly depart from the scientific computing mindset and study the topological structure of the Whitney complex

$$0 \rightarrow \mathcal{P}_1^- \Lambda^0(\mathcal{T}_h) \xrightarrow{d} \mathcal{P}_1^- \Lambda^1(\mathcal{T}_h) \xrightarrow{d} \dots \xrightarrow{d} \mathcal{P}_1^- \Lambda^n(\mathcal{T}_h) \rightarrow 0, \quad (9.3.1)$$

sometimes denoted $(\mathcal{P}_1^- \Lambda^\bullet(\mathcal{T}_h), d)$. The Whitney complex is the simplest of the finite element de Rham complexes introduced in the previous section, if only because its rungs have the lowest dimensions possible. The goal of this section is to show that the cohomology of (9.3.1) is isomorphic to the simplicial cohomology of \mathcal{T}_h . Once we prove this fact, we are able to extend Corollary 9.2.2 to (mostly) arbitrary mesh size parameters. For an alternative discussion, see [18].

Before beginning the mathematical material in earnest, I would like to make some “big picture” remarks on the purely topological significance of (9.3.1). Essentially, the Whitney complex is the simplest cochain complex for which de Rham theory can be developed. Whitney put this point of view forward in his 1957 text on integration theory [89], inspired in part by a 1952 paper of Weil (the relevant citation to Weil is in [89, p. 139]). Using the Whitney complex, Whitney was able to prove a slightly modified version of de Rham’s Theorem [89, Thm. 29a, pp. 142–143]. The idea of his proof is to explicitly build an inverse for the de Rham homomorphism between k -forms and k -cochains

$$\omega \mapsto \left([c] \mapsto \int_c \omega \right),$$

but we do not have enough time to learn the details. Whitney’s approach is similar in spirit to de Rham’s original 1931 proof of the theorem bearing his name [24]. Specifically, de Rham built a family of “easy” differential forms canonically corresponding with simplices in the triangulation, in turn making the required inverse of the de Rham homomorphism trivial to define. However, unlike Whitney, de Rham did not explicitly use the jargon of homological algebra, nor did he work with piecewise-linear forms. For a more comprehensive sketch of de Rham’s proof, see [27, pp. 63–64]. Now, since at least the 1980s, many textbooks and lecture courses have proven de Rham’s Theorem using either sheaf theory or abstract methods from homological algebra. Accordingly, Whitney’s almost combinatorial perspective on de Rham cohomology may come across as novel to the modern reader; one could argue that Whitney’s work (and its modern interpretation in the context of FEM theory) is an interesting demonstration of the continuing applicability of Poincaré’s intuitionistic, partition-focused perspective on topology.

We begin by defining a sort of (slightly degenerate) coordinate system on a triangulation \mathcal{T}_h . Throughout the remainder of this section, we use the notation

$$\Delta_0(\mathcal{T}_h) = \{p_0, \dots, p_N\}.$$

That is, vertices of \mathcal{T}_h are referred to as p_j . We also let σ denote an element of $\Sigma_0(k, N)$, the set of increasing maps $\{0, \dots, k\} \rightarrow \{0, \dots, N\}$.

Definition 9.3.1. *The global barycentric coordinate functions $\lambda_i \in \mathcal{P}_1\Lambda^0(\mathcal{T}_h)$ are defined uniquely by the conditions*

1) $\lambda_i(p_j) = \delta_{ij}$, and

2) For all $p \in |\mathcal{T}_h|$ and all i , $0 \leq \lambda_i(p) \leq 1$.

Since the functions λ_i are defined by prescribing their values on vertices, the traces of λ_i on all faces of \mathcal{T}_h are single-valued. By definition of the DOFs for the Lagrange finite element space, then, we can be sure that each $\lambda_i \in \mathcal{P}_1\Lambda^0(\mathcal{T}_h) = \mathcal{P}_1^-\Lambda^0(\mathcal{T}_h)$. By d-conformity, therefore, the weak derivatives $d\lambda_i$ all exist.

Let $f = [p_{\sigma(0)} \dots p_{\sigma(k)}] \in \Delta_k(\mathcal{T}_h)$. The restrictions of the global barycentric coordinate functions $\lambda_{\sigma(i)}$ ($i = 0, \dots, k$) to f are precisely equal to the local barycentric coordinate functions on f . With this in mind, it is clear that the global barycentric coordinate function λ_i is either

1) an extension of some local barycentric coordinate by zero if $p_i \in \partial\Omega$, or

2) a ‘‘tent-shaped’’ function, equal to 1 at x_i and supported inside $\text{Star}(p_i)$, if $p_i \in \Omega^0$.

Additionally, using the properties of local barycentric coordinates, we see immediately that the global barycentric coordinates yield a sort of ‘‘partition of unity’’ on $|\mathcal{T}_h|$:

$$1 = \sum_j \lambda_j. \tag{9.3.2}$$

Before going any further, we go through a simple example to make some of the above discussion a little more concrete.

Example 9.3.2. Consider the triangulation of $[0, 2]$ given by

$$[0, 2] = [0, 1] \cup [1, 2].$$

We label the three vertices as $0 = p_0$, $1 = p_1$, and $2 = p_2$. The global barycentric coordinate functions with respect to this triangulation are written in terms of the Cartesian coordinate x as

$$\lambda_0 = \begin{cases} 1 - x & \text{if } 0 \leq x \leq 1 \text{ and} \\ 0 & \text{if } 1 \leq x \leq 2; \end{cases}$$

$$\lambda_1 = \begin{cases} x & \text{if } 0 \leq x \leq 1 \text{ and} \\ 2 - x & \text{if } 1 \leq x \leq 2; \end{cases}$$

$$\lambda_2 = \begin{cases} 0 & \text{if } 0 \leq x \leq 1 \text{ and} \\ x - 1 & \text{if } 1 \leq x \leq 2. \end{cases}$$

The graph of λ_1 is tent-shaped: λ_1 starts from zero at each boundary point, rising up to equal one at the interior vertex $x = 1$. Further, we directly verify that $1 = \sum_j \lambda_j$.

Now, we are comfortable enough with the barycentric coordinates to begin presenting some interesting results. First, we define a linear map taking k -cochains to piecewise polynomial k -forms. We soon see that this map essentially allows us to identify the cohomology of $(\mathcal{P}_1^- \Lambda^\bullet(\mathcal{T}_h), d)$ with the simplicial cohomology of \mathcal{T}_h . In the remainder of this section, we let $T \in \Delta_n(\mathcal{T}_h)$ and $c = [p_{\sigma(0)} \dots p_{\sigma(k)}] \in C^k(\mathcal{T}_h)$.

Definition 9.3.3. Define a linear map $\varphi: C^k(\mathcal{T}_h) \rightarrow \mathcal{P}_1^- \Lambda^k(\mathcal{T}_h)$ by the following action on elementary k -cochains $c = [p_{\sigma(0)} \dots p_{\sigma(k)}]$:

$$\varphi(c) = k! \sum_{i=0}^k (-1)^i \lambda_{\sigma(i)} d\lambda_{\sigma(0)} \wedge \dots \wedge \widehat{d\lambda_{\sigma(i)}} \wedge \dots \wedge d\lambda_{\sigma(k)}. \quad (9.3.3)$$

If c is an elementary k -cochain then $\varphi(c)$ is called an **elementary k -form**.

Lemma 9.3.4. For all $c \in C^k(\mathcal{T}_h)$, $\varphi(c) \in \mathcal{P}_1^- \Lambda^k(\mathcal{T}_h)$.

Proof. (Sketch, compare with [4, §4.3]) By linearity it suffices to prove that the elementary k -forms live in $\mathcal{P}_1^- \Lambda^k(\mathcal{T}_h)$. Pick any elementary k -cochain $c = [p_{\sigma(0)} \dots p_{\sigma(k)}]$. Let $T \in \Delta_n(\mathcal{T}_h)$. Clearly, the switch from Cartesian to barycentric coordinates on T is accomplished via an affine transformation, so the radial vector field on T may be expressed as

$$\sum_{i=1}^n x^i \frac{\partial}{\partial x^i} = \sum_{j=1}^n \lambda_j \frac{\partial}{\partial \lambda_j}.$$

Note that we use the convention that the 0th barycentric coordinate is the “dependent one”, as we have done previously. Suppose that $0 \notin \text{range } \sigma$. Then, we have that the action of the Koszul operator on $\text{tr}_T(d\lambda_{\sigma(0)} \wedge \dots \wedge d\lambda_{\sigma(k)})$ is given by

$$\begin{aligned} \kappa(\text{tr}_T(d\lambda_{\sigma(0)} \wedge \dots \wedge d\lambda_{\sigma(k)})) &= \sum_{j=1}^n \left(\lambda_j \frac{\partial}{\partial \lambda_j} \right) \text{tr}_T(d\lambda_{\sigma(0)} \wedge \dots \wedge d\lambda_{\sigma(k)}) \\ &= \sum_{i=0}^k (-1)^i \text{tr}_T(\lambda_{\sigma(i)}) \text{tr}_T(d\lambda_{\sigma(0)} \wedge \dots \wedge \widehat{d\lambda_{\sigma(i)}} \wedge \dots \wedge d\lambda_{\sigma(k)}) \\ &= \frac{1}{k!} \text{tr}_T \varphi([p_{\sigma(0)} \dots p_{\sigma(k)}]). \end{aligned}$$

Now, $\kappa(\text{tr}_T(d\lambda_{\sigma(0)} \wedge \dots \wedge d\lambda_{\sigma(k)})) \in \kappa \mathcal{H}_0 \Lambda^{k+1}(T) \subseteq \mathcal{P}_1^- \Lambda^k(T)$.

If $0 \in \text{range } \sigma$, conversely, then we must have $0 = \sigma(0)$, since σ is increasing. Using the Leibniz rule for contractions [51, Lemma 14.13], we find that

$$\kappa(\text{tr}_T(d\lambda_0 \wedge d\lambda_{\sigma(1)} \wedge \dots \wedge d\lambda_{\sigma(k)})) = \frac{1}{k!} \text{tr}_T \varphi([p_0 p_{\sigma(1)} \dots p_{\sigma(k)}]) - \text{tr}_T(d\lambda_{\sigma(1)} \wedge \dots \wedge d\lambda_{\sigma(k)}).$$

Since $\text{tr}_T(d\lambda_{\sigma(1)} \wedge \dots \wedge d\lambda_{\sigma(k)}) \in \mathcal{P}_0 \Lambda^k(T)$, we know that $\text{tr}_T \varphi([p_0 p_{\sigma(1)} \dots p_{\sigma(k)}]) \in \mathcal{P}_1^- \Lambda^k(T)$. So, in either case, the elementary k -forms belong to the trimmed spaces piecewise.

To complete the proof, we have to show that the DOFs for $\mathcal{P}_1^- \Lambda^k(\mathcal{T}_h)$ are single-valued on each elementary form $\varphi(c)$. If each $\varphi(c)$ is weakly differentiable then by Theorem 6.4.1 the facet traces of $\varphi(c)$ are single-valued and so all DOFs are single-valued on $\varphi(c)$. Now, showing that $\varphi(c)$ has a weak derivative is very simple if we know that d obeys the product rule. However, proving the product rule rigorously requires a good deal of time to fine-tune certain density arguments, so we just take it for granted here; essentially, all of the ingredients necessary to complete the proof are contained in [11, pp. 265–270]. \square

Example 9.3.5. The elementary 0-form corresponding to the vertex p_j is equal to λ_j .

Example 9.3.6. Let $T = [p_0 p_1 p_2]$. We compute the elementary 1-forms on T . Any elementary 1-cochain in T may be represented as $c = [p_i p_j]$ with i, j ranging from 0 to 2. Using (9.3.3), we have

$$\varphi([p_i p_j]) = \lambda_i d\lambda_j - \lambda_j d\lambda_i.$$

Then, we simply plug in all the admissible values of i and j and use the condition $\lambda_0 = 1 - \lambda_1 - \lambda_2$:

$$\begin{aligned} \varphi([p_0 p_1]) &= (1 - \lambda_2) d\lambda_1 + \lambda_1 d\lambda_2, \\ \varphi([p_0 p_2]) &= \lambda_2 d\lambda_1 + (1 - \lambda_1) d\lambda_2, \quad \text{and} \\ \varphi([p_1 p_2]) &= -\lambda_2 d\lambda_1 + \lambda_1 d\lambda_2. \end{aligned}$$

In the notation of Example 6.2.4, we see that $\varphi([p_1 p_2])$ corresponds (upon switching from Cartesian to barycentric coordinates) to the vector field defined by $a_i = 0$ ($i = 1, 2$) and $b = 1$. Similarly, $\varphi([p_0 p_2])$ corresponds to $a_1 = 0, a_2 = 1$, and $b = -1$, and $\varphi([p_0 p_1])$ corresponds to $a_1 = 1, a_2 = 0$, and $b = 1$. Then, the elementary 1-forms are clearly a basis for $\mathcal{P}_1^- \Lambda^1(T)$, which corresponds (upon applying the Hodge star) to the shape space for the lowest Raviart–Thomas element.

Proposition 9.3.7. *Elementary forms constitute a basis of $\mathcal{P}_1^- \Lambda^k(\mathcal{T}_h)$.*

Proof. We essentially show that the elementary forms are the basis of $\mathcal{P}_1^- \Lambda^k(\mathcal{T}_h)$ dual to the DOFs from Definition 8.3.1, up to a scalar multiple. The proof is trivial if $k = n$, so we take $k \leq n - 1$. First, we show that there are $\dim \mathcal{P}_1^- \Lambda^k(T)$ distinct elementary k -forms associated to each $T \in \Delta_n(\mathcal{T}_h)$ (since the DOFs are single-valued on elementary forms, showing this establishes that there are $\dim \mathcal{P}_1^- \Lambda^k(\mathcal{T}_h)$ distinct elementary k -forms). By Proposition 7.3.15, $\dim \mathcal{P}_1^- \Lambda^k(T) = \binom{n+1}{k+1}$. We also know that

$$\dim C^k(T) = |\Delta_k(T)| = \binom{n+1}{k+1}.$$

Clearly, $\dim C^k(T)$ is precisely the number of elementary k -forms associated to T (if two elementary forms are equal then they must be associated to the same elementary cochain). Hence, we at least have that there are enough elementary forms to span $\mathcal{P}_1^- \Lambda^k(T)$.

Next, observe that the only DOF for $\mathcal{P}_1^- \Lambda^k(\mathcal{T}_h)$ associated to $f \in \Delta_k(\mathcal{T}_h)$ is

$$u \mapsto \int_f \text{tr}_f u.$$

Suppose, without loss of generality, that f is positively oriented. We prove that

$$\int_f \text{tr}_f(\varphi(c)) = \begin{cases} k! \text{vol}(f) & \text{if } c = f \text{ and} \\ 0 & \text{otherwise} \end{cases}$$

for all elementary k -cochains c . Write $f = [p_{\sigma(0)} \dots p_{\sigma(k)}]$ with $p_i \in \Delta_0(\mathcal{T}_h)$. Now, the restrictions of $\lambda_{\sigma(0)}, \dots, \lambda_{\sigma(k)}$ to f are simply the local barycentric coordinates on f . Accordingly, we have that

$$\sum_{j=0}^k \text{tr}_f(\lambda_{\sigma(j)}) = 1$$

and so

$$\text{tr}_f(d\lambda_{\sigma(0)}) = - \sum_{j=1}^k \text{tr}_f(d\lambda_{\sigma(j)}).$$

Therefore, for each nonzero i ,

$$\begin{aligned} \text{tr}_f \left(d\lambda_{\sigma(0)} \wedge \dots \wedge \widehat{d\lambda_{\sigma(i)}} \wedge \dots \wedge d\lambda_{\sigma(k)} \right) &= - \sum_{j=1}^k \text{tr}_f(d\lambda_{\sigma(j)} \wedge d\lambda_{\sigma(1)} \wedge \dots \wedge \widehat{d\lambda_{\sigma(i)}} \wedge \dots \wedge d\lambda_{\sigma(k)}) \\ &= (-1)^i \text{tr}_f(d\lambda_{\sigma(1)} \wedge \dots \wedge d\lambda_{\sigma(k)}). \end{aligned}$$

Substituting this result in the definition of $\text{tr}_f \varphi(f)$, we have

$$\begin{aligned} \text{tr}_f(\varphi(f)) &= k! \sum_{i=0}^k \text{tr}_f(\lambda_{\sigma(i)}) \text{tr}_f(d\lambda_{\sigma(1)} \wedge \dots \wedge d\lambda_{\sigma(k)}) \\ &= k! \text{tr}_f(d\lambda_{\sigma(1)} \wedge \dots \wedge d\lambda_{\sigma(k)}). \end{aligned}$$

Then, computing the integral of $\text{tr}_f \varphi(f)$ over f is easy. We have

$$\int_f \text{tr}_f(\varphi(f)) = k! \int_f \text{tr}_f(d\lambda_{\sigma(1)} \wedge \dots \wedge d\lambda_{\sigma(k)}) = k! \text{vol}(f).$$

If $c = [p_{\sigma'(0)} \dots p_{\sigma'(k)}] \neq f$ then at least one of the $\lambda_{\sigma'(i)}$ satisfies $\text{tr}_f \lambda_{\sigma'(i)} = 0$. Accordingly, $\text{tr}_f d\lambda_{\sigma'(i)} = 0$ too. Trivially, then, $\int_f \text{tr}_f(\varphi(c)) = 0$ and the proof is complete. \square

Corollary 9.3.8. $\varphi: C^k(\mathcal{T}_h) \rightarrow \mathcal{P}_1^- \Lambda^k(\mathcal{T}_h)$ is an isomorphism. \square

Before we get to the main theorem of this section, we need a helpful lemma.

Lemma 9.3.9. Let $c = [p_{\sigma(0)} \dots p_{\sigma(k)}]$. Then,

$$\sum_{\substack{j \notin \text{range } \sigma \\ p_j \in \text{Star}(c)}} d\lambda_j \wedge d\lambda_{\sigma(0)} \wedge \dots \wedge \widehat{d\lambda_{\sigma(i)}} \wedge \dots \wedge d\lambda_{\sigma(k)} = (-1)^{i+1} d\lambda_{\sigma(0)} \wedge \dots \wedge d\lambda_{\sigma(k)}. \quad (9.3.4)$$

Proof. Recall that the global barycentric coordinates satisfy

$$1 = \sum_j \lambda_j.$$

Taking the derivative of both sides, we see that

$$0 = \sum_j d\lambda_j. \quad (9.3.5)$$

Expanding out the sum more explicitly, we have

$$0 = \sum_{j \in \text{range } \sigma} d\lambda_j + \sum_{\substack{j \notin \text{range } \sigma \\ p_j \in \text{Star}(c)}} d\lambda_j + \sum_{\substack{j \notin \text{range } \sigma \\ p_j \notin \text{Star}(c)}} d\lambda_j. \quad (9.3.6)$$

The subscript on the third sum is a bit redundant, but it makes the ideas perfectly clear. Now, we take the wedge product of both sides with $d\lambda_{\sigma(0)} \cdots \wedge \widehat{d\lambda_{\sigma(i)}} \cdots \wedge d\lambda_{\sigma(k)}$ to obtain

$$\begin{aligned} 0 &= \sum_{j \in \text{range } \sigma} d\lambda_j \wedge d\lambda_{\sigma(0)} \cdots \wedge \widehat{d\lambda_{\sigma(i)}} \cdots \wedge d\lambda_{\sigma(k)} + \sum_{\substack{j \notin \text{range } \sigma \\ p_j \in \text{Star}(c)}} d\lambda_j \wedge d\lambda_{\sigma(0)} \cdots \wedge \widehat{d\lambda_{\sigma(i)}} \cdots \wedge d\lambda_{\sigma(k)} \\ &+ \sum_{\substack{j \notin \text{range } \sigma \\ p_j \notin \text{Star}(c)}} d\lambda_j \wedge d\lambda_{\sigma(0)} \cdots \wedge \widehat{d\lambda_{\sigma(i)}} \cdots \wedge d\lambda_{\sigma(k)}. \end{aligned}$$

For all j such that $p_j \notin \text{Star}(c)$, $d\lambda_j \wedge d\lambda_{\sigma(0)} \cdots \wedge \widehat{d\lambda_{\sigma(i)}} \cdots \wedge d\lambda_{\sigma(k)}$ is identically zero because $d\lambda_j$ is supported outside $\text{Star}(c)$ while $d\lambda_{\sigma(0)} \cdots \wedge \widehat{d\lambda_{\sigma(i)}} \cdots \wedge d\lambda_{\sigma(k)}$ is supported inside $\text{Star}(c)$. Using this fact together with the antisymmetry of the wedge product, we see that

$$0 = (-1)^i d\lambda_{\sigma(0)} \wedge \cdots \wedge d\lambda_{\sigma(k)} + \sum_{\substack{j \notin \text{range } \sigma \\ p_j \in \text{Star}(c)}} d\lambda_j \wedge d\lambda_{\sigma(0)} \cdots \wedge \widehat{d\lambda_{\sigma(i)}} \cdots \wedge d\lambda_{\sigma(k)},$$

and the proof is complete. \square

The next result is all but stated explicitly in Whitney's treatise [89]. This theorem says that we can completely understand simplicial cohomology using differential forms, even if the forms we deal with are piecewise linear. That is, all topological information contained in the de Rham complex is also contained in the simpler subcomplex $(\mathcal{P}_1^- \Lambda^\bullet, d)$.

Theorem 9.3.10. *The Whitney complex is isomorphic (in the category of cochain complexes) to the simplicial cochain complex by way of the map φ taking elementary k -cochains to elementary k -forms.*

Proof. By Corollary 9.3.8 each φ is an isomorphism. Accordingly, we need only show that φ is a morphism of cochain complexes; that is, the diagram below commutes:

$$\begin{array}{ccc} C^k & \xrightarrow{\partial^*} & C^{k+1} \\ \varphi \downarrow & & \downarrow \varphi \\ \mathcal{P}_1^- \Lambda^k & \xrightarrow{d} & \mathcal{P}_1^- \Lambda^{k+1} \end{array} \quad (9.3.7)$$

As usual, let $c = [p_{\sigma(0)} \cdots p_{\sigma(k)}]$ be any elementary cochain. Then,

$$\begin{aligned} d\varphi(c) &= k! \sum_{i=0}^k (-1)^i d\lambda_{\sigma(i)} \wedge d\lambda_{\sigma(0)} \wedge \cdots \wedge \widehat{d\lambda_{\sigma(i)}} \wedge \cdots \wedge d\lambda_{\sigma(k)} \\ &= k! \sum_{i=0}^k d\lambda_{\sigma(0)} \wedge \cdots \wedge d\lambda_{\sigma(k)}. \end{aligned}$$

Therefore,

$$d\varphi(c) = (k+1)! d\lambda_{\sigma(0)} \wedge \cdots \wedge d\lambda_{\sigma(k)}. \quad (9.3.8)$$

Next, we compute $\varphi(\partial^*c)$. Using the explicit formula for the coboundary operator (5.2.3), we have that

$$\begin{aligned} \varphi(\partial^*c) &= (k+1)! \sum_{\substack{j \notin \text{range } \sigma \\ p_j \in \text{Star}(c)}} \varphi([p_j p_{\sigma(0)} \cdots p_{\sigma(k)}]) \\ &= (k+1)! \sum_{\substack{j \notin \text{range } \sigma \\ p_j \in \text{Star}(c)}} \left[\lambda_j d\lambda_{\sigma(0)} \wedge \cdots \wedge d\lambda_{\sigma(k)} - \sum_{i=0}^k (-1)^i \lambda_{\sigma(i)} d\lambda_j \wedge d\lambda_{\sigma(0)} \wedge \cdots \wedge \widehat{d\lambda_{\sigma(i)}} \wedge \cdots \wedge d\lambda_{\sigma(k)} \right]. \end{aligned}$$

Now, $\lambda_{\sigma(0)}, \dots, \lambda_{\sigma(k)}$ vanish outside of $\text{Star}(c)$. Accordingly, the derivatives of these functions vanish outside $\text{Star}(c)$ as well and we obtain

$$\sum_{\substack{j \notin \text{range } \sigma \\ p_j \in \text{Star}(c)}} \lambda_j d\lambda_{\sigma(0)} \wedge \cdots \wedge d\lambda_{\sigma(k)} = \sum_{j \notin \text{range } \sigma} \lambda_j d\lambda_{\sigma(0)} \wedge \cdots \wedge d\lambda_{\sigma(k)}.$$

Therefore,

$$\begin{aligned} \varphi(\partial^*c) &= (k+1)! \sum_{j \notin \text{range } \sigma} \lambda_j d\lambda_{\sigma(0)} \wedge \cdots \wedge d\lambda_{\sigma(k)} \\ &\quad - (k+1)! \sum_{i=0}^k (-1)^i \lambda_{\sigma(i)} \sum_{\substack{j \notin \text{range } \sigma \\ p_j \in \text{Star}(c)}} d\lambda_j \wedge d\lambda_{\sigma(0)} \wedge \cdots \wedge \widehat{d\lambda_{\sigma(i)}} \wedge \cdots \wedge d\lambda_{\sigma(k)}. \end{aligned}$$

Now, we can apply Lemma 9.3.9 to see that

$$\begin{aligned} \varphi(\partial^*c) &= (k+1)! \sum_{j \notin \text{range } \sigma} \lambda_j d\lambda_{\sigma(0)} \wedge \cdots \wedge d\lambda_{\sigma(k)} + (k+1)! \sum_{i=0}^k \lambda_{\sigma(i)} d\lambda_{\sigma(0)} \wedge \cdots \wedge d\lambda_{\sigma(k)} \\ &= (k+1)! \sum_j \lambda_j d\lambda_{\sigma(0)} \wedge \cdots \wedge d\lambda_{\sigma(k)}. \end{aligned}$$

However, by (9.3.2), the above reduces to

$$\varphi(\partial^*c) = (k+1)! d\lambda_{\sigma(0)} \wedge \cdots \wedge d\lambda_{\sigma(k)}. \quad (9.3.9)$$

Comparing with (9.3.8) with (9.3.9), the proof is complete. \square

Remark 9.3.11. *The factor of $k!$ in the definition of the elementary forms is necessary in proving that φ is a morphism.*

Corollary 9.3.12. *The cohomology of the Whitney complex is isomorphic (in the category of graded vector spaces) to the simplicial cohomology of \mathcal{T}_h .* \square

So, in this section, we have seen that we can study the topology of a simplicial complex using a piecewise linear approximation of the de Rham complex. This required a bit of fiddling with the explicit formula for the coboundary operator, but the end result is rather elegant.

9.4 Two Big Theorems

In this section, we apply our new knowledge of the Whitney complex to obtain more information about the cohomology–mimicking properties of FEMs. We prove a deep result that we call the Christiansen Triple–Decker Theorem [18], telling us that our finite element de Rham complexes have cohomology that is isomorphic to the L^2 de Rham cohomology. Such isomorphisms are independent of the mesh size, drastically improving Corollary 9.2.2.

Christiansen’s theorem is one of the the great pillars of FEEC, and it certainly represents the culmination of all the theory we have developed so far. Before proving this theorem, however, we show that the Whitney complex allows us to re–prove de Rham’s Theorem for polyhedral domains. Thanks to the Whitney complex, both proofs are almost trivial, amounting to little more than drawing simple diagrams and applying some basic facts from algebraic topology and linear algebra.

Theorem 9.4.1. (*de Rham’s Theorem, Finite Element Version*) [5, §5.6, p. 61] *The simplicial cohomology of a polyhedral domain is isomorphic to the L^2 de Rham cohomology.*

Proof. Let φ be the isomorphism taking elementary cochains to elementary forms. We have the following commuting diagram:

$$\begin{array}{ccccccc}
 \dots & \xrightarrow{d} & H\Lambda^{k-1} & \xrightarrow{d} & H\Lambda^k & \xrightarrow{d} & H\Lambda^{k+1} & \xrightarrow{d} & \dots \\
 & & \pi_h^{k-1} \downarrow & & \pi_h^k \downarrow & & \downarrow \pi_h^{k+1} & & \\
 \dots & \xrightarrow{d} & \mathcal{P}_1^- \Lambda^{k-1}(\mathcal{T}_h) & \xrightarrow{d} & \mathcal{P}_1^- \Lambda^k(\mathcal{T}_h) & \xrightarrow{d} & \mathcal{P}_1^- \Lambda^{k+1}(\mathcal{T}_h) & \xrightarrow{d} & \dots \\
 & & \varphi^{-1} \downarrow & & \varphi^{-1} \downarrow & & \downarrow \varphi^{-1} & & \\
 \dots & \xrightarrow{\partial^*} & C^{k-1}(\mathcal{T}_h) & \xrightarrow{\partial^*} & C^k(\mathcal{T}_h) & \xrightarrow{\partial^*} & C^{k+1}(\mathcal{T}_h) & \xrightarrow{\partial^*} & \dots
 \end{array} \tag{9.4.1}$$

Since cohomology is a functor (see Proposition 4.3.4), we may combine Corollary 9.2.2 with Theorem 9.3.10 to see that for sufficiently small h the cohomology of the top row (the de Rham cohomology) is isomorphic to the cohomology of the bottom row (the simplicial cohomology of \mathcal{T}_h). However, after our discussions in Chapter 5, we know that the simplicial cohomology of \mathcal{T}_h is the same regardless of how small h is. The proof is now complete. \square

Now, we discuss the Christiansen Triple–Decker Theorem, originally proved by Christiansen in 2005 [18]. Even though this theorem is one of the most profound results in all of FEEC, the proof is startlingly simple.

Theorem 9.4.2. (*Christiansen Triple–Decker Theorem* [18]) *The finite element de Rham complexes (Λ_h^\bullet, d) have cohomologies that are isomorphic to the L^2 cohomology.*

Proof. We begin by building a 3–level commuting diagram (whence the theorem’s colourful name is derived) using the bounded cochain projections π_h from Theorem 9.2.1 and the projections I_h^k discussed in the

paragraphs directly above Theorem 9.2.1:

$$\begin{array}{ccccccc}
\dots & \xrightarrow{d} & H\Lambda^{k-1} & \xrightarrow{d} & H\Lambda^k & \xrightarrow{d} & H\Lambda^{k+1} \xrightarrow{d} \dots \\
& & \pi_h^{k-1} \downarrow & & \pi_h^k \downarrow & & \downarrow \pi_h^{k+1} \\
\dots & \xrightarrow{d} & \Lambda_h^{k-1} & \xrightarrow{d} & \Lambda_h^k & \xrightarrow{d} & \Lambda_h^{k+1} \xrightarrow{d} \dots \\
& & I_h^{k-1} \downarrow & & I_h^k \downarrow & & \downarrow I_h^{k+1} \\
\dots & \xrightarrow{d} & \mathcal{P}_1^- \Lambda^{k-1}(\mathcal{T}_h) & \xrightarrow{d} & \mathcal{P}_1^- \Lambda^k(\mathcal{T}_h) & \xrightarrow{d} & \mathcal{P}_1^- \Lambda^{k+1}(\mathcal{T}_h) \xrightarrow{d} \dots
\end{array} \tag{9.4.2}$$

Since the forms in each Λ_h^k are piecewise smooth with respect to \mathcal{T}_h , applying the projection operators I_h^k to elements of Λ_h^k indeed makes sense. Now, all of the π_h^k and I_h^k are surjective, hence both of the maps on cohomology induced by these projections are also surjective. By de Rham’s Theorem, the cohomology of the lowest level of the diagram is isomorphic to the cohomology of the highest level. Finishing the proof then amounts to using some basic linear algebra: we cannot have a surjective linear map from a space of dimension n to a space of dimension strictly greater than n . \square

With Christiansen’s theorem, we can confidently say that some (not all) finite element methods preserve de Rham cohomology. Accordingly, we can guarantee that the Hodge–Helmholtz decomposition is roughly preserved in the discrete setting (at least up to the quality of the approximation of the coderivative δ), provided we are careful about what finite element spaces we choose.

9.5 The Periodic Table of the Finite Elements

With all of the theory on cohomology taken care of, we can relax a bit and turn towards some more concrete applications. In this short section, we provide examples of cohomology–preserving choices of finite element spaces in two space dimensions. We also describe how such choices are nicely summarized in the Periodic Table of the Finite Elements [6]. We are essentially describing structure–preserving **mixed FEMs**, so called because they involve using several different finite element spaces at once.

Example 9.5.1. Consider a triangulation \mathcal{T}_h of a polyhedral domain in \mathbb{R}^2 . The “modified” Whitney complex

$$0 \rightarrow \mathcal{P}_1^- \Lambda^0(\mathcal{T}_h) \xrightarrow{d} \star \mathcal{P}_1^- \Lambda^1(\mathcal{T}_h) \xrightarrow{d} \mathcal{P}_1^- \Lambda^2(\mathcal{T}_h) \rightarrow 0$$

has cohomology isomorphic to the de Rham cohomology on Ω (one can show that taking the Hodge star of the 1–forms doesn’t affect the cohomology). By using vector proxies and switching to the “classical” names for the finite element spaces, we can equivalently write the Whitney complex over a domain in \mathbb{R}^2 as

$$0 \rightarrow \text{CG}(1) \xrightarrow{\nabla^\perp} \text{RT}(1) \xrightarrow{\nabla} \text{DG}(0) \rightarrow 0. \tag{9.5.1}$$

Sequences similar to (9.5.1) have appeared in treatments of FEM theory not focused on FEEC (see, for instance, the diagram on [10, p. 116]). In the literature, these are often called **de Rham diagrams** for obvious reasons. Figure 9.51 provides a pictorial representation of this special Whitney complex in terms of the DOF diagrams studied in Chapter 8.

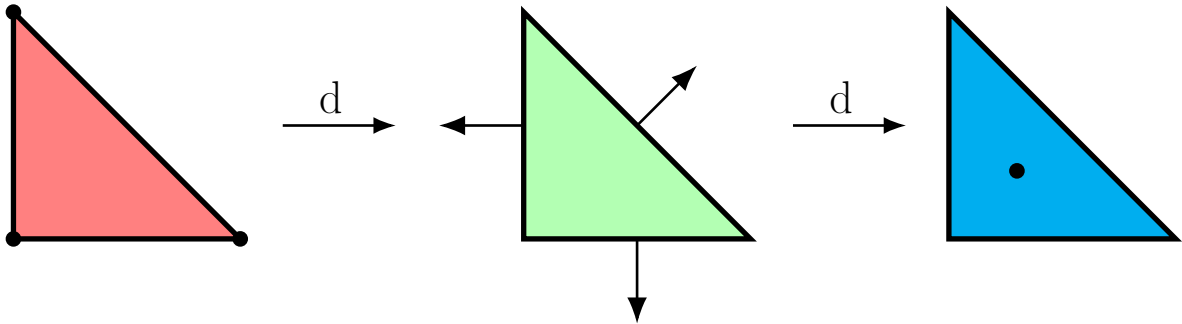


Figure 9.51: The simplest cohomology-preserving choice of finite element spaces built with the trimmed family, corresponding to the Whitney complex.

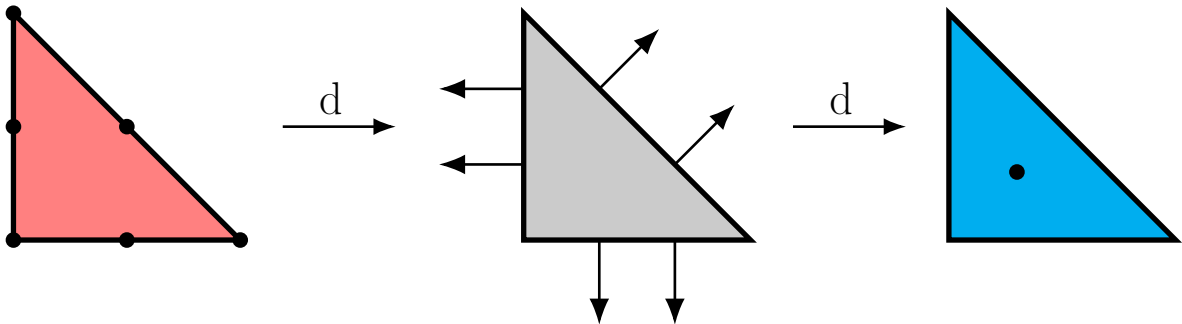


Figure 9.52: The simplest cohomology-preserving choice of finite element spaces built with the regular family.

Example 9.5.2. We can also repeat the construction of the previous example using the regular family, obtaining a finite element de Rham complex like so:

$$0 \rightarrow \mathcal{P}_2\Lambda^0(\mathcal{T}_h) \xrightarrow{d} \star\mathcal{P}_1\Lambda^1(\mathcal{T}_h) \xrightarrow{d} \mathcal{P}_0\Lambda^2(\mathcal{T}_h) \rightarrow 0.$$

Again, notice how in the trimmed case the polynomial degrees stay the same, while in the regular case each rung is one degree lower than the rung to its left. In more customary notation, the above sequence is written

$$0 \rightarrow \text{CG}(2) \xrightarrow{\nabla^\perp} \text{BDM}(1) \xrightarrow{\nabla \cdot} \text{DG}(0) \rightarrow 0. \tag{9.5.2}$$

See Figure 9.52 for an illustration. We re-iterate that the cohomology of (9.5.2) is isomorphic to the de Rham cohomology.

With these two concrete examples in mind, we are in a better position to discuss the Periodic Table of the Finite Elements and how it helps us develop consistent FEMs. I have included a version of the Periodic Table here (split between Figures 9.53 and Figures 9.54), though I encourage the reader to visit femtable.org to look at the interactive web version of the table too (I recommend opening this particular link in Firefox). The table presented here lists most of the finite element spaces compatible with finite

The $\mathcal{P}_r^- \Lambda^k$ family

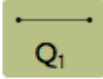
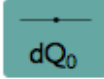
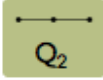
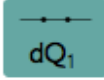
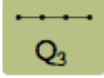
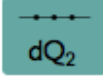





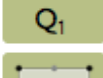

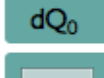


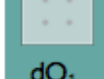
The $\mathcal{P}_r \Lambda^k$ family





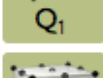

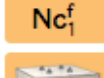


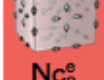
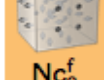
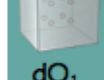


Figure 9.53: The first two columns of the Periodic Table of the Finite Elements. Cohomology-preserving families of finite element spaces are arranged horizontally in the first column and diagonally in the second column.

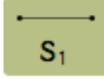
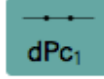
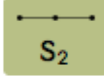
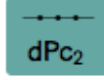
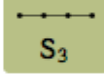
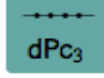
The $Q_r\Lambda^k$ family




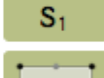

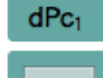
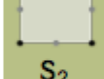

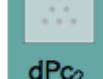
	$k=0$	$k=1$	$k=2$	$k=3$
$n=1$	 Q_1	 dQ_0		
	 Q_2	 dQ_1		
	 Q_3	 dQ_2		

$n=2$	 Q_1	 RTc_1	 dQ_0
	 Q_2	 RTc_2	 dQ_1
	 Q_3	 RTc_3	 dQ_2

$n=3$	 Q_1	 Nc_1^e	 Nc_1^f	 dQ_0
	 Q_2	 Nc_2^e	 Nc_2^f	 dQ_1
	 Q_3	 Nc_3^e	 Nc_3^f	 dQ_2

The $S_r\Lambda^k$ family

	$k=0$	$k=1$	$k=2$	$k=3$
$n=1$	 S_1	 dPc_1		
	 S_2	 dPc_2		
	 S_3	 dPc_3		

$n=2$	 S_1	 $BDMc_1$	 dPc_1
	 S_2	 $BDMc_2$	 dPc_2
	 S_3	 $BDMc_3$	 dPc_3





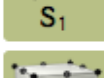

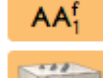

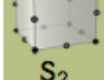

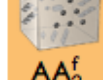
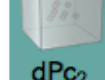
$n=3$	 S_1	 AA_1^e	 AA_1^f	 dPc_1
	 S_2	 AA_2^e	 AA_2^f	 dPc_2
	 S_3	 AA_3^e	 AA_3^f	 dPc_3

Figure 9.54: The third and fourth columns of the Periodic Table of the Finite Elements.

element exterior calculus: recent work of Gillette and Kloefkorn [38] indicates that at least one more column can be added to the table.

Two important remarks on the Periodic Table are in order. First, notice that we only introduced the first two columns (Figure 9.53) of the table in this thesis, namely those corresponding to finite elements defined over simplices. The last two columns (Figure 9.54) describe finite elements defined over quadrilaterals. Readers looking for a brief introduction to the FEEC perspective on quadrilateral finite elements may consult [3] or [38]. Second, we must describe how the table encodes information on cohomology. Looking at the fourth row of the first column, we see that it lists precisely the finite element spaces discussed in Example 9.5.1 (notice, however, that the table ignores the necessity of taking the Hodge star in obtain the RT space from the trimmed space of 1-forms). So, we can obtain a cohomology-preserving trimmed finite element de Rham complex just by reading off the spaces in this row from left to right. Checking the other rows in the first column and using the dictionary between FEEC and classical FEM jargon established in Chapter 8, we see that this pattern of choosing “good” spaces extends to all polynomial degrees and all spatial dimensions. However, if we use the same reasoning in the fourth row of the *second* column, we do not reproduce Example 9.5.2. So, the strategy for the first column is not universally applicable. If we are dealing with the second column, we actually have to choose families of spaces lying on diagonal lines rather than horizontal ones. Indeed, if we start on the $k = 0$ block of the fifth row of the second column and work our way up diagonally, we recover the spaces described in Example 9.5.2 (granted, the $DG(0)$ space isn’t reflected in the second column even though it “should” be there, so there are quite a few things that the table doesn’t remember for us).

So far in this chapter, we have seen some of the big “punchlines” of FEEC, so before going further a thorough review is in order. First, we defined bounded cochain projections in an abstract Hilbert complex setting, using them to make statements about how well we could approximate cohomology when passing to a subcomplex. Then, we introduced finite element de Rham complexes using the two main families of finite element spaces from Chapter 8 and proved a weak result on their capacity to mimic exact de Rham cohomology. From there, we studied the Whitney complex $(\mathcal{P}_1^- \Lambda^\bullet(\mathcal{T}_h), d)$ in great detail, proving that its cohomology is isomorphic to the simplicial cohomology of \mathcal{T}_h . Our examination of the Whitney complex led us to a finite element-theoretic proof of de Rham’s Theorem and then into Christiansen’s Triple Decker Theorem, which says we are guaranteed to reproduce de Rham cohomology on the discrete level, irrespective of a particular triangulation, provided we choose finite element spaces fitting together to give a subcomplex of the de Rham complex. After all of this heavy-lifting with general theory, we concluded by presenting some nice visual ways of thinking about the cohomology of FEMs. This endeavour culminated with a discussion of the Periodic Table of the Finite Elements. We saw that we could use the Periodic Table to find cohomology-preserving choices of finite element spaces by going horizontally in the first column (trimmed family) and diagonally in the second column (regular family).

Remark 9.5.3. *I have developed a computer demo that the reader may consult to see how the choices of finite element spaces advocated in this section can be applied to numerically solve a family of elliptic PDEs. Additionally, the demo includes the option to use finite element spaces not recommended by FEEC. Choosing these “bad” spaces can result in the numerical routine performing quite poorly, especially when compared to the “good” choices informed by FEEC. See Appendix A to learn how to download the demo, and to see a list of the software requirements you’ll need to actually use it.*

Part II

Applications to the Numerical Solution of the Green–Naghdi Equations

Chapter 10

The Green–Naghdi Equations

In this Chapter we derive and investigate some basic properties of the **Green–Naghdi equations (GN)**, a system of PDEs arising in ideal hydrodynamics. The GN generalize the more well-known **shallow water equations (SW)** that describe the vertically averaged motion of a thin layer of ideal fluid by adding some dispersive terms to the evolution equation for fluid velocity. These extra terms scale with the square of the relative thickness of the fluid layer and represent the influence of deviations from hydrostatic balance. Like the SW, the GN can easily be extended to include the effects of a rotating spatial domain, and so can potentially serve as a useful model in geophysical fluid dynamics. When we consider the GN or SW in a rotating reference frame I add an “R” (for “rotating”) to the acronym, so for example “RGN” stands for “rotating Green–Naghdi equations”.

Owing to the fact that the GN allow for non-hydrostatic effects, they should describe a wider range of physical phenomena than the SW and therefore be more realistic. This is indeed the case, but there is a more specific reason why the GN are preferable to the SW from a computational physicist’s perspective. Namely, the tendency for numerical simulations of the SW to develop shocks can be curbed by working with the GN instead: the additional dispersive terms present in GN tend to make solutions “spread out” in space, counteracting nonlinear steepening. The shock smoothing properties are the primary reason why we focus on the GN, rather than the SW, in the numerical simulations described in Chapter 12 .

Remark 10.0.1. *The reader who is skeptical about my claim that the GN are less susceptible to shock formation than the SW is encouraged to wait patiently until Section 12.4 and Appendix A, where we discuss a Firedrake demo allowing the user to easily switch between solving the GN or the SW. Using this demo to simulate the propagation of a single wave over a sudden “jump” in the seafloor, you can very quickly convince yourself that the GN are less likely to develop shocks than the SW.*

The RSW (and, by proxy, the SW) have been used to describe oceanic flows for over a century, appearing in their linearized form at least as early as 1880 in an article by Lord Kelvin [86]. As Kelvin himself indicates, many of the simplifying physical assumptions used to obtain the RSW were used in the eighteenth century by Laplace to derive equations governing the tides. So, in a loose sense, some of the ideas associated to the RSW have been around almost as long as mathematical fluid dynamics itself. Additionally, mid-twentieth-century investigations of the RSW revealed that they provide a relatively simple mathematical description of various physical phenomena investigated by oceanographers in the 1930s and 1940s; by studying the RSW, scientists could better quantify the “profound intuitive insights” [67, p. 58] of

oceanographers like Rossby and Stommel. Hence, the RSW became a ubiquitous mathematical modeling tool in geophysical fluid dynamics.

Conversely, the GN are much younger and much less famous than the SW. The 1D version of the GN first appeared in a 1953 paper of François Serre [78, Equation 22], and was independently rediscovered by Su and Gardner in 1969. The full GN, valid in both one and two space dimensions, were obtained in 1976 by Green and Naghdi [40], who derived the model by imposing certain invariance requirements on the Euler equations of ideal fluid motion. Later, different authors began to investigate the GN from the perspective of Lagrangian (that is, variational) and Hamiltonian classical field theory. Miles and Salmon [58] showed how the RGN could be derived using variational principles, and Camassa et al. [16] subsequently demonstrated that the RGN can be written as a non-canonical Lie–Poisson Hamiltonian system. Both of these publications were largely focused on applications of the RGN to oceanography, though the uses of this model are (potentially) not limited to the terrestrial: Dellar [26] applied the Hamiltonian description of the RGN to derive a model for the dynamics of the solar tachocline, a thin, rotating plasma layer within the Sun. The derivation of the GN and RGN presented in this chapter follows Dellar’s approach, though we do not discuss any of the electromagnetic effects required for applications to heliophysics.

In Section 10.1, we derive the GN using techniques from Hamiltonian fluid dynamics. In Section 10.2, we explain how the Hamiltonian description of the GN can be doctored to accommodate a rotating reference frame. In Section 10.3, we study some geophysically significant solutions to the linearized RGN.

In this chapter, the Einstein summation convention is always in effect, all functionals have unique variational derivatives (we say that such functionals are **smooth**), and unless otherwise stated all smooth vector fields are compactly supported.

10.1 Derivation of the Green–Naghdi Equations

In this section, we derive the **Green–Naghdi equations (GN)** using the machinery of non-canonical Hamiltonian fluid dynamics. The reader seeking a user-friendly introduction to the Hamiltonian viewpoint in fluid mechanics is encouraged to consult [76] or [84].

Let $h: \mathbb{R}^2 \times [0, \infty) \rightarrow [0, \infty)$ be smooth. Consider a fluid occupying the region

$$\Omega = \{(x, y, z) \mid (x, y) \in \mathbb{R}^2, 0 \leq z \leq h(x, y, t)\}.$$

Then, h represents the depth of the fluid (see Figure 10.11). We denote the fluid’s velocity by $\mathbf{u} = (u, v, w) \in \mathfrak{X}(\Omega)$. The simplifying physical constraints imposed on the GN are as follows:

- 1) the fluid is ideal and incompressible;
- 2) the fluid has constant unit density;
- 3) $\partial_z u = \partial_z v = 0$ (the fluid moves in columns).

Assumption 1 immediately implies the no-normal flow boundary condition $w|_{z=0} = 0$ and the divergence-free condition $\partial_x u + \partial_y v + \partial_z w = 0$. The same assumptions are used to derive the **shallow water equations (SW)** in the Hamiltonian setting. However, when deriving the SW we demand that if h_0 is a typical scale

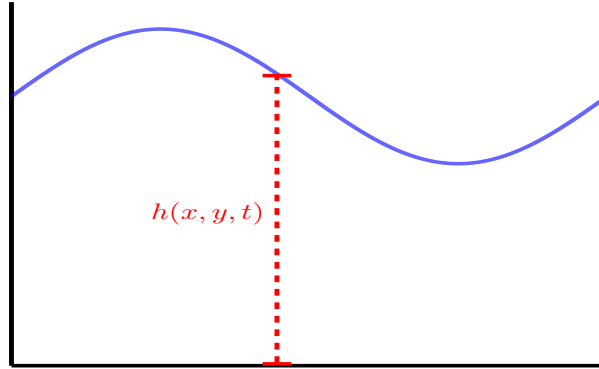


Figure 10.11: Definition of the depth h . The solid curve represents the free surface of the fluid, and the dashed line illustrates the physical meaning of h as the distance between the flat bottom ($z = 0$ here) and the free surface.

for depth and ℓ is a typical scale for horizontal length then the **aspect ratio** h_0/ℓ is small and only terms up to but not including $\mathcal{O}(h_0^2/\ell^2)$ are kept. Accordingly, the GN and SW differ only by a scale assumption.

Remark 10.1.1. *Some of the assumptions stated above can be neglected to derive more general versions of the GN. For instance, the assumption of constant unit density can easily be relaxed to investigate more diverse physical phenomena such as stratification. Additionally, some authors study the GN on a domain with nontrivial bottom topography [16]; we postpone the discussion of topography to Section 10.2 and defer most of the details to [16].*

To obtain the GN via Hamiltonian formalism, we must write down an expression for the energy density $\mathcal{H}_{\text{GN}}(x, y, z)$ of the fluid occupying Ω . Naturally, the energy density is given by

$$\mathcal{H}_{\text{GN}}(x, y, z) = \frac{1}{2}|\mathbf{u}|^2 + gz, \quad (10.1.1)$$

the sum of kinetic energy density and gravitational potential energy density. Integrating \mathcal{H}_{GN} over Ω , we obtain the total energy, or **Hamiltonian**, as

$$H_{\text{GN}} = \frac{1}{2} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \int_0^{h(x,y,t)} dz |\mathbf{u}|^2 + 2gz. \quad (10.1.2)$$

To simplify H_{GN} , we explicitly evaluate the integral with respect to z . Let $\nabla_{\text{H}} = \partial_x \hat{\mathbf{x}} + \partial_y \hat{\mathbf{y}}$ denote the horizontal gradient operator and $\mathbf{u}_{\text{H}} = (u, v)$ denote the horizontal part of the velocity vector. Then, integrating the incompressibility constraint and using assumptions 2 and 3,

$$w = -z \nabla_{\text{H}} \cdot \mathbf{u}_{\text{H}}. \quad (10.1.3)$$

Plugging (10.1.3) into (10.1.2) yields

$$\begin{aligned} H_{\text{GN}} &= \frac{1}{2} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \int_0^{h(x,y,t)} dz \left(u^2 + v^2 + z^2 [\nabla_{\text{H}} \cdot \mathbf{u}_{\text{H}}]^2 + 2gz \right) \\ &= \frac{1}{2} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \left(gh^2 + h|\mathbf{u}_{\text{H}}|^2 + \frac{1}{3}h^3 [\nabla_{\text{H}} \cdot \mathbf{u}_{\text{H}}]^2 \right). \end{aligned}$$

Observe that if U is a typical scale for \mathbf{u}_{H} , then $\nabla_{\text{H}} \cdot \mathbf{u}_{\text{H}} = \mathcal{O}(U/\ell)$. Thus

$$\frac{w^2}{|\mathbf{u}_{\text{H}}|^2} = \frac{(z \nabla_{\text{H}} \cdot \mathbf{u}_{\text{H}})^2}{|\mathbf{u}_{\text{H}}|^2} = \mathcal{O}(h_0^2/\ell^2). \quad (10.1.4)$$

We would therefore throw away the term with the $1/3$ in front if we wanted to derive the SW. However, we do not assume a small aspect ratio, meaning that these terms must now be retained. Dropping the subscripts “H” yields the final expression for the Green–Naghdi Hamiltonian:

$$H_{\text{GN}} = \frac{1}{2} \int dx dy gh^2 + h|\mathbf{u}|^2 + \frac{1}{3}h^3|\nabla \cdot \mathbf{u}|^2. \quad (10.1.5)$$

We have suppressed the limits of integration for notational ease.

Recall that the Lagrangian of a mechanical system is equal to

$$\text{Kinetic Energy} - \text{Potential Energy}.$$

In light of the derivations above, therefore, the Lagrangian for the GN is

$$\ell_{\text{GN}}(\mathbf{u}, h) = \frac{1}{2} \int dx dy h|\mathbf{u}|^2 + \frac{1}{3}h^3|\nabla \cdot \mathbf{u}|^2 - gh^2. \quad (10.1.6)$$

Solutions (\mathbf{u}, h) of the GN correspond to stationary points of the Lagrangian functional [58], but we do not make use of this variational paradigm in the sequel.

Remark 10.1.2. *We could have also assumed that our domain was given by*

$$\Omega = \{(x, y, z) \mid (x, y) \in \Omega' \subseteq \mathbb{R}^2 \text{ with } \Omega' \text{ compact, } 0 \leq z \leq h(x, y, t)\}$$

and obtained the same Hamiltonian (up to changing the domain of integration), provided we demand that the normal component of the horizontal velocity vanishes on $\partial\Omega'$. All of the derivations performed in this chapter can be repeated for bounded domains as long as this boundary condition is satisfied (periodic boundary conditions also work, too).

Having found a convenient expression for the governing Hamiltonian H_{GN} , we may turn to writing down the **non-canonical momentum** $\mathbf{m} = \frac{\delta \ell_{\text{GN}}}{\delta \mathbf{u}}$ and the **Poisson bracket** $\{\cdot, \cdot\}$ governing the equations of motion. We calculate the momentum \mathbf{m} explicitly first, and then turn to [26] to obtain an appropriate form of the Poisson bracket. The reader requiring a primer on variational calculus might want to skim [41, Chapter 2] or [84, pp. 52–55] before proceeding.

Remark 10.1.3. $\{\cdot, \cdot\}$ can be derived from scratch using the theory of Hamiltonian flows on semidirect products, but such a derivation is beyond the scope of this thesis. Interested readers are encouraged to see [44]. The semidirect product approach demonstrates that $\{\cdot, \cdot\}$ is actually a Lie–Poisson bracket arising from a “reduction” of the Lagrangian description of the flow (that is, studying the flow by analyzing the evolution of each little fluid parcel). Accordingly, we get that $\{\cdot, \cdot\}$ actually satisfies the definition of a Poisson bracket immediately, thus avoiding a long, tedious verification of the Jacobi identity [26, §IV].

Proposition 10.1.4. *The non-canonical momentum for the Green–Naghdi Hamiltonian is*

$$\mathbf{m} = h\mathbf{u} - \frac{1}{3}\nabla(h^3\nabla\cdot\mathbf{u}). \quad (10.1.7)$$

Proof. Pick an arbitrary vector field $\delta\mathbf{u} \in \mathfrak{X}_c(\mathbb{R}^2)$. To calculate variational derivatives, we use the L^2 pairing of vector fields,

$$\langle \mathbf{m}, \mathbf{u} \rangle = \int dx dy \mathbf{m} \cdot \mathbf{u} = \int dx dy m_i u_i.$$

Then, by definition of the variational derivative,

$$\begin{aligned} \left\langle \frac{\delta\ell_{\text{GN}}}{\delta\mathbf{u}}, \delta\mathbf{u} \right\rangle &= \frac{1}{2} \lim_{\epsilon \rightarrow 0} \frac{\partial}{\partial\epsilon} \int dx dy \left[-gh^2 + h|\mathbf{u} + \epsilon\delta\mathbf{u}|^2 + \frac{1}{3}h^3|\nabla\cdot\mathbf{u} + \epsilon\nabla\cdot\delta\mathbf{u}|^2 \right] \\ &= \frac{1}{2} \lim_{\epsilon \rightarrow 0} \frac{\partial}{\partial\epsilon} \int dx dy 2\epsilon \left[h\mathbf{u} \cdot \delta\mathbf{u} + \frac{1}{3}h^3(\nabla\cdot\mathbf{u})(\nabla\cdot\delta\mathbf{u}) \right] \\ &= \langle h\mathbf{u}, \delta\mathbf{u} \rangle + \int dx dy \frac{1}{3}h^3(\nabla\cdot\mathbf{u})(\nabla\cdot\delta\mathbf{u}). \end{aligned}$$

If we apply the Divergence Theorem and use the fact that \mathbf{u} is equal to 0 at infinity, we get

$$\left\langle \frac{\delta\ell_{\text{GN}}}{\delta\mathbf{u}}, \delta\mathbf{u} \right\rangle = \langle h\mathbf{u}, \delta\mathbf{u} \rangle - \int \nabla \left[\frac{1}{3}h^3\nabla\cdot\mathbf{u} \right] \cdot \delta\mathbf{u} dx dy.$$

Using the linearity of the L^2 pairing, the above calculations yield

$$\left\langle \frac{\delta\ell_{\text{GN}}}{\delta\mathbf{u}}, \delta\mathbf{u} \right\rangle = \left\langle h\mathbf{u} - \frac{1}{3}\nabla(h^3\nabla\cdot\mathbf{u}), \delta\mathbf{u} \right\rangle.$$

Since the variation $\delta\mathbf{u}$ was chosen arbitrarily,

$$\mathbf{m} = h\mathbf{u} - \frac{1}{3}\nabla(h^3\nabla\cdot\mathbf{u})$$

and the claim holds. \square

Before finding the equations of motion (that is, expressions for $\partial_t\mathbf{m}$ and $\partial_t h$), H_{GN} must be written in terms of \mathbf{m} and h ; \mathbf{u} is now considered to be a function of \mathbf{m} and h , not an independent field variable. Applying the Divergence Theorem and recalling that \mathbf{u} vanishes at infinity, we have

$$H_{\text{GN}} = \frac{1}{2} \int dx dy gh^2 + \mathbf{m} \cdot \mathbf{u}. \quad (10.1.8)$$

Proposition 10.1.5.

$$\frac{\delta H_{\text{GN}}}{\delta \mathbf{m}} = \mathbf{u} \text{ and} \quad (10.1.9a)$$

$$\frac{\delta H_{\text{GN}}}{\delta h} = gh - \frac{1}{2}|\mathbf{u}|^2 - \frac{1}{2}h^2|\nabla \cdot \mathbf{u}|^2. \quad (10.1.9b)$$

Proof. Observe that

$$H_{\text{GN}}(\mathbf{m}, h) = \langle \mathbf{m}, \mathbf{u} \rangle - \ell_{\text{GN}}(h, \mathbf{u}). \quad (10.1.10)$$

With (10.1.10) in mind, we proceed via prodigious application of the variational chain rule and the fact that $\frac{\delta h}{\delta \mathbf{m}} = 0$ (because \mathbf{m} and h are independent). The derivative with respect to momentum is given by

$$\begin{aligned} \frac{\delta H_{\text{GN}}}{\delta \mathbf{m}} &= \frac{\delta \langle \mathbf{m}, \mathbf{u} \rangle}{\delta \mathbf{m}} - \frac{\delta \ell_{\text{GN}}}{\delta \mathbf{m}} \\ &= \left\langle \frac{\delta \mathbf{m}}{\delta \mathbf{m}}, \mathbf{u} \right\rangle + \left\langle \mathbf{m}, \frac{\delta \mathbf{u}}{\delta \mathbf{m}} \right\rangle - \left\langle \frac{\delta \ell_{\text{GN}}}{\delta \mathbf{u}}, \frac{\delta \mathbf{u}}{\delta \mathbf{m}} \right\rangle - \left\langle \frac{\delta \ell_{\text{GN}}}{\delta h}, \frac{\delta h}{\delta \mathbf{m}} \right\rangle \\ &= \mathbf{u} + \left\langle \mathbf{m}, \frac{\delta \mathbf{u}}{\delta \mathbf{m}} \right\rangle - \left\langle \frac{\delta \ell_{\text{GN}}}{\delta \mathbf{u}}, \frac{\delta \mathbf{u}}{\delta \mathbf{m}} \right\rangle - \left\langle \frac{\delta \ell_{\text{GN}}}{\delta h}, 0 \right\rangle \\ &= \mathbf{u} + \left\langle \mathbf{m} - \frac{\delta \ell_{\text{GN}}}{\delta \mathbf{u}}, \frac{\delta \mathbf{u}}{\delta \mathbf{m}} \right\rangle \end{aligned}$$

Since $\mathbf{m} = \frac{\delta \ell_{\text{GN}}}{\delta \mathbf{u}}$, the above calculations imply

$$\frac{\delta H_{\text{GN}}}{\delta \mathbf{m}} = \mathbf{u},$$

hence (10.1.9a) holds.

Remembering that \mathbf{u} depends on h according to (10.1.7), we have

$$\begin{aligned} \frac{\delta H_{\text{GN}}}{\delta h} &= \frac{\delta \langle \mathbf{m}, \mathbf{u} \rangle}{\delta h} - \frac{\delta \ell_{\text{GN}}}{\delta h} - \left\langle \frac{\delta \ell_{\text{GN}}}{\delta \mathbf{u}}, \frac{\delta \mathbf{u}}{\delta h} \right\rangle \\ &= \left\langle \frac{\delta \mathbf{m}}{\delta h}, \mathbf{u} \right\rangle + \left\langle \mathbf{m}, \frac{\delta \mathbf{u}}{\delta h} \right\rangle - \frac{\delta \ell_{\text{GN}}}{\delta h} - \left\langle \frac{\delta \ell_{\text{GN}}}{\delta \mathbf{u}}, \frac{\delta \mathbf{u}}{\delta h} \right\rangle \\ &= \langle 0, \mathbf{u} \rangle + \left\langle \mathbf{m} - \frac{\delta \ell_{\text{GN}}}{\delta \mathbf{u}}, \frac{\delta \mathbf{u}}{\delta h} \right\rangle - \frac{\delta \ell_{\text{GN}}}{\delta h} \\ &= -\frac{\delta \ell_{\text{GN}}}{\delta h}. \end{aligned} \quad (10.1.11)$$

Using (10.1.6), $\frac{\delta \ell_{\text{GN}}}{\delta h}$ is trivially given by

$$\frac{\delta \ell_{\text{GN}}}{\delta h} = -gh + \frac{1}{2}|\mathbf{u}|^2 + \frac{1}{2}h^2|\nabla \cdot \mathbf{u}|^2. \quad (10.1.12)$$

Combining and (10.1.11) and (10.1.12), we obtain (10.1.9b). \square

Remark 10.1.6. *In the jargon of geometric mechanics, (10.1.10) tells us that H_{GN} is the **partial Legendre transform** of ℓ_{GN} [44, p. 27].*

Remark 10.1.7. *A shorter alternative proof involves just computing δH_{GN} and comparing the terms involving $\delta \mathbf{m}$ and δh , analogous to the usual derivation of Hamilton's canonical equations seen in an intermediate mechanics course. The interested reader may consult [43, p. 24] for further details.*

Now, we are ready to write down an expression for the Poisson bracket which generates the GN. Let F and G be functionals of \mathbf{m} and h . Their Poisson bracket is defined by [26]

$$\{F, G\} = \int dx dy \begin{bmatrix} \frac{\delta F}{\delta m_i} & \frac{\delta F}{\delta h} \end{bmatrix} \begin{bmatrix} -(m_j \partial_i + \partial_j m_i) & -h \partial_i \\ -\partial_j h & 0 \end{bmatrix} \begin{bmatrix} \frac{\delta G}{\delta m_j} \\ \frac{\delta G}{\delta h} \end{bmatrix}. \quad (10.1.13)$$

This is the same Poisson bracket used to derive the usual SW using Hamiltonian formalism (see for example [25]). Let's therefore refer to $\{\cdot, \cdot\}$ as the **shallow water bracket**.

Recall that the time rate of change of a functional F as it moves with the Hamiltonian flow generated by H_{GN} is

$$\partial_t F = \{F, H_{\text{GN}}\}.$$

Recall also that, if $\delta((x, y) - (x_0, y_0))$ denotes the Dirac function on $\mathbb{R}^2 \times \mathbb{R}^2$, we have that

$$\frac{\delta m_i(x, y)}{\delta m_j(x_0, y_0)} = \delta_{ij} \delta((x, y) - (x_0, y_0)) \quad \text{and} \quad \frac{\delta h(x, y)}{\delta h(x_0, y_0)} = \delta((x, y) - (x_0, y_0)).$$

Then, combining the above expressions with (10.1.13), (10.1.9a), and (10.1.9b) we obtain the equations of motion as

$$\partial_t h = -\nabla \cdot \left(h \frac{\delta H_{\text{GN}}}{\delta \mathbf{m}} \right) = -\nabla \cdot (h \mathbf{u}) \quad \text{and} \quad (10.1.14a)$$

$$\begin{aligned} \partial_t \mathbf{m} &= -m_j \nabla \left(\frac{\delta H_{\text{GN}}}{\delta m_j} \right) - \partial_j \left(\mathbf{m} \left(\frac{\delta H_{\text{GN}}}{\delta \mathbf{m}} \right)_j \right) - h \nabla \left(\frac{\delta H_{\text{GN}}}{\delta h} \right) \\ &= -m_j \nabla u_j - \partial_j (\mathbf{m} u_j) - h \nabla \left(gh - \frac{1}{2} |\mathbf{u}|^2 - \frac{1}{2} h^2 |\nabla \cdot \mathbf{u}|^2 \right). \end{aligned} \quad (10.1.14b)$$

(10.1.14a) is the **continuity equation** familiar from the theory of the SW (that is, only the evolution equation for momentum changes when we go from SW to GN). If we integrate both sides of (10.1.14a) over a domain $\Sigma \subseteq \mathbb{R}^2$ with unit outward normal \mathbf{n} and then apply the Divergence Theorem, we see that

$$\frac{d}{dt} \int_{\Sigma} h dx dy = - \int_{\partial \Sigma} h \mathbf{u} \cdot \mathbf{n} ds. \quad (10.1.15)$$

Physically, (10.1.15) says that the only changes to the total fluid mass in a given area Σ are due to the transport of fluid in and out of Σ by the velocity field. So, the continuity equation is nothing more than the infinitesimal expression of mass conservation. In Chapter 11, we discuss methods to ensure that the law of mass conservation remains satisfied in numerical simulations of the GN.

10.2 Adding the Effects of Rotation

We now modify the GN to describe a fluid moving through a spatial domain rotating about the z -axis with constant angular velocity. The resulting system of PDEs is referred to as the **rotating Green–Naghdi equations (RGN)**, generalizing the **rotating shallow water equations (RSW)** to (vertically averaged) flows with an arbitrary aspect ratio. To derive the RGN, we utilize methods previously applied in [26, §IV] and [79, §4.4]. First we make a change of field variables, then we appeal to physical intuition in order to derive an alternative Poisson bracket accommodating information about rotation. We call this modified bracket the **rotating shallow water bracket** $\{\cdot, \cdot\}_{\text{rot}}$.

As a “rabbit out of a hat,” we implement the change of variables $\mathbf{m} \mapsto \mathbf{v} = (1/h) \mathbf{m}$, allowing us to rewrite H_{GN} as

$$H_{\text{GN}} = \frac{1}{2} \int h \mathbf{v} \cdot \mathbf{u} + gh^2 \, dx \, dy. \quad (10.2.1)$$

The variable \mathbf{v} does not have a standard name in the fluid dynamics literature. We refer to it as the **pseudovelocity** because, in the small aspect ratio regime, $\mathbf{v} = \mathbf{u}$.

Proposition 10.2.1.

$$\frac{\delta H_{\text{GN}}}{\delta \mathbf{v}} = h \mathbf{u} \text{ and} \quad (10.2.2a)$$

$$\frac{\delta H_{\text{GN}}}{\delta h} = gh - \frac{1}{2} |\mathbf{u}|^2 - \frac{1}{2} h^2 (\nabla \cdot \mathbf{u})^2 + \mathbf{v} \cdot \mathbf{u}. \quad (10.2.2b)$$

Proof. Combine the variational chain rule with Proposition 10.1.5. \square

The next proposition provides motivation for the form of the rotating shallow water bracket.

Proposition 10.2.2. *Let F, G be functionals of $\mathbf{v} = (v_1, v_2)$ and h . Using the variables (\mathbf{v}, h) , the shallow water bracket $\{\cdot, \cdot\}$ is given by*

$$\{F, G\} = - \int \begin{bmatrix} \frac{\delta F}{\delta v_1} & \frac{\delta F}{\delta v_2} & \frac{\delta F}{\delta h} \end{bmatrix} \begin{bmatrix} 0 & -\frac{\hat{\mathbf{z}} \cdot (\nabla \times \mathbf{v})}{h} & \partial_x \\ \frac{\hat{\mathbf{z}} \cdot (\nabla \times \mathbf{v})}{h} & 0 & \partial_y \\ \partial_x & \partial_y & 0 \end{bmatrix} \begin{bmatrix} \frac{\delta G}{\delta v_1} \\ \frac{\delta G}{\delta v_2} \\ \frac{\delta G}{\delta h} \end{bmatrix} \, dx \, dy. \quad (10.2.3)$$

Proof. First, we re-write the shallow water bracket with a 3×3 operator matrix. Explicitly writing out the sums in (10.1.13) and letting the superscript T denote matrix transposition, we have

$$- \{F, G\} = \int \, dx \, dy \begin{bmatrix} \frac{\delta F}{\delta m_1} \\ \frac{\delta F}{\delta m_2} \\ \frac{\delta F}{\delta h} \end{bmatrix}^{\text{T}} \begin{bmatrix} m_1 \partial_x + \partial_x m_1 & m_2 \partial_x + \partial_y m_1 & h \partial_x \\ m_1 \partial_y + \partial_x m_2 & m_2 \partial_y + \partial_y m_2 & h \partial_y \\ \partial_x h & \partial_y h & 0 \end{bmatrix} \begin{bmatrix} \frac{\delta G}{\delta m_1} \\ \frac{\delta G}{\delta m_2} \\ \frac{\delta G}{\delta h} \end{bmatrix}. \quad (10.2.4)$$

The above form of the shallow water bracket is much more convenient for changing to the variables (\mathbf{v}, h) . For the rest of this proof, we use the convention that a subscript following a derivative in parentheses

indicates that the derivative is taken while holding constant the variable in the subscript (such a convention is common in classical thermodynamics). Using the variational chain rule [41, p. 37],

$$\left(\frac{\delta F}{\delta h}\right)_{\mathbf{v}} = \left(\frac{\delta F}{\delta h}\right)_{\mathbf{m}} + \mathbf{v} \cdot \frac{\delta F}{\delta \mathbf{m}} \quad \text{and} \quad \frac{\delta F}{\delta \mathbf{v}} = h \frac{\delta F}{\delta \mathbf{m}}. \quad (10.2.5)$$

In terms of vectors and matrices, (10.2.5) reads

$$\begin{bmatrix} \frac{\delta F}{\delta v_1} \\ \frac{\delta F}{\delta v_2} \\ \left(\frac{\delta F}{\delta h}\right)_{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} h & 0 & 0 \\ 0 & h & 0 \\ v_1 & v_2 & 1 \end{bmatrix} \begin{bmatrix} \frac{\delta F}{\delta m_1} \\ \frac{\delta F}{\delta m_2} \\ \left(\frac{\delta F}{\delta h}\right)_{\mathbf{m}} \end{bmatrix}.$$

Therefore,

$$\begin{bmatrix} \frac{\delta F}{\delta m_1} \\ \frac{\delta F}{\delta m_2} \\ \left(\frac{\delta F}{\delta h}\right)_{\mathbf{m}} \end{bmatrix} = h^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -v_1 & -v_2 & h \end{bmatrix} \begin{bmatrix} \frac{\delta F}{\delta v_1} \\ \frac{\delta F}{\delta v_2} \\ \left(\frac{\delta F}{\delta h}\right)_{\mathbf{v}} \end{bmatrix}. \quad (10.2.6)$$

Plugging (10.2.6) into (10.2.4) (while harmlessly neglecting the $dx \, dy$ in the integral), we get

$$-\{F, G\} = \int \begin{bmatrix} \frac{\delta F}{\delta v_1} \\ \frac{\delta F}{\delta v_2} \\ \left(\frac{\delta F}{\delta h}\right)_{\mathbf{v}} \end{bmatrix}^T \begin{bmatrix} h^{-1} & 0 & -h^{-1}v_1 \\ 0 & h^{-1} & -h^{-1}v_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} (m_1\partial_x + \partial_x m_1) & (m_2\partial_x + \partial_y m_1) & h\partial_x \\ (m_1\partial_y + \partial_x m_2) & (m_2\partial_y + \partial_y m_2) & h\partial_y \\ \partial_x h & \partial_y h & 0 \end{bmatrix} \begin{bmatrix} h^{-1} & 0 & 0 \\ 0 & h^{-1} & 0 \\ -h^{-1}v_1 & -h^{-1}v_2 & 1 \end{bmatrix} \begin{bmatrix} \frac{\delta G}{\delta v_1} \\ \frac{\delta G}{\delta v_2} \\ \left(\frac{\delta G}{\delta h}\right)_{\mathbf{v}} \end{bmatrix}.$$

Performing the matrix multiplications, we see that

$$-\{F, G\} = \int \begin{bmatrix} \frac{\delta F}{\delta v_1} \\ \frac{\delta F}{\delta v_2} \\ \left(\frac{\delta F}{\delta h}\right)_{\mathbf{v}} \end{bmatrix}^T \begin{bmatrix} (v_1\partial_x + h^{-1}\partial_x v_1 - \partial_x h^{-1}v_1 - h^{-1}v_1\partial_x) & (v_2\partial_x + h^{-1}\partial_y v_1 - \partial_x h^{-1}v_2 - h^{-1}v_1\partial_y) & \partial_x \\ (v_1\partial_y + h^{-1}\partial_x v_2 - \partial_y h^{-1}v_1 - h^{-1}v_2\partial_x) & (v_2\partial_y + h^{-1}\partial_y v_2 - \partial_y h^{-1}v_2 - h^{-1}v_2\partial_y) & \partial_y \\ \partial_x & \partial_y & 0 \end{bmatrix} \begin{bmatrix} \frac{\delta G}{\delta v_1} \\ \frac{\delta G}{\delta v_2} \\ \left(\frac{\delta G}{\delta h}\right)_{\mathbf{v}} \end{bmatrix}.$$

Simplifying the entries in the upper 2×2 block of the operator matrix completes the proof. \square

The quantity $\hat{\mathbf{z}} \cdot (\nabla \times \mathbf{v})$, related to the vertical component of the fluid's vorticity (when the aspect ratio is small, this is precisely the vertical component of vorticity), appears very clearly in the shallow water bracket after changing variables. Reasonably, we should be able go from $\{\cdot, \cdot\}$ to $\{\cdot, \cdot\}_{\text{rot}}$ by simply including the rotating frame's contribution to "vorticity" in that part of $\{\cdot, \cdot\}$ depending on "vorticity", namely the upper 2×2 block of the operator matrix in (10.2.3). To this end, we introduce the **potential vorticity**, q : if f denotes the **Coriolis parameter** (a constant equal to twice the angular frequency of the domain's rotation), then

$$q(x, y, t) \doteq \frac{f + \hat{\mathbf{z}} \cdot (\nabla \times \mathbf{v})}{h}. \quad (10.2.7)$$

In keeping with the loose discussion above we see that q takes into account both the effects of rotation and the curl of the velocity field, hence it contains information on the mean total vorticity of the fluid. So, intuitively, we see that $\{\cdot, \cdot\}_{\text{rot}}$ can be obtained by just replacing the quantity $\frac{\hat{\mathbf{z}} \cdot (\nabla \times \mathbf{v})}{h}$ appearing in (10.2.3) with q . Thus, we have argued that

$$\{F, G\}_{\text{rot}} = - \int \begin{bmatrix} \frac{\delta F}{\delta v_1} & \frac{\delta F}{\delta v_2} & \frac{\delta F}{\delta h} \end{bmatrix} \begin{bmatrix} 0 & -q & \partial_x \\ q & 0 & \partial_y \\ \partial_x & \partial_y & 0 \end{bmatrix} \begin{bmatrix} \frac{\delta G}{\delta v_1} \\ \frac{\delta G}{\delta v_2} \\ \frac{\delta G}{\delta h} \end{bmatrix} dx dy. \quad (10.2.8)$$

In the case of a non-rotating frame $f = 0$ and so $\{\cdot, \cdot\}_{\text{rot}}$ reduces to $\{\cdot, \cdot\}$, helping to cement the validity of our intuitive argument.

Using the definition of the Hamiltonian flow generated by $H_{\text{GN}}(\mathbf{v}, h)$ and $\{\cdot, \cdot\}_{\text{rot}}$, the RGN are found to be [26]

$$\zeta = \hat{\mathbf{z}} \cdot (\nabla \times \mathbf{v}), \quad (10.2.9a)$$

$$\mathbf{v} = \mathbf{u} - \frac{1}{3h} \nabla (h^3 \nabla \cdot \mathbf{u}), \quad (10.2.9b)$$

$$\partial_t \mathbf{v} + (f + \zeta) \hat{\mathbf{z}} \times \mathbf{u} + \nabla \left(gh - \frac{1}{2} |\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2} h^2 (\nabla \cdot \mathbf{u})^2 \right) = 0, \text{ and} \quad (10.2.9c)$$

$$\partial_t h + \nabla \cdot (h\mathbf{u}) = 0. \quad (10.2.9d)$$

Observe how we have defined the **pseudovorticity** ζ as an auxiliary variable in an extended system of governing equations. This choice of notation serves to help simplify the finite element discretization of (10.2.9) in Chapter 3.

Remark 10.2.3. *The identity*

$$\partial_t q + \mathbf{u} \cdot \nabla q = 0 \quad (10.2.10)$$

holds (this is justified by symmetry principles in [58]); physically, this equation tells us that we see no change in potential vorticity while moving with a fluid that evolves according to (10.2.9). See [76, pp. 61–63] for a discussion of the physical importance of q in the RSW case.

Note the similarities between (10.2.9c) and the **vector-invariant form** of the RSW, which is precisely the Hamiltonian form of the RSW in (\mathbf{v}, h) variables (recall that in the SW regime $\mathbf{v} = \mathbf{u}$). For future reference, we write down this form of the RSW below:

$$\zeta = \hat{\mathbf{z}} \cdot (\nabla \times \mathbf{u}), \quad (10.2.11a)$$

$$\partial_t \mathbf{u} + (f + \zeta) (\hat{\mathbf{z}} \times \mathbf{u}) + \nabla \left(gh + \frac{1}{2} |\mathbf{u}|^2 \right) = 0, \text{ and} \quad (10.2.11b)$$

$$\partial_t h + \nabla \cdot (h\mathbf{u}) = 0. \quad (10.2.11c)$$

So, there are two main differences between the RGN and the RSW: the presence of the pseudovelocity variable \mathbf{v} , and the specific form of the **Bernoulli function** (the scalar function inside the gradient in (10.2.9c) and (10.2.11b)).

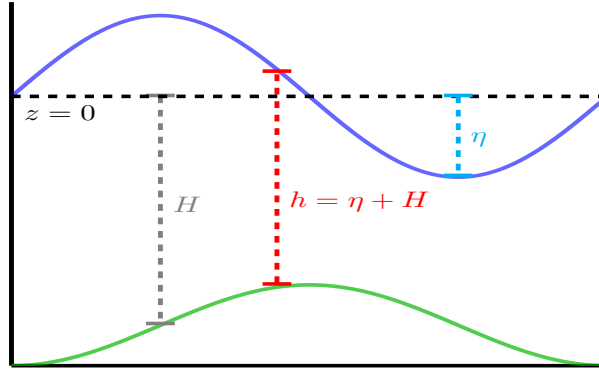


Figure 10.21: Definition of h , H , and η . The green curve represents the seafloor and the blue curve represents the free surface of the fluid.

10.2.1 Topographic Terms

Occasionally, we would also like to add nontrivial bottom topography to RGN. That is, we would like to see how (10.2.9) is altered when our fluid is moving over an uneven seafloor with shape described by $z = -H(x, y)$ (see Figure 10.21). Naturally, the new Hamiltonian is simply

$$H_{\text{GN}} = \frac{1}{2} \int h \mathbf{v} \cdot \mathbf{u} + g(h - H)^2 \, dx \, dy. \quad (10.2.12)$$

Then, we defer to [16, Equations 3.12, 4.16, 4.18] to see that the RGN with topography may be written as

$$\zeta = \hat{\mathbf{z}} \cdot (\nabla \times \mathbf{v}), \quad (10.2.13a)$$

$$\mathbf{v} = \mathbf{u} - \frac{1}{3h} \nabla (h^3 \nabla \cdot \mathbf{u}) - \frac{1}{2h} \nabla (h^2 \mathbf{u} \cdot \nabla H) + \frac{1}{2} (h \nabla \cdot \mathbf{u}) \nabla H + (\mathbf{u} \cdot \nabla H) \nabla H, \quad (10.2.13b)$$

$$\partial_t \mathbf{v} + (f + \zeta) \hat{\mathbf{z}} \times \mathbf{u} + \nabla \left(g(h - H) - \frac{1}{2} |\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2} (h \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) = 0, \text{ and} \quad (10.2.13c)$$

$$\partial_t h + \nabla \cdot (h \mathbf{u}) = 0. \quad (10.2.13d)$$

So, adding bottom topography makes the pseudovelocity much more complicated, while the other equations are mostly unchanged. Sometimes (for instance, in Section 12.4), we use the letter η to denote the difference between depth h and seafloor height H (see Figure 10.21). η is called the **free surface deformation**.

10.2.2 Non-dimensionalization

Before going any further, we take a moment to non-dimensionalize the RGN (including topographic terms) and examine what dimensionless quantities are most important in determining the behaviour of

these equations; understanding these dimensionless quantities is important for appreciating some of the numerical tests featured in future chapters. Let's scale our variables as

$$\begin{aligned} h^* &= h_0 h, \\ t^* &= T t, \\ \mathbf{u}^* &= U \mathbf{u}, \\ \mathbf{v}^* &= U \mathbf{v}, \text{ and} \\ (x^*, y^*) &= L(x, y). \end{aligned}$$

Then, defining dimensionless parameters

$$\begin{aligned} \text{Ro} &= U/fL, \\ \text{Fr} &= U/\sqrt{gh_0}, \quad \text{and} \\ \gamma &= (h_0/L)^2, \end{aligned}$$

(10.2.13) may be written as

$$\zeta = \hat{\mathbf{z}} \cdot (\nabla \times \mathbf{v}), \quad (10.2.14a)$$

$$\mathbf{v} = \mathbf{u} - \gamma \left[\frac{1}{3h} \nabla (h^3 \nabla \cdot \mathbf{u}) - \frac{1}{2h} \nabla (h^2 \mathbf{u} \cdot \nabla H) + \frac{1}{2} (h \nabla \cdot \mathbf{u}) \nabla H + (\mathbf{u} \cdot \nabla H) \nabla H \right], \quad (10.2.14b)$$

$$\partial_t \mathbf{v} + fT (1 + \text{Ro} \zeta) \hat{\mathbf{z}} \times \mathbf{u} + \nabla \left(\frac{1}{\text{Fr}^2} (h - H) + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2} |\mathbf{u}|^2 - \frac{\gamma}{2} (h \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) = 0, \text{ and} \quad (10.2.14c)$$

$$\partial_t h + \nabla \cdot (h \mathbf{u}) = 0. \quad (10.2.14d)$$

The number Ro is known as the **Rossby number**. Large-scale oceanic flows, where the effects of rotation are very important, are characterized by the requirement that Ro is “order one or less” [67, p. 2]. When simulating the RGN on the computer, we often choose our time scale to be $T = 1/f$ so that $\text{Ro} = 1$. The parameter Fr , called the **Froude number**, is a familiar sight in elementary fluid mechanics, representing how fast a disturbance in our fluid moves relative to a typical small-amplitude wave. The parameter γ is simply the aspect ratio squared, hence we might expect that the dimensionless RSW can be recovered in the limit $\gamma \rightarrow 0$. Examining (10.2.14), we see that this is indeed the case. Accordingly, to completely isolate the differences between the behaviour of RSW and RGN, it is useful to keep the Froude and Rossby numbers constant (usually equal to 1) and just vary γ .

Note that, if we consider the non-rotating GN, we set $f = 0$. Therefore, the above scaling in terms of the Rossby number is not valid. The appropriate non-dimensional version of the GN momentum equation in this case is

$$\partial_t \mathbf{v} + \zeta \hat{\mathbf{z}} \times \mathbf{u} + \nabla \left(\frac{1}{\text{Fr}^2} (h - H) + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2} |\mathbf{u}|^2 - \frac{\gamma}{2} (h \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) = 0.$$

Additionally, if we add viscosity to the RGN then the dynamics is also influenced by yet another dimensionless parameter, the famous **Reynolds number**. Recall that, if ν denotes the kinematic viscosity of our fluid, then the Reynolds number is defined by

$$\text{Re} \doteq UL/\nu.$$

We see that the larger the Reynolds number is, the less important viscosity becomes, so the limit $\text{Re} \rightarrow \infty$ corresponds to an ideal fluid. The non-dimensionalized momentum equation with $f = 0$, modified to include viscous dissipation, reads

$$\partial_t \mathbf{v} + \zeta \hat{\mathbf{z}} \times \mathbf{u} + \nabla \left(\frac{1}{\text{Fr}^2} (h - H) + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2} |\mathbf{u}|^2 - \frac{\gamma}{2} (h \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) + \frac{1}{\text{Re}} \Delta \mathbf{v} = 0.$$

10.3 Linearized RGN

The goal of this section is to describe some simple solutions to the linearized RGN, namely **geostrophic balance** and **linear gravity waves**. For the remainder of this section, we eschew the pseudovelocity variable \mathbf{v} in favour of the usual Eulerian velocity \mathbf{u} , in order to make the physical content of the discussion more transparent. Please note, however, that in Chapters 11 and 12 we use the variables (\mathbf{v}, h) exclusively, so the earlier developments of this chapter were not in vain.

Let $\frac{D}{Dt} = \partial_t + \mathbf{u} \cdot \nabla$ denote the **material derivative**. In terms of the variables (\mathbf{u}, h) , the RGN become [26, §III]

$$\partial_t h + \nabla \cdot (h \mathbf{u}) = 0 \text{ and} \tag{10.3.1a}$$

$$\frac{D\mathbf{u}}{Dt} + f \hat{\mathbf{z}} \times \mathbf{u} + g \nabla h = -\frac{1}{3h} \nabla \left(h^2 \frac{D^2 h}{Dt^2} \right). \tag{10.3.1b}$$

Recall that (10.3.1a) is called the continuity equation. We call (10.3.1b) the **momentum equation** because it is the continuum-mechanical analogue of Newton's Second Law for the system.

Now, we are ready to linearize the RGN about the rest state $(\mathbf{u}, h) = ((u, v), h) = ((0, 0), h_0)$, for a constant h_0 . Consider small perturbations of this state in the form (u', v', h') . If terms of second order or higher in the primed variables are neglected, then the RGN reduce to the following simplified linear system after dropping the primes:

$$\partial_t h + h_0 (\partial_x u + \partial_y v) = 0, \tag{10.3.2a}$$

$$\partial_t u - f v + g \partial_x h = -\frac{h_0}{3} \partial_t^2 \partial_x h, \text{ and} \tag{10.3.2b}$$

$$\partial_t v + f u + g \partial_y h = -\frac{h_0}{3} \partial_t^2 \partial_y h. \tag{10.3.2c}$$

10.3.1 Geostrophic Balance

Our first task is to find steady-state solutions to (10.3.2). To do so, we ignore time dependence in the field variable (this amounts to setting $\partial_t = 0$) and isolate u and v . Doing so, we obtain

$$(u, v) = \frac{g}{f} (-\partial_y h, \partial_x h).$$

We may use the operator $\nabla^\perp = -\partial_y \hat{\mathbf{x}} + \partial_x \hat{\mathbf{y}}$ to write the above condition as

$$\mathbf{u} = \frac{g}{f} \nabla^\perp h. \tag{10.3.3}$$

For any smooth depth field $h(x, y)$, then, we may use (10.3.3) to obtain a valid steady solution to the RGN. Such solutions are said to be in **geostrophic balance**, representing states where gravity and the Coriolis effect tend to cancel each other out. Geostrophically balanced solutions are familiar from regular RSW theory [76, Chapter 2].

Remark 10.3.1. *The geostrophic balance condition (10.3.3) just says that \mathbf{u} has a streamfunction, and that this streamfunction is proportional to the depth h .*

Many large-scale oceanic processes can be understood by separating the flow into two parts:

- 1) a geostrophic component, corresponding to prominent quasi-static features such as large vortices, and
- 2) a “fast” component, corresponding to (usually small) rapidly propagating waves perturbing the balance.

Geostrophically balanced velocity fields are divergence-free and, as we shall see in the next subsection, most fast waves have nonzero divergence. Accordingly, the two components of a large-scale oceanic flow can mathematically be distinguished by way of a Helmholtz decomposition of the velocity field \mathbf{u} . In light of Part 1, the importance of Helmholtz decompositions in the theory of large-scale flows immediately suggests that FEEC provides a useful setting for the discretization of the RGN.

10.3.2 Linear Gravity Waves

We now discuss normal mode solutions of (10.3.2). Physically, such normal modes represent **gravity waves**; that is, gravity provides the restoring force leading to oscillations. Let U, V , and H be complex constants with small moduli, let (k, ℓ) be the wave vector, and let ω be the angular frequency. Then, we can form the normal mode ansatz

$$(u, v, h) = \text{Re} [(U, V, H) \exp(i(kx + \ell y - \omega t))]. \quad (10.3.4)$$

By plugging (10.3.4) into (10.3.2) we can find the **dispersion relation**, an implicit expression for ω as a function of k and ℓ . In order to obtain the dispersion relation, however, we must first “compress” the system (10.3.2) to get a single PDE involving only h . First, differentiate both sides of (10.3.2b) with respect to x and both sides of (10.3.2c) with respect to y . Then, adding the resulting equations and letting $\Delta = -\partial_x^2 - \partial_y^2$ denote the Laplacian operator, we have

$$\partial_t(\partial_x u + \partial_y v) - g\Delta h + f(\partial_y u - \partial_x v) = -\frac{h_0}{3}\partial_{tt}\Delta h. \quad (10.3.5)$$

We can simplify (10.3.5) using the linearized continuity equation (10.3.2a) and doing a bit of re-arranging, yielding

$$-gh_0\Delta h - \partial_{tt}\left(\frac{h_0^2}{3}\Delta h + h\right) = fh_0(\partial_x v - \partial_y u). \quad (10.3.6)$$

To completely isolate h in (10.3.6) we take the curl of the momentum equations (10.3.2b) and (10.3.2c), as is customary in geophysics. That is, we compute $\partial_x(10.3.2c) - \partial_y(10.3.2b)$ to obtain an expression for $\partial_x v - \partial_y u$. Doing so, we have

$$\partial_t(\partial_x v - \partial_y u) = \partial_t\left(\frac{f}{h_0}h\right). \quad (10.3.7)$$

Thus

$$\partial_x v - \partial_y u = \frac{f}{h_0} h. \quad (10.3.8)$$

Combining (10.3.6) with (10.3.8), we obtain the desired decoupled wave equation summarizing (10.3.2):

$$f^2 h + g h_0 \Delta h = -\partial_{tt} \left(h + \frac{h_0^2}{3} \Delta h \right). \quad (10.3.9)$$

(10.3.9) is, as readers who have studied the RSW might expect, the Klein–Gordon equation modulo some extra dispersive terms on the order of aspect ratio squared.

Plugging the wave ansatz (10.3.4) into (10.3.9), we obtain the dispersion relation

$$\omega^2 = \frac{f^2 + g h_0 (k^2 + \ell^2)}{1 + \frac{h_0^2}{3} (k^2 + \ell^2)}. \quad (10.3.10)$$

Non-dimensionalizing, the dispersion relation becomes

$$\omega^2 = \frac{1 + \frac{g h_0}{L^2 f^2} (k^2 + \ell^2)}{1 + \frac{\gamma}{3} (k^2 + \ell^2)}, \quad (10.3.11)$$

where ω, k , and ℓ denote the dimensionless versions of the corresponding variables. When $\gamma \rightarrow 0$, we recover the dispersion relation of the usual RSW [76, p. 70]. Hence, the linearized RGN can be viewed as a slight tweak of the linear RSW by a modification of the dispersion relation, though this is not a great surprise. Also, as γ becomes larger (that is, as we leave the regime of validity for the shallow-water model), the dispersion relation indicates that gravity waves have a smaller **phase velocity**

$$\mathbf{c} \doteq \frac{\omega}{\sqrt{k^2 + \ell^2}} (k, \ell).$$

To completely solve for the linear gravity wave solutions, we must find polarization relations connecting the amplitude parameters U, V, H which appear in (10.3.4). To begin, assume that

$$U = H\tilde{U} \text{ and } V = H\tilde{V}.$$

Let us also take $H \in \mathbb{R}$ for simplicity. Then, with the above assumptions in mind, we plug (10.3.4) into the momentum equations (10.3.2b), (10.3.2c) to obtain a system of linear algebraic equations for the two unknowns \tilde{U}, \tilde{V} , given by

$$-\omega\tilde{U} + i f \tilde{V} = -gk + \frac{h_0}{3} k \omega^2, \quad (10.3.12a)$$

$$-\omega\tilde{V} - i f \tilde{U} = -g\ell + \frac{h_0}{3} \ell \omega^2. \quad (10.3.12b)$$

Solving (10.3.12), we have

$$\tilde{U} = \frac{g - \frac{h_0}{3} \omega^2}{\omega^2 - f^2} (\omega k + i f \ell), \text{ and} \quad (10.3.13a)$$

$$\tilde{V} = \frac{g - \frac{h_0}{3} \omega^2}{\omega^2 - f^2} (\omega \ell - i f k). \quad (10.3.13b)$$

Using the above expressions and taking the real parts of (10.3.4), we may finally write down the linear gravity wave solutions to (10.3.2) in terms of a single small amplitude parameter H and the phase $\theta = kx + \ell y - \omega t$ as follows:

$$h(x, y, t) = h_0 + H \cos \theta, \quad (10.3.14a)$$

$$u(x, y, t) = \frac{H \left(g - \frac{h_0}{3} \omega^2 \right)}{\omega^2 - f^2} (\omega \ell \cos \theta + f k \sin \theta), \quad (10.3.14b)$$

$$v(x, y, t) = \frac{H \left(g - \frac{h_0}{3} \omega^2 \right)}{\omega^2 - f^2} (\omega k \cos \theta - f \ell \sin \theta). \quad (10.3.14c)$$

We may easily verify that the divergence of these linear gravity waves is nonzero in general, as promised in the previous subsection (recall that small disturbances to balanced states physically correspond to the fast component of a geophysical flow).

To review, we have used this section to study elementary solutions to the linearized RGN. Specifically, we introduced geostrophically balanced states and linear gravity waves. When investigating numerical models of oceanic flows in future chapters, we sometimes separate the flow into two parts: one component corresponding to the “balanced part” of the flow, and the other component corresponding to the “wave-like part” of the flow. Because balanced states have zero divergence, such separation is nothing more than a Helmholtz decomposition. Since we developed extensive theory on discrete Hodge–Helmholtz decompositions in Part 1, we can be confident that our simulations of the RGN will mostly preserve the separation of a flow into fast and slow parts (that is, a numerical approximation of a balanced state still satisfies a necessary condition for balance), provided that our numerical method is developed with FEEC in mind.

Chapter 11

Two Finite Element Methods for the (R)GN

The main goal of this chapter is to describe two mixed finite element methods for the (R)GN, both of which involve a choice of finite element spaces informed by FEEC. We first formulate the (R)GN **semi-discretely**, leaving time as a continuous variable and discretizing the spatial derivatives using the Galerkin approach described in Chapter 6. In the semi-discrete setting, we suppose that all variables are continuously differentiable with respect to time, even though the spatial derivatives may only exist in the weak sense (sometimes, the spatial derivatives may not even exist at all). Time discretization, accomplished via finite differences, is not mentioned until the penultimate section of this chapter. Also, for the sake of generality, we almost always write down discretizations of the full RGN with topography (10.2.13), even though we do not include rotation and topography in every simulation. Note that we only consider either periodic boundary conditions or no-normal flow conditions (the normal components of \mathbf{u} and \mathbf{v} vanish on the boundary of the domain).

The first FEM I introduce, referred to as the **upwind method**, employs a well-known strategy from numerical PDE theory to handle difficulties with defining the mass flux $h\mathbf{u}$ on interelement boundaries. These difficulties are a consequence of placing the approximate depth h in a finite element space of discontinuous functions. The second method, which we call the $H(\text{div})$ -**flux method**, completely bypasses the aforementioned issues by projecting $h\mathbf{u}$ onto a div-conforming finite element space. The $H(\text{div})$ -flux method also conserves vorticity and energy at the semi-discrete level.

In Section 11.1, I briefly outline why FEEC provides an especially useful context for designing FEMs in geophysical fluid dynamics. In Section 11.2, I introduce FEEC-inspired choices of finite element spaces suitable for use in mixed methods for the (R)GN. We then use these spaces to design an FEM for the linear RGN that ensures balanced states remain steady. In Section 11.3 we derive the upwind method for the (R)GN. In Section 11.4, I introduce the $H(\text{div})$ -flux method and prove that it conserves mass, vorticity, and energy. In Section 11.5, I state the finite difference method used for time discretization. In Section 11.6, I give a brief review of some recent papers on numerical methods for the (R)GN to better contextualize our work and to direct curious readers to alternative approaches.

11.1 Why is FEEC Good for Geophysical Fluid Dynamics?

After reading over 140 pages of theory on FEEC, the reader is certainly justified in asking for an answer to the above question. In this section, we build on our discussion of GN in Chapter 10 to demonstrate that FEEC provides a good framework for discretizing models in geophysical fluid dynamics. For more thorough discussions in this vein, I recommend [21, 22].

First, we briefly review some results from Part 1. We have seen that one of the main “punchlines” of FEEC is the construction of commuting diagrams like the one below:

$$\begin{array}{ccccccc}
 \dots & \xrightarrow{d} & H\Lambda^{k-1} & \xrightarrow{d} & H\Lambda^k & \xrightarrow{d} & H\Lambda^{k+1} \xrightarrow{d} \dots \\
 & & \pi_h^{k-1} \downarrow & & \pi_h^k \downarrow & & \downarrow \pi_h^{k+1} \\
 \dots & \xrightarrow{d} & \Lambda_h^{k-1} & \xrightarrow{d} & \Lambda_h^k & \xrightarrow{d} & \Lambda_h^{k+1} \xrightarrow{d} \dots
 \end{array} \tag{11.1.1}$$

Recall that the upper level of the diagram represents the function spaces where the PDE of interest is (weakly) posed and the lower level represents the finite element spaces in which we look for approximate solutions to the PDE. Using the Triple–Decker Theorem, we know that the cohomology of the upper complex is isomorphic to the cohomology of the lower complex provided we choose finite element spaces carefully. After the discussion at the end of Chapter 9, finding good choices of finite element spaces is easy if we use the Periodic Table (choose spaces on horizontal lines in the first column, and on diagonal lines in the second column). So, one of the most important algebraic properties of a PDE can be easily preserved in a finite element discretization.

Now, what is the significance of diagrams like (11.1.1) in the context of geophysical fluid dynamics? The answer becomes more transparent if we write down a special case of the commuting diagram and use different notation. Suppose $\Omega \subseteq \mathbb{R}^2$ is a polyhedral domain. In this case, $\star H\Lambda^1(\Omega)$ can be identified with $H(\text{div}; \Omega)$, the space of all square–integrable rough vector fields on Ω with weak divergence in $L^2(\Omega)$. Similarly, Hodge duality allows us to identify $L^2(\Omega) \simeq L^2\Lambda^2(\Omega)$. Then, the de Rham complex of Ω can be written as

$$H\Lambda^0(\Omega) \xrightarrow{\nabla^\perp} H(\text{div}; \Omega) \xrightarrow{\nabla \cdot} L^2(\Omega). \tag{11.1.2}$$

So, if we switch to denoting finite element spaces by V^k , we may write (11.1.1) in this simple case as

$$\begin{array}{ccccc}
 H\Lambda^0(\Omega) & \xrightarrow{\nabla^\perp} & H(\text{div}; \Omega) & \xrightarrow{\nabla \cdot} & L^2(\Omega) \\
 \pi_h^0 \downarrow & & \pi_h^1 \downarrow & & \pi_h^2 \downarrow \\
 V^0 & \xrightarrow{\nabla^\perp} & V^1 & \xrightarrow{\nabla \cdot} & V^2
 \end{array} \tag{11.1.3}$$

So, if all of the requirements outlined in Chapter 9 are satisfied, then any vector field $\mathbf{u} \in H(\text{div}; \Omega)$ with zero divergence is interpolated onto a vector field $\pi_h^1 \mathbf{u} \in V^1$ with zero divergence. Now, I stated in Chapter 10 that many large–scale, quasi–static features in geophysical flows can be described as geostrophically balanced solutions to the linearized RGN (or RSW). The velocity field of a balanced state is characterized by being in the range of ∇^\perp . With the commuting diagram (11.1.3) in mind, then, FEEC tells us that the interpolation of a balanced state onto V^1 is in the range of $\nabla^\perp|_{V^0}$. So, just using the commutativity of (11.1.3), we know that an approximation of a balanced state itself satisfies a necessary condition for balance. Actually, in the next section we prove (following [21]) that numerical approximations of balanced states are exactly steady in the linear limit. Therefore, the commuting diagram property of (11.1.3)

guarantees that one of the most important components of a geophysical flow is preserved in special finite element discretizations.

The utility of FEEC in fluid dynamics is not limited to preserving balanced states. Other models in fluid mechanics have important connections to cohomology, usually by way of an elliptic equation appearing in the system of governing PDEs. For example, the Stokes equations (governing the equilibrium behaviour of fluids at low Reynolds numbers) include a vector Poisson equation. In Chapter 9, we discussed at length how important discrete cohomology preservation is when dealing with finite element solutions to vector Poisson equations: if we choose finite element spaces poorly, we can end up with a simulation that appears to be solving the vector Laplace equation instead. So, FEEC may be very helpful in designing and analyzing computational models of fluid flow in many different applications.

11.2 General Framework for Mixed Formulations of (R)GN

In this section we outline a general FEEC–inspired methodology for constructing mixed FEMs for the (R)GN. In particular, I state the variables that should be common to each mixed method, and the finite element spaces in which these variables should live. I then use this general framework to construct a mixed formulation of the linear RGN and prove that balanced flows remain exactly steady in the semi–discrete setting.

Let $\Omega \subseteq \mathbb{R}^2$ be a polyhedral domain. After the developments of the previous section, we must choose three finite element spaces associated to Ω in order to build a discrete de Rham complex. We need two spaces of scalar functions, V^0 and V^2 , and a space of vector fields, V^1 . In practice, we work with either

$$V^0 = \text{CG}(r), V^1 = \text{RT}(r), \text{ and } V^2 = \text{DG}(r - 1)$$

or

$$V^0 = \text{CG}(r + 1), V^1 = \text{BDM}(r), \text{ and } V^2 = \text{DG}(r - 1).$$

The above choices correspond to the trimmed and regular finite element de Rham complexes on Ω , respectively. That is,

$$V^0 = \mathcal{P}_r^- \Lambda^0(\mathcal{T}_h), V^1 = \star \mathcal{P}_r^- \Lambda^1(\mathcal{T}_h), \text{ and } V^2 = \star \mathcal{P}_r^- \Lambda^2(\mathcal{T}_h) = \star \mathcal{P}_{r-1} \Lambda^2(\mathcal{T}_h)$$

in the first case and

$$V^0 = \mathcal{P}_{r+1} \Lambda^0(\mathcal{T}_h), V^1 = \star \mathcal{P}_r \Lambda^1(\mathcal{T}_h), \text{ and } V^2 = \star \mathcal{P}_{r-1} \Lambda^2(\mathcal{T}_h)$$

in the second case, up to proxies.

We therefore have that the diagram

$$V^0 \xrightarrow{\nabla^\perp} V^1 \xrightarrow{\nabla^\cdot} V^2.$$

is true for either choice of spaces. We demand that $\zeta \in V^0$, $\mathbf{u}, \mathbf{v} \in V^1$, and $h \in V^2$. In light of the fact that simulations should preserve the balanced components of a flow, the demand that \mathbf{u} live in a div–conforming space is obvious; if we can’t take the divergence of our velocity, we can’t test the easiest necessary condition for balance. However, the reader may be wondering why we want to choose h to be discontinuous. I intend to make the reasoning behind this choice clear in the next subsection. Anyone who

has studied finite volume methods, however, should need no convincing that h ought to live in a space of discontinuous functions. The choice $\zeta \in V^0$ allows us to prove that the PV equation (10.2.10) is weakly satisfied when the $H(\text{div})$ -flux method is used.

Recall that the spaces V^k contain information about boundary conditions. For example, if periodic boundary conditions are applied to the system, then every function in each V^k must satisfy periodic boundary conditions. Similarly, if no-normal flow boundary conditions are used, then the normal components of every vector field in V^1 must vanish on the boundary. I often do not emphasize the particular boundary conditions encoded in the V^k , largely for the sake of convenience; since I want to undertake all of the analysis in this chapter for both periodic and no-normal flow conditions simultaneously, I do not wish to constantly switch between different notation for different boundary conditions.

Remark 11.2.1. *There are other assignments of variables to finite element spaces that still respect FEEC. For example, in [22] the authors “dualize” the above choices so that both the curl and the divergence of the velocity field can be computed, but we won’t worry about such constructions here. Additionally, one can sometimes get away with using choices of finite element spaces not included in the periodic table [21], provided the V^k still form a subcomplex of the de Rham complex with an associated bounded projection.*

Now, to describe an FEM, we need to specify the spaces where each variable lives and the weak form of the governing equations. We have already taken care of the first task, so most of our effort in the remainder of this chapter is focused on obtaining usable semi-discrete weak forms of the RGN. Before moving on to the fully nonlinear equations, however, I illustrate how to use the choices of finite element spaces outlined above to put the linear RGN in a useful semi-discrete weak form. Following this, we see that the mixed framework keeps balanced states exactly steady. I then explain why choosing h to be discontinuous is helpful. In the linear case, we actually prefer to work with the free surface deformation η rather than the layer depth h (see Figure 10.21), but the idea is exactly the same: we pick $\eta \in V^2$.

Now, the linear RGN with constant ambient depth H may be written as

$$\mathbf{v} = \mathbf{u} - \frac{H^2}{3} \nabla (\nabla \cdot \mathbf{u}), \quad (11.2.1a)$$

$$0 = \partial_t \mathbf{v} + f \hat{\mathbf{z}} \times \mathbf{u} + g \nabla \eta, \quad \text{and} \quad (11.2.1b)$$

$$0 = \partial_t \eta + H \nabla \cdot \mathbf{u}. \quad (11.2.1c)$$

Note that, in the linear case, we don’t bother dealing with the vorticity at all. Recall also that balanced states are defined by

$$\mathbf{u} = \nabla^\perp \left(\frac{g}{f} \eta \right).$$

We first derive the appropriate weak form of (11.2.1) given the choices of finite element spaces V^k outlined above. Using integration by parts, we can quickly check that the semi-discrete weak form of (11.2.1) on $V^1 \times V^1 \times V^2$ is

$$\int_{\Omega} \boldsymbol{\lambda} \cdot \mathbf{v} \, dx \, dy = \int_{\Omega} \boldsymbol{\lambda} \cdot \mathbf{u} + \frac{H^2}{3} (\nabla \cdot \mathbf{u}) (\nabla \cdot \boldsymbol{\lambda}) \, dx \, dy, \quad (11.2.2a)$$

$$0 = \frac{d}{dt} \int_{\Omega} \boldsymbol{\mu} \cdot \mathbf{v} \, dx \, dy + \int_{\Omega} f \boldsymbol{\mu} \cdot (\hat{\mathbf{z}} \times \mathbf{u}) - g (\nabla \cdot \boldsymbol{\mu}) \eta \, dx \, dy, \quad \text{and} \quad (11.2.2b)$$

$$0 = \frac{d}{dt} \int_{\Omega} \alpha \eta \, dx \, dy + \int_{\Omega} H \alpha \nabla \cdot \mathbf{u} \, dx \, dy. \quad (11.2.2c)$$

for all test functions $\boldsymbol{\lambda}, \boldsymbol{\mu} \in V^1, \alpha \in V^2$. In the fully nonlinear case, there is no “preferred” weak form, and some choices have to be made to effectively handle the mass equation.

Now, in the previous section we saw how working in the FEEC mindset allows us to build discretization schemes that automatically satisfy a necessary condition for a balanced velocity field having a balanced approximation. While encouraging, this observation does not guarantee that balanced states remain time-independent at the semi-discrete level. Accordingly, we would like to prove that the weak linear RGN (11.2.2) keep balanced flows exactly steady, thus mimicking a crucial property of the system 10.3.2. To do this, we adapt methods applied in [21, §2.7] to solve the same problem for the linear RSW. Recall that we always assume either periodic or no-normal flow boundary conditions.

Theorem 11.2.2. *Suppose that $(\mathbf{u}, \mathbf{v}, \eta) \in V^1 \times V^1 \times V^2$ satisfies (11.2.2). Further, assume that there exists $\psi \in V^0$ such that $\mathbf{u} = \nabla^\perp \psi$, and that η satisfies the weak form of the geostrophic balance condition,*

$$\frac{g}{f} \int_{\Omega} \alpha \eta \, dx \, dy = \int_{\Omega} \alpha \psi \, dx \, dy \quad \forall \alpha \in V^2. \quad (11.2.3)$$

Then, $\partial_t(\mathbf{u}, \mathbf{v}, \eta) = 0$.

Proof. Since $\nabla \cdot \nabla^\perp = 0$, inputting $\mathbf{u} = \nabla^\perp \psi$ in (11.2.2c) gives

$$0 = \frac{d}{dt} \int_{\Omega} \alpha \eta \, dx \, dy \quad \forall \alpha \in V^2.$$

Therefore, the orthogonal projection of $\partial_t \eta$ onto each basis vector of V^2 is precisely zero. We then have that $\partial_t \eta = 0$.

To show that $\partial_t \mathbf{v} = 0$, we insert $\mathbf{u} = \nabla^\perp \psi$ in (11.2.2b), getting

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} \boldsymbol{\mu} \cdot \mathbf{v} \, dx \, dy &= \int_{\Omega} -f \boldsymbol{\mu} \cdot (\hat{\mathbf{z}} \times (\hat{\mathbf{z}} \times \nabla \psi)) + g(\nabla \cdot \boldsymbol{\mu}) \eta \, dx \, dy \\ &= \int_{\Omega} f \boldsymbol{\mu} \cdot \nabla \psi + g(\nabla \cdot \boldsymbol{\mu}) \eta \, dx \, dy \\ &= \int_{\Omega} -f(\nabla \cdot \boldsymbol{\mu}) \psi + g(\nabla \cdot \boldsymbol{\mu}) \eta \, dx \, dy. \end{aligned} \quad (11.2.4)$$

Our choices of finite element spaces guarantee that $\nabla \cdot \boldsymbol{\mu} \in V^2$ for all $\boldsymbol{\mu} \in V^1$. So, we can choose $\alpha = \nabla \cdot \boldsymbol{\mu}$ in (11.2.3) and plug into (11.2.4) to obtain

$$\frac{d}{dt} \int_{\Omega} \boldsymbol{\mu} \cdot \mathbf{v} \, dx \, dy = \int_{\Omega} -g \nabla \cdot \boldsymbol{\mu} \eta + g(\nabla \cdot \boldsymbol{\mu}) \eta \, dx \, dy = 0.$$

By arguments applied earlier, $\partial_t \mathbf{v} = 0$.

To show that $\partial_t \mathbf{u} = 0$, we first simplify (11.2.2a) using the condition $\nabla \cdot \mathbf{u} = 0$. This yields

$$\int_{\Omega} \boldsymbol{\lambda} \cdot (\mathbf{v} - \mathbf{u}) \, dx \, dy = 0.$$

Differentiating this expression with respect to t gives

$$\int_{\Omega} \boldsymbol{\lambda} \cdot (\partial_t \mathbf{v} - \partial_t \mathbf{u}) \, dx \, dy = 0.$$

Since $\partial_t \mathbf{v} = 0$, picking $\boldsymbol{\lambda} = \partial_t \mathbf{u}$ in the above expression immediately gives us that $\partial_t \mathbf{u} = 0$. \square

So, the mixed formulation of the linear RGN keeps balanced states steady, at least when topography is constant. This makes us confident that the mixed framework for the fully nonlinear RGN is an appropriate choice.

Remark 11.2.3. *Note how the above proof relied on us being able to choose a streamfunction $\psi \in V^0$. So, even though the weak form of the governing PDEs does not include any variables living in V^0 , we still have to define this space (and make sure it obeys the rules of FEEC) to prove statements about the discretization.*

Before concluding this section, we prove that the weak form (11.2.2) discretely conserves mass. In doing so, we see that choosing h (or η) to live in a DG space is a critically important step in making sure that the FEM replicates a vital physical principle underlying the RGN.

Proposition 11.2.4. *Suppose that (11.2.2c) is satisfied. Then, for all elements e ,*

$$\frac{d}{dt} \int_e \eta \, dx \, dy = - \int_{\partial e} H \, \mathbf{u} \cdot \mathbf{n} \, ds. \quad (11.2.5)$$

In particular, mass is globally conserved.

Proof. The result follows immediately upon choosing α to be the indicator function of e in (11.2.2c) and applying the Divergence Theorem. \square

Then, choosing a discontinuous approximation to the depth makes semi-discrete mass conservation trivial. The proofs of mass conservation in the nonlinear case are identical, relying on us being able to choose indicator functions as test functions.

11.3 Upwind Formulation of (R)GN

Now, we move on to describe perhaps the simplest mixed FEM of the fully nonlinear (R)GN. We refer to this FEM as the **upwind method** in the sequel, since we apply a technique from numerical PDE theory called **upwinding** to discretize the mass equation. I have included a brief introduction to upwinding in Subsection 11.3.2 for the benefit of those readers who have not seen this method before. Technically, there are different types of “upwinding” (designed to improve the formal accuracy of the numerical scheme), but for the present purposes we need only focus on the simplest version of this discretization technique.

For the sake of easy reference, I re-write the RGN with topography here:

$$\zeta = \hat{\mathbf{z}} \cdot (\nabla \times \mathbf{v}), \quad (10.2.13a)$$

$$\mathbf{v} = \mathbf{u} - \frac{1}{3h} \nabla (h^3 \nabla \cdot \mathbf{u}) - \frac{1}{2h} \nabla (h^2 \mathbf{u} \cdot \nabla H) + \frac{1}{2} (h \nabla \cdot \mathbf{u}) \nabla H + (\mathbf{u} \cdot \nabla H) \nabla H, \quad (10.2.13b)$$

$$\partial_t \mathbf{v} + (f + \zeta) \hat{\mathbf{z}} \times \mathbf{u} + \nabla \left(g(h - H) - \frac{1}{2} |\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2} (h \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) = 0, \text{ and} \quad (10.2.13c)$$

$$\partial_t h + \nabla \cdot (h \mathbf{u}) = 0. \quad (10.2.13d)$$

11.3.1 Weak Form of Momentum, Pseudovelocity, and Vorticity Equations

We start by deriving the weak form of the definition of \mathbf{v} . We choose a smooth test vector field $\boldsymbol{\lambda}$ (satisfying the same Dirichlet boundary conditions as \mathbf{u}) and take the dot product of both sides of (10.2.13b) with $h\boldsymbol{\lambda}$. Then, integrating by parts, we have

$$\begin{aligned} \int_{\Omega} \boldsymbol{\lambda} \cdot h\mathbf{v} \, dx \, dy &= \int_{\Omega} \boldsymbol{\lambda} \cdot h\mathbf{u} + \frac{1}{3}h^3 (\nabla \cdot \mathbf{u}) (\nabla \cdot \boldsymbol{\lambda}) + \frac{1}{2}h^2 (\mathbf{u} \cdot \nabla H) (\nabla \cdot \boldsymbol{\lambda}) \\ &\quad + \frac{1}{2}h^2 (\boldsymbol{\lambda} \cdot \nabla H) (\nabla \cdot \mathbf{u}) + h(\boldsymbol{\lambda} \cdot \nabla H) (\mathbf{u} \cdot \nabla H) \, dx \, dy. \end{aligned} \quad (11.3.2)$$

Now, we do the same for the definition of pseudovorticity. Choose an arbitrary smooth test function ξ . Multiplying both sides of (10.2.13a) by ξ and integrating by parts, we have that

$$\int_{\Omega} \xi \zeta - \nabla \xi \cdot (\hat{\mathbf{z}} \times \mathbf{v}) \, dx \, dy = 0. \quad (11.3.3)$$

Finally, we derive the weak form of the momentum equation. Choose a test function $\boldsymbol{\mu}$, take the dot product of both sides of (10.2.13c) with $\boldsymbol{\mu}$, and integrate over Ω to obtain

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} \boldsymbol{\mu} \cdot \mathbf{v} \, dx \, dy + \int_{\Omega} (f + \zeta) \boldsymbol{\mu} \cdot (\hat{\mathbf{z}} \times \mathbf{u}) \\ + \boldsymbol{\mu} \cdot \nabla \left(g(h - H) - \frac{1}{2}|\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2}(h\nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) \, dx \, dy = 0. \end{aligned}$$

Using integration by parts, we have

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} \boldsymbol{\mu} \cdot \mathbf{v} \, dx \, dy + \int_{\Omega} (f + \zeta) \boldsymbol{\mu} \cdot (\hat{\mathbf{z}} \times \mathbf{u}) \\ - (\nabla \cdot \boldsymbol{\mu}) \left(g(h - H) - \frac{1}{2}|\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2}(h\nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) \, dx \, dy = 0. \end{aligned} \quad (11.3.4)$$

To obtain the discrete weak forms suitable for input into the computer, we follow the Galerkin approach from Chapter 6 and restrict the weak forms (11.3.2), (11.3.3), and (11.3.4) to the finite element spaces V^0 and V^1 . That is, we look for $\zeta \in V^0$ and $\mathbf{u}, \mathbf{v} \in V^1$ such that (11.3.2), (11.3.3), and (11.3.4) are satisfied for all $\xi \in V^0$, all $\boldsymbol{\lambda} \in V^1$, and all $\boldsymbol{\mu} \in V^1$, respectively.

11.3.2 Introduction to Upwinding

Before discretizing the mass equation, we begin by describing how to apply upwinding in the simplest case possible. For some nonzero constant c , we consider the advection equation

$$\partial_t h + c \partial_x h = 0 \quad (11.3.5)$$

subject to an initial condition and either periodic or homogeneous Dirichlet boundary conditions. As we'll see, the simplicity of the physical ideas underlying upwinding makes the technique very easy to extend to more complicated problems.

Let $\{x_0, \dots, x_J\}$ be a partition of $[0, 1]$, and denote each individual element by $I_j = [x_j, x_{j+1}]$. We demand that, at every time t , the approximate solution h of (11.3.5) lives in the finite element space $\text{DG}(0)$. That is, h should restrict to a constant on every I_j , where each constant “piece” varies smoothly in time. To be able to solve for such an approximate h , we need to re-cast the spatial derivatives in (11.3.5) weakly. However, since our approximate h does not necessarily have a weak derivative, we have to derive the appropriate weak form working entirely in $\text{DG}(0)$; there is no way to obtain a usable weak form of (11.3.5) by “restricting” a weak form on some bigger function space to $\text{DG}(0)$. Thus, we begin by choosing an arbitrary test function $\alpha \in \text{DG}(0)$, multiplying both sides of (11.3.5) by α , and integrating over a single element I_j . This yields

$$0 = \int_{I_j} \alpha(x) \partial_t h(x, t) - c \partial_x \alpha(x) h(x, t) \, dx + [c \alpha(x) h(x)]_{\partial I_j}.$$

Technically, $\partial_x \alpha(x) = 0$ on each piece since α is piecewise constant, but I retain this term for pedagogical reasons. Since q and h are not required to be continuous across inter-element boundaries, the boundary term in the above expression has no canonical value, and we must choose the value it takes. This value is known as a **numerical flux**.

If $c > 0$, the solution of (11.3.5) at x_j must depend on the “upstream” points $x \leq x_j$ because solutions of the advection equation take the form of initial data carried uniformly along the x -axis with velocity c ; the signal ought to be influenced only by regions in space that it has already visited. This reasoning motivates us to define a numerical flux by

$$h(x_j) \doteq \begin{cases} h|_{I_{j-1}}, & c > 0; \\ h|_{I_j}, & c < 0. \end{cases} \quad (11.3.6)$$

The above choice constitutes the **upwind flux**, where the nomenclature is inspired by the discussion above. The practice of using the upwind flux to discretize PDEs is called **upwinding**. Other fluxes, some very complicated, are discussed in [52]. Strictly speaking, we would have to make exceptions to the upwinding rule when computing the fluxes at the domain boundaries, but for homogeneous Dirichlet boundary conditions we can ignore these terms entirely. Of course, when dealing with periodic boundary conditions there is no boundary to worry about.

Before writing down the complete weak form of (11.3.5) with upwinding, I introduce some helpful notation. When working on unstructured meshes, we try to avoid thinking in terms of indexed quantities like x_j ; the new notation is more amenable to such an unstructured paradigm. In this new notation, “ e ” stands for “element”. The superscript “+” is used to indicate that the quantity is evaluated over a particular element, and the superscript “−” is used to indicate that the quantity is evaluated over an element adjacent to the aforementioned particular element; the − element can be on the left or the right of the + element. \mathbf{n} is the unit outward normal to an element at some facet (of course, $\mathbf{n} = \pm 1$ in one dimension), and \tilde{h} is the upwind value of h , in the sense of (11.3.6). We call the quantity $\alpha^+ - \alpha^-$ the **jump** in α over the chosen facet. In the new notation, the weak form over a single facet e is

$$0 = \int_e \alpha \partial_t h - c \partial_x \alpha h \, dx + \left(\alpha^+ \tilde{h} c \cdot \mathbf{n}^+ \right) \Big|_{\text{endpoints}}. \quad (11.3.7)$$

Now, we sum up over each element to obtain the complete weak form. For the purposes of computational efficiency, it is useful to consider the fluxes associated to each interior facet f separately, rather than

calculating boundary terms elementwise. Accordingly, the final weak form is best written as

$$0 = \sum_e \int_e \alpha \partial_t h - c \partial_x \alpha h \, dx + \sum_f (\alpha^+ - \alpha^-) (\tilde{h} c \cdot \mathbf{n}^+). \quad (11.3.8)$$

The generalization to the case $h, \alpha \in \text{DG}(r)$ for arbitrary r should be reasonably clear.

11.3.3 Upwind Discretization of the Mass Equation

Having described what upwinding is and how to apply it to a simple equation, we may now describe how it may be used in the discretization of the RGN. We use the same “element-wise” notation applied in the previous subsection. As described in our simple example, we pick $\alpha \in V^2$, multiply both sides of (10.2.13d), and integrate over a single element e . This yields

$$\int_e \alpha \partial_t h + \alpha \nabla \cdot (h \mathbf{u}) \, dx \, dy = 0.$$

To obtain a more suitable weak form, we integrate the above equation by parts and prescribe the element boundary value of h with upwinding. Since \mathbf{u} represents the fluid’s velocity, we should switch c out for \mathbf{u} in (11.3.7). Doing all of this, we have

$$\int_e \alpha \partial_t h - \nabla \alpha \cdot h \mathbf{u} \, dx \, dy + \int_{\partial e} \tilde{h} \alpha|_e \mathbf{u} \cdot \mathbf{n} \, ds = 0.$$

Summing up over each element and using f to denote the interior facets of our triangulation of Ω , we have

$$\sum_e \left[\frac{d}{dt} \int_e \alpha h \, dx \, dy - \int_e \nabla \alpha \cdot h \mathbf{u} \, dx \, dy \right] + \sum_f \int_f \tilde{h} (\alpha^+ - \alpha^-) \mathbf{u} \cdot \mathbf{n}^+ \, ds = 0. \quad (11.3.9)$$

Note that, since we use either periodic or no-normal flow boundary conditions, there are no contributions from the mesh facets lying on $\partial\Omega$.

If we follow the approach of Proposition 11.2.4, the proof of the next result is trivial:

Proposition 11.3.1. *Suppose that the weak mass conservation equation (11.3.9) is satisfied. Then, for all elements e ,*

$$\frac{d}{dt} \int_e h \, dx \, dy = - \int_{\partial e} \tilde{h} \mathbf{u} \cdot \mathbf{n} \, ds. \quad (11.3.10)$$

In particular, mass is globally conserved.

□

Before concluding this section, we write out the semi-discrete upwind formulation of the RGN all at

once, for ease of reference later on. The weak form is

$$\int_{\Omega} \xi \zeta \, dx \, dy = \int_{\Omega} \nabla \xi \cdot (\hat{\mathbf{z}} \times \mathbf{v}) \, dx \, dy, \quad (11.3.11a)$$

$$\begin{aligned} \int_{\Omega} \boldsymbol{\lambda} \cdot h \mathbf{v} \, dx \, dy &= \int_{\Omega} \boldsymbol{\lambda} \cdot h \mathbf{u} + \frac{1}{3} h^3 (\nabla \cdot \mathbf{u}) (\nabla \cdot \boldsymbol{\lambda}) + \frac{1}{2} h^2 (\mathbf{u} \cdot \nabla H) (\nabla \cdot \boldsymbol{\lambda}) \\ &\quad + \frac{1}{2} h^2 (\boldsymbol{\lambda} \cdot \nabla H) (\nabla \cdot \mathbf{u}) + h (\boldsymbol{\lambda} \cdot \nabla H) (\mathbf{u} \cdot \nabla H) \, dx \, dy, \end{aligned} \quad (11.3.11b)$$

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} \boldsymbol{\mu} \cdot \mathbf{v} \, dx \, dy &= \int_{\Omega} - (f + \zeta) \boldsymbol{\mu} \cdot (\hat{\mathbf{z}} \times \mathbf{u}) \\ &\quad + (\nabla \cdot \boldsymbol{\mu}) \left(g(h - H) - \frac{1}{2} |\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2} (h \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) \, dx \, dy, \quad \text{and} \end{aligned} \quad (11.3.11c)$$

$$\sum_e \frac{d}{dt} \int_e \alpha h \, dx \, dy = \sum_e \int_e \nabla \alpha \cdot h \mathbf{u} \, dx \, dy - \sum_f \int_f \tilde{h} (\alpha^+ - \alpha^-) \mathbf{u} \cdot \mathbf{n}^+ \, ds \quad (11.3.11d)$$

for all test functions $\xi \in V^0$, $\boldsymbol{\lambda}, \boldsymbol{\mu} \in V^1$, and $\alpha \in V^2$.

11.4 $H(\text{div})$ –flux Method

In Chapter 10, we saw that the flow of the (R)GN leaves the total energy H_{GN} constant. Additionally, we have discussed how the flow of the RGN transports the potential vorticity q . Finally, the total mass of fluid occupying Ω is constant in time. Naturally, then, a numerical routine reproducing these three conservation principles is highly desirable. In the work of Cotter and Thuburn [22], the authors develop a mixed FEM for the RSW respecting all of the aforementioned conservation principles. Since the RGN are just a generalization of the RSW to arbitrary aspect ratio, we expect that replicating the approach of [22] leads to high-quality numerical routines for the RGN. The present section is devoted to developing a mass, vorticity, and energy conserving semi-discrete mixed finite element formulation of the (R)GN. We call this formulation the $H(\text{div})$ –flux method since it involves projecting the mass flux $h \mathbf{u}$ onto a div-conforming finite element space in order to avoid introducing auxiliary numerical fluxes.

11.4.1 Weak Form

The $H(\text{div})$ –flux method involves six variables:

- pseudovorticity ζ and potential vorticity q in V^0 ;
- velocity \mathbf{u} , pseudovelocity \mathbf{v} , and mass flux \mathbf{F} in V^1 ;
- layer depth h in V^2 .

Since the depth h is chosen to be discontinuous, discrete mass conservation is almost obvious. Accordingly, energy conservation and PV advection are the more novel mimetic properties of the $H(\text{div})$ –flux method.

To begin our derivation of the weak form, we follow [22] and first re-write the RGN in terms of the mass flux $\mathbf{F} = h\mathbf{u}$ and vorticity flux $\mathbf{Q} = (f + \zeta)\mathbf{u}$:

$$\zeta = \hat{\mathbf{z}} \cdot (\nabla \times \mathbf{v}). \quad (11.4.1a)$$

$$h\mathbf{v} = \mathbf{F} - \frac{1}{3}\nabla(h^3\nabla \cdot \mathbf{u}) - \frac{1}{2}\nabla(h^2\mathbf{u} \cdot \nabla H) + \frac{1}{2}(h^2\nabla \cdot \mathbf{u})\nabla H + h(\mathbf{u} \cdot \nabla H)\nabla H, \quad (11.4.1b)$$

$$\partial_t \mathbf{v} + \hat{\mathbf{z}} \times \mathbf{Q} + \nabla \left(g(h - H) - \frac{1}{2}|\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2}(h\nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) = 0, \text{ and} \quad (11.4.1c)$$

$$\partial_t h + \nabla \cdot \mathbf{F} = 0, \quad (11.4.1d)$$

The idea is to assume we can define discrete versions of \mathbf{Q} and \mathbf{F} ; we demand that the discrete version of \mathbf{F} has a weak divergence. Then, we obtain the weak form of the (R)GN as they are written above by working directly on the finite element spaces. Finally, I explain how \mathbf{Q} and \mathbf{F} are actually computed in terms of q, h , and \mathbf{u} to close the system.

We now turn to expressing potential vorticity q , pseudovorticity ζ , and pseudovelocisty \mathbf{v} weakly. Since the definition of q involves no derivatives, this variable admits the easiest weak formulation. Pick an arbitrary test function $\beta \in V^0$, multiply both sides of (10.2.7) by β , and integrate over Ω to obtain

$$\int_{\Omega} \beta (qh - f - \zeta) \, dx \, dy = 0 \quad \forall \beta \in V^0. \quad (11.4.2)$$

The weak definition of ζ is exactly the same as it is in the upwind method. The derivation of the weak form of (11.4.1b) is nearly identical to that for the upwind method as well, except for the presence of \mathbf{F} . Therefore,

$$\begin{aligned} \int_{\Omega} \boldsymbol{\lambda} \cdot h\mathbf{v} \, dx \, dy &= \int_{\Omega} \boldsymbol{\lambda} \cdot \mathbf{F} + \frac{1}{3}h^3(\nabla \cdot \mathbf{u})(\nabla \cdot \boldsymbol{\lambda}) + \frac{1}{2}h^2(\mathbf{u} \cdot \nabla H)(\nabla \cdot \boldsymbol{\lambda}) \\ &\quad + \frac{1}{2}h^2(\boldsymbol{\lambda} \cdot \nabla H)(\nabla \cdot \mathbf{u}) + h(\boldsymbol{\lambda} \cdot \nabla H)(\mathbf{u} \cdot \nabla H) \, dx \, dy \quad \forall \boldsymbol{\lambda} \in V^1. \end{aligned} \quad (11.4.3)$$

Next, we write the weak versions of the mass and momentum equations. We begin with the mass conservation equation (11.4.1d). Multiply both sides of this equation by a test function $\alpha \in V^2$ vanishing on $\partial\Omega$ and integrate over Ω to obtain

$$\frac{d}{dt} \int_{\Omega} \alpha h \, dx \, dy + \int_{\Omega} \alpha \nabla \cdot \mathbf{F} \, dx \, dy = 0 \quad \forall \alpha \in V^2.$$

Finally, we derive the weak evolution equation for \mathbf{v} . Choose a test vector field $\boldsymbol{\mu} \in V^1$, take the dot product of both sides of (11.4.1c) with $\boldsymbol{\mu}$, and integrate over Ω to obtain

$$\frac{d}{dt} \int_{\Omega} \boldsymbol{\mu} \cdot \mathbf{v} \, dx \, dy + \int_{\Omega} \boldsymbol{\mu} \cdot (\hat{\mathbf{z}} \times \mathbf{Q}) + \boldsymbol{\mu} \cdot \nabla \left(g(h - H) - \frac{1}{2}|\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2}(h\nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) \, dx \, dy = 0.$$

Using integration by parts, we have that the weak form is

$$\begin{aligned} &\frac{d}{dt} \int_{\Omega} \boldsymbol{\mu} \cdot \mathbf{v} \, dx \, dy \\ &+ \int_{\Omega} \boldsymbol{\mu} \cdot (\hat{\mathbf{z}} \times \mathbf{Q}) - (\nabla \cdot \boldsymbol{\mu}) \left(g(h - H) - \frac{1}{2}|\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2}(h\nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) \, dx \, dy = 0 \quad \forall \boldsymbol{\mu} \in V^1. \end{aligned}$$

Now, I explain how to actually compute the discrete mass and vorticity fluxes. The auxiliary flux \mathbf{F} is an approximation to $h\mathbf{u}$ defined by the condition

$$\int_{\Omega} \boldsymbol{\sigma} \cdot \mathbf{F} \, dx \, dy = \int_{\Omega} \boldsymbol{\sigma} \cdot h\mathbf{u} \, dx \, dy \quad \forall \boldsymbol{\sigma} \in V^1. \quad (11.4.4)$$

The purpose of introducing such an auxiliary flux is to “regularize” the discontinuous terms arising from factors of h , meaning that we do not have to worry about choosing numerical fluxes. The discrete vorticity flux is given by $\mathbf{Q} = qF$; the above logic does not require us to store \mathbf{Q} in V^1 .

In summary, the semi-discrete $H(\text{div})$ -flux formulation is defined by

$$\int_{\Omega} \xi \zeta \, dx \, dy = \int_{\Omega} \nabla \xi \cdot (\hat{\mathbf{z}} \times \mathbf{v}) \, dx \, dy, \quad (11.4.5a)$$

$$\int_{\Omega} \beta q h \, dx \, dy = \int_{\Omega} \beta (f + \zeta) \, dx \, dy, \quad (11.4.5b)$$

$$\begin{aligned} \int_{\Omega} \boldsymbol{\lambda} \cdot h\mathbf{v} &= \boldsymbol{\lambda} \cdot \mathbf{F} + \frac{1}{3} h^3 (\nabla \cdot \mathbf{u}) (\nabla \cdot \boldsymbol{\lambda}) + \frac{1}{2} h^2 (\mathbf{u} \cdot \nabla H) (\nabla \cdot \boldsymbol{\lambda}) \\ &\quad + \frac{1}{2} h^2 (\boldsymbol{\lambda} \cdot \nabla H) (\nabla \cdot \mathbf{u}) + h (\boldsymbol{\lambda} \cdot \nabla H) (\mathbf{u} \cdot \nabla H) \, dx \, dy, \end{aligned} \quad (11.4.5c)$$

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} \boldsymbol{\mu} \cdot \mathbf{v} \, dx \, dy &= \int_{\Omega} -\boldsymbol{\mu} \cdot (\hat{\mathbf{z}} \times q\mathbf{F}) \\ &\quad + (\nabla \cdot \boldsymbol{\mu}) \left(g(h - H) - \frac{1}{2} |\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2} (h\nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) \, dx \, dy, \end{aligned} \quad (11.4.5d)$$

$$\int_{\Omega} \boldsymbol{\sigma} \cdot \mathbf{F} \, dx \, dy = \int_{\Omega} \boldsymbol{\sigma} \cdot h\mathbf{u} \, dx \, dy, \quad \text{and} \quad (11.4.5e)$$

$$\frac{d}{dt} \int_{\Omega} \alpha h \, dx \, dy = - \int_{\Omega} \alpha \nabla \cdot \mathbf{F} \, dx \, dy \quad (11.4.5f)$$

for all test functions $\xi, \beta \in V^0$, $\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma} \in V^1$, and $\alpha \in V^2$.

11.4.2 Conservation properties for the $H(\text{div})$ -flux formulation

Following [22], we highlight some mimetic conservation properties of the $H(\text{div})$ -flux formulation of the RGN. The proofs of all results in this subsection, with the notable exception of energy conservation, are identical to the proofs of the analogous results for the RSW presented in [22]. We begin with the statement of mass conservation.

Proposition 11.4.1. *If (11.4.5f) is satisfied, then for all elements e we have*

$$\frac{d}{dt} \int_e h \, dx \, dy = - \int_{\partial e} \mathbf{F} \cdot \mathbf{n} \, ds. \quad (11.4.6)$$

In particular, mass is globally conserved.

□

More interestingly, we have that the equation of potential vorticity advection (10.2.10) is satisfied weakly:

Proposition 11.4.2. *If (11.4.5d) and (11.4.5a) are satisfied, then*

$$\frac{d}{dt} \int_{\Omega} \beta qh \, dx \, dy - \int_{\Omega} \nabla \beta \cdot q\mathbf{F} \, dx \, dy = 0 \quad \forall \beta \in V^0. \quad (11.4.7)$$

Proof. We begin by picking an arbitrary test function $\xi \in V^0$ and differentiating both sides of (11.4.5a) with respect to time. This gives

$$\frac{d}{dt} \int_{\Omega} \xi \zeta \, dx \, dy = \frac{d}{dt} \int_{\Omega} \nabla \xi \cdot (\hat{\mathbf{z}} \times \mathbf{v}) \, dx \, dy = - \int_{\Omega} \nabla^{\perp} \xi \cdot \partial_t \mathbf{v} \, dx \, dy.$$

Then, choosing $\boldsymbol{\mu} = -\nabla^{\perp} \xi$ in (11.4.5d) and using $\nabla \cdot \nabla^{\perp} = 0$, we obtain

$$- \int_{\Omega} \nabla^{\perp} \xi \cdot \partial_t \mathbf{v} \, dx \, dy = \int_{\Omega} \nabla \xi \cdot q\mathbf{F} \, dx \, dy.$$

Putting the above two equations together, we get

$$\frac{d}{dt} \int_{\Omega} \xi \zeta \, dx \, dy - \int_{\Omega} \nabla \xi \cdot q\mathbf{F} \, dx \, dy = 0 \quad \forall \xi \in V^0. \quad (11.4.8)$$

To establish the validity of (11.4.7), pick a test function $\beta \in V^0$. Differentiating (11.4.5b) with respect to time and applying (11.4.8) with $\xi = \beta$, we get

$$\frac{d}{dt} \int_{\Omega} \beta qh \, dx \, dy = \frac{d}{dt} \int_{\Omega} \beta (f + \zeta) \, dx \, dy = \frac{d}{dt} \int_{\Omega} \beta \zeta \, dx \, dy = \int_{\Omega} \nabla \beta \cdot q\mathbf{F} \, dx \, dy.$$

□

Corollary 11.4.3.

$$\frac{d}{dt} \int_{\Omega} \zeta \, dx \, dy = 0.$$

Proof. Choose $\xi = 1$ in (11.4.8). □

With [79, Equation 4.55] and Corollary 11.4.3 in mind, we see that the $H(\text{div})$ -flux method preserves the simplest Casimir of the shallow water bracket.

I remark that the proofs presented so far have actually been independent of the choice of flux vector \mathbf{F} , as observed in [22]. The next two results, however, crucially depend on (11.4.4). Let's define $\text{Proj}_{V^2} : L^2(\Omega) \rightarrow V^2$ by the condition

$$\int_{\Omega} f \alpha \, dx \, dy = \int_{\Omega} \text{Proj}_{V^2}(f) \alpha \, dx \, dy \quad \forall \alpha \in V^2.$$

Lemma 11.4.4.

$$\int_{\Omega} \partial_t \mathbf{u} \cdot h\mathbf{v} \, dx \, dy = \frac{1}{2} \int_{\Omega} \frac{\partial}{\partial t} (h\mathbf{u} \cdot \mathbf{v}) - \partial_t h \, \text{Proj}_{V^2} \left(|\mathbf{u}|^2 + (h\nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) \, dx \, dy. \quad (11.4.9)$$

Proof. Letting $\boldsymbol{\lambda} = \partial_t \mathbf{u}$ in (11.4.5c) and using (11.4.5e), we have

$$\begin{aligned} \int_{\Omega} \partial_t \mathbf{u} \cdot h \mathbf{v} \, dx \, dy &= \int_{\Omega} \partial_t \mathbf{u} \cdot h \mathbf{u} + \frac{1}{3} h^3 (\nabla \cdot \mathbf{u}) (\nabla \cdot \partial_t \mathbf{u}) + \frac{1}{2} h^2 (\mathbf{u} \cdot \nabla H) (\nabla \cdot \partial_t \mathbf{u}) \\ &\quad + \frac{1}{2} h^2 (\partial_t \mathbf{u} \cdot \nabla H) (\nabla \cdot \mathbf{u}) + h (\partial_t \mathbf{u} \cdot \nabla H) (\mathbf{u} \cdot \nabla H) \, dx \, dy \\ &= \frac{1}{2} \int_{\Omega} \frac{\partial}{\partial t} \left(h |\mathbf{u}|^2 + \frac{1}{3} h^3 (\nabla \cdot \mathbf{u})^2 + h^2 (\nabla \cdot \mathbf{u}) (\mathbf{u} \cdot \nabla H) + h (\mathbf{u} \cdot \nabla H)^2 \right) \\ &\quad - \partial_t h \left(|\mathbf{u}|^2 + h^2 (\nabla \cdot \mathbf{u})^2 + 2h (\nabla \cdot \mathbf{u}) (\mathbf{u} \cdot \nabla H) + (\mathbf{u} \cdot \nabla H)^2 \right) \, dx \, dy. \end{aligned}$$

Now, if we let $\boldsymbol{\lambda} = \mathbf{u}$ in (11.4.5c), the above becomes

$$\int_{\Omega} \partial_t \mathbf{u} \cdot h \mathbf{v} \, dx \, dy = \frac{1}{2} \int_{\Omega} \frac{\partial}{\partial t} (h \mathbf{u} \cdot \mathbf{v}) - \partial_t h \left(|\mathbf{u}|^2 + h^2 (\nabla \cdot \mathbf{u})^2 + 2h (\nabla \cdot \mathbf{u}) (\mathbf{u} \cdot \nabla H) + (\mathbf{u} \cdot \nabla H)^2 \right) \, dx \, dy.$$

However, $h \in V^2$ implies that $\partial_t h \in V^2$ as well. Applying the definition of Proj_{V^2} to the above expression completes the proof. \square

Theorem 11.4.5. *The semi-discrete $H(\text{div})$ -flux method conserves the Hamiltonian H_{GN} .*

Proof. Differentiating H_{GN} with respect to time, we have

$$\begin{aligned} \frac{dH_{\text{GN}}}{dt} &= \frac{1}{2} \frac{d}{dt} \int_{\Omega} g(h - H)^2 + h \mathbf{u} \cdot \mathbf{v} \, dx \, dy \\ &= \frac{1}{2} \int_{\Omega} 2g \, \partial_t h (h - H) + (\partial_t h) \mathbf{u} \cdot \mathbf{v} + \partial_t \mathbf{u} \cdot h \mathbf{v} + h \mathbf{u} \cdot \partial_t \mathbf{v} \, dx \, dy. \end{aligned}$$

By the definition of \mathbf{F} , we have

$$\frac{dH_{\text{GN}}}{dt} = \frac{1}{2} \int_{\Omega} 2g \, \partial_t h (h - H) + (\partial_t h) \mathbf{u} \cdot \mathbf{v} + \partial_t \mathbf{u} \cdot h \mathbf{v} + \mathbf{F} \cdot \partial_t \mathbf{v} \, dx \, dy.$$

Now, since $\mathbf{F} \in V^1$, we may let $\boldsymbol{\mu} = \mathbf{F}$ in (11.4.5d). This allows us to re-write the above expression as

$$\begin{aligned} \frac{dH_{\text{GN}}}{dt} &= \frac{1}{2} \int_{\Omega} 2g \, \partial_t h (h - H) + (\partial_t h) \mathbf{u} \cdot \mathbf{v} + \partial_t \mathbf{u} \cdot h \mathbf{v} - \mathbf{F} \cdot (\hat{\mathbf{z}} \times q \mathbf{F}) \\ &\quad + (\nabla \cdot \mathbf{F}) \left(g(h - H) - \frac{1}{2} |\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2} (h \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) \, dx \, dy \\ &= \frac{1}{2} \int_{\Omega} 2g \, \partial_t h (h - H) + (\partial_t h) \mathbf{u} \cdot \mathbf{v} + \partial_t \mathbf{u} \cdot h \mathbf{v} \\ &\quad + (\nabla \cdot \mathbf{F}) \left(g(h - H) - \frac{1}{2} |\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2} (h \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) \, dx \, dy. \end{aligned}$$

Since $\nabla \cdot \mathbf{F} \in V^2$,

$$\begin{aligned} \frac{dH_{\text{GN}}}{dt} &= \frac{1}{2} \int_{\Omega} 2g \, \partial_t h (h - H) + (\partial_t h) \text{Proj}_{V^2} (\mathbf{u} \cdot \mathbf{v}) + \partial_t \mathbf{u} \cdot h \mathbf{v} \\ &\quad + (\nabla \cdot \mathbf{F}) \text{Proj}_{V^2} \left(g(h - H) - \frac{1}{2} |\mathbf{u}|^2 + \mathbf{v} \cdot \mathbf{u} - \frac{1}{2} (h \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) \, dx \, dy. \end{aligned}$$

Using (11.4.5f), we get

$$\begin{aligned} \frac{dH_{\text{GN}}}{dt} &= \frac{1}{2} \int_{\Omega} g \partial_t h (h - H) + \partial_t \mathbf{u} \cdot h \mathbf{v} - \frac{1}{2} (\nabla \cdot \mathbf{F}) \text{Proj}_{V^2} \left(|\mathbf{u}|^2 + (h \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) dx dy \\ &= \frac{1}{2} \int_{\Omega} g \partial_t h (h - H) + \partial_t \mathbf{u} \cdot h \mathbf{v} + \frac{1}{2} \partial_t h \text{Proj}_{V^2} \left(|\mathbf{u}|^2 + (h \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla H)^2 \right) dx dy. \end{aligned}$$

Now, we apply Lemma 11.4.4 to see that

$$\frac{dH_{\text{GN}}}{dt} = \frac{1}{2} \int_{\Omega} g \partial_t h (h - H) + \frac{1}{2} \frac{\partial}{\partial t} (h \mathbf{u} \cdot \mathbf{v}) dx dy = \frac{1}{2} \frac{dH_{\text{GN}}}{dt}.$$

Therefore,

$$\frac{dH_{\text{GN}}}{dt} = 0$$

and the proof is complete. □

Accordingly, any discrepancy in energy associated with the $H(\text{div})$ -flux method arises due to either roundoff error or the temporal discretization scheme.

11.5 Time Discretization

In this section, I briefly introduce the **Implicit Midpoint Rule** (IMR), the finite-difference time discretization scheme used in all forthcoming numerical tests. This discretization scheme is very simple, and in principle more complicated and accurate time-steppers could be used to help numerically integrate the RGN. Accordingly, I certainly do not claim that the IMR is the best available time-stepper, but its performance is certainly acceptable enough for the “proof-of-concept” numerical tests undertaken in the coming chapters.

Suppose that we would like to discretize the system of ODEs

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}) \tag{11.5.1}$$

using finite differences. That is, we attempt to find the value of the solution \mathbf{u} only at discrete times t^n and approximate the derivatives in (11.5.1) by a special difference quotient taken over an interval of size $\Delta t \doteq t^{n+1} - t^n$. The IMR defines a choice of such a difference quotient. Letting $\mathbf{u}^n \doteq \mathbf{u}(t^n)$, the IMR is defined by

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \Delta t \mathbf{f} \left(\frac{1}{2} (\mathbf{u}^{n+1} + \mathbf{u}^n) \right). \tag{11.5.2}$$

Readers who have studied numerical analysis may notice similarities between the IMR and the more famous Crank–Nicolson method [47, §16.3]. While the two methods agree if \mathbf{f} is a linear function of \mathbf{u} , in general they are not the same: in IMR, the averaging between time steps occurs *inside* the function \mathbf{f} , while for Crank–Nicolson the value of \mathbf{f} at t^{n+1} and t^n are averaged.

One of the main reasons I have leaned towards using the IMR as a testbed time-stepper is that this discretization method preserves quadratic invariants of the system (11.5.1) [46, Problem 4.11]. However,

the kinetic energy term in the Green–Naghdi Hamiltonian (given by $\frac{1}{2}h\mathbf{u} \cdot \mathbf{v}$) is actually cubic, so the discrete energy conservation property of this time–stepper does not readily adapt to our situation. So, we cannot a priori expect that energy is conserved at the fully discrete level.

Before finishing this chapter, we recall an important concept from numerical PDE theory appearing throughout the remainder of Part 2. When solving hyperbolic PDEs on the computer, the time step Δt must be chosen so that, if c is the signal speed and h is the mesh size parameter,

$$\frac{c\Delta t}{h} \leq 1. \tag{11.5.3}$$

This ensures that information propagates on the grid no faster than it propagates in space. The dimensionless parameter on the left–hand side of (11.5.3) is called the **Courant number**.

Remark 11.5.1. *For the numerical tests in Chapter 12, we prescribe our time step by fixing the Courant number (often at 0.8), specifying h , and roughly estimating the maximum signal speed on the grid. We make no attempt to determine an “optimal” time step or Courant number through numerical experimentation; for our purposes, satisfying the inequality (11.5.3) is a guiding constraint principle, rather than the main focus of the investigations.*

11.6 Survey of Alternative Numerical Approaches to GN and RGN

In the past decade, several authors [30, 50, 53, 59, 66] have proposed different schemes for the numerical resolution of the GN and RGN, with varying degrees of focus on the mimicry of mass, energy, and PV conservation. In this section, I provide a brief overview of the numerical methodology of each paper. I then describe some of the physical applications that have been addressed and discuss how well each method has been demonstrated to respect mass, vorticity, and energy conservation.

In [53], the authors use a central discontinuous Galerkin (CDG) method in space and a finite difference method in time to simulate the 1D GN with topography. CDG methods are a special type of FEM that take advantage of “dual meshes” to maintain the benefits of using discontinuous approximating functions while avoiding having to choose numerical fluxes. [30] is essentially a sequel to this paper, presenting a similar scheme with performance improvements on the previous solver, though not providing any numerical tests involving topography. Both of these CDG formulations are based on the Hamiltonian description of the GN. [50] presents a Godunov–type method, rather than an FEM, for the GN; test cases in both 1D and 2D are presented. Again, the Hamiltonian description of the GN informs the discretization presented in this paper. [59] describes a continuous finite element approach to the 1D GN, where a solution is sought in a finite–dimensional space of spline functions. Finally, [66] describes a pseudospectral method for the RGN, valid in 1D and 2D.

Remark 11.6.1. *Many of the above papers [30, 50, 53] use the same Hamiltonian formulation of the GN we have applied here to design their methods. This is to be expected: the seemingly unusual variables encountered in the Hamiltonian description of the (R)GN conveniently allow us to reduce the order of the governing PDEs, vastly easing discretization.*

The primary purpose of many of the above papers is the study of wave propagation, and every one of the publications referenced above presents numerical schemes adept at replicating various solitary wave

solutions to the GN. Additionally, the evolution of shock waves in dam–break problems is discussed in [50] and [59], and [53] includes the effects of topography on wave propagation. The only paper featuring a numerical test involving a flow with a prominent balanced component is [66]. So, there does not appear to be a great deal of modern work on numerical simulations of RGN with *both* wave–like and balanced components. Using the mixed FEEC approach to the (R)GN, therefore, one may be able to make a novel contribution to the literature.

With regards to auxiliary conservation principles, both [30, 53] provide mass and energy conservation results for some of their test cases, and the results are rather good. The capacity of the Godunov-type method described in [50] to reproduce energy and vorticity conservation is not mentioned by the authors, though the use of such a method does guarantee mass conservation. Additionally, the finite element procedure implemented in [59] conserves energy exceptionally well, and the pseudospectral routine of [66] respects PV advection spectacularly. However, both of these schemes also have limitations: the pseudospectral routine includes local energy dissipation, and mass conservation is not discussed in [59] (their use of continuous finite element spaces indicates that mass conservation cannot be taken for granted).

Other researchers have written on the numerical solutions of modified Green–Naghdi models. For example, in [32], “local” discontinuous Galerkin methods are used in conjunction with a modified formulation of GN to study wave propagation over topography in the fully 2D regime. The modified version of GN used in this paper involves parameterizing non–hydrostatic corrections to SW as source terms, building off earlier work of the same authors. Similarly to [32], [65] describes a local DG method for another modified version of GN with topography, though the latter paper only describes 1D test cases. The physicist looking for models of non–hydrostatic effects in wave propagation, therefore, might also do well to consult these sources, rather than just focusing on the GN “as they are”.

Chapter 12

Numerical Results

In this chapter, I present several tests of the FEMs introduced in Chapter 11. The focus is largely on the $H(\text{div})$ -flux method, since its energy and vorticity conservation properties make it vastly preferable to upwinding. As discussed in Section 11.6, most of the literature on FEMs for the GN is concentrated on applications to nonlinear and dispersive wave dynamics, with particular attention paid to the simulation of solitary waves. Accordingly, to facilitate comparison between the $H(\text{div})$ -flux method and other FEMs for GN, I present many numerical test cases illustrating how well the method approximates the propagation of solitary waves, both over flat seafloors and variable topography. However, since one of the main reasons for using FEEC for RGN is to preserve important properties of large-scale oceanic flows, I also go through some test cases demonstrating how both the $H(\text{div})$ -flux method and the upwind method fare when describing large-scale dynamics.

Note that, though upwinding is very simple and intuitive, it usually leads to energy dissipation [52, §8.6]. However, numerical experiments I have conducted show that the $H(\text{div})$ -flux method is susceptible to convergence failures in 2D simulations, especially when complicated small-scale behaviour is present. While upwinding lacks many of the nice features of the $H(\text{div})$ -flux method, it seems to be a little more versatile. Accordingly, I do not claim the $H(\text{div})$ -flux method is the ultimate FEM for the (R)GN. Rather, I hope to convince the reader that it provides at least a good starting point for developing conservative numerical simulations of the (R)GN; in future studies, perhaps the $H(\text{div})$ -flux method can be modified (for instance, by sacrificing energy conservation but retaining potential vorticity advection) to develop a more robust routine.

In Section 12.1, we investigate whether or not the $H(\text{div})$ -flux method is able to reliably preserve a geostrophic balance solution to the RGN. In Section 12.2, we use the $H(\text{div})$ -flux method and the upwind method to simulate the propagation of certain solitary wave solutions to the GN by reproducing a test from [50, §6.1.2]. We also compare our results to a solitary wave propagation test from [30]. In Section 12.3, we use the $H(\text{div})$ -flux method to simulate an overtaking collision between two solitary waves, after a similar problem in [30, §5.5]. In Section 12.4, we study solitary waves moving over variable topography in a few simple test cases. In Section 12.5, we test the $H(\text{div})$ -flux method's ability to handle flows with both balanced and divergent components by simulating a wave-vortex interaction in the vein of [49], restricted to the regime of weak nonlinearity in order to prevent shock formation. In Section 12.6, we study the breakdown of an unstable balanced state in 2D using the upwind method. In Section 12.7 I present some preliminary results from test cases involving higher-degree shape functions.

All finite element codes used to implement the test cases were written in Firedrake [72]. The figures were made using Matplotlib [45] and the cmocean colormap library [87].

12.1 Geostrophic Balance

In this section we provide a quantitative analysis of how well the $H(\text{div})$ -flux method reproduces a geostrophic balance solution to the RGN. Since both FEMs from Chapter 11 respect the structure of FEEC, we anticipate that both FEMs introduced in Chapter 11 should reproduce the divergence-free balanced velocity field reasonably closely, provided we work with small enough amplitudes (recall that geostrophic balances are exact solutions to the *linearized* equations). However, since the $H(\text{div})$ -flux method conserves energy semi-discretely while upwinding is typically dissipative, we expect that the $H(\text{div})$ -flux method is more capable of reproducing a balanced state for a longer period of time. Accordingly, we do not even bother running a test with the upwind method. Throughout this section, we choose our time scale to be $T = 1/f$, our Froude number to be 1, and our length scale to be $L = \sqrt{gH}/f$; such choices constrain the Rossby number to be 1. We also assume that the bottom topography is trivial ($H = 0$).

In order to include the effects of rotation while still only having to work on a one-dimensional grid, we use the **1.5-dimensional** version of the RGN. That is, we set $\partial_y = 0$ in (10.2.14c), effectively reducing the problem to a single space dimension. However, both components of velocity and pseudovelocity are still present.

We work on a mesh of the interval $[0, 50]$ consisting of 4000 elements. For simplicity, we use periodic boundary conditions. Our time step is $\Delta t = 0.01$ and we run the routine for 10^4 time steps. Since we work in one space dimension, FEEC tells us that we only have two finite element spaces to choose, summarized in the de Rham diagram

$$\text{CG}(r) \xrightarrow{\partial_x} \text{DG}(r-1).$$

For our simulations we set $r = 1$. The depth h is the only variable that lives in $\text{DG}(0)$; every other function, including the components of the velocity and pseudovelocity, live in $\text{CG}(1)$.

The specific balanced state we want to simulate is given in terms of three constants A , a , and x_0 as

$$h_B(x) = 1 + A e^{-a(x-x_0)^2}, \quad (12.1.1a)$$

$$\mathbf{u}_B(x) = -2(x-x_0)aA e^{-a(x-x_0)^2} \hat{\mathbf{y}}. \quad (12.1.1b)$$

In our simulations, we choose $A = 0.1$, $a = 0.5$, and $x_0 = 25$. Note that we must choose A to be small in order to stay within the linear regime. We study the simulation of the above balanced state for both $\gamma = 0$ and $\gamma = 1$ to illustrate the robustness of our method with respect to changes in aspect ratio (recall that γ is the square of the aspect ratio).

To study how well our $H(\text{div})$ -flux method preserves the steady solution (12.1.1), we plot the L^2 error in the depth, defined by

$$\delta h \doteq \sqrt{\int_0^{50} (h(x,t) - h_B(x))^2 dx},$$

and the relative error in the Hamiltonian H_{GN} for both γ -values. The L^2 errors are shown in Figure 12.11, and the energy errors are shown in Figure 12.12. The error in depth is on the order of 10^{-6} , so the routine preserves the steady solution well. Additionally, the maximum relative error in energy is on

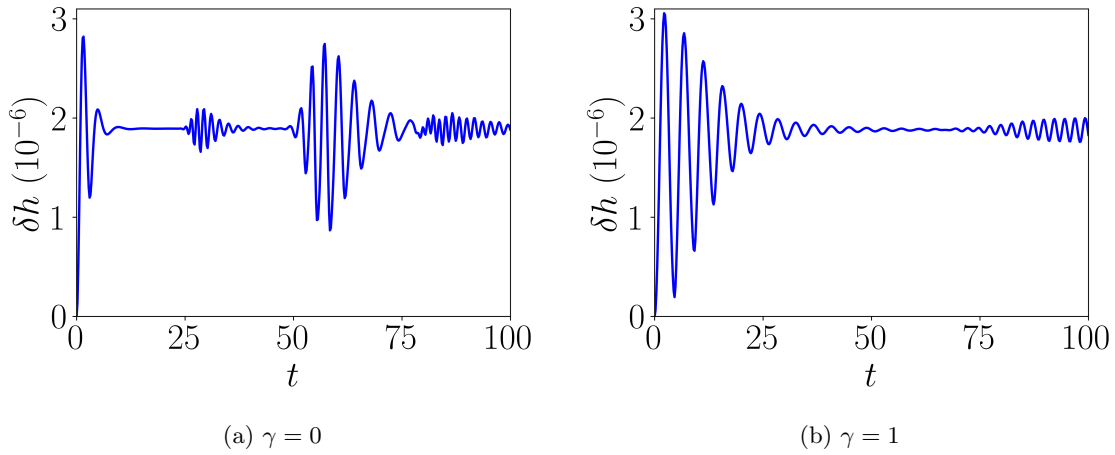


Figure 12.11: L^2 error in approximation of the depth for the geostrophic balance test. The $H(\text{div})$ -flux method has been used.

the order of 10^{-14} . This tells us that, at least in this very simple case, the $H(\text{div})$ -flux method indeed conserves energy, as predicted by Theorem 11.4.5. Note, however, that Theorem 11.4.5 only guarantees energy conservation at the semi-discrete (continuous time) level, so our numerical experiments indicate that our time-stepper does not seem to affect energy conservation.

We also remark that the divergence of the velocity field is precisely 0 for the duration of the simulation. So, our routine preserves the divergence-free condition of the balanced state perfectly. This in turn tells us that we can confidently isolate the unbalanced component of a flow by taking the divergence of the velocity field in more complicated situations.

In summary, the $H(\text{div})$ -flux method simulates the balanced state (12.1.1) rather well, with the L^2 error in the depth field on the order of 10^{-6} . Again, we expect such an exceptionally good approximation both because the numerical method is designed with FEEC in mind (divergence-free vector fields stay divergence-free) and because the numerical method conserves energy semi-discretely.

Remark 12.1.1. *To be thorough, I also performed the tests presented in this section using the upwind method. The L^2 errors in h are effectively the same for both methods, though the energy conservation is on the order of 10^{-11} when upwinding is used. So, energy conservation is clearly worse in this case, but it is certainly not terrible. Indeed, since the number of elements used for the tests is very large, we might expect that the dissipation characteristic of upwinding may not be very severe in this case.*

12.2 Solitary Waves

Now, we reproduce an elementary test from [50] using both the upwind method and the $H(\text{div})$ -flux method. This test involves examining how precisely a special solitary wave solution to the GN is preserved over a long period of time. A similar, but not identical, test is presented in [30], where two DG methods are applied to simulate solitary wave propagation. Neither of these DG methods, however, are FEEC-inspired

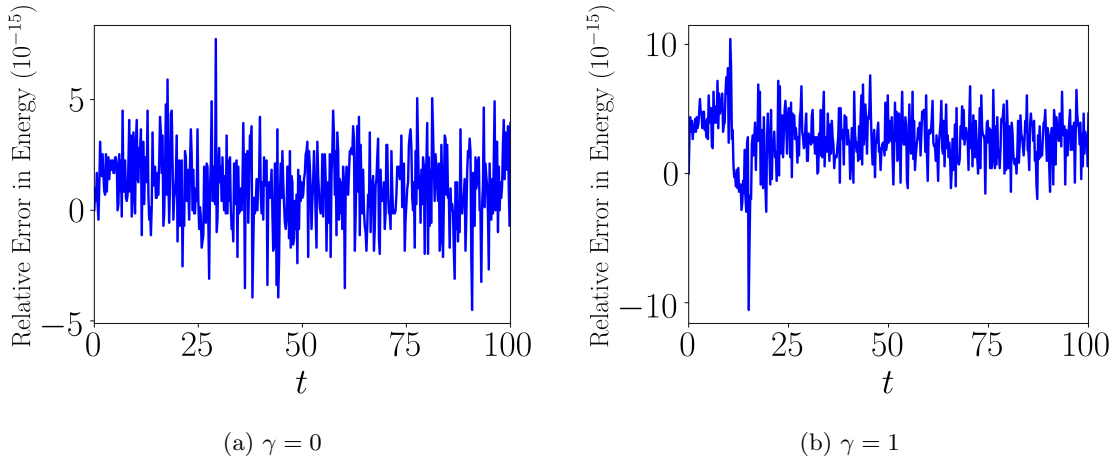


Figure 12.12: Relative error in energy for the geostrophic balance test. The $H(\text{div})$ -flux method has been used.

mixed methods like those presented in Chapter 11. We show that the $H(\text{div})$ -flux method conserves energy quite well and also maintains the shape and speed of the solitary wave. The upwind method, conversely, leads to energy dissipation and poor preservation of the solitary wave's shape over long times.

After [50, 83], we may write down an exact traveling-wave solution to the one-dimensional GN in terms of the parameters below:

- h_∞ is the depth of the fluid far from the wave;
- h_{\max} is the maximum depth of the fluid, that is, the value of h at the peak of the wave;
- $c = \sqrt{gh_{\max}}$ gives the speed of the solitary wave, where g is acceleration due to gravity.

Letting $z = x - ct$, the precise form of the solution is

$$h_{\text{wave}}(z) = h_\infty + (h_{\max} - h_\infty) \operatorname{sech}^2 \left(\frac{z}{2} \sqrt{\frac{3(h_{\max} - h_\infty)}{h_{\max} h_\infty^2}} \right) \quad (12.2.1)$$

$$u_{\text{wave}}(z) = c \left(1 - \frac{h_\infty}{h(z)} \right). \quad (12.2.2)$$

In the simulations, we take $h_\infty = 10m$, $h_{\max} = 22.5m$, and $g = 10ms^{-2}$, so $c = 15ms^{-1}$. Our domain for this problem is the interval $[0, 300m]$. We use a periodic spatial grid with 5000 elements and a Courant number of 0.8. These parameters are chosen in order to precisely replicate the conditions of the solitary wave test in [50].

The purpose of the numerical experiments in this section is to examine how well both FEMs from Chapter 11 preserve the above traveling wave solution over a very long period of time. More specifically, we pick

$$h(x, 0) = h_{\text{wave}}(x - 150) \quad \text{and} \quad u(x, 0) = u_{\text{wave}}(x - 150)$$

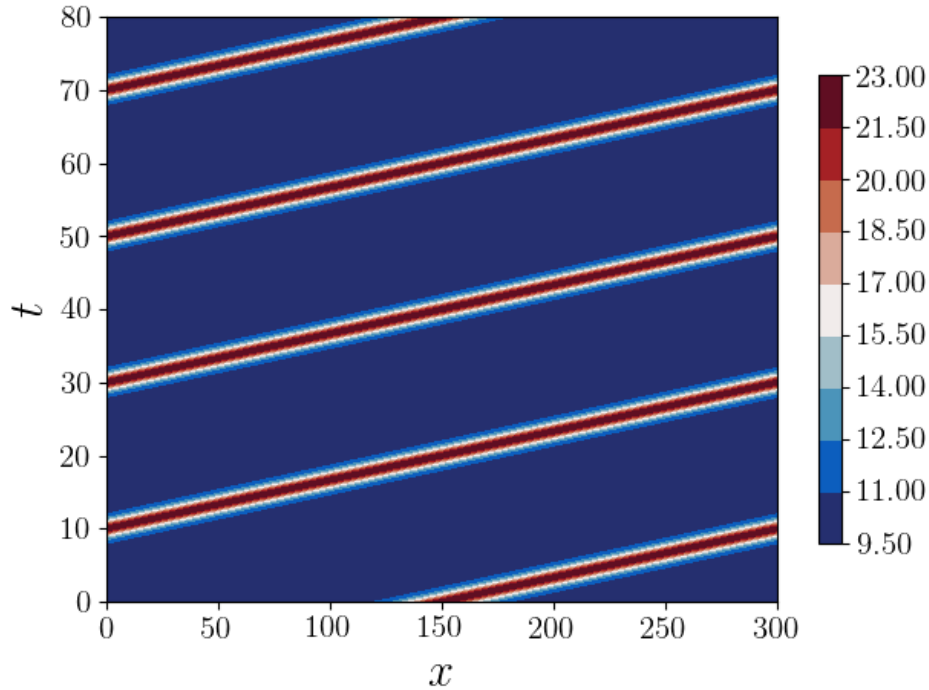


Figure 12.21: Hovmöller plot of the evolution of a solitary wave predicted by the $H(\text{div})$ -flux method, with x in metres and t in seconds. The straight red lines, representing the trajectories of the wave’s peak in the xt -plane, indicate that the wave seems to maintain constant speed.

as our initial conditions and allow the simulations to run for 25000 time steps (80s), comparing the numerical approximation of depth to the exact solution h_{wave} . In order to determine whether or not our FEMs preserve the shape and speed of the solitary wave, we make filled contour plots of the approximate solutions for depth in the xt -plane. Such pictures are called **Hovmöller plots**. Naturally, since the $H(\text{div})$ -flux method conserves energy while upwinding usually leads to numerical dissipation, we expect that it is more capable to accurately reproducing the solitary wave’s behaviour. To test this hypothesis, we produce a plot comparing the exact solution at the final time to the approximations produced by both FEMs at the final time. We also plot the relative error in energy for each method, whence we can unambiguously determine which method more accurately simulates the solitary wave.

In this simple test case, only four of our seven variables (namely, u, v, F , and h) are nonzero. The de Rham diagram of our method is

$$\text{CG}(r) \xrightarrow{\partial_x} \text{DG}(r - 1),$$

and as in the previous section we take $r = 1$. As a reminder, we choose $u, v \in \text{CG}(1)$ and $h \in \text{DG}(0)$ for both methods, and we choose $F \in \text{CG}(1)$ for the $H(\text{div})$ -flux method (of course, F does not appear in the upwind method).

We begin by providing a Hovmöller plot of the depth obtained with the $H(\text{div})$ -flux method (Figure

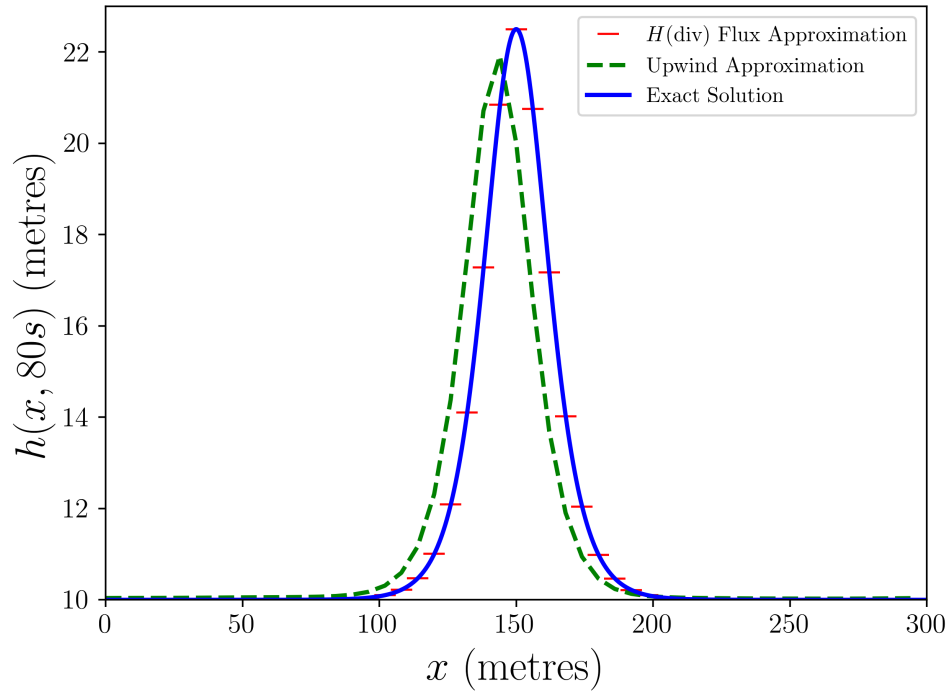
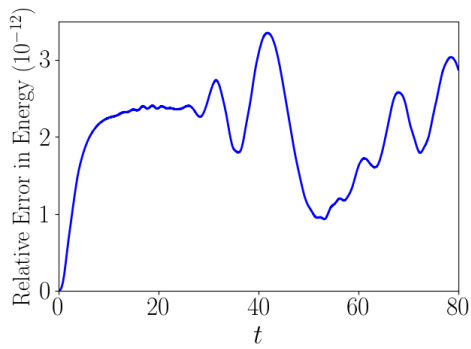
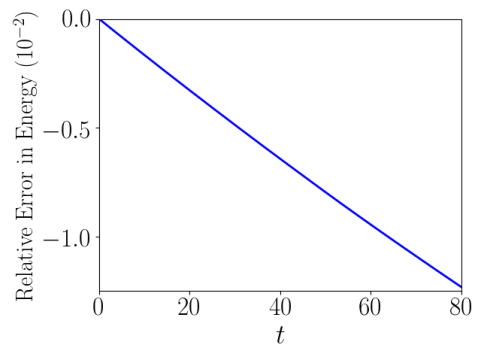


Figure 12.22: Comparison of the two approximate solutions for depth with the exact solution h_{wave} at the end of the simulation.



(a) $H(\text{div})$ -flux method



(b) Upwind method

Figure 12.23: Plot of the relative error in energy for both FEMs when 5000 elements are used. Time is measured in seconds. Note how each graph is scaled differently.

12.21) in order to quickly assess whether or not the method is preserving the shape and speed of the exact solution. The straight lines in the plot confirm that the speed of the solitary waves (namely, $15m/s$) is replicated rather well. Further, we see that the profile of the wave does not seem to change appreciably in time. That is, there is no visible dispersion error at this level of resolution. As it turns out, the same Hovmöller plot obtained with the upwind method is almost indistinguishable from Figure 12.21, so a more thorough analysis is needed in order to determine whether or not one FEM is better than the other.

Figure 12.22 shows how both approximations to depth at time $80s$ compare to the exact solution at time $80s$. We see that the approximation obtained with the $H(\text{div})$ -flux method fits the exact solution excellently, meaning that this method is at least as capable of accurately reproducing solitary waves as the schemes from [30] and [50]. The upwind approximation, however, is markedly different. Specifically, the amplitude of the solution for the upwind method has visibly decreased from its initial value. Additionally, the approximate solution is displaced to the left of the exact solution, indicating that the speed of the wave has decreased. Figure 12.22 therefore suggests that the $H(\text{div})$ -flux method provides a more accurate approximation of h_{wave} . Recall that we hypothesized that this is true because the $H(\text{div})$ -flux method conserves energy semi-discretely. So, by plotting the relative error in energy for both methods (see Figure 12.23), we may also examine whether the reasoning underlying our hypothesis is correct. According to Figure 12.23, this is indeed the case: the relative error in energy for the upwind method is about 10^{10} times higher than the error for the $H(\text{div})$ -flux method.

We remark that the $H(\text{div})$ -flux method conserves energy much better than either of the FEMs discussed in [30, Fig 2.], though the parameters of their test were different than the ones used here. Specifically, they used only 500 elements on a mesh of length 90, picked different values for h_{∞} , h_{max} , and g , and ran the simulation until the wave had travelled over the whole domain 15 times over. To examine if the superior energy conservation of the $H(\text{div})$ -flux method persists for longer times and rougher meshes, we also tested the $H(\text{div})$ -flux method on a mesh with 500 elements up to time $300s$, keeping the other parameters at the same values we've been using thus far; since these parameters only affect the size and speed of the wave, changing them should have no bearing on the quality of the numerical solution. The relative error in energy for this new test is shown in Figure 12.24. The maximum of the error is on the order of 10^{-8} , so the $H(\text{div})$ -flux method achieves energy conservation four orders of magnitude higher than that of the schemes in [30] on a mesh that is over three times coarser.

We conclude that the $H(\text{div})$ -flux method is an excellent routine for simulating the behaviour of solitary wave solutions to the GN, preserving the shape, speed, and energy of a single wave very well. Conversely, the upwind method is dissipative, even for a very fine mesh, and fails to preserve the exact solution over long times. We have seen, as in Section 12.1, that the semi-discrete energy preservation of the $H(\text{div})$ -flux method seems to extend to the fully discrete setting, implying that this FEM is preferable to upwinding for long-time simulations. Given its accuracy and excellent energy conservation, we see that the $H(\text{div})$ -flux method is capable of competing with both the DG methods from [30] and the Godunov-type scheme presented in [50], at least when it comes to simulating solitary waves.

12.3 Collision of Two Solitary Waves

In this section, we numerically study a collision between two solitary waves using the $H(\text{div})$ -flux method, which we found in the previous section to be adept at accurately describing the behaviour of a single solitary wave. The particular collision we investigate involves two solitary waves, one large and fast and the other small and slow, traveling to the right. The large wave begins to the left of the small wave,

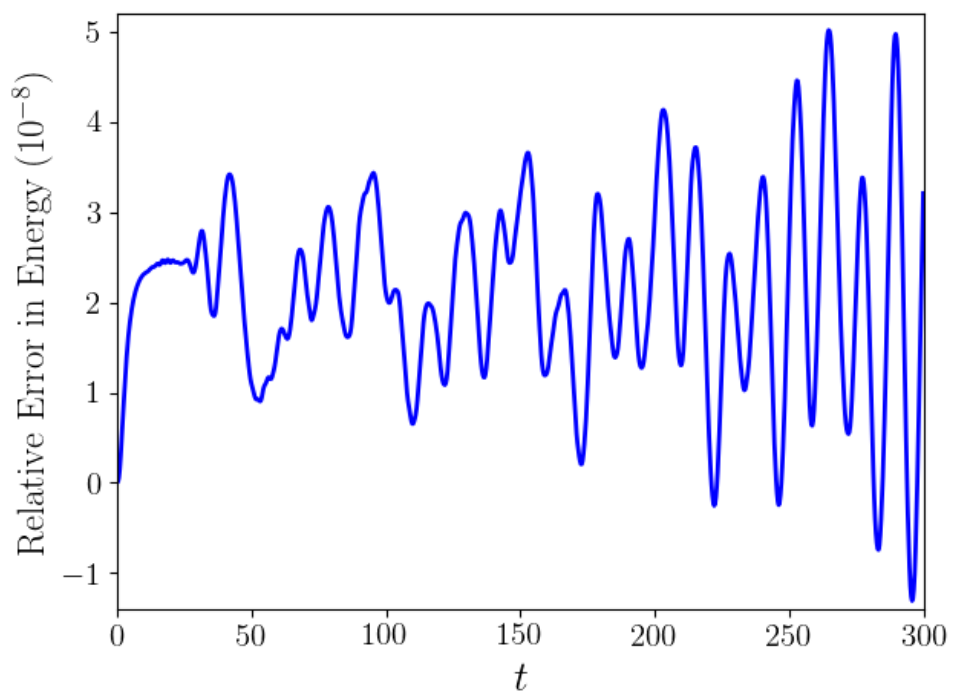


Figure 12.24: Relative error in the energy for solitary wave propagation. To obtain this graph, the $H(\text{div})$ -flux method with 500 elements was used.

slightly overlapping with it initially. As time evolves, the large wave eventually overtakes the small wave completely and ends up on its right side. A similar problem is considered in [30, §5.4], and I have tried to emulate their tests reasonably closely to facilitate comparisons. Since the authors of this paper do not provide precise parameter values, however, I do not endeavour to provide exact comparisons to their results.

Before describing our results, we introduce the precise form of the initial conditions. We need to define some parameters to specify the two solitary waves:

- h_∞ is the depth of the fluid far from the waves;
- h_1 is the the value of h at the peak of the big wave;
- h_2 is the the value of h at the peak of the small wave;
- g is the acceleration due to gravity;
- $c_1 \doteq \sqrt{gh_1}$ is the speed of the big wave;
- $c_2 \doteq \sqrt{gh_2}$ is the speed of the small wave;

Then, our initial conditions are defined by

$$\eta_1(x) \doteq (h_1 - h_\infty) \operatorname{sech}^2 \left(\frac{x - 30.24}{2} \sqrt{\frac{3(h_1 - h_\infty)}{h_1 h_\infty^2}} \right), \quad (12.3.1)$$

$$\eta_2(x) \doteq (h_2 - h_\infty) \operatorname{sech}^2 \left(\frac{x - 41.04}{2} \sqrt{\frac{3(h_2 - h_\infty)}{h_2 h_\infty^2}} \right), \quad (12.3.2)$$

$$h(x, 0) = h_\infty + \eta_1(x) + \eta_2(x), \quad \text{and} \quad (12.3.3)$$

$$u(x, 0) = c_1 \left(1 - \frac{h_\infty}{h_\infty + \eta_1(x)} \right) + c_2 \left(1 - \frac{h_\infty}{h_\infty + \eta_2(x)} \right). \quad (12.3.4)$$

We scale our variables so that $h_\infty = 1$, $h_1 = 1.44$, $h_2 = 1.15$, and $g = 1$. Mimicking [30], we use a periodic mesh of $[0, 72]$ with 1000 elements. The time step is $\Delta t = 0.048$, so the Courant number is 0.8 as in our tests for a single solitary wave. Our choices of finite element spaces are identical to those used for the test of the $H(\operatorname{div})$ -flux method in the previous section.

Figure 12.31 shows a Hovmöller plot of the two-wave interaction, and Figure 12.33 plots the depth h at eight different time steps. Examining the Hovmöller plot, we indeed see the big wave collides with the small wave (the collision becomes quite noticeable around time $t = 20$) before eventually passing through it (the waves begin to separate again just after $t = 60$), as outlined in the beginning of this section. After the collision, the size and shape of both waves do not appear to have changed markedly, but the reader should not take this as evidence that the solitary waves are in fact solitons; we have assumed a basic superposition in the initial conditions, so some nonlinear effects in the collision may be missing. Further, numerical simulations in [30] show that interactions between solitary wave solutions to GN may generate dispersive “tails”, so in general wave shape is not preserved during collisions. Figure 12.33 is similar to [30, Figure 6] but, since we do not use exactly the same parameters, the two figures are slightly different. In particular, our waves seem to be a bit slower than theirs. The qualitative agreement of our results with [30, §5.4] indicates that our simulation is reliably describing the dynamics of the two-wave collision, but a

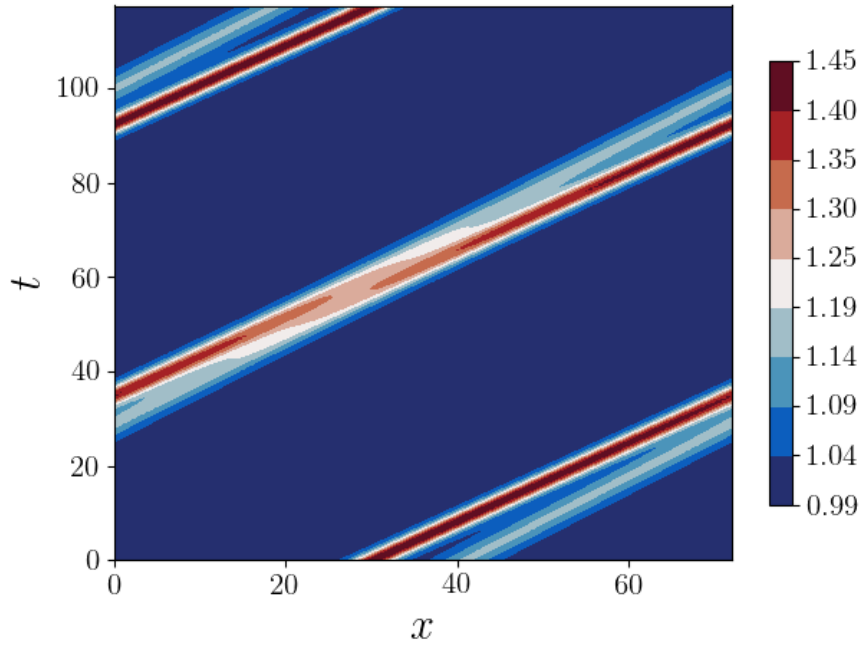


Figure 12.31: Hovmöller plot of the solitary wave collision as simulated by the $H(\text{div})$ -flux method.

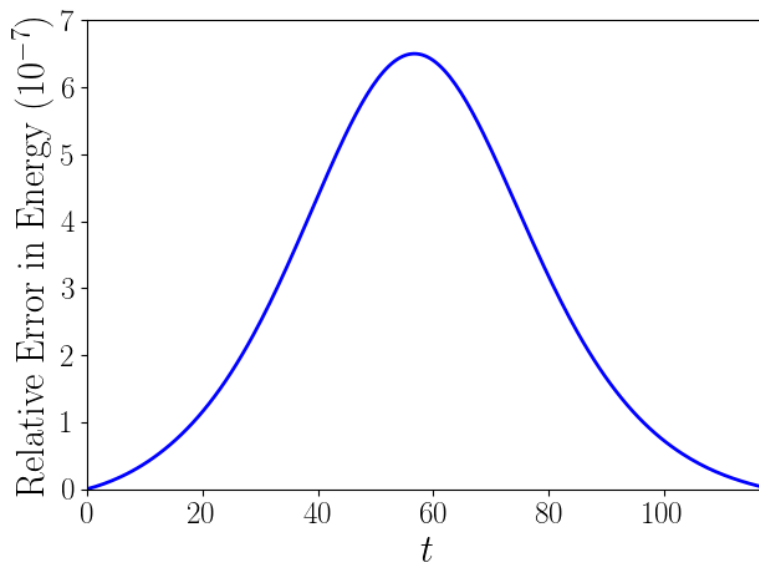


Figure 12.32: Relative error in the energy for the collision problem. The $H(\text{div})$ -flux method has been used.

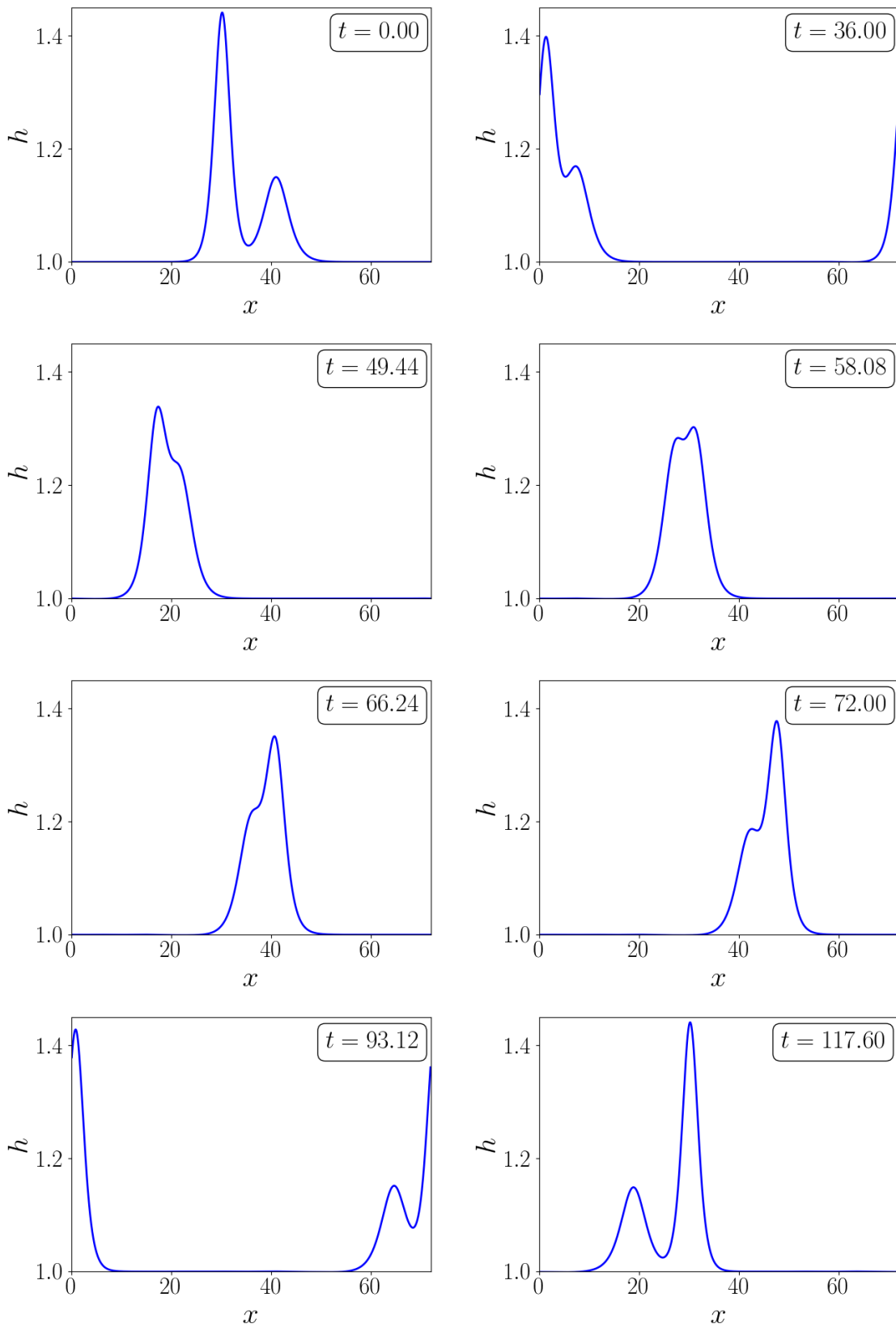


Figure 12.33: Overtaking collision of two solitary waves at eight different times, simulated by the $H(\text{div})$ -flux method.

more concrete measure of physical accuracy is required to cement our confidence. To this end, we discuss discrete energy conservation.

Figure 12.32 plots the relative error in energy for the simulation of the collision. We see that the relative error in energy is on the order of 10^{-7} ; energy conservation for the overtaking collision is not discussed at all in [30]. So, even for this nontrivial collision problem, we can rely on our method discretely conserving energy within an excellent margin of error. Further, the fastest computational runtime for the overtaking collision test in [30] is over thirty minutes, while my Firedrake implementation of the $H(\text{div})$ -flux method can complete the simulation in less than five minutes on a Mac mini. Taking the plots of h , discrete energy conservation, and computational speed all together, we see that the $H(\text{div})$ -flux method is an attractive candidate for simulating collisions between GN solitary waves.

12.4 Flow Over Topography

In the previous two sections, we have seen that the $H(\text{div})$ -flux method provides quality simulations of solitary wave solutions to the 1D GN. Now, we investigate how this numerical scheme fares when simulating solitary waves moving over an uneven seafloor. We use the non-dimensionalization discussed in Subsection 10.2.2. Recall that the dimensionless parameters controlling the dynamics for the non-rotating GN are the Froude number Fr and the squared aspect ratio γ .

We study the 1D GN with topography on a mesh of the interval $[0, L]$ with N elements; the values of L and N vary from test to test. We impose the no-normal flow boundary conditions

$$u(0, t) = u(L, t) = 0 \quad \text{and} \quad v(0, t) = v(L, t) = 0. \quad (12.4.1)$$

Remark 12.4.1. *I chose to impose a homogeneous boundary condition on the pseudovelocity v because, in the small aspect ratio regime, $v = u$. Therefore, if v satisfied a boundary condition different from the one imposed on u , the problem would be self-contradicting when $\gamma = 0$. Additionally, some preliminary numerical tests I performed yielded spurious results when this boundary condition on v was not imposed, even when $\gamma \neq 0$.*

We define the following parameters:

- h_0 is the mean undisturbed depth of the fluid;
- $h_\infty \doteq h_0 + H(0)$;
- h_{crest} is the value of h at the peak of the wave;
- x_0 is the initial position of the wave peak.

The particular values of most these parameters vary between different simulations. The initial peak position x_0 is set to $x_0 = \frac{1}{5}L$. We choose our velocity scale to be the speed of the solitary wave in water of depth h_∞ , so the Froude number is $\text{Fr} = \sqrt{h_{\text{crest}}/h_\infty}$. For all tests, we use an initial free surface deformation and initial velocity defined by

$$\eta(x, 0) = (h_{\text{crest}} - h_\infty) \operatorname{sech}^2 \left(\frac{x - x_0}{2} \sqrt{\frac{3(h_{\text{crest}} - h_\infty)}{h_{\text{crest}} h_\infty^2}} \right),$$

$$u(x, 0) = 1 - \frac{h_\infty}{\eta(x, 0) + h_\infty}.$$

If $H(x) = h_0$ identically, the initial conditions above should result in the simulation outputting a well-preserved solitary wave propagating at constant speed, at least until the wave collides with the solid wall.

We consider two main types of topography, specified by the following choices of the undisturbed depth $H(x)$:

$$\text{Case 1 : } H(x) = h_0 + sx, \tag{12.4.2}$$

$$\text{Case 2 : } H(x) = h_0 \left(1 - A_{\text{bump}} \exp \left(-\sigma \left(x - \frac{L}{2} \right)^2 \right) \right). \tag{12.4.3}$$

Case 1 corresponds to a linear ramp with slope s and Case 2 corresponds to a Gaussian hill centred at $L/2$ with amplitude A_{bump} (the parameter σ controls the width of the hill: as σ gets bigger, the hill gets narrower).

I have developed a user-friendly Firedrake demo to illustrate how the $H(\text{div})$ -flux method is implemented in practice and to allow the reader to further explore flow over topography for themselves. All of the results presented in this section are obtained using the demo. Additionally, the demo allows the user to choose undisturbed depth profiles $H(x)$ different from those presented here. These alternative depth profiles are a hyperbolic tangent and a sine function with a variable number of periods in $[0, L]$. Information on how to obtain the demo is discussed in Appendix A. For clarity, the parameter names in the demo are a little different and more explicit than they are here, but since the topography profiles are so simple this should not cause confusion.

Remark 12.4.2. *At the beginning of Chapter 10, I mentioned that the reader could use the aforementioned demo to see for themselves that solutions to GN are less susceptible to shock formation than solutions of SW. The best way to do this is to pick the demo's parameter `TopType` to be equal to 3, corresponding to a seafloor given by a hyperbolic tangent. Then, run the demo once with $\gamma = 0$ and once with $\gamma = 1$, making sure to save the outputted movies each time (you will need to change the file names to make sure one doesn't get overwritten). The movie for the $\gamma = 0$ flow shows the solitary wave breaking once it passes over the jump in topography, while the movie for $\gamma = 1$ shows the wave getting taller and narrower, but not steepening, after it passes over the jump. You may have to use a finite Reynolds number in order to keep the noise in the SW routine at a manageable level.*

12.4.1 Still-Water Solutions

Some scientists in the nonlinear waves community (see for example [53]) are concerned with developing algorithms preserving the **still-water** steady solutions to the GN, defined by

$$\begin{aligned} u(x, t) &= 0, \quad \text{and} \\ h(x, t) &= H(x). \end{aligned}$$

Note that the second equation is equivalent to $\eta(x, t) = 0$. Before presenting results on solitary wave propagation over topography, then, we should take a moment and verify that such solutions are preserved by the $H(\text{div})$ -flux method. Throughout all tests in this subsection, $H_0 = 1$, $L = 100$, $N = 500$, $\gamma = 1$, and $\text{Fr} = 1$, and the simulation is run up to $T = 200$. The Courant number is 0.8.

We set $h(x, 0) = H(x)$ and $u(x, 0) = 0$ to test still-water solutions for three different topographies: Case 1 with $s = 0$ and $s = -\frac{4h_0}{5L}$, and Case 2 with $A_{\text{bump}} = 0.8$ and $\sigma = 10$. In all three situations, the

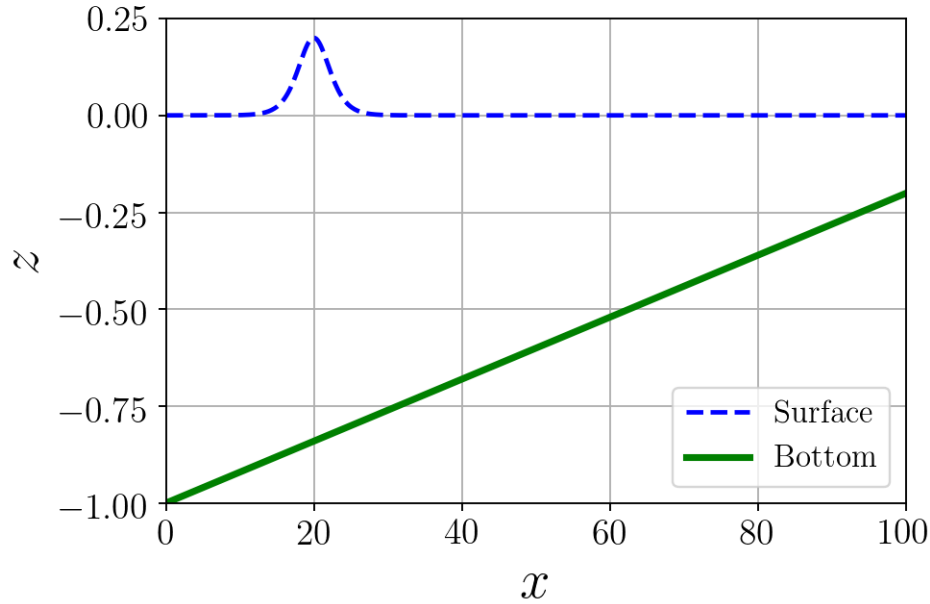


Figure 12.41: Initial conditions for the linear ramp test with $s = -\frac{4h_0}{5L}$.

relative error in energy and the L^2 error in depth are zero, so the $H(\text{div})$ -flux method preserves still-water solutions within machine precision.

12.4.2 Linear Ramps

In this subsection, we study the undisturbed depth profile (12.4.2) with

$$s = \pm \frac{4h_0}{5L}.$$

We choose $L = 100$, $N = 500$, $\gamma = 1$, $h_0 = 1$, and $h_{\text{crest}} = 1.2$. Consequently, $\text{Fr} = 1.22$. We run the simulation up to time $T = 80$ for the negative slope and $T = 60$ for the positive slope. Again, the Courant number is 0.8.

Figure 12.41 plots the initial state of the system with $s = -\frac{4h_0}{5L}$, and Figure 12.42 shows a Hovmöller plot of the solitary wave's propagation for this value of the slope. As the wave moves towards the right, it radiates waves of smaller amplitude, eventually developing a wide tail. The radiated waves all lie above the undisturbed free surface, so η is nonnegative over the course of the entire simulation. Additionally, the wave becomes taller, narrower, and slower as it moves into shallower water. The decrease of wave speed in shallow water is obvious, since the speed of a solitary wave varies proportionally with the square root of the undisturbed depth. On the other hand, the vertical stretching of the wave is an interesting feature. Such wave heightening is observed in nature when tsunamis approach shorelines, suggesting that GN may be a workable model for tsunami propagation (in fact, the non-hydrostatic effects present in GN may

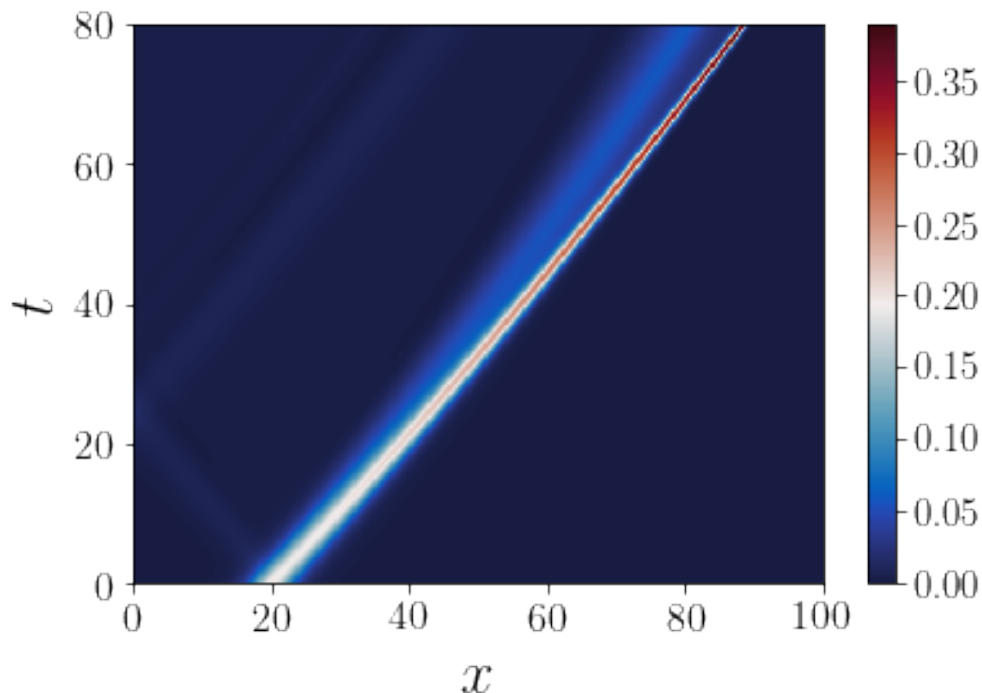


Figure 12.42: Hovmöller plot of a solitary wave traveling over a linearly increasing ramp. The $H(\text{div})$ -flux method has been used.

be valuable to tsunami modelling, after the remarks of [1]). In order to test if the $H(\text{div})$ -flux method provides physically accurate predictions of tsunami behaviour, however, simulations of waves running up onto beaches must be developed and analyzed.

The relative error in energy for this simulation is displayed in Figure 12.43. The maximum in relative error is on the order of 10^{-3} , so the quality of energy conservation has dramatically diminished in passing from $H = 0$ to variable H . This large gain in error indicates that our naïve time-stepping routine does not always maintain discrete energy conservation at a satisfactory level. In Section 13 we briefly discuss possible avenues for developing a method with better energy conservation properties.

Figure 12.44 displays the initial conditions for the system when the slope is given by $s = \frac{4h_0}{5L}$ and Figure 12.45 summarizes the propagation of the wave. The qualitative behaviour of the wave in this case is, as we might expect, exactly the opposite of what we saw in Figure 12.42: as time progresses, the wave widens, becomes shorter, and moves faster. Relative error in energy for this case is shown in Figure 12.46. Interestingly, this energy error curve is markedly different from the error for the uphill ramp (shown in Figure 12.43). First, its maximum is on the order of 10^{-4} , improving on the uphill case by an order of magnitude. Also, the two error curves indicate that energy decreases when the wave moves into shallower water and increases when the wave moves into deeper water. This suggests that the particular topography $H(x)$ strongly affects the quality of energy conservation, as well as the specific behaviour of the energy error curve. In the next section, we investigate a wave moving over non-monotone topography in order to

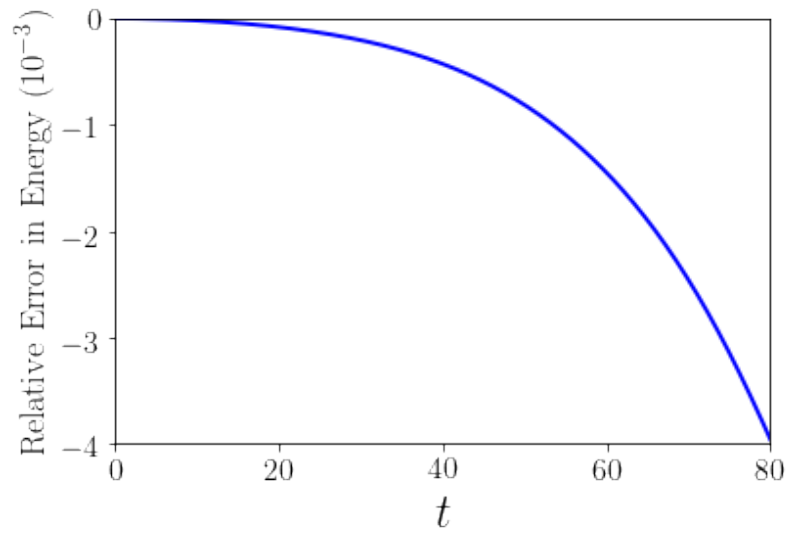


Figure 12.43: Relative error in energy for a solitary wave traveling over a linearly increasing ramp. The $H(\text{div})$ -flux method has been used.

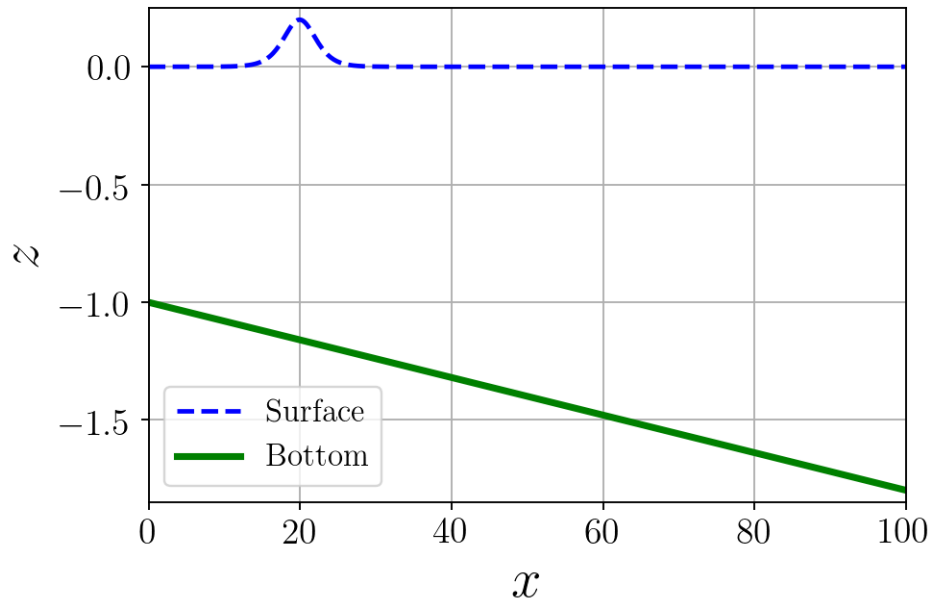


Figure 12.44: Initial conditions for the linear ramp test with $s = \frac{4h_0}{5L}$.

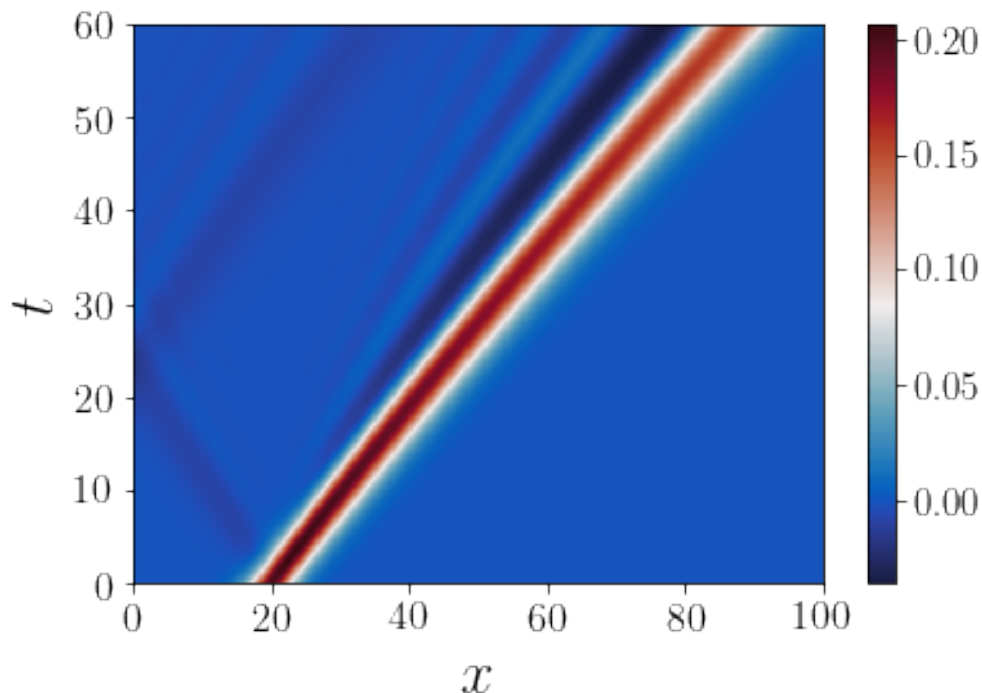


Figure 12.45: Hovmöller plot of a solitary wave traveling over a linearly decreasing ramp. The $H(\text{div})$ -flux method has been used.

verify if these patterns in energy error persist in more general situations.

12.4.3 Gaussian Hill

Now, we study the propagation of a solitary wave over a hill with shape defined by (12.4.3). The parameters of the hill are $A_{\text{bump}} = 0.6$ and $\sigma = 0.1$. As in the previous subsection, we choose $L = 100$, $N = 500$, $\gamma = 1$, $h_0 = 1$, and $h_{\text{crest}} = 1.2$. The Courant number remains at 0.8.

The initial state of the system is shown in Figure 12.47. Figure 12.48 shows a Hovmöller plot of the wave propagation over the hill. The wave maintains its shape and speed until it nears the hill, as we would naturally expect. Upon passing over the hill, a portion of the wave is reflected to the left and the wave develops a dispersive tail. The wave is briefly stretched as it travels over the hill, though it widens and decreases in amplitude once it has cleared the hill. Such behaviour is consistent with the findings of the previous subsection. However, the dispersion appears to be more complicated than we saw with the linear topography tests, with the tail continuing to develop long after the wave has moved over the hill. We discuss ideas for handling topography-induced dispersion more carefully in the next subsection.

The relative error in energy for this problem is shown in Figure 12.49. The energy error is on the order of 10^{-4} , making it an order of magnitude better than the uphill ramp test and roughly the same quality as the downhill ramp test. As we saw in the previous subsection, the relative error in energy becomes more

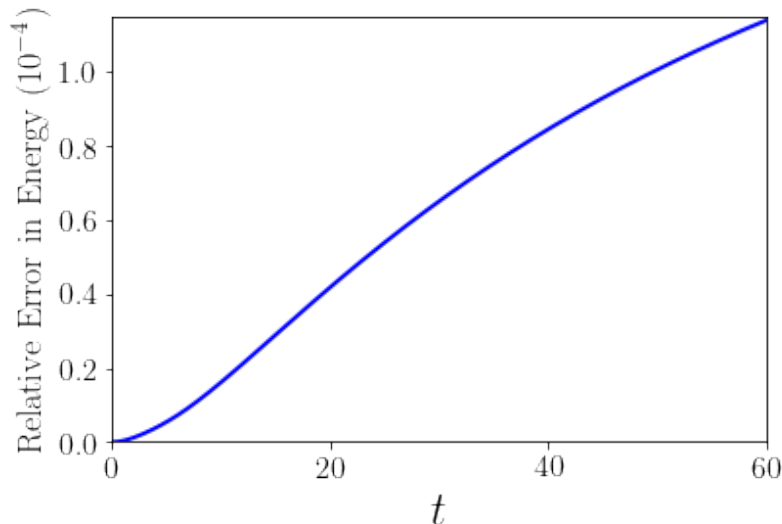


Figure 12.46: Relative error in energy for a solitary wave traveling over a linearly decreasing ramp. The $H(\text{div})$ -flux method has been used.

negative as the wave moves into a shallower region and more positive as the wave moves into deeper water. Indeed, the shape of the energy error curve almost mirrors the shape of the graph of H . We conclude that the error in energy for the $H(\text{div})$ -flux routine depends very strongly on topography.

12.5 Wave–Vortex Interaction

Next, we discuss a more complicated numerical test, inspired by [49], that gives some insight into how the $H(\text{div})$ -flux method handles flows including both divergent and balanced components. We study gravity waves colliding with a geostrophically balanced vortex, investigating how the strength of interaction between the wave and the vortex is affected by changes in aspect ratio, vortex amplitude, and vortex “parity”. Note that we use the term **vortex** to refer to a localized region of relatively high pseudovorticity $\zeta = \hat{\mathbf{z}} \cdot \nabla \times \mathbf{v}$; in particular, we do not deal with point vortices here. Such use of the term “vortex” is consistent with the nomenclature of [49].

We use the same scaling applied in Section 12.1, with the time scale given by $T = 1/f$ and the length scale given by $L = \sqrt{gh}/f$. So, the Rossby and Froude numbers are equal to 1. Then, the only dimensionless parameter appearing in the equations is the aspect ratio γ . In [49], only the case $\gamma = 0$ was considered. We begin by prescribing initial conditions on the field variables \mathbf{u} and h in terms of the parameters below:

- A_V is related to the amplitude of the initial potential vorticity curve;
- A_W is the initial amplitude of the waves;
- x_V is the centre of the balanced vortex;

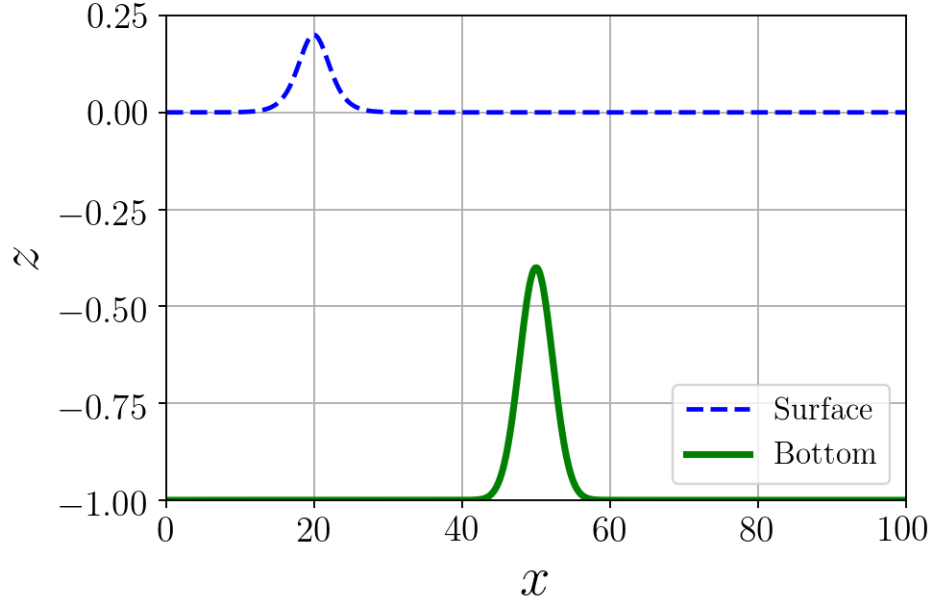


Figure 12.47: Initial conditions for the Gaussian hill test.

- x_W is the initial position of the waves;
- $C = \pm 1$ is called the **vortex parity**; if $C = 1$, we say the vortex is a **cyclone**, and if $C = -1$ then we say the vortex is an **anticyclone**.

We determine initial conditions by solving the system

$$q(x, 0) = 1 + \frac{A_V C}{2} \left[\tanh \left(\frac{x - x_V + \frac{1}{2}}{S_V} \right) + \tanh \left(\frac{-x + x_V + \frac{1}{2}}{S_V} \right) \right], \quad (12.5.1a)$$

$$0 = h_0'' - q h_0 + 1, \text{ and} \quad (12.5.1b)$$

$$\mathbf{u}_0(x) = \left(A_W e^{-(x-x_W)^2}, h_0'(x) \right). \quad (12.5.1c)$$

Figure 12.51 shows the initial potential vorticity, depth, and y -component of velocity for a special set of parameters. If the initial conditions $h_0(x)$, $\mathbf{u}_0(x)$ on $h(x, t)$, $\mathbf{u}(x, t)$ satisfy (12.5.1) then the depth and y -component of velocity initially obey the conditions of geostrophic balance. However, the system is unsteady because of the disturbance to geostrophy present in the x -component of \mathbf{u}_0 . Consequently, starting at time 0, gravity waves are radiated from x_W and, after traveling sufficiently far, these waves begin to interact with the vortex centered about x_V .

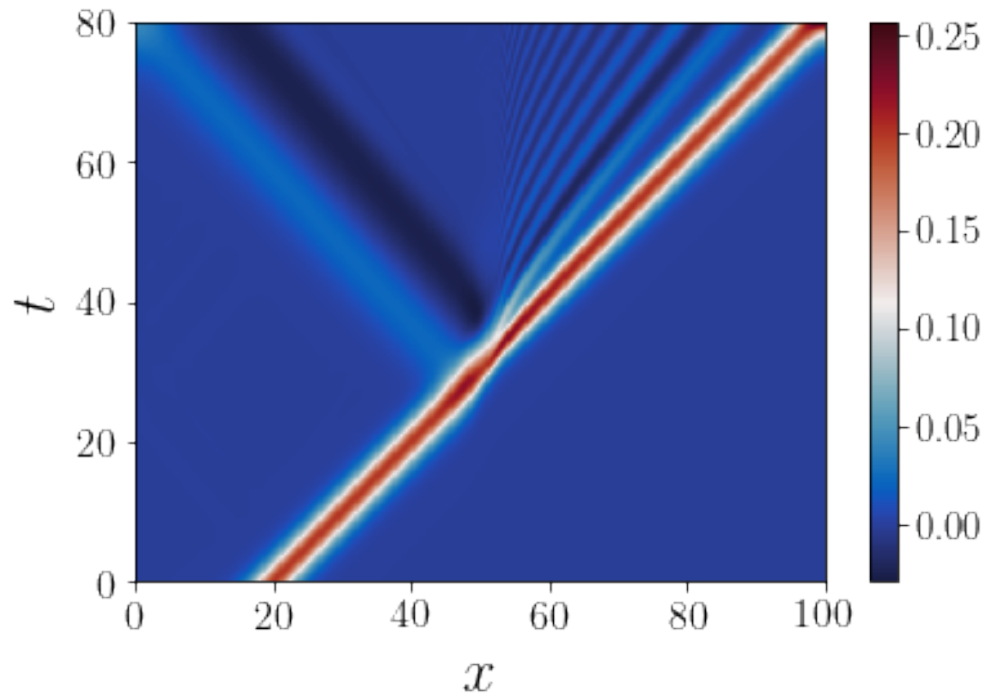


Figure 12.48: Hovmöller plot of a solitary wave traveling over a Gaussian hill. The $H(\text{div})$ -flux method has been used.

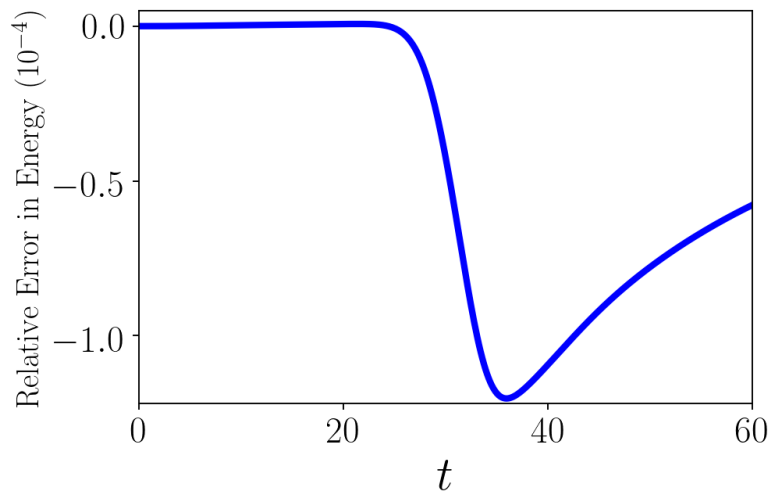


Figure 12.49: Relative error in energy for a solitary wave traveling over a Gaussian hill. The $H(\text{div})$ -flux method has been used.

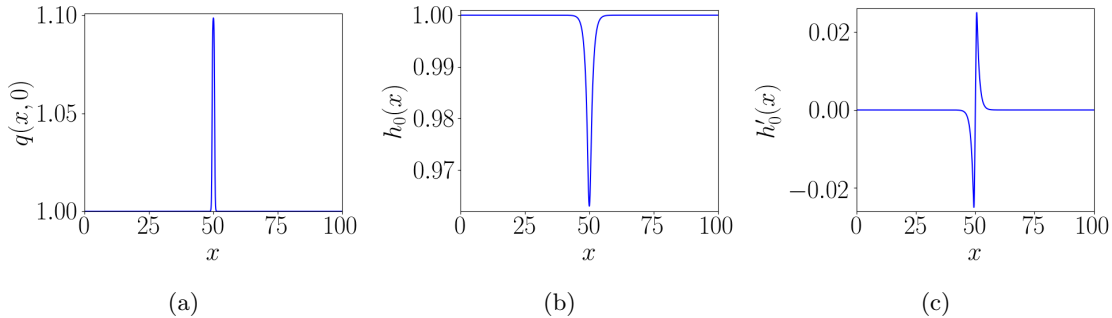


Figure 12.51: Plot of the initial (a) potential vorticity, (b) depth, and (c) y -component of velocity for the wave–vortex interaction test when the parameter values are $A_V = 0.1$, $S_V = 0.2$, $x_V = 50$, and $C = 1$.

In [49], the authors found that the “strength” $a(t)$ of the wave–vortex interaction, defined by

$$a(t) \doteq \frac{\int_{x_V-3}^{x_V+3} |q(x, t) - 1| \, dx}{\int_{x_V-3}^{x_V+3} |q(x, 0) - 1| \, dx}, \quad (12.5.2)$$

exhibits weak dependence on the vortex amplitude A_V and the cyclone parity C . More specifically, they concluded that changing either of these parameters does not affect the shape of the curve $a(t)$ much (see Figure 10 of [49]), but the amplitude of $a(t)$ is larger for $C = 1$ and larger A_V . By extending the tests from [49] to the regime of arbitrary aspect ratio, we would like to understand how $a(t)$ changes as a function of γ as well. Additionally, we can verify whether or not the aforementioned weak dependence on A_V and C holds for nonzero γ . However, we have to make a certain compromise that was unnecessary in [49]; we must restrict ourselves to $A_W = \mathcal{O}(10^{-1})$ in order to prevent shock formation, which remains an impedance to numerical solutions even for nonzero γ . Of course, adding a viscous diffusive term to the momentum equation provides a simple remedy for this problem at the cost of energy conservation, though one must be careful to discern the changes in the central vortex due to wave interaction from the changes due to viscosity–driven diffusion. A more complicated but perhaps more satisfying avenue for finding solutions in the highly nonlinear case is to try a hybrid of the $H(\text{div})$ –flux and upwind FEMs, perhaps in conjunction with a slope limiter, in order to retain PV advection at the discrete level. The development of such an amalgam FEM is outlined in [22], though I do not know if a working implementation of this method for the RGN exists at the time of writing.

To investigate the wave–vortex interaction described above, we run the 1.5D RGN model on a mesh of the interval $[0, 100]$ with periodic boundary conditions and 2000 elements. The Courant number was held at 0.8 for all tests. The assignments of variables to finite elements spaces is identical to that from Section 12.1. Recall that our goal is to study how the wave–vortex interaction strength $a(t)$ varies with A_V , C ,

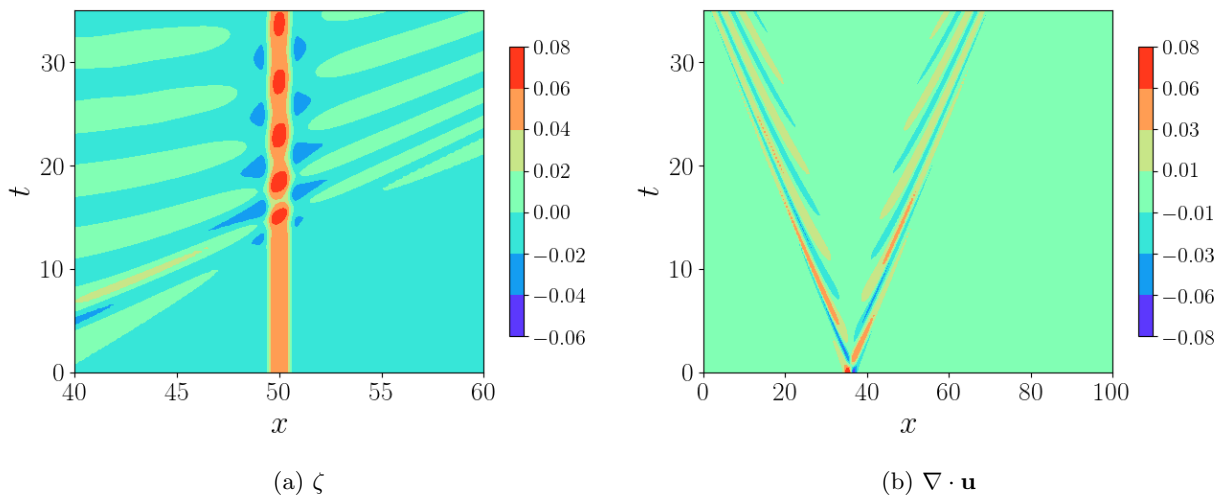


Figure 12.52: Hovmöller plot of (a) ζ and (b) the divergence of velocity with $\gamma = 0$ when $A_V = 0.1$ and $C = 1$. The $H(\text{div})$ -flux method has been used.

and γ . All other parameters are kept constant at the following values for the remainder of this section:

$$\begin{aligned} S_V &= 0.2, \\ x_V &= 50, \\ A_W &= 0.1, \text{ and} \\ x_W &= x_V - 14. \end{aligned}$$

We provide Hovmöller plots of the vorticity and divergence for two different γ -values to illustrate the qualitative changes to the wave-vortex interaction resulting from a change in aspect ratio. From there, we show various interaction strength curves $a(t)$ to investigate whether or not the results of [49] continue to hold in the regime of arbitrary aspect ratio.

Figure 12.52 shows two Hovmöller plots summarizing the evolution of the system when $\gamma = 0$, $A_V = 0.1$, and $C = 1$. Figure 12.53 does the same, except $\gamma = 1$ and the system has been allowed to evolve for a longer time. The plots shown in 12.52a and 12.53a allow us to specifically see how the vortex is affected by the incident gravity waves as the system evolves. In particular, we see that the qualitative evolution of the pseudovorticity ζ near the centre of the domain differs only slightly as we increase the aspect ratio, though the two different graphs are still clearly distinguishable. Further, by plotting the divergence alone, we isolate the propagation of gravity waves generated by the initial perturbation; since our finite element spaces are chosen within a FEEC framework, we can rely on the numerical divergence of the balanced component being zero. These plots indicate a qualitative correspondence with the dispersion relation (10.3.10), which predicts that gravity waves travel more slowly as γ increases. The relative error in energy is on the order of 10^{-9} for $\gamma = 0$ and 10^{-10} for $\gamma = 1$.

Now that we have a visual representation of the broad features of the dynamics, we describe the wave-vortex interaction as measured by the number $a(t)$. Note that the potential vorticity stayed localized around x_V in our simulations, so we could actually compute the integrals in (12.5.2) over the whole

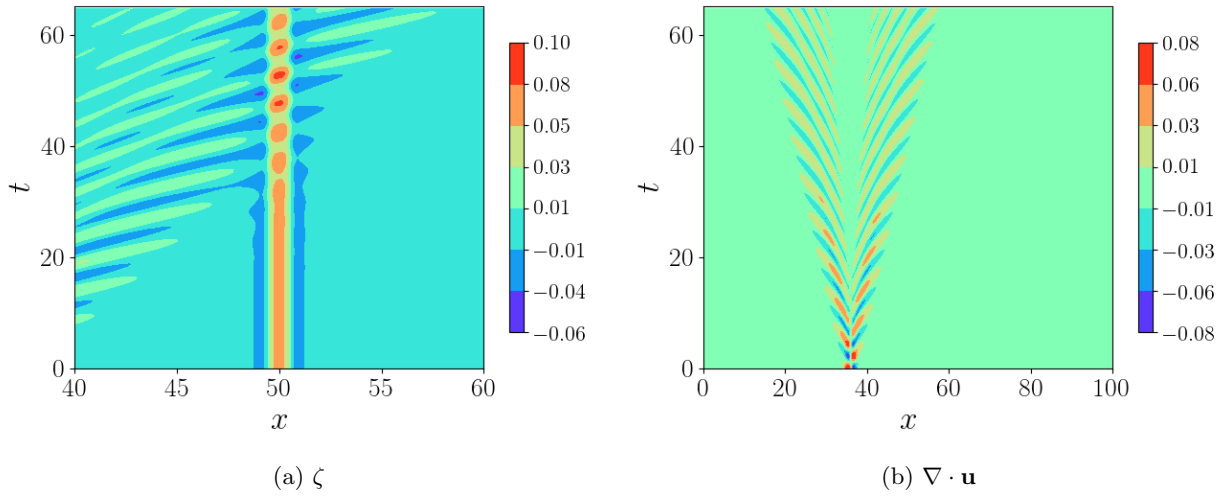


Figure 12.53: Hovmöller plot of the evolution of (a) ζ and (b) the divergence of velocity with $\gamma = 1$ when $A_V = 0.1$ and $C = 1$. The $H(\text{div})$ -flux method has been used.

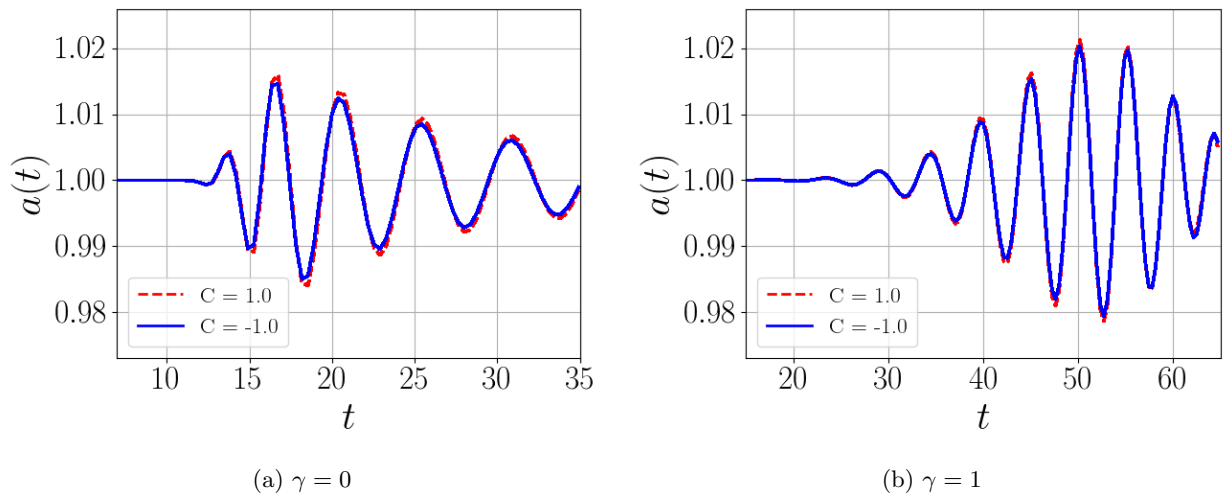


Figure 12.54: Evolution of the wave-vortex interaction parameter $a(t)$ for (a) $\gamma = 0$ and (b) $\gamma = 1$ when $A_V = 0.1$. In both cases, the solid blue curve displays the evolution when $C = -1$, and the dotted red curve represents $C = 1$. The $H(\text{div})$ -flux method has been used.

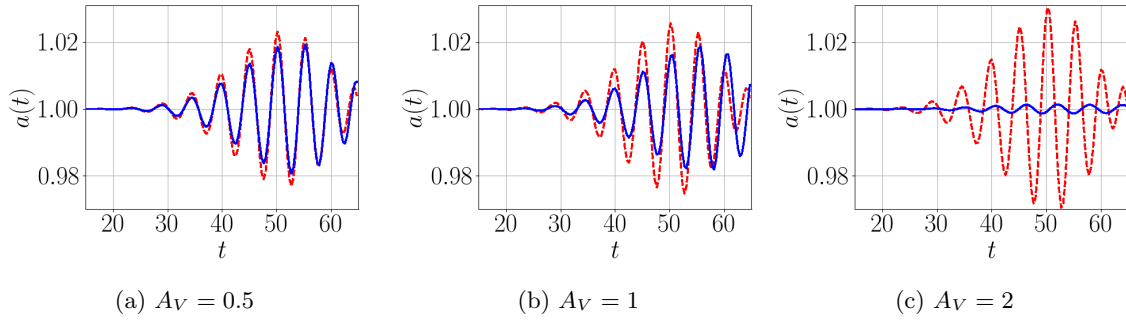


Figure 12.55: Evolution of the wave–vortex interaction parameter $a(t)$ for $A_V = 0.5$ (a), $A_V = 1$ (b), and $A_V = 2$ (c) when $\gamma = 1$. The solid blue curve represents the case $C = -1$, and the dotted red curve represents $C = 1$. The $H(\text{div})$ -flux method has been used.

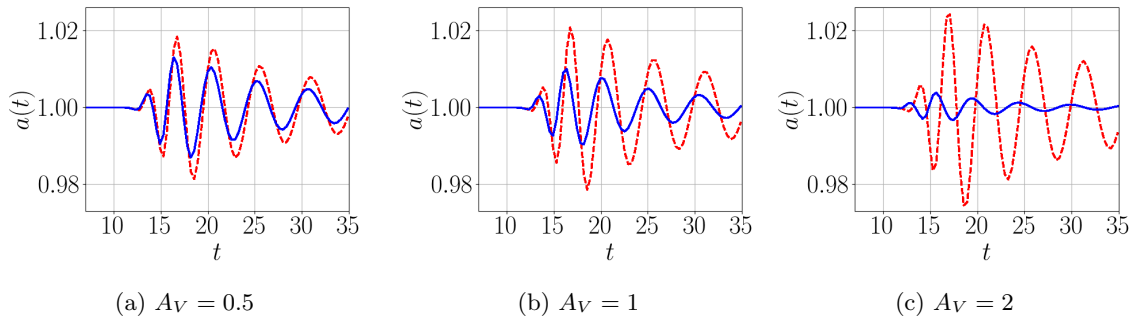


Figure 12.56: The same quantities plotted in Figure 12.55, except with $\gamma = 0$.

domain without changing $a(t)$; in Firedrake, computing such a global integral is easier than computing a local integral. Figure 12.54 plots the evolution of $a(t)$ for $A_V = 0.1$ with two curves representing $C = 1$ and $C = -1$. Comparing these curves to the Hovmöller plots (Figures 12.52a and 12.53a), we see that $a(t)$ begins to change only when the gravity waves from x_W approach the region close to x_V , cementing the reliability of $a(t)$ as a useful measure of the magnitude of the wave–vortex interaction.

Figure 12.55 plots $a(t)$ for $\gamma = 1$ and $A_V = 0.5, 1$, and 2 . Each plot contains two curves: a solid blue one for $C = -1$, and a dotted red one for $C = 1$. As A_V increases, the interaction becomes slightly stronger if $C = 1$, and slightly weaker if $C = -1$. So, an anticyclone with a large amplitude is not affected much by the fast perturbation. This shows that the weak dependence of $a(t)$ on A_V and C presented in [49] persists for nonzero γ .

Figure 12.55 demonstrates that, for $\gamma = 1$, $a(t)$ only increases by about 0.8% when A_V is increased from 0.5 to 2. In order to get a better idea of how γ affects the amplitude of $a(t)$, we reproduced Figure 12.55 for $\gamma = 0$ (Figure 12.56). Note that we cannot reliably compare Figure 12.55 quantitatively with the results of [49], since we must use a much smaller perturbation amplitude A_W than the one used in this paper. Figure 12.56 shows that, at $\gamma = 0$, the change in amplitude of $a(t)$ when A_V goes from 0.5 to 2 is about 0.4%. So, it appears that the wave–vortex interaction becomes slightly more pronounced as γ increases.

To summarize, we see that the amplitude of $a(t)$ increases with γ . In agreement with [49], we also have that cyclonic vortices (corresponding to $C = 1$) interact more strongly with the incident gravity waves than do the anticyclonic ($C = -1$) vortices. Further, the interaction becomes stronger with A_V as long as $C = 1$, while for $C = -1$ the opposite is true. These insights into the asymmetry between cyclones and anticyclones may have applications in studying the stability of balanced states; a balanced cyclone “feels” fast perturbations more than a balanced anticyclone and so, at least naïvely, it seems that a cyclone is more susceptible to instability than an anticyclone. However, the numerical results of [49] seem to suggest that (at least in the realm of the 1.5D RSW) balanced vortices like those studied in this section are, over long times, essentially unaffected by incident gravity waves. Since the dependence of $a(t)$ on aspect ratio is somewhat weak, I expect that a similar asymptotic stability result holds for the 1.5D RGN too, though developing a numerical routine that remains accurate in the strongly nonlinear regime ($A_W = \mathcal{O}(1)$ or greater) is required before this conjecture can be conclusively addressed. All in all, in order to conclusively probe the relationship between vortex stability and cyclone parity more thoroughly, one may likely have to study a 2D wave–vortex interaction, if only because in 1.5D the system is “too stable”.

Remark 12.5.1. *The general problem of cyclone–anticyclone asymmetry in geophysical flows has attracted some attention in the computational physics community (see for example [70, 71]). I have avoided an extended discussion of this subject because any proper treatment of the antisymmetry problem must necessarily involve some heavy machinery from turbulence theory. However, I remark that “preferences” for cyclones or anticyclones do exist in nature. For example, the direction of a hurricane’s rotation depends on the hemisphere where the hurricane originates.*

12.6 Breakdown of an Unstable Balanced State

I now present a test case for the upwind method in 2D involving the breakdown of an unstable balanced state. This test is inspired by [66, §5], where the authors used the same initial conditions to demonstrate the capability of their pseudospectral method to maintain advection of the potential vorticity q at the

discrete level. Our goal here is simply to demonstrate that the mixed framework outlined in the previous chapter can at least yield methods capable of resolving complicated 2D dynamics. As we show, however, there is still some work to be done before the mixed framework can be considered competitive in 2D.

Remark 12.6.1. *The $H(\text{div})$ -flux method suffered from convergence failures when I attempted to perform this test, so I was forced to switch to upwinding in order to get any results. This convergence failure is another prominent strike against the $H(\text{div})$ -flux method, alongside the energy conservation issues brought up in Section 12.4 and the inability to effectively deal with high-amplitude gravity waves in Section 12.5. I have referenced the idea of fusing the upwind and $H(\text{div})$ -flux methods while maintaining PV advection to solve some of these issues in Section 12.5, but I re-iterate that a working implementation of this combined method is unavailable at the time of writing.*

Remark 12.6.2. *In [66], fourth-order dissipative terms are added to the governing PDEs in order to quell small-scale noise in the simulation. Accordingly, we expect that grid-scale numerical dissipation might be necessary in order to resolve the complicated motion present in this instability test. While I do not intend to present a complete proof of the conjecture that our upwind method dissipates energy on small scales (such a proof would likely involve much fussing with spectral decompositions), we can naively expect that this is true: in the simplest upwind finite-difference scheme for the linear advection equation*

$$\partial_t u + c \partial_x u = 0,$$

the leading-order error term depends on $\partial_x^2 u$, hence numerical dissipation is more important on smaller length scales.

For this instability test, we assume the seafloor is flat. The initial conditions are

$$h(x, y, 0) = \frac{\tanh(\pi)}{\pi}(y - \pi) - \tanh(y - \pi) + 1 + \left[0.01 \exp \left(-\frac{25(x - \pi)^2}{\pi^2} - \frac{25(y - \pi)^2}{\pi^2} \right) \right], \quad (12.6.1)$$

$$\mathbf{u}(x, y, 0) = \left(\text{sech}^2(y - \pi) - \frac{\tanh(\pi)}{\pi} \right) \hat{\mathbf{x}}. \quad (12.6.2)$$

If the term in the square brackets is ignored, then the system is in geostrophic balance. The bracketed term, however, spoils the balance slightly and, as we soon see, actually causes the balanced state to completely break down. We set Fr , Ro , and $\gamma = 1$, following [66]. We use a 60×60 periodic mesh of the square with side length 2π , and the time step is

$$\Delta t = \frac{2\pi}{75 \left(1.01 + \frac{\tanh(\pi)}{\pi} \right)} \approx 0.063.$$

This time step is chosen because, very roughly, we can estimate the typical gravity wave speed in the fluid as $1.01 + \frac{\tanh(\pi)}{\pi}$, and we wish to keep the Courant number at about 0.8. For the particular results shown, the de Rham diagram of the method is

$$\text{CG}(2) \xrightarrow{\nabla^\perp} \text{BDM}(1) \xrightarrow{\nabla} \text{DG}(0).$$

We could have also, in principle, used the trimmed complex.

Stills of the layer depth at various times are displayed in Figure 12.61. For a while, the differences between the solution from the balanced state are slight, but eventually small-scale motion in the region

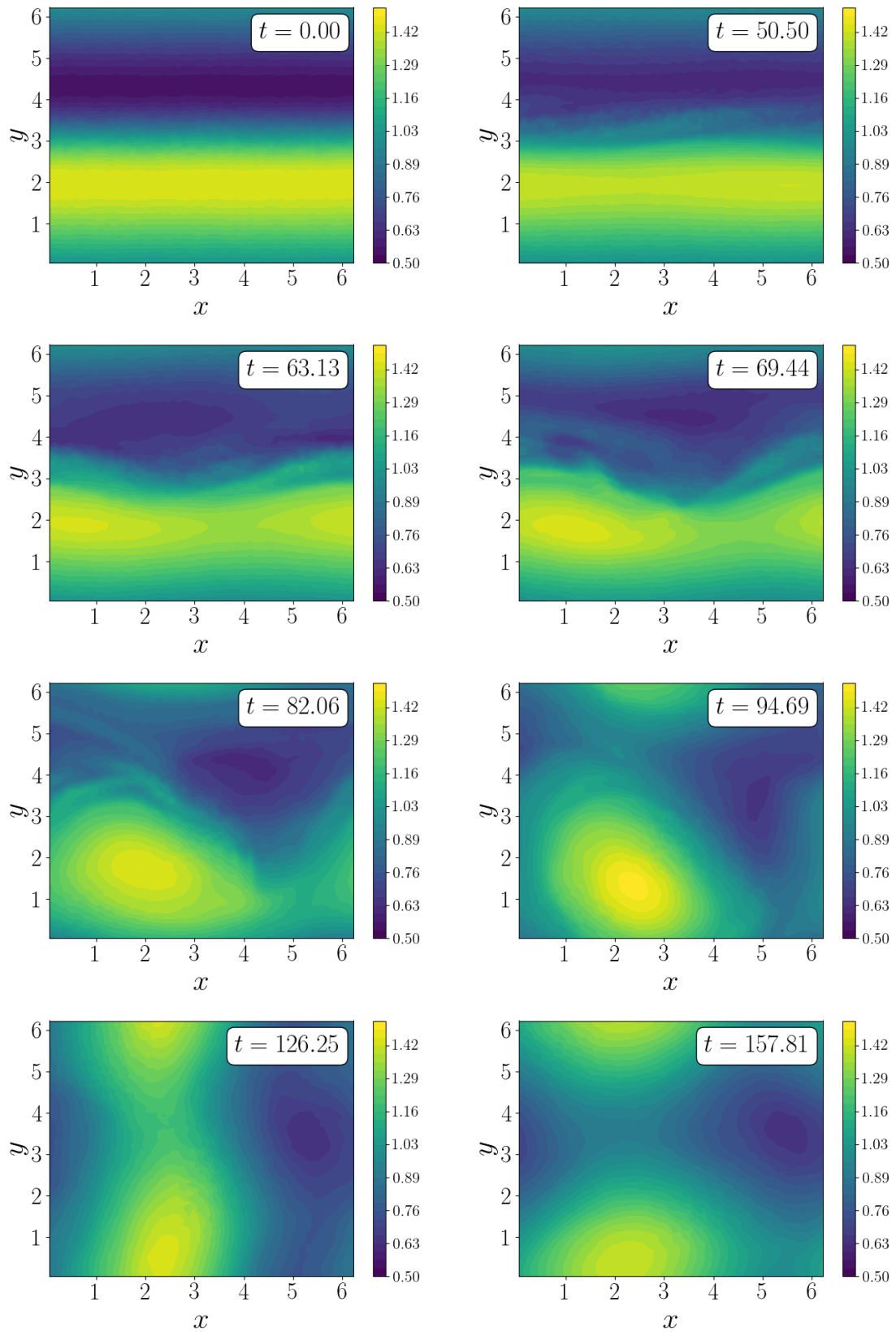


Figure 12.61: Filled contour plots of the layer depth h for the instability test at eight different times. The upwind method has been used.

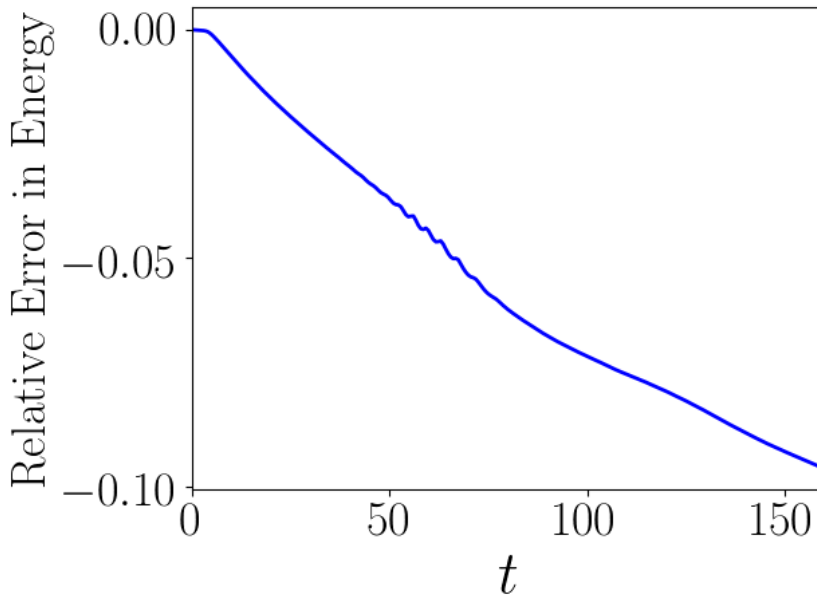


Figure 12.62: Relative error in energy for the instability test. The upwind method has been used.

$2 \leq y \leq 4$ starts to influence the system more strongly, creating more pronounced differences. The structure of the initial depth field has completely broken down by $t = 82.06s$. So, the upwind method seems to be capable of resolving the complicated breakdown of the balanced state to at least a modest degree.

To assess the quality of the simulation more quantitatively, we investigate how well the simulation conserves energy. The relative error in energy is shown in Figure 12.62. The conservation quality here is the poorest seen so far in this chapter, with the maximum error on the order of 10^{-1} . We conclude that, even though the complicated small-scale motions do not wreak havoc on the solver, there is more work to be done before the method can be earnestly relied upon.

12.7 Higher-Order Shape Functions: Preliminary Results

All of the numerical tests presented in Chapter 12 used the lowest-order shape function degrees possible; my focus was more on assessing the quality of energy conservation and preserving the separation between balanced and fast components of a flow, rather than formally analyzing the accuracy of the numerical schemes. However, I have completed some rudimentary tests with higher-order polynomials and would like to briefly highlight my findings here. I re-performed the balanced state test from Section 12.1 only using the $H(\text{div})$ -flux method and $\gamma = 1$, as well as a version of the solitary wave test from Section 12.2 using both the $H(\text{div})$ -flux method and the upwind method. The de Rham diagram associated to both of these problems is

$$\text{CG}(r) \xrightarrow{\partial_x} \text{DG}(r-1)$$

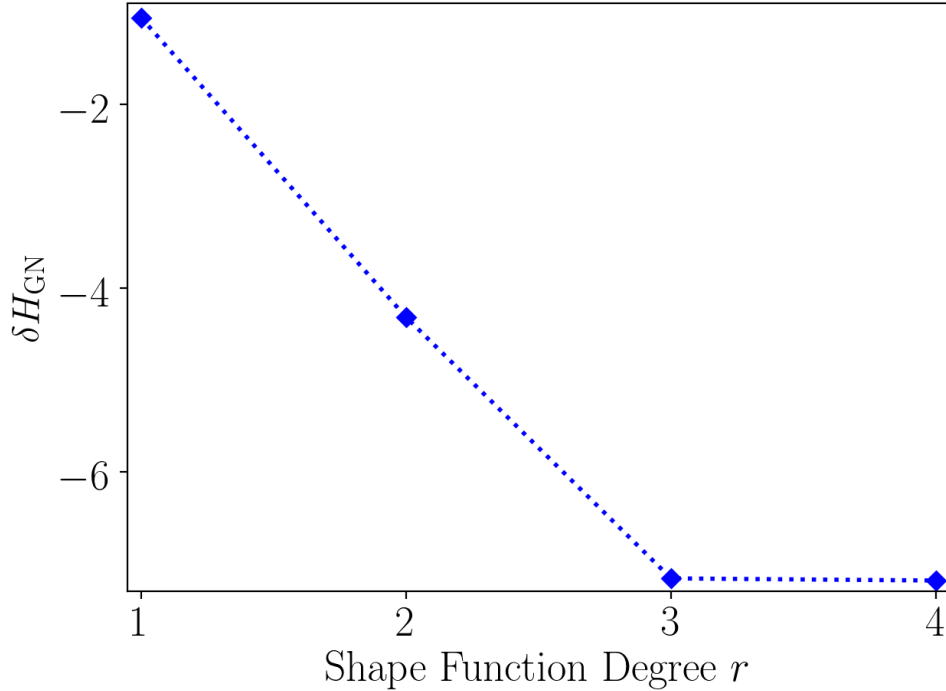


Figure 12.71: Changes in $\delta H_{GN} \doteq \log_{10} \left(\max \frac{H_{GN}(t) - H_{GN}(0)}{H_{GN}(0)} \right)$ with respect to shape function degree r for simulations of solitary wave propagation by the upwind method. A mesh with 510 elements was used, and the Courant number was 0.8.

and I considered the cases $r = 2, 3, 4$.

For the balanced state tests, the L^2 error in h decreases as r increases, eventually reaching the order of 10^{-12} when $r = 4$. However, the magnitude of relative error in energy increases slightly as r increases. For all r , the maximum energy error is on the order of 10^{-14} , though as r gets bigger these maxima in error appear much more frequently.

For solitary waves, the situation is more complicated. Using a mesh of $[0, 300]$ with 500 elements and $r = 3, 4$, the $H(\text{div})$ -flux method could run well up to time 80s. However, if the routine was allowed to run up to around 140s, numerical instability would develop and spoil the simulation (decreasing the time step does not appear to resolve this problem). When $r = 2$, the simulation developed spurious dispersion very quickly. Accordingly, the $H(\text{div})$ -flux method does not seem reliable when $r > 1$. The upwind method, however, suffers from no such problems when r is increased. Indeed, when the upwind method is used on a mesh of $[0, 300]$ with 510 elements and $r \geq 3$, the relative error in energy decreases dramatically as r is increased (see Figure 12.71): when $r = 3$, the error drops down to the order of 10^{-7} . Therefore, anyone wanting to use higher-order shape functions in the mixed framework for the (R)GN should probably stick to the upwind method.

Chapter 13

Conclusions and Future Directions

13.1 Summary of Results and Conclusions

In this section, I summarize some of the main results presented in this chapter and present some general conclusions about the quality of the numerical methods used. First, we saw that the $H(\text{div})$ -flux method preserves balanced flows very well, with L^2 error in depth on the order of 10^{-6} and relative error in energy on the order of 10^{-14} . As I argued earlier, this quality was to be expected considering both the FEEC-informed framework for choosing finite element spaces and the mimetic conservation properties of the $H(\text{div})$ -flux method. We then saw that solitary wave propagation is also well-resolved by the $H(\text{div})$ -flux method, with relative errors in energy on the order of 10^{-12} for a mesh of $[0, 300]$ with 5000 elements and on the order of 10^{-8} for a mesh with 500 elements. Upwinding performed comparatively poorly in this test case, with the wave noticeably losing amplitude and slowing down by the end of the simulation. The $H(\text{div})$ -flux method also resolved solitary wave motion satisfactorily during simulations of wave-wave and wave-topography interactions, though errors in energy do increase markedly (sometimes up to the order of 10^{-3}) when the topography is nontrivial. We also tested the $H(\text{div})$ -flux method by simulating a wave-vortex interaction (that is, a flow with both a balanced and a wave-like component). The simulation preserved the separation between the balanced and wave-like components very well and energy conservation was again seen to be very good, with relative errors on the order of 10^{-9} or less. Additionally, we saw that the strength of the wave-vortex interaction displayed a “cyclone-anticyclone asymmetry” for both $\gamma = 0$ and $\gamma = 1$, slightly extending results presented in [49]. We then used the upwind method to simulate the breakdown of an unstable balanced state in 2D; the $H(\text{div})$ -flux method was not able to resolve the complex small-scale motion of the fluid in this situation. While energy conservation is poor for this test (relative error is on the order of 10^{-1} at the end of the simulation), we at least have a first step towards applying the FEEC-informed mixed formulation of RGN to turbulent flows. Finally, we saw that the $H(\text{div})$ -flux method performs poorly when compared to upwinding as shape function degree is increased, at least for the simple preliminary tests highlighted in Section 12.7.

To conclude, the $H(\text{div})$ -flux method with low-order shape functions provides good simulations of many one-dimensional problems, often with exceptional energy conservation (especially when compared with the upwind scheme). The method is also adept at preserving the balanced components of flows in rotating domains. Energy conservation does become less impressive when bottom topography is present, however. Additionally, the $H(\text{div})$ -flux scheme can fail to converge when small-scale motion is complicated. When

such complex motion is present, the upwind method can be used to resolve the dynamics, though further work is needed before this routine can be considered competitive in 2D.

13.2 Limitations of the Mixed Methods and Suggestions for Further Work

In this section, we review some of the main problems with both the upwind and $H(\text{div})$ -flux methods, suggesting possible paths towards solutions to be pursued in future research. Since the main issue with upwinding, namely poor energy conservation, is well-known in numerical analysis, we focus primarily the limitations of the $H(\text{div})$ -flux method that we discovered in our numerical tests. These issues include convergence failures in 2D, loss of high-quality energy conservation in the presence of topography, and lack of a thorough analysis of numerical dispersion error.

I have already said that a good “first step” in trying to fix the convergence failures might involve using an amalgam of both mixed methods with the goal of maintaining PV advection while decreasing energy; I first brought up this idea in Section 12.5 as a way of getting around the issues with high-amplitude gravity waves we encountered when reproducing the tests from [49]. This idea of a combined method was first presented by Cotter and Thuburn in [22], and it is currently is being pursued in the context of the RSW by Shipton and Cotter [80]. However, a practical implementation of the amalgam method may be rather difficult; in the winter of 2018 I attempted to implement this method in Firedrake with the help of Jemma Shipton, but I was unable to get a working simulation of the instability test from Section 12.6 and ultimately abandoned the attempt. However, this new method may work with simpler test cases and may be worth another try in the future.

In Section 12.4, we saw that the impressive energy conservation of the $H(\text{div})$ -flux method witnessed in the flat-bottom solitary wave tests does not persist for general topography profiles. In particular, the magnitude of the relative error in energy increases as waves move over very large hills or troughs in the seafloor. Accordingly, alternative time-stepping methods ought be investigated to ensure that the semi-discrete energy conservation property of the $H(\text{div})$ -flux method is maintained at the fully discrete level; of course, we know that the time-stepper is the problem since the semi-discrete version of the $H(\text{div})$ -flux method conserves energy exactly. Since the Green-Naghdi Hamiltonian is cubic rather than quadratic, finding a useful conservative time-stepper may be difficult. Several energy-conserving algorithms for ODEs are discussed in [42], so this may be an ideal source in which to start looking for appropriate time discretization schemes. In the literature on numerical solutions of the GN (or modified versions of the GN), finite-difference time discretizations are often accomplished with strong stability-preserving Runge-Kutta time-steppers (see for example [32, 53]), indicating that such methods may also be worth a try in improving the performance of the $H(\text{div})$ -flux method. Alternatively, one could attempt to preserve energy conservation by doing away with the $H(\text{div})$ -flux method entirely and replacing it with a discretization of the GN that somehow “respects” their variational structure. Some work on variational finite element discretizations for ideal fluids has been initiated recently in [63].

Of course, energy conservation is not the only issue with flow over topography that is worth investigating. Namely, we have seen that the dispersion initiated by wave-topography interaction can become a bit complicated, even in very simple situations such as flow over a Gaussian hill. Now, since dispersion is a very important feature of the GN, we would be foolish not to expect to see dispersion as a very important feature in some of our simulations. However, since we can no longer rely on excellent energy conservation while keeping our mesh reasonably coarse, we can no longer say for sure that all features of the simulation

are physical. This leads us to conclude that an analysis of the dispersion error in the scheme is necessary before more complicated simulations of flow over topography are undertaken, let alone relied upon. I recommend that such an error analysis should follow the approach of Cotter and Shipton [21], which is focused on understanding how specific choices of finite element spaces may lead to spurious oscillations in the depth field. These spurious modes may even occur when finite element spaces are chosen according to the rules of FEEC, so unfortunately the theory from Part 1 does not solve every problem for us. Once the dispersion error in the numerical scheme is better understood, steps may be taken to optimize the method's performance, perhaps by changing the choices of finite element spaces as in [21]. At this point, more complicated numerical tests, such as the flow over a non-smooth shoal discussed in [53, 65], can be performed.

References

- [1] Alexey Androsov, Sven Harig, Annika Fuchs, Antonia Immerz, Natalja Rakowsky, Wolfgang Hiller, and Sergey Danilov, *Tsunami Wave Propagation*, Wave Propagation: Theories and Applications (Yi Zheng, ed.), InTechOpen, 2013, pp. 43–72.
- [2] Martin S. Alnæs, Jan Blechta, Johan Hake, August Johansson, Benjamin Kehlet, Anders Logg, Chris Richardson, Johannes Ring, Marie E. Rognes, and Garth N. Wells, *The FEniCS Project Version 1.5*, Archive of Numerical Software **3** (2015), no. 100.
- [3] Douglas N. Arnold, *Spaces of finite element differential forms*, Analysis and numerics of partial differential equations, Springer INdAM Ser., vol. 4, Springer, Milan, 2013, pp. 117–140.
- [4] Douglas N. Arnold, Richard S. Falk, and Ragnar Winther, *Finite element exterior calculus, homological techniques, and applications*, Acta Numer. **15** (2006), 1–155.
- [5] ———, *Finite element exterior calculus: from Hodge theory to numerical stability*, Bull. Amer. Math. Soc. (N.S.) **47** (2010), no. 2, 281–354.
- [6] Douglas N. Arnold and Anders Logg, *Periodic Table of the Finite Elements*, SIAM News **47** (2014), no. 9.
- [7] Vladimir I. Arnol'd, *Mathematical methods of classical mechanics*, Graduate Texts in Mathematics, vol. 60, Springer-Verlag, New York, [1989]. Translated from the 1974 Russian original by K. Vogtmann and A. Weinstein.
- [8] ———, *Lectures on partial differential equations*, Universitext, Springer-Verlag, Berlin; Publishing House PHASIS, Moscow, 2004. Translated from the second Russian edition by Roger Cooke.
- [9] Satish Balay, Shrirang Abhyankar, Mark F. Adams, Jed Brown, Peter Brune, Kris Buschelman, Lisandro Dalcin, Victor Eijkhout, William D. Gropp, Dinesh Kaushik, Matthew G. Knepley, Dave A. May, Lois Curfman McInnes, Karl Rupp, Patrick Sanan, Barry F. Smith, Stefano Zampini, Hong Zhang, and Hong Zhang, *PETSc Users Manual*, Argonne National Laboratory, 2017.
- [10] Daniele Boffi, Franco Brezzi, and Michel Fortin, *Mixed finite element methods and applications*, Springer Series in Computational Mathematics, vol. 44, Springer, Heidelberg, 2013.

- [11] Haim Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Universitext, Springer, New York, 2011.
- [12] Franco Brezzi, Jim Douglas Jr., Ricardo Durán, and Michel Fortin, *Mixed finite elements for second order elliptic problems in three variables*, Numer. Math. **51** (1987), no. 2, 237–250.
- [13] Franco Brezzi, Jim Douglas Jr., and L. Donatella Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math. **47** (1985), no. 2, 217–235.
- [14] Jochen Brüning and Matthias Lesch, *Hilbert complexes*, J. Funct. Anal. **108** (1992), no. 1, 88–132.
- [15] Roberto A. Camassa and Darryl D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett. **71** (1993), 1661–1664.
- [16] Roberto A. Camassa, Darryl D. Holm, and C. David Levermore, *Long-time effects of bottom topography in shallow water*, Phys. D **98** (1996), no. 2–4, 258–286. Nonlinear phenomena in ocean dynamics (Los Alamos, NM, 1995).
- [17] Jeff Cheeger, *On the Hodge theory of Riemannian pseudomanifolds*, Geometry of the Laplace Operator (Proc. Sympos. Pure Math., Univ. Hawaii, Honolulu, Hawaii, 1979), Proc. Sympos. Pure Math., XXXVI, Amer. Math. Soc., Providence, R.I., 1980, pp. 91–146.
- [18] Snorre H. Christiansen, *Stability of Hodge decompositions in finite element spaces of differential forms in arbitrary dimension*, Numer. Math. **107** (2007), no. 1, 87–106.
- [19] Snorre H. Christiansen and Ragnar Winther, *Smoothed projections in finite element exterior calculus*, Math. Compp. **77** (2008), no. 262, 813–829.
- [20] Philippe G. Ciarlet, *The finite element method for elliptic problems*, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.
- [21] Colin J. Cotter and Jemma Shipton, *Mixed finite elements for numerical weather prediction*, J. Comput. Phys. **231** (2012), no. 21, 7076–7091.

- [22] Colin J. Cotter and John Thuburn, *A finite element exterior calculus framework for the rotating shallow-water equations*. part B, *J. Comput. Phys.* **257** (2014), 1506–1526.
- [23] Richard Courant, *Variational methods for the solution of problems of equilibrium and vibrations*, *Bull. Amer. Math. Soc.* **49** (1943), 1–23.
- [24] Georges de Rham, *Sur l'analysis situs des variétés à n dimensions*, Doctoral Thesis (Université de Paris), 1931.
- [25] Paul Dellar, *Hamiltonian and symmetric hyperbolic structures of shallow water magnetohydrodynamics*, *Phys. Plasmas* **9** (2002), no. 4, 1130–1136.
- [26] ———, *Dispersive shallow water magnetohydrodynamics*, *Phys. Plasmas* **10** (2003), no. 3, 581–590.
- [27] Jean Dieudonné, *A history of algebraic and differential topology. 1900–1960*, Birkhäuser Boston, Inc., Boston, MA, 1989.
- [28] Manfredo P. do Carmo, *Differential forms and applications*, Universitext, Springer-Verlag, Berlin, 1994. Translated from the 1971 Portuguese original.
- [29] Vít Dolejší and Miloslav Feistauer, *Discontinuous Galerkin method: analysis and applications to compressible flow*, Springer Series in Computational Mathematics, vol. 48, Springer, 2015.
- [30] Haiyun Dong and Maojun Li, *A reconstructed central discontinuous Galerkin-finite element method for the fully nonlinear weakly dispersive Green-Naghdi model*, *Appl. Numer. Math.* **110** (2016), 110–127.
- [31] David S. Dummit and Richard M. Foote, *Abstract algebra*, 3rd ed., John Wiley & Sons, Inc., Hoboken, NJ, 2004.
- [32] Antonio Duran and Fabien Marche, *A discontinuous Galerkin method for a new class of Green-Naghdi equations on simplicial unstructured meshes*, *Appl. Math. Model.* **45** (2017), 840–864.
- [33] Howard C. Elman, David J. Silvester, and Andrew J. Wathen, *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, 2nd ed., Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2014.

- [34] FFmpeg development team, *FFmpeg*, <https://www.ffmpeg.org/>.
- [35] Harley Flanders, *Differential forms with applications to the physical sciences*, Academic Press, New York-London, 1963.
- [36] Robert Ghrist, *Elementary Applied Topology*, Createspace, 2014.
- [37] ———, *Three examples of applied and computational homology*, *Nieuw Arch. Wiskd.* (5) **9** (2008), no. 2, 122–125, available at <https://www.math.upenn.edu/~ghrist/preprints/nieuwarchief.pdf>.
- [38] Andrew Gillette and Tyler Kloefkorn, *Trimmed serendipity finite element differential forms*, *Mathematics of Computation* **to appear** (2018), available at <https://arxiv.org/abs/1607.00571>.
- [39] Sergei K. Godunov, *A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics*, *Mat. Sb. (N.S.)* **47 (89)** (1959), 271–306 (Russian).
- [40] Albert E. Green and Paul M. Naghdi, *A derivation of equations for wave propagation in water of variable depth*, *J.Fluid Mech.* **78** (1976), no. 2, 237–246.
- [41] Walter Greiner and Joachim Reinhardt, *Field quantization*, Springer-Verlag, Berlin, 1996. Translated from the German; With a foreword by D. A. Bromley.
- [42] Ernst Hairer, Christian Lubich, and Gerhard Wanner, *Geometric numerical integration*, 2nd ed., Springer Series in Computational Mathematics, vol. 31, Springer-Verlag, Berlin, 2006.
- [43] Darryl D. Holm, *Ideal shallow water dynamics in a rotating frame*, MPE CDT (2016). Lectures delivered in Autumn 2016.
- [44] Darryl D. Holm, Jerrold E. Marsden, and Tudor S. Ratiu, *The Euler-Poincaré equations and semidirect products with applications to continuum theories*, *Adv. Math.* **137** (1998), no. 1, 1–81.
- [45] J. D. Hunter, *Matplotlib: A 2D graphics environment*, *Computing In Science & Engineering* **9** (2007), no. 3, 90–95.

- [46] Arieh Iserles, *A first course in the numerical analysis of differential equations*, 1st ed., Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 1996.
- [47] ———, *A first course in the numerical analysis of differential equations*, 2nd ed., Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 2009.
- [48] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter Development Team, *Jupyter Notebooks- a publishing format for reproducible computational workflows*, Positioning and Power in Academic Publishing: Players, Agents and Agendas (F. Loizides and B. Schmidt, eds.), IOS Press, 2016, pp. 87–90.
- [49] Allen C. Kuo and Lorenzo M. Polvani, *Wave-vortex interaction in rotating shallow water. I. One space dimension*, *J. Fluid Mech.* **394** (1999), 1–27.
- [50] Olivier Le Métayer, Sergey Gavriluk, and Sarah Hank, *A numerical scheme for the Green-Naghdi model*, *J. Comput. Phys.* **229** (2010), no. 6, 2034–2045.
- [51] John M. Lee, *Introduction to smooth manifolds*, 2nd ed., Graduate Texts in Mathematics, vol. 218, Springer, New York, 2013.
- [52] Randall J. LeVeque, *Finite volume methods for hyperbolic problems*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 2002.
- [53] Maojun Li, Philippe Guyenne, Fengyan Li, and Liwei Xu, *High order well-balanced CDG-FE methods for shallow water waves by a Green-Naghdi model*, *J. Comput. Phys.* **257 Part A** (2014), 169–192.
- [54] Anders Logg, Garth N. Wells, and Kent-Andre Mardal (eds.), *Automated Solutions of Differential Equations by the Finite Element Method: The FEniCS Book*, Lecture Notes in Computational Science and Engineering, Springer, New York, 2012.
- [55] Saunders Mac Lane, *Topology becomes algebraic with Vietoris and Noether*, *J. Pure Appl. Algebra* **39** (1986), no. 3, 305–307.

- [56] Ricardo Mañé, *Ergodic theory and differentiable dynamics*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], vol. 8, Springer-Verlag, Berlin, 1987. Translated from the Portuguese by Silvio Levy.
- [57] James Clerk Maxwell, *Remarks on the Mathematical Classification of Physical Quantities*, Proceedings of the London Mathematical Society **s1-3** (1869), no. 1, 224–233.
- [58] John Miles and Rick Salmon, *Weakly dispersive nonlinear gravity waves*, J. Fluid Mech. **157** (1985).
- [59] Dimitrios Mitsotakis, Boaz Ilan, and Denys Dutykh, *On the Galerkin/finite-element method for the Serre equations*, J. Sci. Comput. **61** (2014), no. 1, 166–195.
- [60] Michael Monastyrsky, *Topology of gauge fields and condensed matter*, Plenum Press, New York; Mir Publishers, Moscow, 1993. Translated from the Russian by Oleg Efimov.
- [61] James R. Munkres, *Elements of algebraic topology*, Addison-Wesley Publishing Company, Menlo Park, CA, 1984.
- [62] Mikio Nakahara, *Geometry, topology and physics*, 2nd ed., Graduate Student Series in Physics, Institute of Physics, Bristol, 2003. MR2001829
- [63] Andrea Natale and Colin J Cotter, *A variational $\mathbf{H}(\text{div})$ finite-element discretization approach for perfect incompressible fluids* (2017), preprint, available at <https://arxiv.org/abs/1606.06199v4>.
- [64] Jean–Claude Nédélec, *Mixed finite elements in \mathbb{R}^3* , Numer. Math. **35** (1980), no. 3, 315–341.
- [65] Nishant Panda, Clint Dawson, Yao Zhang, Andrew B. Kennedy, Joannes J. Westerink, and Aaron S. Donahue, *Discontinuous Galerkin methods for solving Boussinesq–Green–Naghdi equations in resolving non-linear and dispersive surface water waves*, J. Comput. Phys. **273** (2014), 572–588.
- [66] J. D. Pearce and J. Gavin Esler, *A pseudo-spectral algorithm and test cases for the numerical solution of the two-dimensional rotating Green–Naghdi shallow water equations*, J. Comput. Phys. **229** (2010), no. 20, 7594–7608.

- [67] Joseph Pedlosky, *Geophysical Fluid Dynamics*, 2nd ed., Springer-Verlag, New York, 1987.
- [68] Rainer Picard, *An elementary proof for a compact imbedding result in generalized electromagnetic theory*, Math. Z. **187** (1984), no. 2, 151–164.
- [69] Henri Poincaré, *Papers on topology: Analysis situs and its five supplements*, History of Mathematics, vol. 37, American Mathematical Society, Providence, RI; London Mathematical Society, London, 2010. Translated and with an introduction by John Stillwell.
- [70] Lorenzo M. Polvani, James C. McWilliams, Michael A. Spall, and R. Ford, *The coherent structures of shallow-water turbulence: Deformation-radius effects, cyclone/anticyclone asymmetry and gravity-wave generation*, Chaos **4** (1994), no. 2, 177–186.
- [71] Francis J. Poulin and Glenn R. Flierl, *The nonlinear evolution of barotropically unstable jets*, J. Phys. Oceanogr. **33** (2003), no. 10, 2173–2192.
- [72] Florian Rathgeber, David A. Ham, Lawrence Mitchell, Michael Lange, Fabio Luporini, Andrew T. T. McRae, Gheorghe-Teodor Bercea, Graham R. Markall, and Paul H. J. Kelly, *Firedrake: automating the finite element method by composing abstractions*, ACM Trans. Math. Software **43** (2017), no. 3, 24:1–24:27.
- [73] Pierre–Arnaud Raviart and Jean–Marie Thomas, *A mixed finite element method for 2nd order elliptic problems*, Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975), Springer, Berlin, 1977, pp. 292–315. Lecture Notes in Math., Vol. 606.
- [74] Michael Reed and Barry Simon, *Methods of modern mathematical physics. I. Functional analysis*, Academic Press, 1972.
- [75] Jonathan Rosenberg, *Applications of analysis on Lipschitz manifolds*, Miniconferences on harmonic analysis and operator algebras (Canberra, 1987).
- [76] Rick Salmon, *Lectures on geophysical fluid dynamics*, Oxford University Press, New York, 1998.
- [77] Günter Schwarz, *Hodge decomposition—a method for solving boundary value problems*, Lecture Notes in Mathematics, vol. 1607, Springer–Verlag, Berlin, 1995.

- [78] François Serre, *Contribution à l'étude des écoulements permanents et variables dans les canaux*, La Houille Blanche **6** (1953), 830–872 (French).
- [79] Theodore G. Shepherd, *Symmetries, conservation laws, and Hamiltonian structure in geophysical fluid dynamics*, Adv. Geophys. **32** (1990).
- [80] Jemma Shipton and Colin J. Cotter, *Higher-order compatible finite element schemes for the nonlinear rotating shallow water equations on the sphere* (2017), preprint, available at <https://arxiv.org/abs/1707.00855>.
- [81] Harold Simmons, *An introduction to category theory*, Cambridge University Press, Cambridge, 2011.
- [82] Gilbert Strang and George J. Fix, *An analysis of the finite element method*, Prentice-Hall Series in Automatic Computation, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1973.
- [83] C. H. Su and C. S. Gardner, *Korteweg-de Vries Equation and Generalizations. III. Derivation of the Korteweg-de Vries Equation and Burgers Equation*, J. Math. Phys. **10** (1969), no. 3, 536–539.
- [84] Gordon E. Swaters, *Introduction to Hamiltonian fluid dynamics and stability theory*, Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics, vol. 102, Chapman & Hall/CRC, Boca Raton, FL, 2000.
- [85] Nicolae Teleman, *The index of signature operators on Lipschitz manifolds*, Inst. Hautes Études Sci. Publ. Math. **58** (1983), 39–78 (1984).
- [86] William Thomson, *On Gravitational Oscillations of Rotating Water*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **10** (1880), no. 60, 109–116.
- [87] Kristen M. Thyng, Chad A. Greene, Robert D. Hetland, Heather M. Zimmerle, and Steven F. DiMarco, *True Colors of Oceanography: Guidelines for Effective and Accurate Colormap Selection*, Oceanography **29** (2016), 9–13.
- [88] Christopher L. Tsee, *Computation and visualization of geometric partial differential equations*, Doctoral Thesis (University of California San Diego), 2015.

- [89] Hassler Whitney, *Geometric integration theory*, Princeton University Press, Princeton, N. J., 1957.
- [90] Joseph Wloka, *Partial differential equations*, Cambridge University Press, Cambridge, 1987. Translated from the German by C. B. Thomas and M. J. Thomas.
- [91] Kōsaku Yosida, *Functional analysis*, Classics in Mathematics, Springer-Verlag, Berlin, 1995. Reprint of the sixth (1980) edition.
- [92] Afra J. Zomorodian, *Topology for computing*, Cambridge Monographs on Applied and Computational Mathematics, vol. 16, Cambridge University Press, Cambridge, 2005.

Appendix A

Supplementary Firedrake Demos

In the body of the thesis, I make reference to two computer demos that the reader can use to supplement their understanding of the content. In this appendix, I briefly summarize what the demos are, where they are mentioned in the thesis, how they may be downloaded, and what software is needed to run them. The two demos are

- 1) `MixedSL.ipynb`: a routine showing how different choices of finite element spaces strongly affect the quality of numerical solutions to a Sturm–Liouville PDE, mentioned briefly in Section 9.5;
- 2) `GN_Topography.ipynb`: a routine showing how the GN describe the influence of bottom topography on solitary waves, discussed in Section 12.4 (all of the numerical results in this section were obtained using this demo, though the demo also includes the option to use topography profiles not discussed in the thesis).

Both of these demos are Jupyter notebooks [48] written in Firedrake [72]. They are mostly self-explanatory, so I do not provide a detailed description of how to use them here. To obtain the demos, simply visit git.uwaterloo.ca/a24morga/Thesis_Demos and download the required files.

If you have University of Waterloo online access credentials, you can run both demos simply by using the University’s online Jupyter server at uwaterloo.syzygy.ca. If you are not affiliated with the University of Waterloo, you must download and install some software before you can access the demos. Note that Firedrake is not tested on Windows, so if you are using Windows your best option is probably to install Ubuntu before starting. Additionally, if you are using a Mac, I recommend that you make sure Homebrew is up-to-date, and that you have also installed both Xcode and Command Line Developer Tools (otherwise, you may have to stop the Firedrake installation and obtain these packages before re-starting). Here are the steps you need to take to download the required software:

- 1) install Firedrake from firedrakeproject.org/download.html;
- 2) activate the Firedrake virtual environment in the terminal, then install Jupyter from jupyter.org/install.html;
- 3) if you want to use the topography demo, you also need to install both the cmocean colormaps (you can do this by typing `pip install cmocean` into the terminal) and FFmpeg [34] (see ffmpeg.org/download.html

for download instructions);

4) type `jupyter notebook` into the terminal to start;

If you encounter bugs in either demo, please contact me at adam.morgan@uwaterloo.ca.