

# ON INVERSE PROBLEMS AND FINITE ELEMENTS IN SHAPE ANALYSIS

Andreas Bock  
Department of Mathematics  
Imperial College London

SUBMITTED FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY OF IMPERIAL COLLEGE LONDON  
AND THE DIPLOMA OF IMPERIAL COLLEGE

## ABSTRACT

Shapes are nonlinear, multifarious objects and it is a nontrivial task to attach meaning to statements about their comparison. A central problem in the field of *shape analysis* is therefore to make sense of the notion of a metric on an abstract shape space. Through the lens of infinite-dimensional Riemannian geometry one can endow such spaces with different metrics stemming from nonrigid group actions to compute so-called geodesic matchings between shapes that share diffeomorphic or homeomorphic orbits. Owing to shape analysis being a relatively new field in the history of mathematics, several open questions require answering to reconcile theoretical development with numerical approximation. This project sets out to unite the framework of diffeomorphometry with linear functional analysis amenable to finite element methods. Shapes in this project will be images, curves or even measures, and we derive and compute convergent numerical approximations to the geodesic equations related to the motion on the aforementioned groups.

I hereby certify that the material in this thesis is my own and that the work of others has been referenced or otherwise acknowledged.

Andreas Bock

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

## ACKNOWLEDGEMENTS

I warmly thank Professor Colin Cotter for his supervision, funding and endless patience. I also thank Alexis Arnaudon for his constructive collaboration, and the Firedrake team for their help. In particular Lawrence Mitchell and Rob Kirby. I would like to express my deepest gratitude to my friends and family for their support and dedicate this work to my brother and my parents.

## LIST OF TABLES

|     |                                                                 |     |
|-----|-----------------------------------------------------------------|-----|
| 4.1 | Manufactured solutions to problem 4 where $d = 1$ . . . . .     | 99  |
| 4.2 | Convergence rates for example 0 from table 4.1. . . . .         | 100 |
| 4.3 | Convergence rates for example 1 from table 4.1. . . . .         | 101 |
| 4.4 | Convergence rates for example 2 from table 4.1. . . . .         | 102 |
| 4.5 | Convergence rates for example 3 from table 4.1. . . . .         | 103 |
| A.1 | Manufactured solutions to problem 4 where $d = 2$ . . . . .     | 171 |
| A.2 | Convergence rates for the examples in table A.1. . . . .        | 171 |
| A.3 | Convergence rates for example 1 from table A.1. . . . .         | 172 |
| A.4 | Convergence rates for example 2 from table A.1. . . . .         | 172 |
| B.1 | Branch and revision information for the Firedrake components. . | 173 |
| B.2 | Convergence results for problem (B.1) . . . . .                 | 174 |
| B.3 | Convergence results for problem (B.2) . . . . .                 | 174 |
| B.4 | Convergence results for problem (B.3) . . . . .                 | 174 |
| B.5 | Convergence results for problem (B.4) . . . . .                 | 175 |

## LIST OF FIGURES

|     |                                                                                                               |     |
|-----|---------------------------------------------------------------------------------------------------------------|-----|
| 1.1 | Landmark matching. . . . .                                                                                    | 11  |
| 1.2 | Matching closed planar curves. . . . .                                                                        | 11  |
| 1.3 | Matchings between different kinds of fishes. . . . .                                                          | 12  |
| 1.4 | A template and target where a non-diffeomorphic match is required. .                                          | 12  |
| 2.1 | Schematic of diffeomorphic shape matching. . . . .                                                            | 18  |
| 2.2 | Schematic of metamorphosis. . . . .                                                                           | 20  |
| 3.1 | Smoothing effect of $(\mathbf{id} - \alpha\Delta)^{-1}\delta_{x_0}$ . . . . .                                 | 45  |
| 3.2 | Plot of the discrete curve defined by $q_{0h} \circ S^1$ . . . . .                                            | 49  |
| 3.3 | Evolution of the circle from figure 3.2 obtained by forward<br>integration of the system (3.11). . . . .      | 83  |
| 3.4 | Matching of curves for different parameterisations of the template<br>(bottom) with the target (top). . . . . | 86  |
| 4.1 | Density matches computed from the manufactured solutions to<br>problem 4 in table 4.1. . . . .                | 100 |
| 4.2 | Convergence rates for the examples in table 4.1. . . . .                                                      | 101 |

|     |                                                                                                                            |     |
|-----|----------------------------------------------------------------------------------------------------------------------------|-----|
| 4.3 | Results for the outer problem for various template and target pairs in one dimension using $\sigma^{-2} = 1$ . . . . .     | 123 |
| 4.4 | Results for the outer problem for various template and target pairs in one dimension using $\sigma^{-2} = 10^5$ . . . . .  | 123 |
| 4.5 | Results for the outer problem for various template and target pairs in two dimensions using $\sigma^{-2} = 1$ . . . . .    | 124 |
| 4.6 | Results for the outer problem for various template and target pairs in two dimensions using $\sigma^{-2} = 10^5$ . . . . . | 125 |
| 5.1 | Landmark matching in the plane with diffeomorphic and metamorphic models . . . . .                                         | 134 |
| 5.2 | Landmark matching in the plane with selective metamorphosis. . . . .                                                       | 136 |
| 5.3 | Heatmap of the functional (5.11) as a function of the position of a single centroid. . . . .                               | 140 |
| 5.4 | pCN results for the configuration in the left panel of figure 5.1 . . . . .                                                | 142 |
| 5.5 | pCN results for the configuration in the right panel of figure 5.1 . . . . .                                               | 143 |
| A.1 | Convergence rates for the examples in table A.1. . . . .                                                                   | 172 |

# TABLE OF CONTENTS

|       |                                                |            |
|-------|------------------------------------------------|------------|
| 1     | INTRODUCTION                                   | <b>9</b>   |
| 1.1   | ORGANISATION . . . . .                         | 13         |
| 2     | BACKGROUND                                     | <b>15</b>  |
| 2.1   | DIFFEOMORPHOMETRY . . . . .                    | 15         |
| 2.1.1 | GROWTH . . . . .                               | 19         |
| 2.1.2 | INVERSE PROBLEMS . . . . .                     | 21         |
| 2.2   | VARIATIONAL PROBLEMS . . . . .                 | 23         |
| 2.2.1 | ELEMENTARY FUNCTIONAL ANALYSIS . . . . .       | 23         |
| 2.2.2 | FINITE ELEMENT METHODS . . . . .               | 27         |
| 3     | CURVE MATCHING                                 | <b>40</b>  |
| 3.1   | GEODESIC EQUATIONS . . . . .                   | 41         |
| 3.2   | BUILDING AN OUTER METRIC . . . . .             | 42         |
| 3.3   | NONCONFORMING DISCRETISATION . . . . .         | 47         |
| 3.3.1 | ALTERNATIVE APPROACHES & PERSPECTIVE . . . . . | 78         |
| 3.3.2 | COMPUTING THE FORWARD MAP . . . . .            | 80         |
| 3.4   | INVERSE PROBLEM . . . . .                      | 84         |
| 3.5   | SUMMARY . . . . .                              | 85         |
| 4     | SPACE-TIME METAMORPHOSIS                       | <b>87</b>  |
| 4.1   | INNER PROBLEM . . . . .                        | 89         |
| 4.1.1 | ANALYTICAL SOLUTION . . . . .                  | 91         |
| 4.1.2 | PRIMAL FORMULATION . . . . .                   | 93         |
| 4.1.3 | DUAL FORMULATION . . . . .                     | 102        |
| 4.2   | OUTER PROBLEM . . . . .                        | 119        |
| 4.2.1 | GRADIENT DESCENT . . . . .                     | 119        |
| 4.3   | SUMMARY . . . . .                              | 125        |
| 5     | SELECTIVE METAMORPHOSIS                        | <b>128</b> |
| 5.1   | LANDMARKS . . . . .                            | 128        |
| 5.1.1 | PREVIOUS WORK . . . . .                        | 130        |
| 5.1.2 | SELECTIVE LANDMARK METAMORPHOSIS . . . . .     | 135        |
| 5.2   | BAYESIAN FRAMEWORK . . . . .                   | 138        |
| 5.3   | IMAGES . . . . .                               | 143        |
| 5.3.1 | SINGULAR SOLUTIONS . . . . .                   | 146        |
| 5.4   | SUMMARY . . . . .                              | 148        |
| 6     | CLOSURE                                        | <b>150</b> |
| 6.1   | SUMMARY . . . . .                              | 150        |
| 6.2   | OPEN PROBLEMS . . . . .                        | 151        |

|                                                                      |     |
|----------------------------------------------------------------------|-----|
| REFERENCES                                                           | 154 |
| A INNER PROBLEM CONVERGENCE RESULTS FOR $d = 2$                      | 171 |
| B CONVERGENCE RESULTS FOR THE ROBUST XU-WU ELEMENT                   | 173 |
| C TOOLS AND TECHNIQUES FOR ANALYSIS OF NONCONFORMING<br>ELEMENTS     | 176 |
| C.1 PROOF SKETCHES OF (3.22a) AND (3.22b) . . . . .                  | 176 |
| C.2 ERROR ANALYSIS FOR NONCONFORMING $\mathcal{P}^1$ FINITE ELEMENTS | 177 |
| D SCALING ARGUMENTS                                                  | 180 |
| E LIMITING PROCEDURE                                                 | 182 |

# CHAPTER 1

---

## INTRODUCTION

---

Shape analysis is the mathematical study of shapes. These are abstract objects such as closed curves and surfaces, functions or even densities representing data such as outlines or images of morphology scale anatomy. The field encompasses the theoretical description of shapes, their evolution and comparison as well as the development of algorithms and engineering of computer programmes for their manipulation. Differential geometry lies at the heart of this subject and is the source of many equations of interest in shape analysis. This thesis focuses on the comparison of shapes i.e. how to continuously deform, or evolve, a *template* shape into a desired *target* shape whereby obtaining a natural notion of distance between them measured as the energy required for such a deformation. This is called a *matching problem* which we shall see amounts to metrising a *shape space*. The aim of this thesis is to utilise the tools of linear functional analysis to derive and place the equations of motion for shapes evolution in the frame of Hilbert spaces to leverage finite element theory in order to describe well-posed convergent and numerically tractable methods. This chapter describes the historical evolution of our shape analysis to motivate our place in the field.

The Scottish scholar D’Arcy Thompson pioneered the mathematical study of biological shapes in his book *On growth and form* [216] in 1942. See also [12] which provides a very brief but comprehensive historical introduction. D’Arcy was interested in morphogenesis: the creation of structure and growth in biology. He conjectured that the laws of physics that govern the orbits of planets must be the same that dictate mechanical biological movement and that arterial motion, for instance, must be of the same character as any inorganic motion. He argues that many biological phenomena are explained by careful physical consideration and draws intriguing parallels between hexagonal symmetries of the plates of a tortoise’s shell, cracks in drying mud and cellular patterns appearing in varnish [216, Chapter 7]. The pioneering work of D’Arcy did immense service to the understanding of mathematical biology, and therefore shape analysis, since many applications of shape analysis are derived from morphology scale biology. As a result of this the field also sometimes carries the moniker *computational anatomy*. To study biological shapes such as silhouettes

of fish or images of Macaque monkey brains with mathematics these objects require a mathematical representation. In an attempt to provide a general theoretical framework for the study and representation of abstract shapes and images, Grenander and his collaborators developed *pattern theory* at Brown University [110] in 1993. This theory provides the fundamental vocabulary for describing knowledge and patterns, striving to condense the variability and structure of information about patterns into a mathematical theory. Pattern theory is an incredibly powerful and equally abstract setting so we shall not endeavour to describe it in full and only bring in the key concepts on which we rely while acknowledging the origin of our field of study. What is essential is that interpreting certain matching problems appears completely natural from the vantage of pattern theory. We refer to [112] for a more contemporary tutorial exposition linking pattern theory and computational anatomy. Fields Medalist David Mumford's monograph [179] also provides interesting insight into the motivation behind structuring and classifying knowledge by way of pattern theory. In particular, pattern theory lead to the development of the *group action model* spear-headed by Grenander and his team and has since generated an incredibly rich literature. In this model, so-called *deformable templates* (i.e. shapes) are acted upon by different elements of a group, so that finding a match between two images is equivalent to finding the right action in some group of transformations so that the application of the group translates between the two images. Focusing on modelling the *group* abstracts away the representation of the underlying shape, or deformable template, on which it acts. As we shall see shortly, a particular strength of the group action framework is its ability to handle the nonlinear nature of shapes. The main group with which we are concerned is the group of diffeomorphisms: smooth invertible mappings. These are excellent candidates for matching problems as they naturally give rise to smooth transformations between different shapes. We now present three examples of different shape objects to make the notions above concrete. The goal is to use the vocabulary above to articulate the main problems we aim to tackle.

**Landmarks:** a canonical example of a problem in this field is that of *landmark matching* [141]. Landmarks are measures, or positions in space, typically representing points of interest of some underlying higher-dimensional shape space as exemplified in figure 1.1. In practice, such landmarks could represent the part of the brain called the *Corpus Callosum* [188, 156]. Here, the template (left) and target (right) clearly outline some a priori known curve or image, and the goal is to find a group acting on the landmarks so that each point  $x$  is mapped to  $x'$ . A linear interpolation between the pairs  $(x, x')$  is indeed one such candidate for the example given. However, if we were to swap the location of  $a'$  and  $d'$  in this example and compute a linear match between the landmarks, the paths of the landmarks would intersect. Such a result would appear unnatural or counter-intuitive if, for instance, the landmarks were to represent some underlying morphology scale biology where matter cannot necessarily intersect or overlap. This illustrates the importance of computing

*continuous paths.*

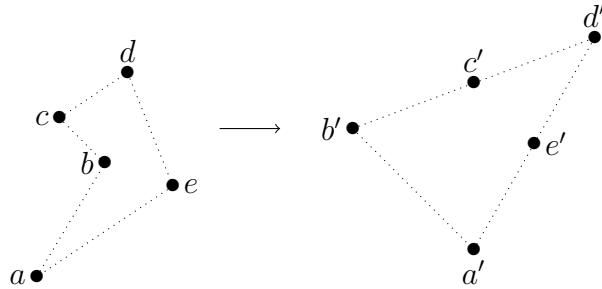


FIGURE 1.1: Matching between two sets of landmarks with knowledge of pairwise correspondence i.e  $x \mapsto x'$ . The dashed lines are only there to show that the landmarks outline some underlying geometry.

**Closed curves:** The landmarks in figure 1.1 were sampled from the infinite-dimensional objects depicted in figure 1.2. One representation of these closed curves are functions of  $S^1$  into  $\mathbb{R}^2$ . We may easily imagine the template curve

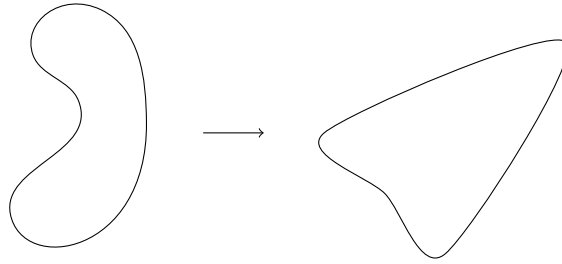


FIGURE 1.2: Matching closed planar curves.

on the left moving to the right to align with the arrowhead target, and our aim is to describe this mathematically. We are interested in studying the group of diffeomorphic transformations allowing us to describe this matching mathematically, and to derive efficient algorithms that act on these heavily nonlinear objects. We are especially interested in diffeomorphic group actions that preserve the topology of this curve: we allow for no self-intersections or breaking of the curve. From an applications point of view numerical methods for the evolution of discrete closed curves are obviously of great interest. Closed curves can represent outlines of biology in the form of cells or traces of cardiac images [27].

**Images:** these are central objects in shape analysis and the ability to compute matches between arbitrary images have a myriad of applications in everything from medical imaging [167, 30, 31, 228, 84] to music videos [32, 15]. Chapter 17 of *On growth and form* describes a theory of transformations whereby examples of diffeomorphic deformation of the ambient space transforms between different types of fishes, see figure 1.3. Numerical matching of images

was, to the best of the author's knowledge, first carried out in this pattern theoretic setting by Grenander and Miller in [113]. Images are also incredibly interesting mathematically because, depending on the application, one does not always want a diffeomorphic match. Indeed, there exists no diffeomorphism matching the two images in figure 1.4. We therefore also need to make sense of *homeomorphic* matches. This is useful in medical imaging where images of the human anatomy is available in the form of magnetic resonance imaging which produces a sequence of two-dimensional images representing various segments of the brain. These images may be combined pair-wise in order to reconstruct and subsequently study a three-dimensional model [94]. A homeomorphic match could be particularly useful to model *tumour evolution* which from a mathematical point of view is similar to what is depicted in figure 1.4.

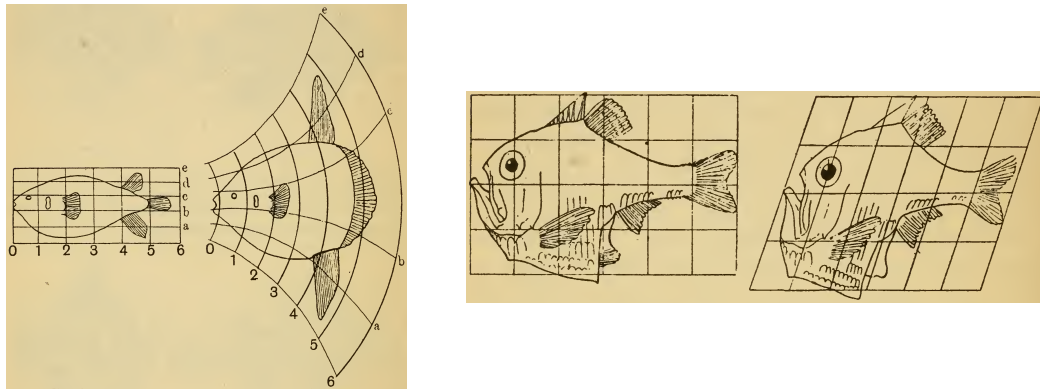


FIGURE 1.3: Left: Matching a *Diodon* to an *Orthogoriscus Mola* by a diffeomorphism. Right: Matching an *Argyropelecus Olfersi* to a *Sternoptyx diaphana* by a simple shear. Taken from [216, Fig. 517, 518, 525, 526]. Such drawings are in spirit of Albrecht Dürer, a renaissance thinker who influenced D'Arcy Thompson.

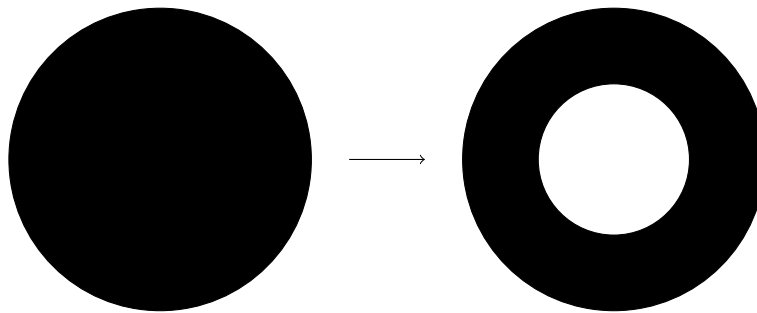


FIGURE 1.4: A template and target where a non-diffeomorphic match is required in order to create a hole in the circle.

The group action model isolates the study of shape matching by instead studying the group acting upon them. We mentioned before that *comparison* of different shapes means endowing some metric structure on the shape space in question. Metrics provide a quantitative measure of how close two objects

are stemming from the energy of group action that transforms between the objects. As we shall see, the groups we will be studying also automatically provide a meaningful way of interpolation (in some sense) between shapes so as to provide a concrete visual transformation between shapes. It will be useful to think of a template shape as the match at time  $t = 0$ , and the target as the match at time  $t = 1$ . One of our central aims is exactly to make sense of the match at time  $0 < t < 1$  with an intuitive interpolation between the template and the target. The power of the group action model is that by leveraging the vocabulary of differential geometry we can make meaningful statements about the *shortest path* between two shapes measured in the metric induced by the deformation. This sets the stage for this thesis: we wish to explore the use of finite element methods (FEM) [89] to efficiently compute *deformations* that transform templates of various data types into prescribed targets.

The main tools with which we want to solve the problems above are primarily differential geometry, functional analysis and FEM. Shape analysis is naturally intrinsically related to differential geometry; we can attempt to understand nonlinear objects such as shapes by attaching differential structure to them and inspecting linearised, infinitesimal changes to them. Riemannian geometry allows us to do exactly this. Then, via integration, can we make sense of the structure of the nonlinear objects with which we are concerned. Linearisation also leads to variational principles (and variational equations), and tools from functional analysis are especially suited to deal with these linearised structures and make statements about well-posedness. The study of numerical approximation of these systems is also convenient through the lens of functional analysis via FEM. Our objective is therefore to solve the different problems we have described above by means of efficient and convergent numerical computation.

## 1.1 ORGANISATION

This rest of this thesis is comprised of five chapters with the following content:

- Chapter 2 provides the background material for this thesis. We provide an in-depth account of matching using diffeomorphisms in the group action model. This section also explains how matching problems in shape analysis can be viewed as inverse problems where the objective is to infer the parameter (i.e. group action) that generates smooth deformation between shapes. This chapter concludes with the prerequisite technical tools of variational and functional analysis needed to treat the systems we aim to derive in our later content chapters, as well as a selection of results from finite element theory.
- Chapter 3 describes a novel method for planar curve matching such as those depicted in figure 1.2. We successfully discretise the Euler-Lagrange equations associated with this problem using nonconforming finite element to generate a diffeomorphic curve evolution. This chapter

provides extensive novel theoretical results for general nonconforming finite element methods with a kind of measure-valued source term and shows convergence under relatively mild assumptions.

- In Chapter 4 we examine the Euler-Lagrange equations associated with non-diffeomorphic image matching such as that required in figure 1.4. We treat images as functions and present a Hilbert space framework determining the optimal trajectory between a template and target. We solve the problem analytically and show how interpreting the problem in least-squared variational setting allows us to obtain several theoretical results. We then derive a conforming finite element method and show numerical experiments demonstrating convergence under certain conditions. We also describe the theoretical foundations of a discontinuous Petrov-Galerkin (DPG) method allowing for fully discontinuous image matching.
- Chapter 5 explores landmark matching and develops a framework in which we can interpolate between diffeomorphic and non-diffeomorphic matching in a meaningful way. This leads to qualitative models that shed new light diagnostic methods for growth models within the traditional frameworks.
- Chapter 6 contains a discussion of our work and an outlook on future research.

# CHAPTER 2

---

## BACKGROUND

---

### 2.1 DIFFEOMORPHOMETRY

This section covers diffeomorphometry, the study of distances between shapes via actions of a diffeomorphism group. The contemporary textbook by Younes [234] describes or references most of what we present here. We begin with a general and abstract explanation of the group action model.

$\mathcal{I}$  shall denote an abstract *shape space* containing the shapes we wish to study such as measures, images, functions or curves. For the purpose of this exposition it is not yet necessary to be more precise. We wish to match two members  $i_0$  and  $i_1$  in  $\mathcal{I}$  with a diffeomorphism  $\varphi$  in the group  $\text{Diff}(\mathcal{I})$  of diffeomorphism of the shape space  $\mathcal{I}$  so that  $i_1$  is in some sense close to  $\varphi.i_0$  (where  $\cdot$  is the group action). We define the *orbit*  $\mathcal{I}_i$  at  $i \in \mathcal{I}$  of the diffeomorphism group  $\text{Diff}(\mathcal{I})$  on  $\mathcal{I}$  by as:

$$\mathcal{I}_i = \{\varphi.i \mid \varphi \in \text{Diff}(\mathcal{I})\}.$$

The strategy is therefore to approximate a *target shape*  $i_1$  by an element in the diffeomorphic orbit of a *template*  $i_0$ . At this stage we are not particularly interested in the definition of the group action and concentrate on the group elements in  $\text{Diff}(\mathcal{I})$  themselves. This is precisely the power of the Grenander's group action model mentioned in the introduction. The goal here is to focus the modelling effort on the map  $\varphi$  to endow the space of images with a metric  $d_{\mathcal{I}} : \mathcal{I} \times \mathcal{I} \rightarrow \mathbb{R}$ . Abstractly we can formulate this distance using the notation seen thus far:

$$d_{\mathcal{I}}(i_0, i_1) = \inf_{\varphi \in \text{Diff}(\mathcal{I})} \{d_{\text{Diff}(\mathcal{I})}(\mathbf{id}, \varphi) \mid i_0 = \varphi.i_1\}, \quad (2.1)$$

where  $\mathbf{id}$  is the identity operator and  $d_{\text{Diff}(\mathcal{I})}(\cdot, \cdot)$  is a metric on the  $\text{Diff}(\mathcal{I})$ . This definition should sit well with our intuition, since the distance between a pair of images is determined by the least amount of deformation required in order to match one shape with another under the group action. Our shape spaces will therefore inherit a notion of a distance via a Riemannian metric  $d_{\text{Diff}(\mathcal{I})}(\cdot, \cdot)$  endowed on  $\text{Diff}(\mathcal{I})$ . We highly recommend the following two papers [24, 48] to readers wanting an introduction to shape spaces and the diffeomorphism group

and Riemannian geometry for shape analysis. In order for symmetry of  $d_{\mathcal{I}}$  we request that the metric  $d_{\text{Diff}(\mathcal{I})}$  is invariant in the following sense:

$$d_{\text{Diff}(\mathcal{I})}(h.h_1, h.h_2) = d_{\text{Diff}(\mathcal{I})}(h_1, h_2), \quad \text{for } h \in \text{Diff}(\mathcal{I}).$$

This should also make sense intuitively; should we want to measure the difference between two diffeomorphisms  $h_1$  and  $h_2$  then a reparameterisation by a diffeomorphism  $h$  does not change the distance between them. Using this invariance we can see that e.g.  $d_{\text{Diff}(\mathcal{I})}(a^{-1}, \mathbf{id}) = d_{\text{Diff}(\mathcal{I})}(a, \mathbf{id})$  and we get the required symmetry<sup>1</sup>. This group is extremely useful and plays a central role in due to its Riemannian structure. Arnold's seminal work on perfect fluids [11] established a relationship between a group of diffeomorphisms, now denoted  $\text{Diff}_{\mathfrak{X}}(\mathcal{I})$ , and the space of vector fields  $\mathfrak{X}$ . Indeed, for  $u_t \in \mathfrak{X}$ ,  $t \in [0, 1]$  this relationship is described via the following ordinary differential equation (ODE) over the time interval  $[0, 1]$  where we use the notation  $\dot{g} = \frac{\partial g}{\partial t}$ :

$$\dot{\varphi}_t = u_t \circ \varphi_t, \quad (2.2a)$$

$$\varphi_0 = \mathbf{id}, \quad (2.2b)$$

posed in Lagrangian coordinates. This ODE describes a curve on  $\text{Diff}_{\mathfrak{X}}(\mathcal{I})$  which we can also interpret in the Eulerian frame for a flow tracer as follows. Let  $\Omega$  be a closed connected subset of  $\mathbb{R}^d$ ,  $d = 1, 2, 3$  over which we define some initial tracer  $\mathbf{X}$ . We let  $\rho_t$  describe the tracer  $\mathbf{X}$  at time  $t$  by evolution of  $\Omega$  by  $\varphi$  via composition. That is, points  $x_0 \in \Omega$  at time  $t = 0$  are transported to positions  $x_t = \varphi_t(x_0)$  at time  $t \in (0, 1]$ , so  $\rho_t(x_t) = \mathbf{X} \circ \varphi_t^{-1}(x_t)$ . Since  $\varphi_t$  is a diffeomorphism the tracer  $\rho$  must be invariant with respect to time:

$$0 = \frac{d}{dt}\rho_t(x_t) = \frac{d}{dt}\rho_t \circ \varphi_t(x_0) = \frac{\partial \rho_t}{\partial t} \circ \varphi_t(x_0) + \nabla \rho_t \circ \varphi_t(x_0) \cdot \frac{\partial \varphi_t}{\partial t}(x_0).$$

Using (2.2) we can therefore say:

$$\frac{\partial \rho_t}{\partial t} \circ \varphi_t(x_0) + \nabla \rho_t \circ \varphi_t(x_0) \cdot u_t \circ \varphi_t(x_0) = 0,$$

for any  $x_0 \in \Omega$ . To summarise, we recognise now the Eulerian transport equation for  $\rho_t$  known from continuum mechanics which holds for all  $t \in [0, 1]$ :

$$\dot{\rho}_t + u_t \cdot \nabla \rho_t = 0, \quad (2.3a)$$

$$\rho_0 = \mathbf{X}. \quad (2.3b)$$

The problem of finding a diffeomorphic match between the template  $i_0$  and target  $i_1$  is therefore equivalent to finding a family  $\{u_t\}_{0 \leq t \leq 1}$  of vector fields  $u_t \in \mathfrak{X}$  so that  $\varphi_1.i_0 \approx i_1$ . This is incredibly powerful as the curve (2.2) then also provides a curve  $\varphi_t.i_0$  on  $\mathcal{I}$  giving rise to meaningful diffeomorphic interpolation in shape space. That is, we consider  $\text{Diff}_{\mathfrak{X}}(\mathcal{I})$  a subgroup of diffeomorphisms

---

<sup>1</sup> Take  $h = a^{-1}$ ,  $h_1 = \mathbf{id}$  and  $h_2 = a$

generated by smooth vector fields<sup>2</sup>. This approach was, to the best of the author's knowledge, first used for matching problems in [59, 218]. As mentioned the link between diffeomorphisms and velocity fields can be explicitly described using notions of geometry. Indeed, for a certain  $\varphi \in \text{Diff}_{\mathfrak{X}}(\mathcal{I})$  we can equip the tangent space  $T_{\varphi}\text{Diff}_{\mathfrak{X}}(\mathcal{I})$  with a metric via (2.2):

$$\|\dot{\varphi}_t\|_{T_{\varphi}\text{Diff}_{\mathfrak{X}}(\mathcal{I})} = \|\dot{\varphi}_t \circ \varphi_t^{-1}\|_{\mathfrak{X}} = \|u_t\|_{\mathfrak{g}}. \quad (2.4)$$

The last equality relies on the fact that the tangent space of  $\text{Diff}_{\mathfrak{X}}(\mathcal{I})$  pulled back to the identity coincides with the Lie algebra  $\mathfrak{g}$  of this group [234, Appendix B.4]. Supposing  $\psi, \psi' \in \text{Diff}_{\mathfrak{X}}(\mathcal{I})$ , then we can finally define a distance  $d_{\text{Diff}_{\mathfrak{X}}(\mathcal{I})} : \text{Diff}_{\mathfrak{X}}(\mathcal{I}) \times \text{Diff}_{\mathfrak{X}}(\mathcal{I}) \rightarrow \mathbb{R}$  using the motion of  $\varphi$  defined in (2.2):

$$d_{\text{Diff}_{\mathfrak{X}}(\mathcal{I})}(\psi, \psi')^2 = \inf_{u \in \mathfrak{X}} \left\{ \int_0^1 \|u_t\|_{\mathfrak{g}}^2 dt \mid \varphi_0 = \psi, \varphi_1 = \psi' \right\}. \quad (2.5)$$

We shall also sometimes refer to the following quantity as the *kinetic energy* of a deformation:

$$E(u) = \int_0^1 \|u_t\|_{\mathfrak{g}}^2 dt. \quad (2.6)$$

The distance on  $\mathcal{I}$  in equation (2.1) is then completely well-defined [174]. We call the diffeomorphism associated with a minimiser of (2.5) a *geodesic*. This defines the "shortest path" between two shapes  $i_0$  and  $i_1$  in this metric. The key point to note here is that we define the metric between shapes via a metric on a group acting on said shapes, which in turn is *intrinsically related* to equipping the Lie algebra with the right kind of inner product on  $\mathfrak{g}$ . The last statement is contained in the equation (2.4). Figure 2.1 provides a diagram portraying the interplay of these objects. Not only does this framework provide a metric  $d_{\mathcal{I}}$  cf. (2.1) but also a way of smoothly interpolating between  $i_0$  and  $i_1$  images using the evolution of  $\varphi$ .

An early application of the theory of Trouné [218] was presented in the sequel [219] demonstrating numerical results for results for matchings with the diffeomorphism group. Based on Trouné's work a practical way to compute a match between two images is through the following variational optimisation problem for  $u_t$ :

$$\inf \int_0^1 \|u_t\|_{\mathfrak{g}}^2 dt + \|\varphi_1 \cdot i_0 - i_1\|_{\mathcal{I}}^2 \quad (2.7)$$

$$\text{subject to} \quad (2.2), \quad (2.8)$$

where  $\|\cdot\|_{\mathcal{I}}$  is some known norm. This can be viewed as a so-called *inverse problem*, as the goal is to find the optimal family of parameters  $u$  for which the solution of the ODE (2.2) provides the answer to the matching problem.

---

<sup>2</sup> Not every diffeomorphism admits a characterisation akin to (2.2) via a velocity, see e.g. [148].

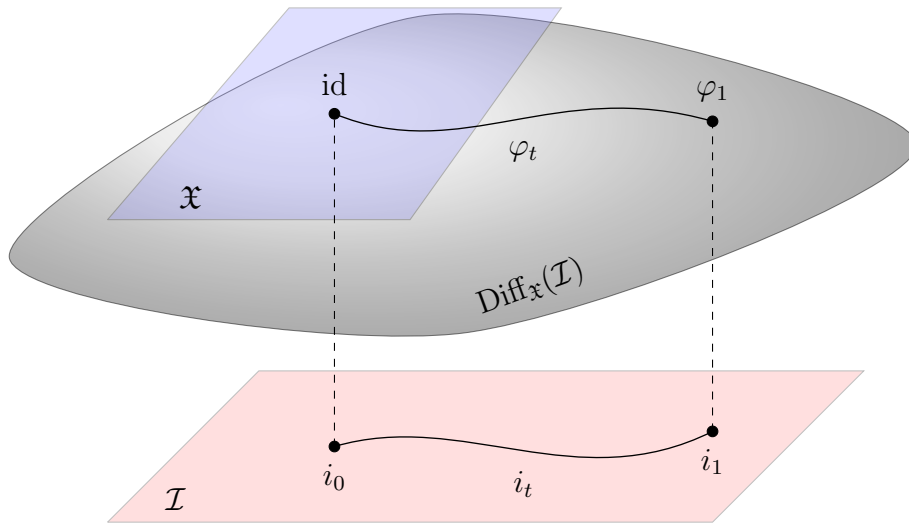


FIGURE 2.1: Schematic of diffeomorphic shape matching.

This approach is popular in the literature and carries the moniker *Large Deformation by Diffeomorphic Metric Mapping* (LDDMM). The *Deformation* (resp. *Large*) part of this label describes the vector field  $u_t$  which generates the diffeomorphism (resp. for all times). Here, the penalty is useful as it represents the condition that we wish to approximate  $i_1$  by the orbit of  $\text{Diff}_{\mathfrak{X}}(\mathcal{I})$  at  $i_0$ .

By virtue of classic Picard-Lindelöf theory  $u_t$  must satisfy a pointwise Lipschitz condition for well-posedness of (2.2). Since the space of such functions is only Banach it is impractical to model from a numerical point of view. It is therefore convenient to let  $\mathfrak{g}$  be a Hilbert space to which we can attach some nice structure. Dupuis et al. [86] shows that a weaker condition than pointwise Lipschitz can ensure well-posedness of the flow map equation (2.2) which roughly states that  $\|\cdot\|_{\mathfrak{g}}$  dominates the first three derivatives (in sense of Sobolev) of the velocity field. Details of this embedding are discussed at length later - the key is that we can in practice work with Hilbert spaces instead of the space of Lipschitz functions. After publication of Dupuis et al., Grenander and Miller published a paper titled *Computational Anatomy: an emerging discipline* [111] which documents in great detail the aggregate work in this field up to that point in time.

Extremising the energy (2.6) subject to the equation leads to a variational equation for the variable  $u_t$  i.e. the associated *Euler-Lagrange* equations known as *EPDiff* (Euler-Poincaré for the diffeomorphism group) [11, 127]. These have been studied extensively - we highlight the work of Mumford and Michor [180] to further elucidate the connection between suitably smooth families of vector fields and diffeomorphisms. [224] takes a Hamiltonian approach to shape spaces in the diffeomorphic framework. A more recent and excellent pedagogical exposition of the Hamiltonian viewpoint in computational anatomy

with applications to real data can be found in [172].

**Remark 1.** *For completeness we also mention an alternative matching framework using elasticity methods [16, 15, 103]. Elastic deformations can be applied to an object such as the ambient space of an image to deform its features visually. We shall not favour these methods as the deformation to an object depends explicitly on the stress that is already present in the material. In other words, it requires more energy to bend an object that is already under stress. In this thesis we focus on methods using diffeomorphism similar to (2.2) owing to the success of the method, its firm theoretical foundation in differential geometry and its simplicity. Other frameworks also exist, see the references in [197].*

The LDDMM framework described here provided the fundamentals for shape matching by finding a suitable diffeomorphism expressed via a curve of velocity fields. Since diffeomorphisms are smooth invertible mappings they cannot, for instance, account for photometric variability in the context of image matching. In other words, the diffeomorphism group in question cannot accommodate changes in the underlying topology of the object studied and express the growth of tumours as alluded to by figure 1.4 in the previous chapter. To cater for such solutions we must make our search space of deformation must larger than  $\text{Diff}_{\mathbf{x}}(\mathcal{I})$ . This challenge is dealt with in the next section where we discuss a mathematical framework allowing for topological alterations in shape matching, primarily in the context of images.

### 2.1.1 GROWTH

From what precedes we have a well-defined notion of a smooth invertible deformation between two shapes that share a diffeomorphic orbit. If we request a *homeomorphism* between two shapes for the purpose of growth simulation, the diffeomorphic orbit can be viewed as a limitation. Applications for such a method can be found in computational anatomy [113] where the growth may represent a tumour appearing on an medical image over the course of time. Here we describe an extension of the LDDMM approach. This is often called the *metamorphosis* metric framework which carries its moniker from [221].

Miller and Younes [174] provide a modified group action model based on the semidirect product space  $\mathcal{A} = \text{Diff}_{\mathbf{x}}(\mathcal{I}) \times \mathcal{I}$ , so that infinitesimal changes to the shape space is represented as a velocity and an element of  $\mathcal{I}$ . See also the sequel paper [173]. [220] takes a more geometric point of view and describes in great detail the Riemannian construction. We provide another schematic for metamorphosis in figure 2.2 to explain these concepts.

**Remark 2** (Images as shapes). *To exemplify the idea we shall explain metamorphosis in the context of images where  $\mathcal{I}$  is a space of functions defined over a square domain whose evaluation at a point returns the grey-scale image*

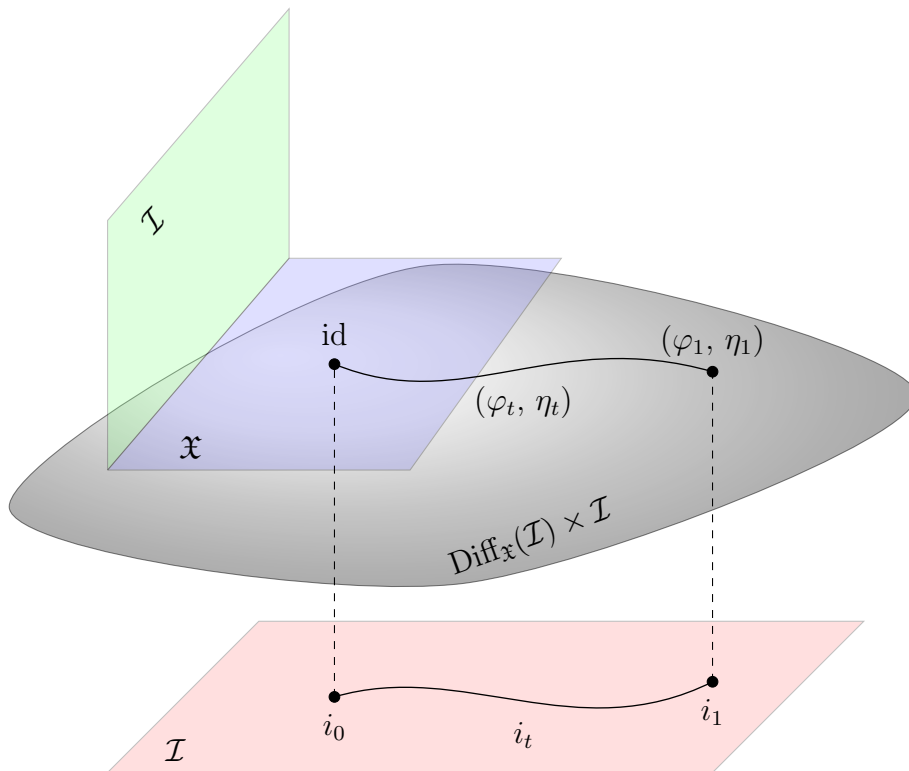


FIGURE 2.2: Schematic of metamorphosis. The image  $i_t$  is given by  $\varphi_t \cdot \eta_t$ , with the boundary conditions  $\eta_0 = i_0$  and  $\varphi_t \cdot \eta_1 = i_1$ . Note that metamorphosis does not always reduce to the semidirect product structure but here we shall use it to articulate the main ideas, see [221] for details.

*intensity at the location. The diffeomorphism part of the  $\mathcal{A}$  acts on the support of the image by stretching or deforming the underlying square, while the metamorphosis part of the group action modifies image intensities.*

Let  $L^2(\Omega)$  denote the space of square-integrable functions over some closed subset  $\Omega$  of  $\mathbb{R}^d$ ,  $d = 1, 2, 3$  with norm  $\|\cdot\|_{0,\Omega}$ . Using the idea from the previous remark we let  $\mathcal{I}$  denote some smooth space of functions so that for some curve  $i$  on this space  $\frac{\partial i_t}{\partial t} + u_t \cdot \nabla i_t \in L^2(\Omega)$ . Recalling that a distance measure on the space of shapes can be written in the form of an optimisation problem we can express metamorphosis in Eulerian coordinates as the geodesics associated with the following energy:

$$\inf \int_0^1 \langle u_\varphi, u_\varphi \rangle_{\mathfrak{g}} + \sigma^{-2} \left\| \frac{\partial i_t}{\partial t} + u_t \cdot \nabla i_t \right\|_{0,\Omega}^2 dt \quad (2.9a)$$

$$\text{s.t. } i_0 \text{ and } i_1 \text{ fixed.} \quad (2.9b)$$

In the Eulerian frame, the equation of motion for the image is then given by the following partial differential equation (PDE):

$$\frac{\partial i_t}{\partial t} + u_t \cdot \nabla i_t = z_t,$$

where  $z_t$  represents the deformation to the shape by a curve on  $\mathcal{I}$ , similar to  $\eta$  in figure 2.2. In light of this problem we emphasise the main difference between metamorphosis and the purely diffeomorphic setting we saw before. Formally speaking we can think of the (2.7) as kind of an initial value problem for  $i_0$  as it seeks a best approximation of  $i_1$  in its diffeomorphic orbit, where  $\|i_0 \circ \varphi_1 - i_1\|_{\mathcal{I}}$  may be a positive quantity. In similar vein, metamorphosis can be thought of as a boundary value problem because we *impose* the template and target  $i_0$  and  $i_1$  and allow for the diffeomorphic *equation* (2.3) to be transgressed. Here we have chosen to cast the problem in Eulerian coordinates simply as an example. In other words, we introduce a source term in the mapping described by (2.3) in order to describe violations of pure advection and thereby permit topological changes in the image. The point here is that if we prescribe an initial as well as a final condition for the image solution, then there may not exist *any* diffeomorphism  $\varphi$  for such that  $i_1 = \varphi.i_0$ . For instance, consider  $i_0$  and  $i_1$  being a circle and ring torus in  $\mathbb{R}^2$ , respectively (i.e. the situation in figure 1.4). However, the advection term appearing in equation (2.9) can be non-zero to cater for such topological changes.

The geodesics associated with the metamorphosis problem from [221] are in [129] recast in the Euler-Poincaré framework of [128]. The relationship between diffeomorphisms and velocities is clearly nonlinear and diffeomorphisms generated by separate vector fields are not additive. In this paper it is explained how a variation with respect to the diffeomorphism is reflected in the family velocity field, and its connection with the Lie bracket over the field. Comparisons between the resulting equations and those of complex fluids is investigated and these are extended to cater for metamorphoses of measures. A very general and in depth exposition of the geometric dynamics of optimisation with metamorphosis as a central application is provided in [102]. For recent and successful results on general growth problems in computational anatomy we refer to the work of Kaltenmark and Trounev [144, 147, 145, 146].

We have outlined the main ideas behind two powerful frameworks using diffeomorphometric mappings: the LDDMM approach, where we seek a purely diffeomorphic mapping between two shapes and the generalisation, metamorphosis, where the purely diffeomorphic orbit implied by (2.2) can be violated to model topological differences between two shapes. In both cases the matching is parameterised by a family of time-dependent velocity fields. We view these matching problems as so-called *inverse problems* for the family of velocity fields which is the topic of the next section.

### 2.1.2 INVERSE PROBLEMS

An inverse problem in its most abstract form seeks to recover the input parameter  $u \in X$  that is mapped to a known state  $q \in Q$  by some typically known *observation operator*  $G : X \rightarrow Q$ . The emphasis here is that the parameter of

the partial differential equation is now the unknown. This means inverting, in some sense,  $G$  so that  $u = G^{-1}(q)$ . In practice, we are often simply looking for an approximation:  $u \approx G^{-1}(q)$ , and the inverse may not be unique or classically defined. This lines up well with matching problems in shape analysis. One can, for instance, view shape matching as the inverse problem of finding the velocity  $u$  such that an initial shape  $i_0$  is mapped to  $i_1 \approx G(u)$  where  $G$  is the solution to the geodesic equations of (2.7) and (2.9). This thesis is focused on developing an efficient and convergent numerical approximations of  $G$  to which we can apply a solution method for the inverse problem. In a very crude sense, we can view the goal of this thesis as bounding  $\|i_1 - G(u)\|_{\mathcal{I}}$  by the following triangle inequality:

$$\begin{aligned} \|i_1 - G(u)\|_{\mathcal{I}} &\leq \|i_1 - G_h(u)\|_{\mathcal{I}} + \|G_h(u) - G(u)\|_{\mathcal{I}} \\ &\leq \|i_1 - G_h(u)\|_{\mathcal{I}} + \|G_h - G\|_{\mathcal{I}^*} \|u\|_X. \end{aligned}$$

where  $G_h$  is a numerical approximation of  $G$  and  $\|\cdot\|_{\mathcal{I}^*}$  is a formal dual norm. The aim is to force  $\|i_1 - G_h(u)\|_{\mathcal{I}}$  to zero by finding a suitable  $u$  by a numerical solution to the inverse problem, and to provide a convergent discretisation  $G_h$  of  $G$  so that  $\|G_h - G\|_{\mathcal{I}^*} \rightarrow 0$ . Our main contributions are related to questions of convergence of this term for various shape spaces. The remainder of this section discusses the former in context of the literature.

A popular method for solving inverse problems in shape analysis is called *shooting*, and we highlight here some central references explaining this approach. Shooting employs a Newton method to solve the nonlinear equation  $G_h(u) - i_1 = 0$  [171, 226]. Since the velocity field occupies a linear space  $X$  it is feasible to construct this method on  $u$  rather than shooting in the nonlinear space of diffeomorphisms. Beg [30] popularised this for LDDMM and adopts a gradient-based optimisation scheme to find geodesics between images using the Euler-Lagrange equations of the problem described by (2.7). [233, 235] studies the evolution equations obtained through perturbations of a diffeomorphism by its vector field for use in a gradient descent algorithm based on an Euler scheme.

We can also view  $G(u) - i_1 = 0$  as a *Bayesian inverse problem* [211, 71] where the aim is to find a distribution of candidate functions  $u$  that minimise the associated kinetic energy of a transformation [65, 163, 162]. This point of view can be advantageous when the inversion of  $G$  requires a more general interpretation such as a probabilistic one. Such a parameterisation requires a probability distribution over the infinite-dimensional space  $X$ , which under suitable conditions allows one to carry over the notion of maximum likelihood estimators from the finite-dimensional setting. The general framework of Bayesian inversion is based on the idea that we can cast optimisation problems in a probabilistic framework where minimisers, roughly speaking, correspond to modes of a certain distribution over function space. Viewing a solution to an inverse problem in this *global* distributional sense can be appropriate depending on the well-posedness of  $G$ . A *local* gradient-based method may not adequately

reveal local minima of the functional from which the geodesic equations are derived. Although it is not a subject of our investigation in this thesis we mention that there exists a rich literature on computational statistics studying the variability of biological population shape data through diffeomorphisms [175, 95, 222, 190]. Considering these applications therefore highly motivates the development of novel theoretical and numerical results for both the forward operator and the inverse problems in shape matching.

## 2.2 VARIATIONAL PROBLEMS

### 2.2.1 ELEMENTARY FUNCTIONAL ANALYSIS

We recall some basic functional analysis to set up the notation for the thesis, essentially quoting definitions and results from [89, Appendix A and B].

Let  $X$  denote a real-valued Banach space where  $X^*$  denotes the dual of  $X$  and  $(\cdot|\cdot)_{X^*,X}$  the pairing between elements of  $X^*$  and  $X$ , and when  $g \in X^*$  and  $v \in X$ , we also sometimes write  $g(v) := (g|v)_{X^*,X}$ . When  $X$  is a Hilbert space we let  $\langle \cdot, \cdot \rangle_X$  denote its canonical inner product. For a real-valued Banach space  $Y$  and a linear operator  $\mathcal{B} : X \rightarrow Y^*$  and  $f \in Y^*$ , we pose the following abstract problem for  $u \in X$  where  $\mathcal{B}u \in Y^*$ :

$$\mathcal{B}u = f \quad \text{in } Y^*,$$

or, in *weak* or *variational* form:

$$b(u, v) = f(v), \quad \forall v \in Y, \tag{2.10}$$

where the bilinear form  $b : X \times Y \rightarrow \mathbb{R}$  is defined by the relation:

$$b(u, v) = \mathcal{B}u(v),$$

typically defined via partial integration.  $\mathcal{B}$  will usually be a differential operator so (2.10) represents a weak form partial differential equation (PDE). We say that the bilinear form  $b$  is continuous if and only if there exists a constant  $c_0 > 0$  such that:

$$|b(u, v)| \leq c_0 \|u\|_X \|v\|_Y,$$

which implies continuity of  $\mathcal{B}$ . Next we define the surrogates for  $X$  and  $Y$  used in this thesis and quote the main results for solving equations such as those above, primarily in the context of Hilbertian Sobolev spaces [90].

The space  $C^k(O, \mathbb{R}^d)$  denotes the space of continuous functions over the domain  $O$  with  $k$  continuous derivatives taking value in  $\mathbb{R}^d$ . When the latter is omitted, the range is  $\mathbb{R}$  e.g.  $C^0(O)$  is the space of continuous functions taking values in  $\mathbb{R}$ .  $L^p(O)$ ,  $p \geq 1$  and integer denotes the usual Lebesgue spaces over  $O$  with  $L^\infty(O)$  being the Banach space of essentially bounded functions on  $O$ . For

$k \geq 0$  (integer or fraction) we let  $W^{k,p}(O)$  define the usual Sobolev spaces, see [90]. Note that  $W^{k,\infty}(O)$  is the Banach space of scalar  $k$ -Lipschitz continuous functions<sup>3</sup> over the convex polygonal Lipschitz domain  $O$  taking values in  $\mathbb{R}$ . We will be dealing with vector and tensor-valued functions as well, and therefore when necessary use superscripts e.g.  $L^\infty(O)^d$  to denote vector-valued functions  $\mathbb{R}^d \rightarrow \mathbb{R}^d$ , or  $W^{k,\infty}(O)^{d \times d}$  denotes the the space of tensor-valued  $k$ -Lipschitz functions on  $O$ . This should be understood in the sense that each component considered as a scalar function satisfies the regularity of the given space. We sometimes use standard multi-index notation i.e. an  $d$ -dimensional multi-index  $\alpha$ , with  $|\alpha| = \sum_{i=1}^d \alpha_i$ , and we define  $\partial^\alpha$ , the partial derivative with respect to  $\alpha$  as  $\partial^\alpha = \frac{\partial^{|\alpha|}}{\partial^{\alpha_1} x^1 \dots \partial^{\alpha_d} x^d}$ . In this thesis  $dx$  - sometimes also  $d\hat{x}$  - denotes the  $d$ -dimensional Lebesgue measure [42, Section 1.1], where  $d$  will be clear from context.

In this thesis we make a clear distinction in our notation depending on the value of  $p$ . This is reflected in the following norms when  $p = 2$  or  $p = \infty$ , respectively:

$$\|f\|_{k,O}^2 = \sum_{|\alpha| \leq k} \|\partial^\alpha f\|_{0,O}^2, \quad (2.11a)$$

$$\|f\|_{W^{k,\infty}(O)} = \sum_{|\alpha| \leq k} \|\partial^\alpha f\|_{\infty,O}, \quad (2.11b)$$

where  $\|\partial^\alpha f\|_{0,O}^2 = \int_O |\partial^\alpha f|^2 dx$  is the usual  $L^2$  norm over  $O$  and  $\|\partial^\alpha f\|_{\infty,O} = \text{ess sup}_O |\partial^\alpha f|$  is the essential supremum norm. Note in particular that due to the presence of the Euclidean norm  $|\cdot|$  here, the norms in (2.11) are well-defined for scalar, vector and tensor-valued objects.

When  $p = 2$  we also use the notation  $H^k(O) := W^{k,2}(O)$  with  $H_0^k(O)$  consisting of such functions that vanish on  $\partial O$  in sense of traces. We also denote by  $|\cdot|_{k,O}$  the usual semi-norm over this space. The inner product on  $H^k(O)$  is denoted by  $\langle \cdot, \cdot \rangle_{k,O}$ .

Sobolev spaces play a central role in this thesis and we therefore highlight several results of which we shall make extensive use.

**Theorem 3** (Trace theorem [89, B.52]). *Let  $O$  be an open  $\mathbb{R}^d$  with Lipschitz boundary. Then there exists a surjective operator  $\gamma_O$  such that:*

$$\gamma_O : H^1(O) \rightarrow H^{1/2}(\partial O),$$

where  $\ker(\gamma_O) = H_0^1(O)$  i.e. the space of functions in  $H^1(O)$  that vanish on  $\partial O$ .

---

<sup>3</sup> We remark that the Lipschitz class of functions does not always coincide with  $W^{k,\infty}(O)$  depending on the nature of  $O$ . Given that we are always dealing with convex domains we characterise the Lipschitz class with the Sobolev spaces  $W^{k,\infty}(O)$  (see [122, Theorem 4.1]) and mention this no further.

Further, when  $\gamma_O(u) = u_\Gamma$ , the following bounds hold:

$$\|u_\Gamma\|_{0,\partial O} \lesssim \|u\|_{1,O},$$

with a constant depending the outward normal to  $O$ , see [114, Theorem 1.5.1.10] for details.

**Theorem 4** (Poincaré [89, B.61]). *Let  $O$  be an open Lipschitz set of  $\mathbb{R}^d$ . Then there exists a constant  $C > 0$  depending only on  $O$  such that:*

$$\|u\|_{0,O} \leq C \|\nabla u\|_{0,O}, \quad u \in H_0^1(O).$$

One of the main tools we use is the embedding theorem below from which we deduce that continuous representatives of Sobolev equivalence classes are members of smooth spaces.

**Theorem 5** (Sobolev Embedding Theorem [239, p. 2.5.1]). *Let  $O$  be a bounded domain in  $\mathbb{R}^d$ . If  $kp > d + mp$ ,  $W_0^{k,p}(O)$  is compactly embedded in  $C^m(\bar{O})$  which we denote by  $W_0^{k,p}(O) \hookrightarrow C^m(\bar{O})$ .*

See also [1] for more details.

**Theorem 6** (Fréchet-Riesz [89, A.28]). *The action of every linear functional  $v' \in X^*$  on a Hilbert space  $X$  can be represented uniquely as an inner product:*

$$v'(w) = \langle u, w \rangle_X, \quad \forall w \in X.$$

Further,  $\|v'\|_{X^*} = \|u\|_X$ .

An important consequence of this is that every Hilbert space is reflexive [89, Proposition A.29].

We can now look at how to solve abstract variational equations such as (2.10). What we present is often called Babuška theory [101]. For an operator  $\mathcal{B} : X \rightarrow Y^*$  we recall that for a given  $x \in X$  we can compute the operator norm in  $Y^*$  which is given by:

$$\|\mathcal{B}x\|_{Y^*} = \sup_{0 \neq y \in Y} \frac{|\mathcal{B}x(y)|}{\|y\|_Y}.$$

Moreover, define the *transpose*  $\mathcal{B}^* : Y \rightarrow X^*$  by the following relation:

$$(\mathcal{B}x|y)_{Y^*,Y} = (\mathcal{B}^*y|x)_{X^*,X} = b(x,y).$$

The next result establishes conditions for invertibility of an operator in the form of inf-sup conditions and can be found in most textbooks on functional analysis.

**Theorem 7.** *For reflexive Banach spaces  $X$  and  $Y$ , let  $\mathcal{B} : X \rightarrow Y^*$ . Then  $\mathcal{B}$  is an isomorphism if and only if:*

*i.  $\mathcal{B}$  is linear,*

*ii.  $x \in X$ ,  $\mathcal{B}x(y) = 0$  for all  $y \in Y$  implies  $x = 0$  (injectivity),*

*iii.  $\mathcal{B}$  satisfies an inf-sup condition with some bound  $\alpha_E$ :*

$$0 < \alpha_E = \inf_{0 \neq x \in X} \sup_{0 \neq y \in Y} \frac{\mathcal{B}x(y)}{\|x\|_X \|y\|_Y}.$$

Condition **ii** corresponds to a uniqueness condition, which can also be written in the form of **iii** as:

$$0 < \alpha_U = \inf_{0 \neq y \in Y} \sup_{0 \neq x \in X} \frac{\mathcal{B}x(y)}{\|x\|_X \|y\|_Y}.$$

The inf-sup condition **iii** corresponds to an existence result. Indeed, **iii** means that the transpose  $\mathcal{B}^*$  is *bounding*:

$$\text{iii} \Leftrightarrow \forall x \in X, \quad \sup_{0 \neq y \in Y} \frac{\mathcal{B}x(y)}{\|x\|_X} = \|\mathcal{B}^*y\|_{X^*} \geq \alpha_E \|y\|_Y,$$

which means that  $\mathcal{B}^*$  is injective so  $\mathcal{B}$  is surjective. Note that we can also verify condition **iii** by the following problem:

$$\forall y \in Y, \quad \text{find } x \in X \text{ such that } \mathcal{B}x(y) \geq \alpha_E \|x\|_X \|y\|_Y. \quad (2.12)$$

For many trivial problems where  $X = Y$  the existence and uniqueness conditions of theorem 7 are satisfied via the property is called *coercivity*:

**Definition 8** (Coercivity). *A bilinear form  $b : X \times X \rightarrow \mathbb{R}$  defined over a Banach space  $X$  is called coercive if and only if there exists a constant  $\alpha_0 > 0$  such that:*

$$\alpha_0 \|v\|_X^2 \leq b(v, v), \quad \forall v \in X.$$

Such a bilinear form is easily seen to satisfy condition **iii** of theorem 7 when  $a(u, v) = \mathcal{B}u(v)$ . This is sometimes called the Lax-Milgram theorem [43, Corollary 5.8].

Finally the following definition will be useful later on.

**Definition 9.** *For a subset  $Z \subset X$  of a Hilbert space  $X$  we define the annihilator  $Z^\circ$  of  $Z$  by:*

$$Z^\circ = \{f \in X^*, f(z) = 0, \forall z \in Z\},$$

*and the orthogonal space  $Z^\perp$  by:*

$$Z^\perp = \{x \in X, \langle x, z \rangle_X = 0, \forall z \in Z\}.$$

These can be understood as objects that are orthogonal to the space  $X$  but in different spaces.

### 2.2.2 FINITE ELEMENT METHODS

This section covers finite element methods - most of the theory here is covered in the textbooks [42, 89, 39].

The finite element method works by replacing the spaces in (2.10) above by discrete counterparts that we can realise numerically. These spaces are typically piecewise polynomial spaces although other methods do exist e.g. the generalised FEM [210, 209], the extended FEM (see [33] for a review) or Trefftz FEM [124]. The work [14] provides an excellent motivation for non-polynomial approximation spaces in the context of the generalised FEM.

Our aim is to communicate the path from the preceding infinite-dimensional setting to the finite-dimensional one here in order to prove convergence results for discrete analogues of (2.10). We are not focused on the implementation or numerical linear algebra related to finite element methods i.e. approximating integration by quadrature, linear solves or the like. The interested reader is referred to [89, Part III] for an overview of the challenges here. Throughout this thesis our numerical results are obtained using the Firedrake framework [191].

In this thesis  $\Omega$  will generally denote a convex connected subset of  $\mathbb{R}^d$  with Lipschitz polygonal boundary  $\partial\Omega$  representing our computational domains (the definition being specific to each setting). For a domain  $O$  we let  $\mathring{O}$  denote the interior of  $O$ . To establish a FEM we construct a triangulation over such a domain.

**Definition 10** (Triangulation). *We use the textbook definition [42, Definition 3.3.11] of a triangulation: a triangulation of a polygonal domain  $O$  is a finite collection  $O_h$  of triangular element domains  $\{K_j\}_j$  such that*

1.  $\mathring{K}_j \cap \mathring{K}_i = \emptyset, i \neq j$
2.  $\overline{O} = \cup_i K_i$

where no vertex of any triangle lies in the interior of an edge of another triangle. We also refer to a triangulation  $O_h$  as the mesh. For a triangulation  $O_h$  we define the mesh resolution, or mesh size, as  $h = \max_{K \in O_h} h_K$ , where  $h_K = \text{diam}(K)$ , see [42, 60]. As we will be studying convergence of finite element methods we introduce the notion of shape-regularity [89, Definition 1.107]. A family of meshes  $\{O_h\}_{h>0}$  is said to be shape-regular if there exists  $c_0 > 0$  independent of  $h$  such that

$$\max_{K \in O_h} \frac{h_K}{\rho_K} \leq c_0, \quad \forall O_h \in \{O_h\}_{h>0}, \quad (2.13)$$

where  $\rho_K$  is the diameter of the largest circle inscribed in  $K$ . We also assume that such families are quasi-uniform [89, Definition 1.140] i.e. there exists a

constant  $c > 0$  such that:

$$\forall h, \forall K \in O_h, \quad h_K \geq ch.$$

When speaking of finite element convergence (i.e. statements to the effect “as  $h \rightarrow 0$ ”), we always mean in the context of a shape-regular family of triangulations. Since  $O_h$  is a collection of sets we shall use the notation  $\int_{O_h} = \sum_{K \in O_h} \int_K$ , and elements of a triangulation  $O_h$  are called elements. See also appendix D for standard results on scaling arguments and mesh-related constants.

When it is clear from context we also refer to triangulations as *meshes*, and members of  $\Omega_h$  are sometimes called *elements*.

**Definition 11** (Skeleton). *Given a triangulation  $O_h$  as above we define its interior facets (or edges, when  $d = 2$ )  $\mathring{\mathcal{E}}_h$  by  $e \in \mathring{\mathcal{E}}_h$  if  $\exists K, K' \in O_h$  such that  $e = K \cap K'$ . We denote by  $\partial O_h$  the edges of  $\partial K$ ,  $K \in O_h$  that trace  $\partial O$ , and we say that  $\mathcal{E}_h = \mathring{\mathcal{E}}_h \cup \partial O_h$  is the mesh skeleton of  $O_h$ . Sometimes we explicitly denote the dependence of the skeleton on its associated triangulation i.e.  $\mathcal{E}_h(O_h)$ . We also use the notation  $\int_{\mathcal{E}_h} = \sum_{e \in \mathcal{E}_h} \int_e$  with similar interpretations for  $\int_{\partial O_h}$  and  $\int_{\mathring{\mathcal{E}}_h}$  and we have:*

$$\int_{\mathcal{E}_h} = \int_{\mathring{\mathcal{E}}_h} + \int_{\partial O_h}.$$

We can describe finite-dimensional function space on such triangulations.

**Definition 12** (Finite Element Space). *Let  $X$  be a Hilbert space over a bounded connected  $\Omega \subset \mathbb{R}^d$  equipped with a triangulation  $\Omega_h$ . We define a finite element space  $X_h$  over  $\Omega_h$  as a function space such that the restriction of  $X_h$  to  $K \in \Omega_h$  is a piecewise polynomial function i.e.  $X_h|_K \subset \mathcal{P}^j(K)$  where the latter is the space of polynomials of order  $j \geq 0$  on  $K$ . If  $X_h \subset X$  we say that  $X_h$  is a conforming finite element space, otherwise it is a nonconforming finite element space [202, 61].*

An abstract discrete version of (2.10) seeks a  $u_h \in X_h$  (possibly nonconforming) such that:

$$b_h(u_h, v_h) = f(v_h) \quad \forall v_h \in X_h. \quad (2.14)$$

Here we have emphasised that the bilinear form and source term act on finite element functions by using a subscript i.e.  $b_h$ , as it may be that this differs from  $b$  upon discretisation. Most of the time, and in particular for conforming methods,  $b_h = b$ . When (2.14) is well-posed (e.g. by coercivity of  $b_h$ ) the following result lets us estimate the difference  $u - u_h$ :

**Lemma 13** (Strang [208]). *Let  $u$  and  $u_h$  solve (2.10) and (2.14), respectively, and that the definition of  $b_h$  can be extended so that  $b_h(u, v_h)$  is well-defined for  $u \in X$  and  $v_h \in X_h$ . Then the following estimate holds:*

$$\|u - u_h\|_{X_h} \lesssim \inf_{w_h \in X_h} \|u - w_h\|_{X_h} + \sup_{v_h \in X_h} \frac{f(v_h) - b_h(u, v_h)}{\|v_h\|_{X_h}}. \quad (2.15)$$

*Proof.* For  $w_h, v_h \in X_h$  the triangle inequality states:

$$\|u - u_h\|_{X_h} \leq \|u - w_h\|_{X_h} + \|u_h - w_h\|_{X_h}.$$

We bound the second term on the right-hand side,  $\|u_h - w_h\|_{X_h}$ , in terms of  $u$  as follows. Since (2.14) is well-posed ( $b_h$  satisfies an inf-sup condition), there exists a constant  $\alpha > 0$  such that:

$$\alpha \|u_h - w_h\|_{X_h} \leq \sup_{v_h \in X_h} \frac{b_h(u_h - w_h, v_h)}{\|v_h\|_{X_h}}. \quad (2.16)$$

Now observe:

$$\begin{aligned} b_h(u_h - w_h, v_h) &= b_h(u_h + u - u - w_h, v_h) \\ &= b_h(u_h - u, v_h) + b_h(u - w_h, v_h) \\ &= f(v_h) - b_h(u, v_h) + b_h(u - w_h, v_h). \end{aligned}$$

Using this in (2.16):

$$\begin{aligned} \alpha \|u_h - w_h\|_{X_h} &\leq \sup_{v_h \in X_h} \frac{f(v_h) - b_h(u, v_h) + b_h(u - w_h, v_h)}{\|v_h\|_{X_h}} \\ &\lesssim \sup_{v_h \in X_h} \frac{f(v_h) - b_h(u, v_h)}{\|v_h\|_{X_h}} + \|u - w_h\|_{X_h}. \end{aligned}$$

Summarising we have:

$$\|u - u_h\|_{X_h} \lesssim \|u - w_h\|_{X_h} + \sup_{v_h \in X_h} \frac{f(v_h) - b_h(u, v_h)}{\|v_h\|_{X_h}},$$

and since  $w_h$  was arbitrary the result follows. See [39, Section 2.2.3] for more details.  $\square$

The first term on the right-hand side of (2.15) is called the *approximation* term and the second a *consistency* term. Consistency of the discretisation means estimating to what extent, as a function of  $h$ , a continuous solution satisfies the discrete equation. We shall return to this topic later on. The first term on the right-hand side of (2.15) can be controlled by an interpretation of the Bramble-Hilbert lemma [42, Lemma 4.3.8] using finite element *interpolation operators*. Here we only provide the key properties of the usual interpolation operator.

**Definition 14** (Local interpolant [42, Lemma 4.4.1 and Theorem 4.4.4]). *For a convex simplex  $K \subset \mathbb{R}^d$  and non-negative integers  $l$  and  $m$  we define the usual local interpolant  $\mathcal{I}_K : C^l(\bar{K}) \rightarrow H^m(K)$  satisfying the following estimate:*

$$\|\mathcal{I}_K u\|_{m,K} \lesssim \|u\|_{C^l(\bar{K})}.$$

When  $m - l - d/2 > 0$  we have the following bound:

$$|u - \mathcal{I}_K u|_{i,K} \lesssim h_K^{m-i} |u|_{m,K}, \quad 0 \leq i \leq m,$$

with a constant that depends on  $m, d$  and  $K$ .

This extends trivially to vector or tensor-valued objects. We can therefore, under the right circumstances, approximate smooth functions on  $K$  by a polynomial. For nonsmooth functions one must seek other interpolants since the nodal basis may not support point evaluation i.e. when the space  $\mathcal{C}^l(\bar{K})$  in definition 14 is replaced by a weaker space like  $L^2(\bar{K})$ . Local averages are often taken as surrogates for the nodal basis in this case, see e.g. the Clément [62], or for interpolation operator satisfying boundary conditions see Scott-Zhang [200]. For more information on this see the discussion in [42, Section 4.8], and [85] for general polynomial approximation of Sobolev spaces. A *central task* in finite element theory is the construction of the nodal basis and the basis functions to cater for convergent approximations of some infinite-dimensional problem. We shall therefore leave these undefined here and return to their construction when necessary in the coming chapters.

From definition 14 we can construct a global interpolant  $\mathcal{I}_h : X \rightarrow X_h$  by:

$$\mathcal{I}_h v|_K = \mathcal{I}_K v, \quad K \in \mathcal{O}_h, \quad (2.17)$$

where  $\mathcal{O}_h$  is a triangulation of  $O \subset \mathbb{R}^d$  over which the spaces  $X$  and  $X_h$  are defined. Global interpolation estimates requires the notion of a *broken* Hilbert-Sobolev norm given by the following definitions:

$$\|u\|_{k, \mathcal{O}_h}^2 := \sum_{K \in \mathcal{O}_h} \|u\|_{k, K}^2, \quad (2.18)$$

with the associated broken inner product:

$$\langle u, v \rangle_{k, \mathcal{O}_h} := \sum_{K \in \mathcal{O}_h} \langle u, v \rangle_{k, K}.$$

The following approximation estimates are then given in terms of broken Sobolev norms.

**Lemma 15** (Approximability). *Suppose  $X = H^m(O)$  where  $O$  is a Lipschitz domain in  $\mathbb{R}^d$  and let  $\mathcal{I}_h$  be the global interpolant from (2.17). Then for  $u \in H^{m+s}(O)$ :*

$$\|u - \mathcal{I}_K u\|_{i, \mathcal{O}_h} \lesssim h_K^{m+s-i} |u|_{m+s, \mathcal{O}_h}, \quad 0 \leq i \leq m, \quad (2.19a)$$

$$\lim_{h \rightarrow 0} \inf_{v_h \in X_h} \|v - v_h\|_{X_h} = 0, \quad \forall v \in X, \quad (2.19b)$$

and we sometimes say  $X_h$  approximates  $X$ .

*Proof.* (2.19a) follows from the definition of  $\mathcal{I}_h$  and definition 14. (2.19b) follows from continuity of the interpolant a standard density argument; see e.g. the proof of [230, Theorem 2.1].  $\square$

Results such as lemma 15 are often easily obtained and are sufficient to bound the approximation term in (2.15). Estimation of the second term on the right-hand side of (2.15) depends on the properties of  $X_h$ . Supposing  $X_h$  is a conforming space then this term vanishes since  $v_h \in X_h \Rightarrow v_h \in X$  whereby the following *Galerkin orthogonality* result applies which states:

$$b_h(u_h - u, v) = 0 \Rightarrow f(v) - b_h(u, v) = 0.$$

Given this orthogonality result, Strang's lemma simplifies to Céa's lemma.

**Lemma 16** (Céa [89, Lemma 2.28]). *Suppose  $u$  and  $u_h$  solve (2.10) and (2.14) for some  $X$  and  $X_h$  such that  $X_h \subset X$ . Then:*

$$\|u - u_h\|_X \leq \inf_{v \in X_h} \frac{c_0}{\alpha_0} \|u - v\|_X,$$

where  $c_0$  and  $\alpha_0$  are the continuity and coercivity constants of  $b(\cdot, \cdot)$ , respectively.

*Proof.* Subtracting (2.10) and (2.14) and letting  $v \in X_h$ :

$$\begin{aligned} b(u - u_h, u - u_h) &= b(u - u_h + v - v, u - u_h) \\ &= b(u - v, u - u_h) + b(v - u_h, u - u_h) \\ &= b(u - v, u - u_h). \end{aligned}$$

Letting  $c_0$  be the continuity constant of  $b$  and  $\alpha_0$  its coercivity constants we derive:

$$\|u - u_h\|_X \leq \frac{c_0}{\alpha_0} \|u - v\|_X.$$

The result follows since  $v$  was chosen arbitrarily.  $\square$

The final two sections discuss more advanced approximation spaces when  $X_h$  is nonconforming. Section 2.2.2.1 is more relevant to chapter 3, while section 2.2.2.2 covers the background for the chapter 4.

### 2.2.2.1 NONCONFORMING METHODS

Consistency, in sense of Strang's result in lemma 13, does not immediate follow, or necessarily make sense, for nonconforming elements. Generally speaking the choice of the local degrees of freedom of a finite element affects the properties of the global approximation space. For instance, when  $X_h$  is constructed from linear polynomials i.e. for  $K \in \Omega_h$ ,  $X_h|_K = \mathcal{P}^1(K)$  and in addition, functions in  $X_h$  are linear interpolates of the nodal values at the vertices of  $K$ , then it is easy to see that  $X_h \subset C^0(\Omega)$ . A nonconforming linear space also exists where the local degrees of freedom are chosen at the midpoint of the edges of  $\partial K$  see e.g. the construction in [42, Section 10.3]. This results in a *locally continuous, globally discontinuous* finite element and  $X_h \not\subset X$ . While the continuity of the

former can be advantageous in light of lemma 16 it is sometimes inconvenient, or even computationally infeasible to design conforming finite element spaces e.g. when the continuous space  $X$  is some higher-order Sobolev space. Generally speaking, for a  $2m^{\text{th}}$ -order PDE a conforming space must be a  $C^{m-1}$ -conforming subspace<sup>4</sup>. This poses a challenge for the design and implementation of finite element software, in particular for practitioners simply wishing run simulations. For instance, automation of code generation for  $C^1(\Omega)$ -conforming finite elements was only very recently solved [149, 150]. Nonconforming FEMs are therefore attractive in developing numerical schemes for higher-order PDEs. Nonconforming spaces can nevertheless recover adequate convergence and are easier to implement due to their local nature. Note that a nonconforming finite element space need not be globally discontinuous such as in the previous example. It may simply be that higher-order derivatives no longer satisfy a global continuity property.

We return to Strang's result for nonconforming finite elements. Convergence for nonconforming methods are usually established in norms such as (2.18) owing to their local nature. Approximation and consistency conditions for nonconforming finite elements can now be formulated using more suitable vocabulary, see [213] or [230, Section 3.1] for the following definition.

**Definition 17** (Consistent approximation). *A finite element space  $X_h$  over a bounded Lipschitz domain  $O$  is said to be a consistent approximation of  $X = H^m(O)$ ,  $m \geq 1$ , if and only if:*

1.  $\lim_{h \rightarrow 0} \inf_{v_h \in X_h} \|v - v_h\|_{X_h} = 0, \quad \forall v \in X$  (approximation).
2. For any sequence  $\{v_{h_k}\}_k$  with  $v_{h_k} \in X_{h_k}$  and  $h_k \rightarrow 0$  as  $k \rightarrow \infty$  such that  $\{\partial_{h_k}^\alpha v_{h_k}\}_{h_k}$  is  $L^2(O)$ -weakly convergent to some  $v^\alpha \in X$  for all  $|\alpha| \leq m$ , it holds that  $v_0 \in X$  and  $v^\alpha = \partial^\alpha v_0$  for all  $|\alpha| \leq m$  (consistency).

*The first condition can be derived from (2.19b) of lemma 15 via a canonical interpolant. The second condition can be interpreted as a kind of weak compactness condition.*

The second condition in the definition above is nonstandard and may appear somewhat convoluted in nature. Chapter 3 provides a context where this definition becomes natural, so it may not yet be necessary for the reader to fully appreciate this definition. The rest of this section discusses how to verify these conditions for a given nonconforming element. It is not obvious how to verify the compactness condition in definition 17 for an arbitrary element. On the other hand, the form in which this condition is stated will be extremely useful later on.

---

<sup>4</sup> The weak form equation would require  $m$  weak derivatives which implies the existence of  $m - 1$  continuous derivatives.

In practice we employ so-called *patch tests* which are sufficient conditions for a consistency condition but motivated by the fact that they are much simpler to prove and sometimes even verify by simple numerical computation. While point-wise inter-element continuity does not hold for nonconforming elements, patch tests often boil down to checking whether the integral moments of derivatives on vanish which is called a *weak continuity* property. We shall navigate the literature without delving deep into the specifics of the different patch tests. The reason for this is that, in general, the design of a finite element requires a bespoke patch test strategy. Historically the first attempt at providing an automated solution to this issue was in engineering by Irons et al. [29] in 1965 but proved to be neither necessary nor sufficient by Stummel [212]. The generalised patch test (GPT) of Stummel [213] states necessary and sufficient conditions leading to convergence of nonconforming elements. Generally speaking, the design of a finite element leads to some weak continuity of which the GPT does not make use in its original form. However, using the form of the GPT as stated in [229, Equation 4.4],  $X_h$  passes the GPT if and only if the following condition stated is satisfied:

$$\lim_{h \rightarrow 0} \sup_{\substack{v_h \in X_h \\ \|v_h\|_{X_h} \leq 1}} |T_{\alpha,i}(\psi, v_h)| = 0, \quad |\alpha| < m, \quad 1 \leq i \leq d, \quad \forall \psi \in C^\infty(\bar{O}), \quad (2.20)$$

where:

$$T_{\alpha,i}(\psi, v_h) = \sum_{K \in O_h} \int_K \left[ \psi \frac{\partial}{\partial x_i} \partial^\alpha v_h + \frac{\partial \psi}{\partial x_i} \partial^\alpha v_h \right] dx.$$

Equivalently, where  $\eta_K$  is the outward normal of  $K \in O_h$ :

$$T_{\alpha,i}(\psi, v_h) = \sum_{K \in O_h} \int_{\partial K} \psi \partial^\alpha v_h \eta_K^i ds. \quad (2.21)$$

A condition similar to (2.20) is used in chapter 3. Alternatives were also proposed, see for instance F-E-M test [203] and the IPT test [131]. A major step forward was taken in [229] where a conjecture was proved saying that given suitable approximation properties (in spirit of lemma 15) and weak continuity, a *weak patch test* (WPT) can be designed which is extremely simple and convenient. We refer the reader to [202] for an excellent overview of the nonconforming FEM literature for second and fourth order PDEs.

### 2.2.2.2 DISCONTINUOUS & PETROV-GALERKIN METHODS

In addition to conforming and nonconforming methods *discontinuous Galerkin* (DG) methods refer to finite element methods employing fully discontinuous spaces. Reed and Hill first introduced the DG method in 1973 [192] in connection with the neutron transport equation. We refer to the pioneering works [83, 17] for DG interior penalty methods e.g. the latter presents a method for imposing global  $C^1$  continuity using only continuous elements. The success of

this method for hyperbolic problems has generated an incredibly rich literature. While the nonconforming approach from the previous section guaranteed consistency via weaker notions of inter-elemental continuity and a priori passing of a patch test, DG methods impose consistency by adding penalty terms to the bilinear form. [10] provides a historical overview of DG methods (in particular the review in Section 2.2) and unifies these in a single framework in the context of elliptic problems. We provide a simple example of this method for the Poisson equation from this reference to highlight the connection with mixed finite element methods and move on to Petrov-Galerkin methods which are the main focus of this section.

We exemplify by posing a mixed system for the Poisson equation  $-\Delta u = f$  on  $O \subset \mathbb{R}^d$  with homogeneous Dirichlet boundary conditions from the example given in [10]:

$$\begin{aligned}\sigma &= \nabla u, \\ -\nabla \cdot \sigma &= f,\end{aligned}$$

which is a form amenable to the DG method [81]. Multiplying this system by test functions  $\tau$  and  $v$ , respectively, and integrating by parts on  $K \in O_h$  we find:

$$\begin{aligned}\int_K \sigma \cdot \tau \, dx &= - \int_K u \nabla \cdot \tau \, dx + \int_{\partial K} u \eta_K \cdot \tau \, ds, \\ \int_K \sigma \cdot \nabla v \, dx &= \int_K f v \, dx + \int_{\partial K} \sigma \cdot \eta_K v \, ds,\end{aligned}$$

where  $\eta_K$  is the outward normal to  $K$ . A DG method introduces the piecewise scalar and vector finite element spaces:

$$V_h = \{u \in L^2(O) \mid u|_K \in V(K) = \mathcal{P}^p(K), K \in O_h\}, \quad (2.22a)$$

$$\Sigma_h = \{v \in L^2(O)^d \mid v|_K \in \Sigma(K) = \mathcal{P}^q(K)^d, K \in O_h\}. \quad (2.22b)$$

for some  $p, q \geq 0$ . The DG version of the Poisson problem is therefore to find  $(u_h, \sigma_h) \in V_h \times \Sigma_h$  such that for all  $K \in O_h$ :

$$\int_K \sigma_h \cdot \tau_h \, dx = - \int_K u_h \nabla \cdot \tau_h \, dx + \int_{\partial K} \hat{u}_h \eta_K \cdot \tau_h \, ds, \quad \forall \tau_h \in \Sigma(K) \quad (2.23a)$$

$$\int_K \sigma_h \cdot \nabla v_h \, dx = \int_K f v_h \, dx + \int_{\partial K} \hat{\sigma}_h \cdot \eta_K v_h \, ds, \quad \forall v_h \in V(K), \quad (2.23b)$$

where  $\eta_K$  is the outward normal to  $K$ . This is not sufficient to specify a DG method, since the *formal traces*  $\hat{u}_h$  and  $\hat{\sigma}_h$  on the mesh skeleton of  $O_h$  are not uniquely specified on the interior facets due to the discontinuous nature of the spaces in (2.22). The abstract system (2.23) forms the basis for virtually all DG methods for the Poisson problem. For the moment we look past the degrees of freedom  $p, q \geq 1$  in (2.22) since it is the choice of the traces (also

sometimes called *numerical fluxes*) that leads to different methods. To give some examples of these fluxes we require some additional notation. The jump  $\llbracket \cdot \rrbracket$  and average  $\{ \cdot \}$  of a function across a boundary are given by:

$$\begin{aligned} \llbracket v \rrbracket &= \begin{cases} v^+ \eta^+ + v^- \eta^- & \text{if } v \text{ is scalar,} \\ v^+ \cdot \eta^+ + v^- \cdot \eta^- & \text{if } v \text{ is vector-valued,} \end{cases} \\ \{v\} &= \frac{1}{2}(v^+ + v^-). \end{aligned}$$

On  $\partial O_h$ ,  $\llbracket v \rrbracket = \eta v$  and  $\{v\} = v$  where  $\eta$  is an outward normal. From this we can derive the identity  $\llbracket fg \rrbracket = \{f\} \llbracket g \rrbracket + \llbracket f \rrbracket \{g\}$ . A natural choice for the fluxes is:

$$\begin{aligned} \hat{u}_h &= \{u_h\} \quad \text{on} \quad \mathring{\mathcal{E}}_h(O_h), \quad \hat{u}_h|_{\partial O_h} = 0, \\ \hat{\sigma}_h &= \{\sigma_h\} \quad \text{on} \quad \mathcal{E}_h(O_h), \end{aligned}$$

leading to the method introduced in [22]. Another method [45] chooses:

$$\begin{aligned} \hat{u}_h &= \{u_h\} \quad \text{on} \quad \mathring{\mathcal{E}}_h(O_h), \quad \hat{u}_h|_{\partial O_h} = 0, \\ \hat{\sigma}_h &= \{\nabla u_h\} - \mu \llbracket u_h \rrbracket \quad \text{on} \quad \mathcal{E}_h(O_h), \end{aligned}$$

for some weighting function  $\mu : O_h \rightarrow \mathbb{R}$ , which if typically chosen as a function of the mesh resolution  $h$ , see e.g. Nitsche's method for imposing weak boundary condition [143]. The consistency, approximation and stability properties (the latter referring to coercivity of the discrete bilinear form on the approximation space) of a method are directly related to the choice of flux and are discussed at length in e.g. [10].

**Remark 18.** *A mixed method is not necessarily a DG method, and vice versa.*

Next we look at the class of discontinuous Petrov-Galerkin methods which were conceived in 2010 by Demkowicz and Gopalakrishnan [80, 77, 79, 240] in the context of transport equations. The *Petrov-Galerkin* in this method refers to the fact that the spaces  $X_h$  and  $Y_h$  are different and we shall shortly see the consequences of this choice. This method was proposed as an alternative to the standard upwind DG methods such as the previously mentioned work of Reed and Hill or [140, 160] and streamline diffusion methods e.g. [136, 47]. To communicate this we drawn upon the work [76, 78]. See also [81] which attempts to establish a more theoretical footing for the general DPG method with an application of the Poisson equation.

In the continuous setting it has been established that the bilinear operator in  $b(\cdot, \cdot)$  associated with the operator  $\mathcal{B}$  must satisfy an inf-sup condition and be injective in its first variable in order for a unique solution of (2.10) to exist. A main challenge for DG methods is that the strict positivity of the inf-sup constant is no longer guaranteed for a finite-dimensional approximation

space. For the moment we consider conforming finite element spaces  $X_h$  and  $Y_h$ . Taking a look at this condition for the equation (2.14):

$$0 < \alpha_h \|x_h\|_{X_h} \leq \sup_{0 \neq y_h \in Y_h} \frac{b(x_h, y_h)}{\|y_h\|_{Y_h}}, \quad \forall x_h \in X_h,$$

which reveals that the supremum is taken over a much smaller space so there is no way of knowing  $\alpha_h$  exists, or if it does, does not degenerate as  $h \rightarrow 0$ . The DPG class emphasises that the discrete analogues of  $X$  and  $Y$  in the (closed range) theorem 7 above are represented by different polynomial spaces and provides a way of reasoning about the existence of a *discrete inf-sup* constant. To this end we define the operator  $T : X \rightarrow Y$  by:

$$\langle Tx, y \rangle_Y = b(x, y), \quad \forall y \in Y.$$

which means that a discrete version of (2.10) with a strictly positive inf-sup constant can be posed as:

$$b(x_h, y) = f(y), \quad \forall y \in Y := \{Tx_h \mid x_h \in X_h\}.$$

To see why this is so, recall that the canonical (Riesz-Fréchet) isometry  $R : Y \rightarrow Y^*$  implicit in lemma 6 is defined by  $\langle Ru, v \rangle_{Y^*, Y} = \langle u, v \rangle_Y$ . Then observe that using the definition of the operator norm of  $\mathcal{B}x_h : Y \rightarrow \mathbb{R}$ ,  $x_h \in X_h$ :

$$\begin{aligned} \sup_{0 \neq y \in Y} \frac{|b(x_h, y)|}{\|y\|_Y} &= \|\mathcal{B}x_h\|_{Y^*} \\ &= \|R^{-1}\mathcal{B}x_h\|_Y \\ &= \frac{\langle R^{-1}\mathcal{B}x_h, R^{-1}\mathcal{B}x_h \rangle_Y}{\|R^{-1}\mathcal{B}x_h\|_Y} \\ &= \frac{\langle \mathcal{B}x_h, R^{-1}\mathcal{B}x_h \rangle_{Y^*, Y}}{\|R^{-1}\mathcal{B}x_h\|_Y} \\ &= \frac{b(x_h, R^{-1}\mathcal{B}x_h)}{\|R^{-1}\mathcal{B}x_h\|_Y}. \end{aligned} \tag{2.24}$$

So if we populate our test space with functions on the form  $R^{-1}\mathcal{B}x_h$  for  $x_h \in X_h$  the discrete problem satisfies a discrete inf-sup condition. This is exactly what the operator  $T := R^{-1}\mathcal{B}$  aims to accomplish so we have successfully moved the problem to computing, from  $X_h$ , the discrete space  $Y$  for the test functions.

**Remark 19** (Minimum residual and mixed problems). *We can formally cast (2.14) as the following optimisation problem:*

$$\min_{x_h \in X_h} \frac{1}{2} \|\mathcal{B}x_h - f\|_{Y^*}^2. \tag{2.25}$$

*Because it is impractical to deal with  $Y^*$  in (2.25) we use the Riesz isometry to formulate (2.25) as:*

$$\min_{x_h \in X_h} \frac{1}{2} \|R^{-1}[\mathcal{B}x_h - f]\|_Y^2. \tag{2.26}$$

Now setting the Gateaux derivative of (2.26) to zero we obtain:

$$\langle \mathbf{R}^{-1}[\mathcal{B}x_h - f], \mathbf{R}^{-1}\mathcal{B}\delta x_h \rangle_Y = 0, \quad \forall \delta x_h \in X_h, \quad (2.27)$$

which, using the definition of  $\mathbf{R}$  can be expressed as:

$$\langle \mathcal{B}x_h - f, \mathbf{R}^{-1}\mathcal{B}\delta x_h \rangle_{Y^*, Y} = 0. \quad (2.28)$$

We can therefore formulate (2.28) as:

$$b(x_h, y_h) = f(y_h), \quad y_h = \mathbf{R}^{-1}\mathcal{B}\delta x_h, \quad \forall \delta x_h \in X_h. \quad (2.29)$$

We of course arrive at the same conclusion that we now have to invert  $\mathbf{R}$  owing to the second equation which can be written as:

$$\langle \mathbf{T}\delta x_h, y \rangle_Y := \langle \mathbf{R}^{-1}\mathcal{B}\delta x_h, y \rangle_Y = b(\delta x_h, y), \quad \forall \delta x_h \in X_h, \forall y \in Y.$$

Instead of defining  $y_h = \mathbf{R}^{-1}\mathcal{B}\delta x_h$ , let us define the residual in  $Y$  as  $\psi = \mathbf{R}^{-1}[\mathcal{B}x_h - f]$ . Then we can write (2.27) as:

$$\langle \psi, \delta\psi \rangle_Y - b(x_h, \delta\psi) = -f(\delta\psi) \quad \forall \delta\psi \in Y, \quad (2.30a)$$

$$b(\delta x_h, \psi) = 0 \quad \forall \delta x_h \in X_h, \quad (2.30b)$$

for  $x_h \in X_h$  and  $\psi \in Y$ , representing the same system as (2.29).

Due to the remark above and the isometry in (2.24) the test functions derived from the Riesz isometry are called *optimal test functions*. Since  $\mathbf{T}$  is an infinite-dimensional operator we cannot practically evaluate it, let alone discuss its inversion. The mixed structure of (2.30) does however provide a fundamental basis for the design of a tractable numerical approximation of the original problem. For completeness we permit a slight digression from the DPG literature and briefly describe a roadmap towards solving mixed systems such as this using what is typically called Brezzi theory<sup>5</sup>. Some of what we present can also be found in e.g. [214] where a mixed form of the Stokes equation is considered.

**Roadmap 20** (Abstract mixed problems). *Consider the abstract mixed problem posed on the Hilbert spaces  $X \times Y$ :*

$$\begin{aligned} (x, y) &\in X \times Y \\ a(x, v) + b(v, y) &= l(v) \quad \forall v \in X, \end{aligned} \quad (2.31a)$$

$$b(x, w) = 0 \quad \forall w \in Y, \quad (2.31b)$$

for  $l \in X^*$  and continuous bilinear forms  $a : X \times X \rightarrow \mathbb{R}$ ,  $b : X \times Y \rightarrow \mathbb{R}$ . What we aim to provide here is an abridged version of the analysis carried out in [39, Chapter 4] tailored to the setting in this thesis.

<sup>5</sup> The equivalence of Brezzi and Babuška theory is discussed in the expository report [75].

1. First note any  $x \in \ker(\mathcal{B})$  satisfies (2.31b).
2. Restrict the test space in (2.31a) to  $\ker(\mathcal{B}) \subset X$ , since this implies  $b(v, y) = 0$  and simplifies the problem. (2.31a) becomes:

$$a(x, v) = l(v), \quad \forall v \in \ker(\mathcal{B}).$$

Usually,  $a(\cdot, \cdot)$  is assumed to be coercive on  $\ker(\mathcal{B})$  so Lax-Milgram finds a unique solution  $x \in \ker(\mathcal{B})$ .

3. Given a solution  $x$  we can prove that there exists a unique  $y \in Y$  such that (2.31b) is solved. Indeed, given that  $x$  is fixed this equation can be written as

$$b(v, y) = l(v) - a(x, v) = l_x(v). \quad (2.32)$$

Since  $l_x$  annihilates the kernel of  $\mathcal{B}$  on  $X$  i.e.  $l_x \in \ker(\mathcal{B})^\circ$  so to solve (2.32) we need to show that  $\mathcal{B}^*$  is an isomorphism from  $Y$  to  $\ker(\mathcal{B})^\circ$ . Usually this is done by appealing to the closed range theorem for surjective operators [39, Theorem 4.1.4 and Proposition 4.1.5] which would imply that the image  $\text{Im}(\mathcal{B})$  of  $\mathcal{B}$  is equal to  $Y^*$  which is trivially closed. This can be formulated as an inf-sup condition.

Specifically, the following inf-sup condition is required in step 2 and 4:

$$0 < \inf_{0 \neq y \in Y} \sup_{0 \neq x \in X} \frac{b(x, y)}{\|x\|_X \|y\|_Y}. \quad (2.33)$$

It is important to note that in the above, the trial function in (2.32) occupies the second "slot" of  $b$ . That is why we take the supremum over test functions  $x \in X$  in (2.33). While this is mathematically consistent we make a point to note this as a convenience to the reader.  $\square$

Most of the conditions of roadmap 20 are easily met for bilinear forms stemming from elliptic operators. The biggest challenge in mixed finite element systems is ensuring that a *discrete* inf-sup condition is satisfied. The objective is to approximate  $\mathbb{T}$  by a discrete analogue  $\mathbb{T}_h$  that ensures unique solutions to (2.30) as well as convergence under mesh refinement. In other words we wish to build a discrete test function space  $\mathbb{T}_h X_h$  that in some sense approximates  $Y$  leading to the so-called *practical DPG method*. To this end one typically uses a tool named after Michel Fortin:

**Theorem 21** (Fortin operator [97]). *Let (2.33) hold and consider conforming spaces  $X_h \subset X$  and  $Y_h \subset Y$ . Suppose there exists a linear operator  $\mathcal{F} : X \rightarrow X_h$  satisfying the following conditions:*

1.  $\forall x \in X, b(x - \mathcal{F}x, y_h) = 0, \quad \forall y_h \in Y_h,$
2.  $\|\mathcal{F}x\|_X \leq C\|x\|_X, \quad \forall x \in X$

with a constant  $C$  that does not degenerate as  $h \rightarrow 0$ . Then the inequality (2.33) holds with  $X_h$  and  $Y_h$  instead of  $X$  and  $Y$ .

*Proof.* Let  $y_h \in Y_h$ . Then:

$$\begin{aligned} \|y_h\|_{Y_h} &\lesssim \sup_{x \in X} \frac{b(x, y_h)}{\|x\|_X} \\ &= \sup_{x \in X} \frac{b(\mathcal{F}x, y_h)}{\|x\|_X} \\ &\leq \frac{1}{C} \sup_{x \in X} \frac{b(\mathcal{F}x, y_h)}{\|\mathcal{F}x\|_X}. \end{aligned}$$

□

Construction of Fortin operators is central to the DPG method. [109] is the main source of information on the design of such operators and practical DPG methods, showing examples for Laplace and linear elasticity equations. We make a point to emphasise that Fortin operators exist independently of the DPG method in the context of conventional DG methods. [38] provides Fortin design applicable to the Stokes problem. See also [166, 196] for more recent references. The MINI element in [9] is a good starting point for understanding how the design of an element is made in order to have good Fortin properties. For general second order problems we refer to [181]. A comprehensive exposition of DPG methods with various formulations of Maxwell's equations is shown in [52].

It may seem that the computational load of mixed methods such as the DPG method may be higher by expanding a single equation into a larger system. This is not always the case as the attractive block structure can result from DPG systems, and in particular the local nature of discontinuous spaces can localise computations and aid in parallel mass matrix assembly. Our sphere of investigation concerning DPG methods is limited to theory, and we do not discuss this interface with linear algebra any further, referring instead to [39, Section 7.3]. Later chapters describe numerical aspects when they become relevant.

# CHAPTER 3

---

## CURVE MATCHING

---

Closed curve matching is a central problem in shape analysis where the goal is to find a diffeomorphism bringing into alignment two closed curves in  $\text{Emb}(S^1, \mathbb{R}^d)$  called the *template* and the *target*, where  $\text{Emb}(S^1, \mathbb{R}^d)$  is the space of embeddings of the periodic domain  $S^1 := [0, 2\pi]$  into  $\mathbb{R}^d$  for  $d = 2$ . The motion of the shape - and therefore the motion of the diffeomorphism - is uniquely characterised by a one-parameter family of velocity field via the Lie algebra whose kinetic energy is measured as the norm of these velocity fields. It is therefore natural to define the *geodesic motion* of the curves as the path whose kinetic energy is minimised and solve the curve matching problem as an optimisation problem. This chapter is concerned with finding a suitable discrete analogue of a one-parameter curve of velocities that generates evolutions of a discrete template curve by forward integration in time. We show here that we can identify this entire family of velocities with a single conjugate momentum defined as a sufficiently smooth function over  $S^1$ . We therefore aim to describe a so-called *forward map*, taking as input the momentum and providing the curve of diffeomorphisms that match the template to the target curve. Obtaining an efficient finite element discretisation of this map then allows us to treat the curve matching problem as an inverse problem for the momentum.

Although curve matching has been studied before our contributions here are, to the best of our knowledge, novel. The early work of [232] geometrised curve matching and defined Riemannian distances on this space using the square root velocity framework; see also [176, 142, 206, 151]. As we mentioned before, previous work in this field also includes the LDDMM approach [108], where a gradient approach is taken to minimising a variational optimisation problem yielding smooth curve solutions. See the references therein for a wide array of practical applications of curve matching, in particular in the medical fields [27]. The application of finite element methods for matching purposes has also been done before [23], see also [26] for a detailed numerical framework for Sobolev metrics on the space of curves. See [170] for an overview of Riemannian metrics on the space of curves in the Hamiltonian frame.

The shape space used here is a subset of  $\text{Emb}(S^1, \mathbb{R}^d)$  and is studied in detail in

[170], but it is not the only option, see [24] for an overview. Typically, one would favour  $\text{Emb}(S^1, \mathbb{R}^d)/\text{Diff}(S^1)$  to identify curves in a parameter-independent manner but leave this as future work and return to this point later in the chapter. Indeed, varifold methods have gained traction as an elegant way of letting the metric over the shape space subsume reparameterisations [28, 55]. More recently, optimal transport methods have appeared in this context as well [25].

We recall that  $\text{id}$  denotes the identity operator and its interpretation will be clear from context. We shall use  $ds$  to denote the measure over time intervals as well as integral over the mesh skeleton when no confusion between the two can arise.

### 3.1 GEODESIC EQUATIONS

Throughout this chapter we set  $m = 3$  and  $d = 2$  to ease the notation, where the former denotes an order of differentiation explained later and the latter the dimension of the curves. Curves are elements of  $\text{Emb}(S^1, \Omega)$ , where  $\Omega$  is a connected convex subset of  $\mathbb{R}^d$ , with polygonal Lipschitz boundary  $\partial\Omega$ . We study *paths*, or maps,  $q \in Q = H^1(S^1)^d \hookrightarrow C^0(S^1)^d$  (where  $\hookrightarrow$  denotes an embedding) from a template curve  $\Gamma_0 \in \text{Emb}(S^1, \Omega)$  to a target curve  $\Gamma_1 \in \text{Emb}(S^1, \Omega)$  whose motion is restricted by the differential equation:

$$\dot{q}_t = u_t \circ q_t, \quad (3.1)$$

where  $u_t$ ,  $t \in [0, 1]$  is a family of time-dependent vector fields on  $\Omega$  with some prescribed spatial smoothness. A *geodesic path* between two such parameterised curves  $\Gamma_0$  and  $\Gamma_1$  is defined as a path minimising the associated kinetic energy in  $u$ :

$$\frac{1}{2} \int_0^1 \|u_t\|_V^2 dt,$$

where  $V$  is a Hilbert space specified later on such that  $\|u\|_V^2 = \int_\Omega u \cdot Bu \, dx$  for some linear differential operator  $B$ . We shall introduce the *conjugate momentum* (Lagrange multiplier)  $p_t \in Q^*$  which we assume has enough regularity so that it has at least Fréchet-Riesz representer in  $H^1(S^1)^d$  (also denoted  $p_t$ , with some abuse of notation). This is done so that  $p_t$  supports point evaluation so a strong form equation for  $p_t$  can be written down later on, and the Sobolev regularity also gains us important error bounds in the numerical analysis of our discretisations in this chapter. This reasoning also applies to the curve  $q$ . We extremise the following the functional:

$$S = \int_0^1 \frac{1}{2} \|u_t\|_V^2 + (p_t | \dot{q}_t - u_t \circ q_t)_{Q^*, Q} dt,$$

where  $(h|g)_{Q^*, Q} = \int_{S^1} h \cdot g \, d\theta$ . Taking variations i.e.  $\delta S = 0$  leads to what we shall call Hamilton's equations for curve matching for  $u_t \in V$ ,  $q_t \in Q$  and

$p_t \in H^1(S^1)^d$  for  $t \in [0, 1]$ :

$$(\delta p \mid \dot{q}_t - u_t \circ q_t)_{Q^*, Q} = 0, \quad \forall \delta p \in H^1(S^1)^d, \quad (3.2a)$$

$$(\dot{p}_t + \nabla u_t^\top \circ q_t p_t \mid \delta q)_{Q^*, Q} = 0, \quad \forall \delta q \in Q, \quad (3.2b)$$

$$\int_{\Omega} B u_t \cdot \delta u \, dx - (p_t \mid \delta u \circ q_t)_{Q^*, Q} = 0, \quad \forall \delta u \in V, \quad (3.2c)$$

for  $t \in [0, 1]$ , where  $[\nabla u_t^\top]^{ij} = \frac{\partial u_t^j}{\partial x_i}$ . In fact, since  $u_t$  is supported on  $\Omega$  it generates a diffeomorphism  $\varphi \in \text{Diff}(\Omega)$  [234] of the entire ambient space via:

$$\dot{\varphi}_t = u_t \circ \varphi_t, \quad \varphi_0 = \text{id}, \quad (3.3)$$

whose motion restricted to the curve  $q_0 \circ S^1$  equals the  $q_t \circ S^1$  at time  $t \in [0, 1]$ . As the kinetic energy measures distances between two elements of  $\text{Emb}(S^1, \Omega)$  via velocity defined over the entire field  $\Omega$ , we refer to this associated distance measure as an *outer* metric on the shape space  $\text{Emb}(S^1, \Omega)$ . This is in contrast to inner metrics intrinsically defined on the embedded shape [23]. We refer to initial value problem (3.2) as the *forward problem* for curve matching. Because the initial momentum  $p_0$  can be seen to uniquely encode the motion of  $u_t$  and  $q_t$  we therefore refer to *geodesic shooting* as the general problem of finding a suitable momentum such that we bring into alignment  $q_1$  and the target curve  $\Gamma_1$ .

In section 3.2 we discuss how to design norms preserving the key Lipschitz property of our velocity field and relate our strategy to the existing literature. Next, section 3.3 describes and proves convergence for a nonconforming finite element method [42, Section 10.3] for Hamilton's equations and contains the main results of this chapter. Finally section 3.4 describes the inverse problem of determining the optimal momentum matching two curves, and 3.5 summarises this chapter.

### 3.2 BUILDING AN OUTER METRIC

The operator  $B$  present in (3.2c) must be chosen carefully in order for the Banach space ODEs (3.2b) and (3.2a) governing  $p$  and  $q$ , respectively, to admit unique solutions for all time. Via classic Picard-Lindelöf analysis it is clear that for well-posedness of such ODEs require a pointwise Lipschitz on  $u_t$  and  $\nabla u_t$ , respectively. As such,  $u_t$  must occupy at least  $W^{2,\infty}(\Omega)^d$  when  $p_0 \in L^2(S^1)^d$  and  $q_0 \in L^\infty(S^1)^d$ , see [65, Theorem 5] (see also corollary 7 in this reference for other host spaces for  $p$  and  $q$ ). Dupuis [86] establishes sufficient conditions accomplishing the same in a Hilbertian setting. The Hilbertian setting is much more convenient for linear functional analysis and general well-posedness, in particular for finite element methods. This is in contrast with e.g.  $W^{1,\infty}(\Omega)^d$  which only a Banach space which, to the best of the author's ability, is not easy to approximate numerically. [157] approximates by means of a fixed point linearisation solutions to the nonlinear  $\infty$ -harmonic equation [21]. While the authors present a useful and convergent method, we shall not pursue this direction

as the finite element spaces themselves do not admit uniform mesh-independent Lipschitz constants. We also wish to avoid the nonlinearity of the operator as we rely heavily on a fully linear Hilbertian setting later on. The work of Oberman [185] shows some promise for a direction using the  $\infty$ -harmonic equation at least at the discrete level. However, owing to the nonlinearity of the  $\infty$ -Laplacian, integrating this into a Hamiltonian framework poses an analytical difficulty as the variational setting effectively requires derivatives of the  $\infty$ -Laplacian (in some appropriate sense). It is unclear what kind of regularity we can expect for the geodesics in this case, which, as we shall see, is central to forward integration of the geodesic equations. Linear operators preserve their structure in the variational framework and are therefore seen as most compatible for the problems in this chapter. The  $\infty$ -harmonic case is left as future research.

We request an operator  $B$  in such a way that a solution to (3.2c) ensures that these conditions are met, which in turn implies global existence and uniqueness of (3.2) by the references above. For  $d = 2$  the Sobolev embedding theorem reveals that  $H_0^3(\Omega)$  is contained in  $C^1(\bar{\Omega})$  and so is Lipschitz on the interior [239, Theorem 2.5.1]. As such, we want to describe a discretisation of (3.2c) ensuring a type of  $H^3$  regularity as the follow theorem shows. The following theorem states that a globally continuous piecewise  $H^3$  function is Lipschitz, which we will use to motivate a nonconforming finite element method for the velocity field.

**Theorem 22.** *Let  $O$  be a convex bounded Lipschitz domain in  $\mathbb{R}^d$  with polygonal boundary and  $O_h$  a triangulation thereof satisfying the regularity requirements from section 2.2.2. Suppose further that  $u \in C(\bar{O})^d$ ,  $u|_K \in H^3(K)^d$  for  $K \in O_h$ . Then  $u \in W^{1,\infty}(O)^d$ .*

*Proof.* The embedding theorem for homogeneous Sobolev spaces (i.e. with zero traces) into the space  $C^j(\bar{O})$  is well-known. However, since the trace  $\gamma_K u$  of  $u$  on  $\partial K$ ,  $K \in O_h$  may not be zero we appeal to a slightly different albeit standard result. By [1, Theorem 5.4], however,  $H^3(K) \hookrightarrow C_B^1(K)$ , where:

$$C_B^1(K) = \{u \in C^1(K) \mid D^\alpha u \text{ is bounded on } K, |\alpha| \leq 1\}.$$

This means any  $H^3(K)$  function has a continuous representative with almost everywhere bounded first derivatives on  $K$ . Since  $u \in C^0(\bar{O})$ ,  $u$  is a continuous function with its first derivative a.e. bounded, implying a Lipschitz condition.  $\square$

As we described in the beginning of this section,  $B$  must be chosen in such a way as to guarantee some Lipschitz regularity of the velocity so that the ODEs in (3.2) are well-posed. Critically, this must also hold at a discrete level. In light of the theorem above we approximate the space of velocity fields by a nonconforming finite element space where global high-order regularity across elements of the triangulation does not hold. This way we can guarantee the

necessary Lipschitz properties of our functions without having to impose higher-order *global* continuity of the finite-dimensional solution spaces. For instance, even  $C^1$ -conforming methods are not trivial to implement and automation of code generation for e.g. Argyris elements was only recently solved [150, 149]. To the best of the author's knowledge we have no access to a  $C^2$  finite element such as the one presented in [236] (also described in the more recent work [117, Section 2.1]). This is an example of a  $C^2$ -conforming finite element space with 25 local degrees of freedom which is practically infeasible to implement without a certain degree of automation. A globally  $H^1$ -conforming (in the context of finite elements this means  $C^0$ -conformity) piecewise  $H^3$  finite element space satisfying the theorem above is presented in [231, Section 4]. Recalling that we set  $m = 3$  throughout this chapter, based on the existence of this finite element we now choose  $B : H^{2m}(\Omega)^d \rightarrow L^2(\Omega)^d$  as follows:

$$\langle Bu, v \rangle_{0,\Omega} = \sum_{i=1}^d \int_{\Omega} \langle (\mathbf{id} - b\Delta)^{2m} u^i, v^i \rangle dx, \quad (3.4)$$

for some length-scale  $b \in \mathbb{R}^+$ .  $\langle \cdot, \cdot \rangle$  is a Euclidean dot product (acting on either scalars, vectors or tensors depending on the context). We summarise why our choice of  $B$  is defined in this manner. Since we are interested in integrating the system (3.2) forward in time we know that  $u_t$  and  $\nabla u_t$  must be Lipschitz in space for  $t \in [0, 1]$ . As explained before, this is clear by inspection of the ODEs (3.2a) and (3.2b). This can be put succinctly by saying we require  $u_t \in W^{2,\infty}(\Omega)^d$  for  $t \in [0, 1]$ . By the Sobolev embedding theorem,  $H^{2m}(\Omega)^d$  is contained in  $W^{2,\infty}(\Omega)^d$ . Since Sobolev spaces are easier to work with owing to their Hilbert structure, we therefore choose the domain of  $B$  to be in  $H^{2m}(\Omega)^d$ . The "iterated Helmholtz" structure in (3.4) is attractive for qualitative reasons explained in the following remark.

**Remark 23.** *The length-scale controls, in some sense, the magnitude of the map from momentum to velocity. A low value of  $b$  localises the velocity field around the curve and decaying rapidly away from it. Conversely, higher values of  $b$  allows the diffeomorphism induced by the velocity to move the domain further away from the curve. In applications we tune this to the size of the domain and the magnitude of the momenta. An example of this is given in figure 3.1. In the context of curve matching we can think of the momentum as being a measure-valued object defined on a discrete initial curve which via  $B^{-1}$  is extended to the whole domain  $\Omega$  to generate the outer metric.*

Note that from a theoretical point of view we could have chosen  $B$  as follows:

$$\langle Bu, v \rangle_{0,\Omega} = \sum_{i=1}^d \int_{\Omega} \langle -\Delta^{2m} u^i, v^i \rangle dx,$$

subject to additional boundary conditions closing the non-trivial kernel if we were to use it in (3.2c). This would still yield a velocity with the desired regularity, so the definition in (3.4) is a modelling choice. However, the length-scale

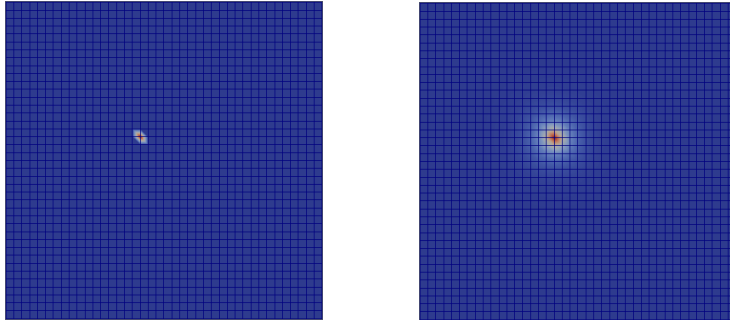


FIGURE 3.1: Left: a Dirac delta  $\delta_{x_0}$  centered around a single point  $x_0$  (interpolated to a continuous space to be visible). Right:  $(\mathbf{id} - \alpha\Delta)^{-1}\delta_{x_0}$  for  $\alpha = 0.05$ . The smoothing effect is clearly seen. Here we use a uniform resolution with  $h = 0.025$

$b$  is introduced here to control, in a sense, how "close" the operator  $B$  is to the identity.

Using this we define the bilinear form:

$$a(u, v) = \sum_{i=1}^d \int_{\Omega} \sum_{j=0}^m b_j \binom{m}{j} \langle D^j u^i, D^j v^i \rangle dx, \quad (3.5)$$

where  $D^0 = \mathbf{id}$ , and

$$D^j = \begin{cases} \nabla D^{j-1} & j \text{ is odd,} \\ \nabla \cdot D^{j-1} & j \text{ is even,} \end{cases}$$

and  $b_0 = 1$ ,  $b_j = b^j$ . For the operator  $A : H^m(\Omega)^d \rightarrow L^2(\Omega)^d$  defined by  $\langle Au, v \rangle_{0,\Omega} = a(u, v)$  we have the equivalence:

$$\langle A^2 u, v \rangle_{0,\Omega} = \langle Bu, v \rangle_{0,\Omega},$$

for sufficiently smooth functions  $u$  and  $v$  with vanishing  $j^{\text{th}}$  order derivatives on  $\partial\Omega$ , for  $j = 0, \dots, m-1$ . While  $H^{2m}(\Omega)^d \hookrightarrow W^{2,\infty}(\Omega)^d$  we restrict the domain of  $B$  to the smaller space  $H^{2m}(\Omega)^d$  for reasons made clear in the following section. Equipped with the nature of the operator  $B$  (via (3.5)) we can pose the abstract Hamiltonian system (3.2) using the function spaces and definitions introduced above. Indeed, the following problem rewrites (3.2) in a way that uses the notation above which will be more convenient once we discretise later on:

**Problem 1 (Continuous Hamiltonian system)**

Find  $u_t \in H_0^m(\Omega)^d$ ,  $w_t \in H^m(\Omega)^d$ ,  $\sigma_t \in H^{m-1}(\Omega)^{d \times d}$ ,  $q_t \in Q$  and  $p_t \in H^1(S^1)^d$  for  $t \in [0, 1]$  such that:

$$a(u_t, v) = \langle w_t, v \rangle_{0, \Omega} \quad \forall v \in H_0^3(\Omega)^d, \quad (3.6a)$$

$$a(w_t, v) = \int_{S^1} p_t \cdot v \circ q_t \, d\theta \quad \forall v \in H^3(\Omega)^d, \quad (3.6b)$$

$$\dot{q}_t = u_t \circ q_t \quad \text{on } S^1, \quad (3.6c)$$

$$\dot{p}_t + \sigma_t \circ q_t p_t = 0 \quad \text{on } S^1, \quad (3.6d)$$

$$\langle \sigma_t - \nabla u_t^\top, v \rangle_{m-1, \Omega} = 0 \quad \forall v \in H^{m-1}(\Omega)^{d \times d}, \quad (3.6e)$$

$$\begin{aligned} (\sigma_t - \nabla u_t)^{i,j}|_{\partial\Omega} \\ = \eta_{\partial\Omega} \cdot \nabla (\sigma_t - \nabla u_t)^{i,j}|_{\partial\Omega} = 0, \quad i, j = 1, d, \end{aligned} \quad (3.6f)$$

where  $\eta_{\partial\Omega}$  is the outward normal to  $\partial\Omega$ . The last condition simply prescribes Dirichlet and Neumann conditions for each component of  $\sigma_t$ . While  $\sigma_t$  defined in (3.6e) is simply a surrogate for  $\nabla u_t^\top$  (and that the boundary conditions for  $\sigma_t$  are in fact homogeneous), this formulation will be useful when we discretise this system. The velocity from (3.6a) generates the diffeomorphism  $\varphi_t$  of the domain, where the curve is its motion restricted to the initial condition  $q_0$ . As usual this diffeomorphism is the identity when  $t = 0$ .

**Remark 24.** Recall that  $d = 2$  and suppose  $g \in C^0(S^1)^d$  forms an embedded curve  $G = g \circ S^1$  in a sufficiently large bounded domain  $\Omega$  (away from the boundary  $\partial\Omega$ ), and let  $\Omega_t^-$ ,  $\Omega_t^+$  denote the inside and outside of the curve. Suppose further that  $g$  is such that we can define trace operators  $\gamma_G^- : \Omega_t^- \rightarrow H^{1/2}(G)$  and  $\gamma_G^+ : \Omega_t^+ \rightarrow H^{1/2}(G)$ . A functional of the form:

$$H^1(\Omega) \ni w \mapsto \int_{S^1} \gamma_G^-[w] \circ g(\theta) \, d\theta,$$

occupies  $H^{-1}(\Omega)$ , and, a fortiori, in  $H^{-3}(\Omega)$ . It is of course equivalent to use the trace  $\gamma_G^+$  in the example above. As such, these cannot be represented as inner products in  $L^2(\Omega)$ . The  $H^{-1}(\Omega)$  regularity prohibits global higher-order elliptic regularity beyond  $H^3(\Omega)$  of the solution  $w_t$  to (3.6b). We refer any interested readers to [115, 153] for a priori regularity estimates for general elliptic equations.

The following theorem makes the observation that the momentum is fully characterised by its initial condition and the flow map.

**Theorem 25.** Assume problem 1 is well-posed, that  $\varphi$  solves (3.1) and  $p_0 \in H^1(S^1)^d$ . Then the solution to (3.6d) is given by  $p_t = J_t^{-\top} p_0$ , where  $J_t = \nabla \varphi_t \circ q_0$ .

*Proof.* First we note that for all  $t \in [0, 1]$ , the smoothness of  $u_t$  implies that  $\varphi_t$  is a  $\mathcal{C}^1$  diffeomorphism so it has a well-defined non-zero Jacobian  $\nabla\varphi_t$  which means its inverse is well-defined. This follows from the inverse mapping theorem; we recommend [234, Chapter 8] for an exposition of this theory in the context of shape analysis. Note also that  $q_0 \in Q$ , which is a differentiable map. Recall that the momentum satisfies the strong form equation:

$$\dot{p}_t + \nabla u_t^\top \circ q_t p_t = 0, \quad \text{on } S^1.$$

Note also that  $\varphi_t \circ q_0 = q_t$  as  $\varphi_0 \circ q_0 = q_0$  and both  $\varphi$  and  $q$  are transported by the same velocity (cf. (3.1) and (3.6c)). We verify the ansatz by simple computation:

$$\begin{aligned} \dot{p}_t + \nabla u_t^\top p_t &= \dot{J}_t^{-\top} p_0 + \nabla u_t^\top \circ q_t J_t^{-\top} p_0 \\ &= -J_t^{-\top} d(J_t^\top) J_t^{-\top} p_0 + \nabla u_t^\top \circ q_t J_t^{-\top} p_0 \\ &= -J_t^{-\top} (dJ_t)^\top J_t^{-\top} p_0 + \nabla u_t^\top \circ q_t J_t^{-\top} p_0 \\ &= -J_t^{-\top} (\nabla u_t \circ q_t J_t)^\top J_t^{-\top} p_0 + \nabla u_t^\top \circ q_t J_t^{-\top} p_0 \\ &= -J_t^{-\top} J_t^\top \nabla u_t^\top \circ q_t J_t^{-\top} p_0 + \nabla u_t^\top \circ q_t J_t^{-\top} p_0 \\ &= -\nabla u_t^\top \circ q_t J_t^{-\top} p_0 + \nabla u_t^\top \circ q_t J_t^{-\top} p_0 \\ &= 0. \end{aligned}$$

In the first step we used the formula for the derivative of the inverse of a matrix [201].  $\square$

### 3.3 NONCONFORMING DISCRETISATION

Integration of a discretised version of (3.6) requires a suitable finite-dimensional analogue of the initial condition  $q_0$ . We approximate it by a continuous piecewise linear curve,  $q_{0h}$ , and establish the following relationship which holds throughout this chapter:

**Assumption 26.** *We assume that there exists a discrete piecewise affine initial curve  $q_{0h}$  approximating the continuous curve with linear speed, i.e.  $\lim_{h \rightarrow 0} \|q_0 - q_{0h}\|_{0,S^1} = 0$  and  $\|q_0 - q_{0h}\|_{0,S^1} \leq h \|q_0\|_{1,S^1}$ , where  $h$  denotes a mesh resolution parameter.*

We present a moving mesh method where we generate a sequence of linear diffeomorphisms  $\varphi_{th}$  so that the evolution of the curve is described by  $q_{th} = \varphi_{th} \circ q_{0h}$ ,  $t \in [0, 1]$ . We shall assume that there exists, for each  $t \in [0, 1]$ , a triangulation  $\Omega_{th}$  (part of a shape-regular, quasi-uniform family) such that for some set  $E_h \subset \mathring{\mathcal{E}}_h(\Omega_{th})$ ,  $\cup_{e \in E_h} e = q_{th} \circ S^1$ . This simply means that  $q_{th}$  is aligned with the interior of the mesh skeleton. The linear diffeomorphism  $\varphi_{th}$  is such that  $\varphi_{th}|_{\partial\Omega_{th}} = \text{id}$  so that the boundaries are fixed. The task is now to compute this family of linear diffeomorphisms.

We shall use a finite element space from [231] to discretise the two sixth order problems in (3.6a) and (3.6b) and we briefly summarise its construction. First we recall some notation. For a triangle  $K$ , let  $\mathcal{P}_i(K)$  denote the space of  $i^{\text{th}}$  order polynomials on  $K$ . Let  $q_K$  denote the *bubble function* on  $K$  which is a non-negative cubic polynomial that is zero on  $\partial K$ ; see e.g. [89] for more details. In the following,  $\frac{\partial}{\partial \nu_e}$  denotes a derivative in the direction of the outward normal  $\nu_e$  on an edge  $e$  in the mesh skeleton of a triangulation. We define the following shape function space on a triangle  $K \in \Omega_{th}$ :

$$\tilde{P}_K^{(3,2)} := \mathcal{P}_3(K) + q_K \mathcal{P}_1(K) + q_K^2 \mathcal{P}_1(K).$$

By [231, Lemma 4.1], the following degrees of freedom determine a function  $v \in \tilde{P}_K^{(3,2)}$ :

$$\frac{1}{|e|} \int_e \frac{\partial^2 v}{\partial \nu_e^2} ds \quad e \text{ is an edge of } \partial K, \quad (3.7a)$$

$$\nabla v(a) \quad a \text{ is a vertex of } \partial K, \quad (3.7b)$$

$$\frac{1}{|e|} \int_e \frac{\partial v}{\partial \nu_e} ds \quad e \text{ is an edge of } \partial K, \quad (3.7c)$$

$$v(a) \quad a \text{ is a vertex of } \partial K. \quad (3.7d)$$

We now define the scalar finite element space  $\tilde{V}_h$  defined over  $\Omega_{th}$  as consisting of all functions  $v_h|_K \in \tilde{P}_K^{(3,2)}$ ,  $K \in \Omega_{th}$ , with the degrees of freedom above being equal to zero if either the edge or vertex lies on the boundary of  $\Omega_{th}$  (see also [231, Equation 4.8] for the definition of  $\tilde{V}_h$ ): This element is locally to  $K$  a seventh-order polynomial. It is designed in such a way that makes it possible to describe convergent  $H^3$ -nonconforming FEM. Our vector-valued finite element space  $\tilde{V}_{th}^d$  is then the space whose components occupy the scalar finite element space  $\tilde{V}_h$ . As shown in the following corollary, this is a continuous finite element space.

**Corollary 27.**  $\tilde{V}_h \hookrightarrow C^0(\Omega)$ .

*Proof.* Let  $K, K'$  be two elements of  $\Omega_{th}$  that share an edge  $e$ . As shown in [231], there are 8 degrees of freedom on this edge. Let  $v$  and  $v'$  be the restriction of a function  $w \in \tilde{V}_h$  to  $\partial K$  and  $\partial K'$ , where these functions occupy  $\mathcal{P}^3(\partial K)$  and  $\mathcal{P}^3(\partial K')$ , respectively. Values and derivative at the vertices are continuous in  $\tilde{V}_h$  in the sense that they are unique, so  $v$  and  $v'$  are order 3 polynomials that agree on 4 values which is only possible if  $v = v'$  everywhere on  $e$ . Since  $e$  was arbitrary,  $\tilde{V}_h \hookrightarrow C^0(\Omega)$ .  $\square$

Although the finite element space above is continuous, this does not extend to higher-order derivatives. Nevertheless, the following central result highlights its *weak* continuity properties:

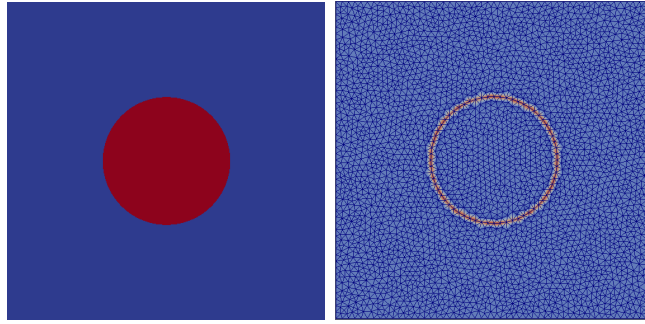


FIGURE 3.2: Left: plot of an indicator function of the interior of the initial curve defined by  $q_{0h} \circ S^1$ . Right: the triangulated domain  $\Omega_{0h}$  where we have highlighted the initial template curve  $q_{0h} \circ S^1$ .

**Lemma 28** (Weak continuity [231, Lemma 4.2]). *Let  $\Omega_{th}$  denote the mesh and  $e \in \mathring{\mathcal{E}}_{th}$  be an interior edge of a triangle  $K$ . For any component  $u^i$ ,  $i = 1, 2$  of  $u \in \tilde{V}_{th}^d$  and  $K' \in \Omega_{th}$  such that  $e \cap K' \neq \emptyset$ ,*

$$\int_e \partial^\alpha u^i|_K ds = \int_e \partial^\alpha u^i|_{K'} ds, \quad |\alpha| = 1, 2.$$

When  $e \in \partial\Omega_{th}$ ,

$$\int_e \partial^\alpha u^i|_K ds = 0.$$

**Remark 29.** *This importance of this result cannot be understated. Weak continuity is the exact reason why we will be able to talk about convergence for nonconforming methods. A naïve piecewise  $H^3$  finite element space constructed from piecewise cubic polynomials functions  $\mathcal{P}^3(K)$ ,  $K \in \Omega_{th}$  does not lead to the same results.*

We equip  $\tilde{V}_{th}^d$  with the broken  $H^3$  norm  $\|\cdot\|_{m, \Omega_{th}}$  dominating all derivatives up to third order. We shall also define a matrix-valued finite element space  $\tilde{V}_{th}^{d \times d}$  whose components occupy  $\tilde{V}_h$ , see the aforementioned reference. Appendix B verifies the implementation of the finite element space  $\tilde{V}_{th}^d$  used to produce the results in this thesis.

**Definition 30** (Linear diffeomorphism of mesh-aligned curves). *Provided the existence of a family of discrete velocities  $u_{th} \in \tilde{V}_{th}^d$  for  $t \in [0, 1]$  we define:*

$$\dot{\varphi}_{th} = \mathbb{I}_\Omega^{1,d}[u_{th}], \quad (3.8)$$

with  $\varphi_{0h} = \text{id}$ , where  $\mathbb{I}_\Omega^{1,d} : \mathbf{C}^0(\Omega)^d \rightarrow \mathbf{P}_{\Omega_{th}}^{1,d}$  is the interpolation operator of continuous vector fields onto the piecewise linear  $\mathbf{C}^0(\Omega)$ -conforming finite element space  $\mathbf{P}_{\Omega_{th}}^{1,d}$ . Restrictions of functions in  $\mathbf{P}_{\Omega_{th}}^{1,d}$  to  $K \in \Omega_{th}$  are therefore linear. Note here that the initial mesh  $\Omega_{0h}$  is such that the union of a subset of its mesh skeleton traces out the initial curve  $q_{0h}$ , see e.g. figure 3.2. What is key is that for all time, owing to (3.8),  $\Omega_{th}$  is such that  $q_{th} \circ S^1$  forms a closed curve of  $\mathcal{E}_{th} \setminus \partial\Omega_{th}$ .

The discrete curves  $q_{th}$  must therefore be piecewise affine and means that the load vector in our numerical method comes from integration over a discrete piecewise linear curve  $q_{th} \circ S^1$  rather than  $S^1$  itself. We discuss this restriction in section 3.3.1 and now continue presenting the proposed method.

**Theorem 31.** *Lipschitz constants of functions in  $\tilde{V}_{th}^d$  are preserved by  $\mathbb{I}_\Omega^{1,d}$  independently of  $h$ .*

*Proof.* Let  $v \in \tilde{V}_{th}^d$ . We want show the following holds with a constant independent of  $h$ :

$$\|\mathbb{I}_\Omega^{1,d}[v]\|_{W^{1,\infty}(\Omega)} \lesssim \|v\|_{W^{1,\infty}(\Omega)}.$$

It is sufficient to bound  $\|\mathbb{I}_K^{1,d}[v]\|_{1,K}$  by  $\|v\|_{W^{1,\infty}(\Omega)}$ ,  $K \in \Omega_{th}$  because the objects we are studying are finite-dimensional. For  $j = 0, 1$  we have:

$$|\mathbb{I}_\Omega^{1,d}[v]|_{j,K} \leq |v|_{j,K} + |v - \mathbb{I}_\Omega^{1,d}[v]|_{j,K}.$$

When  $j = 0$ , the semi-norm coincides with the usual  $L^2$  norm on  $K \in \Omega_{th}$ . By standard interpolation error results stated in e.g. Brenner & Scott's textbook [42, Theorem 4.4.4] we have

$$|v - \mathbb{I}_\Omega^{1,d}[v]|_{j,K} \lesssim |v|_{j,K} \quad j = 0, 1,$$

with a constant independent of  $h_K$ , so it follows that

$$|\mathbb{I}_\Omega^{1,d}[v]|_{j,K} \lesssim 2|v|_{j,K}, \quad j = 0, 1.$$

Squaring, summing this over  $j = 0, 1$  and taking the root we can say:

$$\|\mathbb{I}_\Omega^{1,d}[v]\|_{W^{1,\infty}(\Omega)} \simeq \|\mathbb{I}_\Omega^{1,d}[v]\|_{1,\Omega_{th}} \lesssim \|v\|_{W^{1,\infty}(\Omega)}.$$

□

**Definition 32.** Let  $\mathcal{J}$  be a finite index set of cardinality  $N_{\mathcal{J}}$  so that  $S_i^1$ ,  $i \in \mathcal{J}$  denotes a segment of  $S^1$  corresponding to a linear segment  $q_{0h} \circ S_i^1$  of the piecewise linear curve  $q_{0h} \circ S^1 = \cup_{i \in \mathcal{J}} q_{0h} \circ S_i^1$ .

**Assumption 33.** We assume that there exists a positive number  $C_0$  such that for all  $t \in [0, 1]$ :

$$C_0^{-1} \leq N_{\mathcal{J}} \min_{i \in \mathcal{J}} h_{q_{th} \circ S_i^1},$$

where  $h_{q_{th} \circ S_i^1}$  denotes the length of the edge  $q_{th} \circ S_i^1$ .

This assumption means that the smallest edge of the discrete curve  $q_{th} \circ S^1$  is proportional to the number of edges of which it is comprised. Based on this we set up some notation about linear interpolation and orthogonal projection.

**Definition 34** (Interpolation and projection on  $S^1$ ). *We define  $\mathbf{S}_h^q$  as the globally continuous space of linear polynomials on  $S^1$  given by the usual "hat" functions in one dimension so that  $\mathbf{S}_h^q \hookrightarrow \mathbf{C}^0(S^1)$  and functions in  $\mathbf{S}_h^q$  are linear on each  $S_i^1$ ,  $i \in \mathcal{J}$  ( $S^1$  is a periodic 1-dimensional domain, so we simply have  $S_i^1 \subset [0, 2\pi]$ ). We let  $\mathbf{S}_h^p$  denote the space of piecewise constant vector fields on  $S^1$ .*

We let  $\mathbb{P}_{\mathbf{S}_h^p} : L^2(S^1)^d \rightarrow \mathbf{S}_h^p$  be the canonical orthogonal projection operator onto the space  $\mathbf{S}_h^p$ . We let  $\mathbb{I}_{\mathbf{S}_h^p} : \mathbf{C}^0(S^1)^d \rightarrow \mathbf{S}_h^p$  denote the canonical nodal interpolant. The spaces and operators here depend on  $N_{\mathcal{J}}$  but we suppress the explicit dependence to ease notation.

To pose a discrete version of (3.6) we replace the spaces  $Q$  and  $Q^*$  with affine-invariant finite element spaces over  $S^1$  for the momentum and curve. As implied by the notation, we choose  $\mathbf{S}_h^q$  for  $Q$  and  $\mathbf{S}_h^p$  for  $Q^*$ . Next we write the broken analogue  $a_h$  of 3.5, where we explicitly omit the time dependence of the domain for ease of notation:

$$a_h(u, v) = \sum_{i=1}^d \sum_{K \in \Omega_{th}} \int_K \sum_{j=0}^m b_j \binom{m}{j} \langle D^j u^i, D^j v^i \rangle dx, \quad (3.9)$$

This defines the following *energy norm* as well:

$$\|\cdot\|_{m, \Omega_{th}}^2 = a_h(\cdot, \cdot),$$

and we observe the following norm equivalences based on boundedness of  $b$ :

$$\sqrt{a_h(\cdot, \cdot)} \simeq \|\cdot\|_{m, \Omega_{th}}, \quad \sqrt{a(\cdot, \cdot)} \simeq \|\cdot\|_{m, \Omega}.$$

We remark that a Poincaré-type lemma holds for functions in  $\tilde{V}_h$ :

**Lemma 35** (Poincaré [231, Lemma 4.3]). *It holds that:*

$$\|v_h\|_{m, \Omega_{th}} \lesssim \|v_h\|_{0, \Omega_{th}} + |v_h|_{m, \Omega_{th}}, \quad \forall v_h \in \tilde{V}_h. \quad (3.10)$$

Using (3.10) for each component  $i$  of some  $v_h \in \tilde{V}_{th}^d$  we obtain:

$$\|v_h^i\|_{m, \Omega_{th}}^2 \lesssim \|v_h^i\|_{0, \Omega}^2 + |v_h^i|_{m, \Omega_{th}}^2,$$

so summing over the components  $i$  we get:

$$\|v_h\|_{m, \Omega_{th}}^2 \lesssim \langle v_h, v_h \rangle_{0, \Omega_{th}} + \langle v_h, v_h \rangle_{m, \Omega_{th}} \quad \forall v_h \in \tilde{V}_{th}^d.$$

and as such, the operator  $B$  could have been chosen as simply  $I - \Delta^3$ . However, the scaled inner product-type nature of  $a(\cdot, \cdot)$  (resp.  $a_h(\cdot, \cdot)$ ) is convenient for computation as it offers the notion of a length-scale as explained above.

The discrete version of problem 1 is defined by the following:

**Problem 2 (Discrete Hamiltonian system)**

Find  $u_{th} \in \tilde{V}_{th}^d$ ,  $w_t \in \tilde{V}_{th}^d$ ,  $\sigma_{th} \in \tilde{V}_{th}^{d \times d}$ ,  $q_{th} \in \mathbf{S}_h^q$  and  $p_{th} \in \mathbf{S}_h^p$  for  $t \in [0, 1]$  such that:

$$a_h(u_{th}, v_h) = \langle w_{th}, v_h \rangle_{0,\Omega}, \quad \forall v_h \in \tilde{V}_{th}^d, \quad (3.11a)$$

$$a_h(w_{th}, v_h) = \int_{S^1} p_{th} \cdot v_h \circ q_{th} \, d\theta, \quad \forall v_h \in \tilde{V}_{th}^d, \quad (3.11b)$$

$$\dot{q}_{th} = \mathbb{I}_\Omega^{1,d}[u_{th}] \circ q_{th}, \quad (3.11c)$$

$$\dot{p}_{th} + \mathbb{P}_{\mathbf{S}_h^p}[\sigma_{th} \circ q_{th} p_{th}] = 0, \quad (3.11d)$$

$$\langle \sigma_{th} - \nabla u_{th}^\top, v_h \rangle_{m-1, \Omega_{th}} = 0, \quad \forall v_h \in \tilde{V}_{th}^{d \times d}, \quad (3.11e)$$

$$\begin{aligned} (\sigma_{th} - \nabla u_{th})^{i,j}|_{\partial\Omega_{th}} \\ = \eta_{\partial\Omega_{th}} \cdot \nabla(\sigma_{th} - \nabla u_{th})^{i,j}|_{\partial\Omega_{th}} = 0, \quad i, j = 1, d. \end{aligned} \quad (3.11f)$$

Again here we take  $\varphi_{th}$  as the identity when  $t = 0$ . Note that evaluation of elements of  $\tilde{V}_{th}^d$  on the curve  $q_{th}$  should be understood in sense of the traces as a true restriction is not defined in a pointwise sense.

**Remark 36.** *The Hamiltonian structure of the continuous system (3.6) is such that the norm of the velocity field is a conserved quantity [65]. There are several reasons why this is not the case for the discrete system (3.11). First, because we use a moving mesh method we cannot expect the same to happen as there is an implicit time dependence in the coordinate field in (3.9). That is, for any solution  $u_{th}$  to (3.11a), the quantity  $a_h(u_{th}, u_{th})$  is not conserved by solving the equations above. Secondly, a conservative discrete method would also require the test function in (3.11b) to be replaced by its interpolant to even begin to recover correspondence with a discrete functional. Evaluation of such a source term is, however, not supported from a computational point of view. We return to this point in section 3.3.1.*

**Corollary 37.** *Provided the right-hand sides of the problems (3.6b) and (3.11b) are well-defined, there exist unique solutions  $w_t \in H^3(\Omega)^d$  and  $w_{th} \in \tilde{V}_{th}^d$  to these equations, respectively.*

*Proof.* The proof is trivial since the continuous bilinear forms associated with  $B$  (resp.  $B_h$ ) controls all  $j^{\text{th}}$  order terms for  $j = 0, \dots, 3$  as is therefore naturally coercive on their respective spaces.  $\square$

**Theorem 38.** *There exists an  $\epsilon > 0$  such that the system (3.11) admits unique solutions in the time interval  $[0, \epsilon]$ .*

*Proof.* We verify local Lipschitz dependencies which by Picard-Lindelöf is sufficient for existence and uniqueness, see e.g. [152]. Solutions (3.11a) clearly have a local Lipschitz dependence of the right-hand sides. Let  $w^i$  denote a solution of (3.11b) given by fixed momentum and curve  $p^i, q^i, i = 1, 2$ . In this proof the superscript simply denotes different variables and does not denote components. We can use the fact that  $w^1 - w^2$  are Lipschitz functions to show a Lipschitz dependence of the variable  $w$  on the curve and momentum:

$$\begin{aligned} \|w^1 - w^2\|_{m, \Omega_{th}}^2 &= \int_{S^1} p^1 \cdot (w^1 - w^2) \circ q^1 d\theta - \int_{S^1} p^2 \cdot (w^1 - w^2) \circ q^2 d\theta \\ &\leq \|w^1 - w^2\|_{m, \Omega_{th}} \left( \int_{S^1} p^1 \cdot q^1 d\theta - \int_{S^1} p^2 \cdot q^2 d\theta \right), \end{aligned}$$

implying:

$$\|w^1 - w^2\|_{m, \Omega_{th}} \leq \|p^1\|_{0, S^1} \|q^1 - q^2\|_{0, S^1} + \|q^2\|_{0, S^1} \|p^1 - p^2\|_{0, S^1}. \quad (3.12)$$

In similar vein it is easy to check that  $q_{th}$  is Lipschitz in  $u_{th}$  by using the equation (3.11c):

$$\begin{aligned} \|\mathbb{I}_\Omega^{1,d}[u^1] \circ q^1 - \mathbb{I}_\Omega^{1,d}[u^2] \circ q^1\|_{0, S^1} &\leq \|\mathbb{I}_\Omega^{1,d}[u^1] \circ q^1 - \mathbb{I}_\Omega^{1,d}[u^1] \circ q^2\|_{0, S^1} \\ &\quad + \|\mathbb{I}_\Omega^{1,d}[u^1] \circ q^2 - \mathbb{I}_\Omega^{1,d}[u^2] \circ q^1\|_{0, S^1} \\ &\lesssim \|u^1\|_{m, \Omega_{th}} \|q^1 - q^2\|_{0, S^1} + \|u^1 - u^2\|_{m, \Omega_{th}}, \end{aligned}$$

where we have used that  $\mathbb{I}_\Omega^{1,d}$  preserves the Lipschitz constant of its argument (theorem 31), so we have a local Lipschitz condition for the projected velocities. Next, because  $\mathbb{P}_{S_h^p}$  is a bounded linear operator:

$$\begin{aligned} \|\mathbb{P}_{S_h^p}[\sigma^1 \circ q^1 p^1] - \mathbb{P}_{S_h^p}[\sigma^2 \circ q^2 p^2]\|_{0, S^1} &\leq \|\sigma^1 \circ q^1 p^1 - \sigma^2 \circ q^2 p^2\|_{0, S^1} \\ &= \|\sigma^1 \circ q^1 p^1 + \sigma^2 \circ q^2 p^1 \\ &\quad - \sigma^2 \circ q^2 p^1 - \sigma^2 \circ q^2 p^2\|_{0, S^1}. \end{aligned}$$

Now using boundedness of  $\sigma^1$  and  $\sigma^2$  and adding and subtracting  $\sigma^1 \circ q^2$ :

$$\begin{aligned} \|\mathbb{P}_{S_h^p}[\sigma^1 \circ q^1 p^1] - \mathbb{P}_{S_h^p}[\sigma^2 \circ q^2 p^2]\|_{0, S^1} &\leq \|\sigma^1 \circ q^1 - \sigma^2 \circ q^2\|_{\infty, S^1} \|p^1\|_{0, S^1} \\ &\quad + \|\sigma^2 \circ q^2\|_{\infty, S^1} \|p^1 - p^2\|_{0, S^1} \\ &\leq (\|\sigma^1 \circ (q^1 - q^2)\|_{\infty, S^1} \\ &\quad + \|(\sigma^1 - \sigma^2) \circ q^2\|_{\infty, S^1}) \|p^1\|_{0, S^1} \\ &\quad + \|\sigma^2 \circ q^2\|_{\infty, S^1} \|p^1 - p^2\|_{0, S^1}. \end{aligned}$$

Finally, using the Sobolev embedding theorem and theorem 22 we can write this bound in terms of the  $\|\cdot\|_{m, \Omega_{th}}$  norm:

$$\begin{aligned} \|\mathbb{P}_{S_h^p}[\sigma^1 \circ q^1 p^1] - \mathbb{P}_{S_h^p}[\sigma^2 \circ q^2 p^2]\|_{0, S^1} &\lesssim (\|\sigma^1\|_{m, \Omega_{th}} \|q^1 - q^2\|_{0, S^1} \\ &\quad + \|\sigma^1 - \sigma^2\|_{m, \Omega_{th}}) \|p^1\|_{0, S^1} \\ &\quad + \|\sigma^2 \circ q^2\|_{\infty, S^1} \|p^1 - p^2\|_{0, S^1}, \end{aligned}$$

with a constant depending on the Sobolev embedding theorem. As such, a local Lipschitz condition is verified here. Finally, solutions of (3.11e) also clearly have a Lipschitz dependence of the gradients of  $u_{th}$ , which can be bounded by the energy norm.  $\square$

Although  $H^4$  is sufficient in the continuous setting we use  $H^6$  velocities as we can split inversion of a sixth order form into two third order solves which is convenient given an implementation of  $\tilde{V}_{th}^d$ . Theorem 38 also highlights the need for the auxiliary variable  $\sigma_{th}$ : this is a Lipschitz variable while  $\nabla u_{th} \in L^2(\Omega)$  is not (owing to  $u_{th}$  only being  $H^1$ -conforming) which means only the former can be used to integrate forward the momentum equation. As we shall see shortly, *spatial* convergence of  $\sigma_{th}$  to  $\nabla u_{th}$  is sufficient for convergence of the discrete curve.

**Remark 39.** *Note that gradient recovery operators [158] could provide us with a finite element gradient that is uniformly bounded by  $\nabla u_{th}$ , the latter of which is uniformly bounded by some constant. Since at least  $H^4(\Omega)$  is required by the continuous problem anyway we decide to further restrict the velocity space to  $H^6(\Omega)$  and, at a discrete level decompose the associated operator into two  $H^3$  operator inversions in  $\tilde{V}_{th}^d$ .*

Given  $\Omega_{th}$  we introduce some additional notation that we use in the forthcoming results.

- Recall  $\Omega_{th}^+$  (resp.  $\Omega_{th}^-$ ) denotes the collection of  $K \in \Omega_{th}$  lying on the outside (resp. inside) of the curve  $q_{th} \circ S^1$ .
- We let  $\mathcal{K}^+$  (resp.  $\mathcal{K}^-$ ) denote the collection of  $K \in \Omega_{th}^+$  (resp.  $K \in \Omega_{th}^-$ ) such that  $K \cap q_{th} \circ S^1 \neq \emptyset$ .
- For  $e \in \mathcal{E}_h$ ,  $\mathcal{K}_e$  denotes the collection of  $K \in \Omega_{th}$  such that  $K \cap e \neq \emptyset$ .

**Assumption 40.** *Let  $\sigma_{th} \in \tilde{V}_{th}^{d \times d}$ ,  $t \in [0, 1]$  solve (3.11e) with the boundary conditions (3.11f). There exists a mesh resolution  $h^* > 0$  such that for all  $0 < h < h^*$  we have:*

$$\|\sigma_{th}\|_{m, \Omega_{th}} \leq M_{h^*}, \quad t \in [0, 1],$$

for some  $M_{h^*} > 0$  independent of  $h > 0$  and  $t$ .

This assumption means that the discrete velocity is *a priori* conserved for a sufficiently fine mesh. This bounded property can in practice be inspected *a posteriori*. It is inexpensive to numerically compute this norm in for instance a time-stepping loop, so theorem 38 does in fact provide a useful result.

In preparation for a curve convergence result we introduce the following auxiliary problem where we seek  $\hat{u}_t \in H_0^m(\Omega)^d$ ,  $\hat{w}_t \in H^m(\Omega)^d$  for  $t \in [0, 1]$ :

$$a(\hat{u}_t, v) = \langle \hat{w}_t, v \rangle_{0, \Omega}, \quad \forall v \in H_0^m(\Omega)^d, \quad (3.13a)$$

$$a(\hat{w}_t, v) = \int_{S^1} p_{th} \cdot v \circ q_{th} d\theta, \quad \forall v \in H^m(\Omega)^d. \quad (3.13b)$$

We recall the Strang [208] lemma 13 from chapter 2 (see also [39, Section 2.2.3]). In the present context we can state it as follows:

**Lemma 41** (Strang). *Let  $w_{th}$  be the solution of (3.11b) and  $\hat{w}_t$  be the solution of (3.13b). Then,*

$$\|w_{th} - \hat{w}_t\|_{m, \Omega_{th}} \lesssim \inf_{\mathbf{v} \in \tilde{V}_{th}^d} \|\hat{w}_t - \mathbf{v}\|_{m, \Omega_{th}} + \sup_{\mathbf{v} \in \tilde{V}_{th}^d} \frac{|\int_{S^1} p_{th} \circ q_{th} \cdot \mathbf{v} \, d\theta - a_h(\hat{w}_t, \mathbf{v})|}{\|\mathbf{v}\|_{m, \Omega_{th}}}, \quad (3.14)$$

where the first term on the right-hand side is called the approximation term and the second a consistency term.

*Proof.* Let  $\mathbf{v}, \mathbf{y}_h \in \tilde{V}_{th}^d$ :

$$\begin{aligned} a_h(u_{th} - \mathbf{y}_h, \mathbf{v}) &= a_h(u_{th} - \hat{w}_t, \mathbf{v}) + a_h(\hat{w}_t - \mathbf{y}_h, \mathbf{v}) \\ &= \int_{S^1} p_{th} \circ q_{th} \cdot \mathbf{v} \, d\theta - a_h(\hat{w}_t, \mathbf{v}) + a_h(\hat{w}_t - \mathbf{y}_h, \mathbf{v}). \end{aligned}$$

Dividing by  $\|\mathbf{v}\|_{m, \Omega_{th}}$  and using coercivity  $a_h(\cdot, \cdot)$  we can therefore claim:

$$\|\hat{w}_t - \mathbf{y}_h\|_{m, \Omega_{th}} \leq \sup_{\mathbf{v} \in \tilde{V}_{th}^d} \frac{|\int_{S^1} p_{th} \circ q_{th} \cdot \mathbf{v} \, d\theta - a_h(\hat{w}_t, \mathbf{v})|}{\|\mathbf{v}\|_{m, \Omega_{th}}} + \|\hat{w}_t - \mathbf{y}_h\|_{m, \Omega_{th}},$$

which together with the triangle inequality yields the result.  $\square$

We say that a discretisation for which the terms on the right-hand side of (3.14) vanish as  $h \rightarrow 0$  is an *approximate* and *consistent* discretisation. Approximation is guaranteed via classic element-wise interpolation estimates - we briefly state the main results from [231] adapted to the current setting.

**Definition 42** (Canonical interpolation operator). *For a given  $k \in \Omega_{th}$ ,  $t \in [0, 1]$  let us denote by  $\Pi_K : H^m(K) \rightarrow \tilde{P}_K^{(3,2)}$ , the canonical interpolation operator where  $\tilde{P}_K^{(3,2)} \subset \mathcal{P}^7(K)$  is a 7<sup>th</sup> order polynomial space defined in [231, Equation 4.1]. The nature of the definition of  $\tilde{P}_K^{(3,2)}$  is not relevant here as the goal is simply to construct abstract approximation results for these polynomials. Section 4 of [231] describes this space in great detail. A technical result [231, Lemma 4.1] determines the local degrees of freedom that show unisolvency of  $\tilde{P}_K^{(3,2)}$  i.e. those that uniquely determine any  $v \in \tilde{P}_K^{(3,2)}$ . This unisolvency property allows us to sketch the definition of  $\Pi_K$ :*

$$\Pi_K v = \sum_{j=1}^J l_{K,j} n_{K,j}(v), \quad \forall v \in H^m(K),$$

where  $\{l_{K,j}\}_{j=1}^J$  is a local basis and  $\{n_{K,j}\}_{j=1}^J$  (for some positive integer  $J$ ) is a nodal basis satisfying  $n_{K,i}(l_{K,k}) = \delta_{ik}$ , where  $\delta_{ik}$  is the Kronecker delta. See section 2.3 of the aforementioned reference for an example of a constructing of such a local interpolant.

By standard interpolation theory ([42], [231, Lemma 2.4]) we have:

**Lemma 43.** *For  $s \in [0, 1]$  and  $k$  such that  $0 \leq k \leq m$  and  $K \in \Omega_{th}$ :*

$$|v - \Pi_K|_{k,K} \lesssim h_K^{m+s-k} |v|_{m+s,K}, \quad \forall v \in H^{m+s}(K).$$

The *global* operator  $\pi_h$  on  $H^m(\Omega)^d$  into  $\tilde{V}_{th}^d$  is defined by  $(\pi_h v)|_K = \Pi_h|_K(v|_K)$  for  $K \in \Omega_{th}$  and satisfies the following standard result:

**Theorem 44.** *For  $v \in H^{m+s}(\Omega)$ ,  $s \geq 0$  we have:*

$$\begin{aligned} \|v - \pi_h v\|_{m,\Omega_{th}} &\leq h^s |v|_{m+s,\Omega}, \\ \lim_{h \rightarrow 0} \|v - \pi_h v\|_{m,\Omega_{th}} &= 0. \end{aligned}$$

*Proof.* The first property is obvious from lemma 43. For the second see [231, Theorem 2.5].  $\square$

This states that globally  $m + s$  weakly differentiable functions can be approximated in the norm  $\|\cdot\|_{m,\Omega_{th}}$ . However, by remark 24, global higher-order regularity of solutions to (3.6a) and (3.13b) is not possible. Local regularity, may however be recovered by the following lemma.

**Lemma 45** (Hypoellipticity). *Let  $w_t, \hat{w}_t$  be the solutions to (3.6a) and (3.13b), and define by  $\Omega_t^-, \Omega_t^+$ , (resp.  $\Omega_{th}^-, \Omega_{th}^+$ ) the connected open sets bounded by  $q_t \circ S^1$  and  $\partial\Omega$  (resp.  $\partial\Omega_{th}$  and  $q_{th} \circ S^1$ ) on the inside and outside of the curve. Then the restriction of  $w_t$  (resp.  $\hat{w}_t$ ) to the sets  $\Omega_t^-, \Omega_t^+$  (resp.  $\Omega_{th}^-, \Omega_{th}^+$ ) is  $C^\infty$  in each component.*

*Proof.* We recall that a differential operator  $\mathfrak{D}$  is hypoelliptic [43, Remark 25] if for any open set  $\omega$ ,  $C^\infty(\omega) \ni \mathfrak{D}z \Rightarrow z \in C^\infty(\omega)$  [96, Theorem 6.33]<sup>1</sup>. Moreover, any constant coefficient elliptic operator is in fact hypoelliptic [96, Corollary 6.34]. We observe that the operator  $B$  associated to the bilinear form  $a(\cdot, \cdot)$  is hypoelliptic on open sets, and so in fact, for any open set  $\omega \subset \Omega$  such that  $\bar{\omega} \subset \Omega$  is compact and such that  $\omega \cap q_t \circ S^1 = \emptyset$  and  $t \in [0, 1]$ ,  $Bu|_\omega = 0 \in C^\infty(\omega)$  (resp. for  $q_{th}$ ). In the case of the curve  $q_{th} \circ S^1$  it is easy to see that for any  $K \in \Omega_{th}$  we can solve an elliptic problem with smooth data on the interior and recover the local estimate. The key property here is convexity of  $K$ , see e.g. [114, Chapter 4].  $\square$

**Remark 46.** *Interestingly, local regularity provides a smooth velocity almost everywhere, so one could in theory select e.g.  $B = \text{id} - \Delta$ . This, however, is insufficient to generate integrable geodesics of (3.6) for all time because the Lipschitz norm is not conserved since this choice of  $B$  only controls the  $H^1$  norm so that Lipschitz norm can become arbitrarily large.*

<sup>1</sup> For completeness we highlight that although this result stated for distributions it is also valid for the kind of Hilbert-Sobolev spaces we consider here and we refer to the author's comment on this extension near the end of part C of Chapter 6 of [96].

However, local regularity can be recovered from lemma 45 to state the following nonconforming estimate using theorem 44 applied in an element-wise manner.

**Corollary 47.** *Let  $v \in H^m(\Omega)$  be such that  $v|_K \in H^{m+s}(K)^d$ ,  $\forall K \in \Omega_{th}$ , with  $m = d + 1$ ,  $s \geq 0$ . Then:*

$$\begin{aligned} \|v - \Pi_h v\|_{m, \Omega_{th}} &\leq h^s |v|_{m+s, \Omega_{th}}, \\ \lim_{h \rightarrow 0} \|v - \Pi_h v\|_{m, \Omega_{th}} &= 0. \end{aligned}$$

Corollary 47 establishes an important local approximation property for the discretisation (3.11). We recall that consistency of the discretisation means estimating to what extent, as a function of  $h$ , the continuous solution (3.11b) satisfies the equation (3.13b). This property can be verified by e.g. the generalised patch test (GPT) of Stummel [213]. This essentially asks that the nonconforming finite element spaces satisfy a type of compactness condition (a more modern version of this condition is stated in [230], namely the one we presented in chapter 2).

The proof of consistency of  $\tilde{V}_h$  was omitted in [231] so for completeness we apply an argument similar to [230, Theorem 3.2] using the weak continuity lemma 28 to show this. This property states that *integrals* of derivatives of functions in  $\tilde{V}_{th}^d$  are smooth across interior edges of a triangulation, rather than smoothness in a pointwise sense.

**Remark 48** (Bibliographic note). *Weak continuity is indeed sufficient (but not necessary) for the “SPT” condition required in the aforementioned Theorem 3.2. The following theorems are not novel and are simply included to benefit any unfamiliar readers with the literature. While we can easily verify the GPT by the information we provided above, in the proof of the following we use what we regard as a simpler method given by [229] provided a weak continuity property.*

**Theorem 49.**  $\tilde{V}_{th}^d$  is a consistent approximation of  $H^3(\Omega)^d$ .

*Proof.* As stated in [213, Section 1.2, Assumption V], we verify coercivity and continuity of  $a(\cdot, \cdot)$  and  $a_h(\cdot, \cdot)$  by corollary 37 as required for the GPT [213, Section 1.4, Equation 7]. Using the form of the test as stated in [229, Equation 4.4],  $\tilde{V}_h$  passes the GPT if and only if the following condition stated is satisfied:

$$\lim_{h \rightarrow 0} \sup_{\substack{v_h \in \tilde{V}_h \\ \|v_h\|_{m, \Omega_{th}} \leq 1}} |T_{\alpha, i}(\psi, v_h)| = 0, \quad |\alpha| < m, \quad 1 \leq i \leq d, \quad \forall \psi \in C^\infty(\bar{\Omega}), \quad (3.15)$$

where:

$$T_{\alpha, i}(\psi, v_h) = \sum_{K \in \Omega_{th}} \int_K \left[ \psi \frac{\partial}{\partial x_i} \partial^\alpha v_h + \frac{\partial \psi}{\partial x_i} \partial^\alpha v_h \right] dx.$$

Equivalently, where  $\eta_K$  is the outward normal of  $K$ :

$$T_{\alpha,i}(\psi, v_h) = \sum_{K \in \Omega_{th}} \int_{\partial K} \psi \partial^\alpha v_h \eta_K^i ds. \quad (3.16)$$

For an edge  $e \in \mathring{\mathcal{E}}_{th}$  between  $K, K' \in \Omega_{th}$  and let  $\eta_K$  (resp.  $\eta_{K'}$ ) denote the outward normal to  $K$  (resp.  $K'$ ). We can then write (3.16) as:

$$\begin{aligned} T_{\alpha,i}(\psi, v_h) &= \sum_{e \in \mathring{\mathcal{E}}_{th}} \int_e \psi \partial^\alpha v_h|_K \eta_K^i + \psi \partial^\alpha v_h|_{K'} \eta_{K'}^i ds + \sum_{e \in \partial \Omega_{th}} \int_e \psi \partial^\alpha v_h \eta_{\partial \Omega_{th}}^i ds \\ &= \sum_{e \in \mathring{\mathcal{E}}_{th}} \int_e \psi [\partial^\alpha v_h|_K - \partial^\alpha v_h|_{K'}] \eta_{K'}^i ds + \sum_{e \in \partial \Omega_{th}} \int_e \psi \partial^\alpha v_h \eta_{\partial \Omega_{th}}^i ds \\ &\lesssim \sum_{e \in \mathring{\mathcal{E}}_{th}} \|\psi\|_{L^\infty(e)} \int_e [\partial^\alpha v_h|_K - \partial^\alpha v_h|_{K'}] ds \\ &\quad + \sum_{e \in \partial \Omega_{th}} \|\psi\|_{L^\infty(e)} \int_e \partial^\alpha v_h ds \\ &= 0. \end{aligned}$$

Here we have used lemma 28 stating that facet integral moments of first and second order match between cells, so the consistency condition is satisfied so  $\tilde{V}_h$  passes the GPT, which implies the same result for  $\tilde{V}_{th}^d$  which is therefore a consistent approximation of  $H^3(\Omega)^d$ .  $\square$

The approximation and consistency of the GPT results imply a convergence result. However, this result only estimates the difference between  $w_{th}$  and  $\hat{w}_t$  when the momentum  $p_{th}$  and the curve  $q_{th}$  are *fixed*. In other words, if we refine around the fixed curve  $q_{th} \circ S^1$ , some convergence can be obtained for  $w_{th}$ . To reiterate, the following result therefore discusses convergence when the resolution for the bilinear form in (3.13b) and for  $w_{th}$  is refined. To ease notation, however, we still use  $p_{th}$  and  $q_{th}$ . To state the convergence result we need to prove a technical lemma:

**Lemma 50.** *Let  $\mathbf{v} \in H^3(\Omega)^d$  and the sequence  $\{\mathbf{v}_h\}_h$  in  $\tilde{V}_{th}^d$  be such that  $\lim_{h \rightarrow 0} \|\mathbf{v} - \mathbf{v}_h\|_{m, \Omega_{th}} = 0$ . Then, for any  $q_{th} \in C^0(S^1)^d$  such that  $q_{th} \circ S^1$  is comprised of edges  $e \in \mathring{\mathcal{E}}_{th}$ ,  $(\mathbf{v}_{h_k} - \mathbf{v}) \circ q_{th}$  converges a.e. pointwise to zero on  $S^1$  (with respect to the 1-dimensional Lebesgue measure).*

*Proof.* Pointwise convergence on  $S^1$  is implied by convergence in the sup norm. Using definition 32 we have:

$$\begin{aligned} \text{ess sup}_{\theta \in S^1} |(\mathbf{v}_{h_k} - \mathbf{v}) \circ q_{th}|(\theta) &= \text{ess sup}_{x \in q_{th} \circ S^1} |\mathbf{v}_{h_k} - \mathbf{v}|(x) \\ &= \max_{i \in \mathcal{J}} \text{ess sup}_{x \in q_{th} \circ S_i^1} |\mathbf{v}_{h_k} - \mathbf{v}|(x). \end{aligned}$$

Since  $\mathbf{v}_{h_k} - \mathbf{v}$  is continuous everywhere,  $\mathbf{v}_{h_k} - \mathbf{v} = \{\mathbf{v}_{h_k} - \mathbf{v}\}$  on edges of the interior mesh skeleton  $\mathring{\mathcal{E}}_{th}$ . Therefore, for any two elements  $K_i^+$  and  $K_i^-$  joining an edge  $q_{th} \circ S_i^1$  we can say:

$$\begin{aligned} \max_{i \in \mathcal{J}} \operatorname{ess\,sup}_{x \in q_{th} \circ S_i^1} |\mathbf{v}_{h_k} - \mathbf{v}|(x) &= \max_{i \in \mathcal{J}} \operatorname{ess\,sup}_{x \in q_{th} \circ S_i^1} |\{\mathbf{v}_{h_k} - \mathbf{v}\}|(x) \\ &\leq \max_{i \in \mathcal{J}} \left[ \operatorname{ess\,sup}_{x \in q_{th} \circ S_i^1} |\mathbf{v}_{h_k}|_{K_i^+} - \mathbf{v}|(x) \right. \\ &\quad \left. + \operatorname{ess\,sup}_{x \in q_{th} \circ S_i^1} |\mathbf{v}_{h_k}|_{K_i^-} - \mathbf{v}|(x) \right]. \end{aligned}$$

By the Sobolev embedding theorem in one dimension we have for any bounded open set  $U$ :  $f \in H^1(U) \Rightarrow f \in C^0(U)$ , so:

$$\operatorname{ess\,sup}_{\theta \in S^1} |(\mathbf{v}_{h_k}|_{K_i^+} - \mathbf{v}) \circ q_{th}|(\theta) \leq \max_{i \in \mathcal{J}} \|\mathbf{v}_{h_k}|_{K_i^+} - \mathbf{v}\|_{1, q_{th} \circ S_i^1}.$$

By the trace theorem we have for any  $i \in \mathcal{J}$ :

$$\|\mathbf{v}_{h_k}|_{K_i^+} - \mathbf{v}\|_{1, q_{th} \circ S_i^1} \lesssim \|\mathbf{v}_{h_k} - \mathbf{v}\|_{2, \Omega_{th}}.$$

Since  $\|\mathbf{v}_{h_k} - \mathbf{v}\|_{2, \Omega_{th}}$  converges by assumption we conclude the proof by applying the same argument to the restriction of  $\mathbf{v}_{h_k}$  to  $K_i^-$ .

Another way to see this result is to recall theorem 22 indicating that  $\tilde{V}_{th}^d|_K \subset L^\infty(K)$ ,  $K \in \Omega_{th}$ . Since  $K$  is convex [207, Theorem 5] tells us that an extension result obtains and we can therefore ascribe meaning to the values of a function in  $\tilde{V}_{th}^d|_K$  on  $\partial K$ .  $\square$

**Corollary 51.** *Let  $u_{th}$  and  $\hat{u}_t$  solve (3.11a) and (3.13a), and let  $w_{th}$  and  $\hat{w}_t$  solve (3.11b) and (3.13b). Letting the resolution for  $q_{th}$  and  $p_{th}$  be fixed as a function of  $h$ , then for all  $t \in [0, 1]$  we have:*

$$\lim_{h \rightarrow 0} \|u_{th} - \hat{u}_t\|_{m, \Omega_{th}} = \lim_{h \rightarrow 0} \|w_{th} - \hat{w}_t\|_{m, \Omega_{th}} = 0.$$

*Proof.* The strategy is to use the definition of a consistent approximation to show convergence of the discrete bilinear form to the continuous one, and the same for the discrete source term. It is completely essential that the source term in (3.11b) and (3.13b) are considered fixed in the momentum and the curve. The reader should think about this corollary as a convergence result if we refine the mesh, and resolution of the velocity field *around* a fixed embedded curve ( $q_{th}$ ) for some polynomial momentum ( $p_{th}$ ). The convergence of ( $q_{th}$  and  $p_{th}$ ) are dealt with using Grönwall's lemma).

We note  $a(\hat{u}_t, \hat{u}_t) = a_h(\hat{u}_t, \hat{u}_t)$ , so:

$$a_h(\hat{u}_t - u_{th}, \hat{u}_t - u_{th}) = \langle \hat{w}_t - w_{th}, \hat{u}_t - u_{th} \rangle_{0, \Omega_{th}},$$

so by Cauchy-Schwartz:

$$\|\hat{u}_t - u_{th}\|_{m, \Omega_{th}} \leq \|\hat{w}_t - w_{th}\|_{0, \Omega_{th}}.$$

We know that  $\|\hat{w}_{th}\|_{m, \Omega_{th}} \lesssim \|p_{th}\|_{0, S^1}$  since  $\tilde{V}_{th}^d$  is Lipschitz-conforming cf. theorem 22, so  $\{\hat{w}_{th}\}_h$  is a bounded sequence. Using the definition of consistency in [213, 230] we say that  $\tilde{V}_{th}^d$  is a consistent approximation of  $H^3(\Omega)^d$  by theorem 49 and [230, Section 3.1] so we can extract a subsequence  $\{\hat{w}_{th_k}\}_{k \geq 0}$  of  $\{\hat{w}_{th}\}_h$  and a function  $\mathbf{w}' \in H^3(\Omega)^d$  such that  $\partial^\alpha \hat{w}_{th_k} \rightarrow \partial^\alpha \mathbf{w}'$  weakly in  $L^2(\Omega)$  for all  $|\alpha| \leq m$ . We know by corollary 47 that given any  $\mathbf{v} \in H^3(\Omega)^d$  there exists a sequence  $\{\mathbf{v}_h\}_h$  such that  $\lim_{h \rightarrow 0} \|\mathbf{v} - \mathbf{v}_h\|_{m, \Omega_{th}} = 0$ . Then we can say:

$$\begin{aligned} |a_{h_k}(\hat{w}_{th_k}, \mathbf{v}_{h_k}) - a(\mathbf{w}', \mathbf{v})| &\leq |a_{h_k}(\hat{w}_{th_k}, \mathbf{v}_{h_k} - \mathbf{v})| + |a_{h_k}(\hat{w}_{th_k} - \mathbf{w}', \mathbf{v})| \\ &\leq \|\hat{w}_{th_k}\|_{m, \Omega_{th}} \|\mathbf{v}_{h_k} - \mathbf{v}\|_{m, \Omega_{th}} + |a_{h_k}(\hat{w}_{th_k} - \mathbf{w}', \mathbf{v})| \end{aligned}$$

which vanishes as  $k \rightarrow \infty$ . Next we show:

$$\int_{S^1} p_{th} \cdot \mathbf{v}_{h_k} \circ q_{th} \, d\theta \rightarrow \int_{S^1} p_{th} \cdot \mathbf{v} \circ q_{th} \, d\theta \quad \text{as } h_k \rightarrow 0, \quad (3.17)$$

which essentially amounts to showing that we can pass the limit inside the integral of  $S^1$ . By the dominated convergence theorem this requires at least pointwise convergence of  $(\mathbf{v}_{h_k} - \mathbf{v}) \circ q_{th}$  and pointwise boundedness by some integrable function. The latter is trivial since  $\mathbf{v}_{h_k}$  and  $\mathbf{v}$  are Lipschitz. The information we have is strong convergence of  $\|\mathbf{v}_{h_k} - \mathbf{v}\|_{m, \Omega_{th}}$  by the weak approximation property, so by lemma 50 (3.17) is verified.

In light of this we must have  $a(\mathbf{w}, \mathbf{v}) = \int_{S^1} p_{th} \circ \mathbf{v} \circ q_{th} \, d\theta$  and so corollary 37 implies  $\mathbf{w}' \equiv \hat{w}_t$ .

Next we show strong convergence. Let  $\mathbf{y}_h \in \tilde{V}_{th}^d$  be a sequence such that  $\|\hat{w}_t - \mathbf{y}_h\|_{m, \Omega_{th}} \rightarrow 0$  as  $h \rightarrow 0$ . Then,

$$\begin{aligned} \|\hat{w}_t - w_{th}\|_{m, \Omega_{th}}^2 &\lesssim \|\hat{w}_t - \mathbf{y}_h\|_{m, \Omega_{th}}^2 + a_h(\mathbf{y}_h - w_{th}, \mathbf{y}_h - w_{th}) \\ &\lesssim \|\hat{w}_t - \mathbf{y}_h\|_{m, \Omega_{th}}^2 + a(\hat{w}_t, \hat{w}_t) - 2a_h(\hat{w}_t, w_{th}) + a_h(w_{th}, w_{th}) \\ &\lesssim \|\hat{w}_t - \mathbf{y}_h\|_{m, \Omega_{th}}^2 + a(\hat{w}_t, \hat{w}_t) - 2a_h(\hat{w}_t, w_{th}) \\ &\quad + \int_{S^1} p_{th} \circ w_{th} \circ q_{th} \, d\theta. \end{aligned}$$

Now using what we just derived above:

$$\begin{aligned} &a(\hat{w}_t, \hat{w}_t) - 2a_h(\hat{w}_t, w_{th}) + \int_{S^1} p_{th} \circ w_{th} \circ q_{th} \, d\theta \\ \rightarrow &a(\hat{w}_t, \hat{w}_t) - 2a_h(\hat{w}_t, \hat{w}_t) + \int_{S^1} p_{th} \circ \hat{w}_t \circ q_{th} \, d\theta = 0 \end{aligned}$$

as  $h \rightarrow 0$ . □

**Remark 52** (Comparison with  $L^2(\Omega)$  source terms). *The standard setting is for the source term to be defined in terms of some  $f \in L^2(\Omega)^d$  on the form  $\langle f, \mathbf{v}_{h_k} \rangle_{0,\Omega}$ . Weak  $L^2(\Omega)^d$  convergence therefore trivially completes the argument in this case since:*

$$\langle f, \mathbf{v}_{h_k} \rangle_{0,\Omega} \rightarrow \langle f, \mathbf{v} \rangle_{0,\Omega},$$

*is by definition, weakly convergent. This argument is ill-suited to the setting above due to the singular nature of the  $S^1$  integral as a source term. This is the reason for the use of lemma 50 in corollary 51.*

It is *hugely* important to note that corollary 51 provides a convergence result for a fixed curve and momentum. Here we commit a slight abuse of notation by keeping the subscript  $h$  on all our variables, as the notation otherwise comes in the way of the exposition of the result. We briefly elaborate on this here. The key thing to note in the result above is that the implicit conforming shape-regular family of triangulations  $\{\Omega_{th}\}_{h_j}$  are such that for any  $h_j \rightarrow 0$  as  $j \rightarrow \infty$  there is a subset of  $\mathring{\mathcal{E}}_{th_j}$  tracing out  $q_{th}$ . One can think of corollary 51 as saying that if we refine the mesh "around the curve"  $q_{th}$  our discrete solution converges to the continuous one. Mesh refinement for the spaces  $\mathbf{S}_h^q$  and  $\mathbf{S}_h^p$  for  $q_{th}$  and  $p_{th}$  is therefore not a factor here. Keeping this in mind we proceed to prove a convergence rate. To this end we introduce the notion of *conforming relatives* first proposed in [41].

**Proposition 53** (Conforming relatives). *There exists an  $H^m$ -conforming finite element space  $V_h^{c,d} \subset H_0^m(\Omega)^d$  and an operator  $\Pi_h^c : \tilde{V}_{th}^d \rightarrow V_h^{c,d}$  such that  $\forall v \in \tilde{V}_{th}^d$ :*

$$\sum_{j=0}^{m-1} h^{2(j-m)} |v_h^i - [\Pi_h^c v_h]^i|_{j,\Omega_{th}}^2 + |[\Pi_h^c v_h]^i|_{m,\Omega_{th}}^2 \lesssim |v_h^i|_{m,\Omega_{th}}^2. \quad (3.18)$$

for each component  $v_h^i$  of  $v_h$ .

This result can be seen in e.g. [135] or [134, Lemma 3.2]<sup>2</sup>. The proof we provide next uses these operators as well as the hypoellipticity result in lemma 45.

**Corollary 54.** *The convergence rate for the expression  $\lim_{h \rightarrow 0} \|\hat{u}_t - u_{th}\|_{m,\Omega_{th}} = 0$  is almost linear in  $h$  via the following bound:*

$$\begin{aligned} \|\hat{u}_t - u_{th}\|_{m,\Omega_{th}} &\lesssim h \left( \sum_{i=1}^2 \|D^3 \hat{w}_t^i\|_{0,\Omega_{th}} + \|D^5 \hat{w}_t^i\|_{0,\Omega_{th}} + \|D^4 \hat{w}_t^i\|_{0,\Omega_{th}} \right. \\ &\quad \left. + h \left( \sum_{e \in q_{th} \circ S^1} \| [D^3 \hat{w}_t^i] \|_{0,e} + \| [D^5 \hat{w}_t^i] \|_{0,e} \right) + 2 \|D^4 \hat{w}_t^i\|_{0,\Omega_{th}} \right) \\ &\quad + \sum_{e \in q_{th} \circ S^1} \| [D^3 \hat{w}_t^i - P_{\mathcal{K}_e} [D^3 \hat{w}_t^i]] \|_{0,e} \\ &\quad + h^s |\hat{w}_t|_{m+s,\Omega_{th}}, \end{aligned}$$

<sup>2</sup> The reference given here shows under which conditions the conforming relatives operator exists, for arbitrary  $m \geq 1$ . These are exactly the weak continuity properties provided by [231].

for any integer  $s \geq 0$ , where  $P_{\mathcal{K}_e}$  is the projection operator on  $L^2(\mathcal{K}_e)$  defined by:

$$P_{\mathcal{K}_e} v = \left( \sum_{K \in \mathcal{K}_e} |K| \right)^{-1} \sum_{K \in \mathcal{K}_e} \int_K v \, dx, \quad v \in L^2(\Omega).$$

*Proof.* We first note  $a(\hat{u}_t, \hat{u}_t) = a_h(\hat{u}_t, \hat{u}_t)$ , so since  $\hat{u}_t$  and  $u_{th}$  satisfy:

$$\begin{aligned} a_h(\hat{u}_t, \hat{u}_t) &= \langle \hat{w}_t, \hat{u}_t \rangle_{0, \Omega_{th}}, \\ a_h(u_{th}, u_{th}) &= \langle w_{th}, u_{th} \rangle_{0, \Omega_{th}}. \end{aligned}$$

Subtracting these we obtain:

$$a_h(\hat{u}_t - u_{th}, \hat{u}_t - u_{th}) = \langle \hat{w}_t - w_{th}, \hat{u}_t - u_{th} \rangle_{0, \Omega_{th}},$$

so by Cauchy-Schwartz:

$$\|\hat{u}_t - u_{th}\|_{m, \Omega_{th}} \leq \|\hat{w}_t - w_{th}\|_{0, \Omega_{th}}.$$

We proceed to bound the right-hand side. For arbitrary  $y_h \in \tilde{V}_{th}^d$  we seek to bound the following expression in terms of  $h$ :

$$a_h(\hat{w}_t - w_{th}, y_h) = a_h(\hat{w}_t, y_h) - \int_{S^1} p_{th} \cdot y_h \circ q_{th} \, d\theta.$$

We use the conforming relatives operator to say:

$$\begin{aligned} a_h(\hat{w}_t - w_{th}, y_h) &= a_h(\hat{w}_t, y_h) - \int_{S^1} p_{th} \cdot y_h \circ q_{th} \, d\theta \\ &= a_h(\hat{w}_t, y_h - \Pi_h^c y_h) - \int_{S^1} p_{th} \cdot (y_h - \Pi_h^c y_h) \circ q_{th} \, d\theta, \end{aligned}$$

since  $a_h(\hat{w}_t, \Pi_h^c y_h) - \int_{S^1} p_{th} \cdot \Pi_h^c y_h \circ q_{th} \, d\theta = 0$ .

First we estimate  $\int_{S^1} p_{th} \cdot (y_h - \Pi_h^c y_h) \circ q_{th} \, d\theta$  and show linear convergence in  $h$ . The key tools are the Sobolev embedding theorem and (3.18):

$$\begin{aligned} \int_{S^1} p_{th} \cdot (y_h - \Pi_h^c y_h) \circ q_{th} \, d\theta &\leq \|p_{th}\|_{0, S^1} \| (y_h - \Pi_h^c y_h) \circ q_{th} \|_{0, S^1} \\ &\lesssim \|p_{th}\|_{0, S^1} \| (y_h - \Pi_h^c y_h) \circ q_{th} \|_{\infty, S^1} \\ &= \|p_{th}\|_{0, S^1} \|y_h - \Pi_h^c y_h\|_{\infty, q_{th} \circ S^1} \\ &\lesssim \|p_{th}\|_{0, S^1} \|y_h - \Pi_h^c y_h\|_{1, q_{th} \circ S^1} \text{ (embedding)} \\ &\lesssim \|p_{th}\|_{0, S^1} \|y_h - \Pi_h^c y_h\|_{2, \Omega_{th}} \end{aligned}$$

Now using the bound in (3.18):

$$\int_{S^1} p_{th} \cdot (y_h - \Pi_h^c y_h) \circ q_{th} \, d\theta \lesssim h \|p_{th}\|_{0, S^1} \|y_h\|_{m, \Omega_{th}}. \quad (3.19)$$

As a result we have linear convergence in  $h$  if we suppose  $p_{th}$  and  $q_{th}$  are fixed as functions of  $h$ .

Second we estimate  $a_h(\hat{w}_t, y_h - \Pi_h^c y_h)$  by writing the variational problem for  $\hat{w}_t$  in a strong form. For the sake of exposition we recall the standard integration by parts identities on elements  $K$  of  $\Omega_{th}$  in our notation:

$$\begin{aligned} \int_K \langle D\hat{w}_t^i, Dy_h^i \rangle dx &= - \int_K \langle D^2\hat{w}_t^i, y_h^i \rangle dx + \int_{\partial K} \langle D\hat{w}_t^i, \eta_K y_h^i \rangle ds, \\ \int_K \langle D^2\hat{w}_t^i, D^2y_h^i \rangle dx &= \int_K \langle D^4\hat{w}_t^i, y_h^i \rangle dx + \int_{\partial K} \langle \eta_K D^2\hat{w}_t^i, Dy_h^i \rangle ds \\ &\quad - \int_{\partial K} \langle D^3\hat{w}_t^i, \eta_K y_h^i \rangle ds, \\ \int_K \langle D^3\hat{w}_t^i, D^3y_h^i \rangle dx &= - \int_K \langle D^6\hat{w}_t^i, y_h^i \rangle dx + \int_{\partial K} \langle D^5\hat{w}_t^i, \eta_K y_h^i \rangle ds \\ &\quad - \int_{\partial K} \langle \eta_K D^4\hat{w}_t^i, Dy_h^i \rangle ds + \int_{\partial K} \langle D^3\hat{w}_t^i, \eta_K D^2y_h^i \rangle ds, \end{aligned}$$

Recall that  $\Omega_{th}^-$  (resp.  $\Omega_{th}^+$ ) denotes the parts of domain  $\Omega_{th}$  restricted to the inside (resp. outside) of the curve traced by  $q_{th} \circ S^1$ .

**Remark 55** (Strong form equation). *The strong form version of (3.13b) can be written by using the identities above:*

$$\begin{aligned} B\hat{w}_t &= 0, \quad \text{on } \Omega_{th}^- \text{ and } \Omega_{th}^+, \\ \int_{S^1} p_{th} \cdot v \circ q_{th} d\theta &= \sum_{i=1}^d \sum_{e \in q_{th} \circ S^1} \int_e b_2 \binom{m}{2} \langle \llbracket D^3\hat{w}_t^i \rrbracket, v^i \rangle - b_3 \binom{m}{3} \langle \llbracket D^5\hat{w}_t^i \rrbracket, v^i \rangle \\ &\quad + b_3 \binom{m}{3} \langle \llbracket D^4\hat{w}_t^i \rrbracket, Dv^i \rangle - b_3 \binom{m}{3} \langle \llbracket D^3\hat{w}_t^i \rrbracket, D^2v^i \rangle ds, \\ &\quad \forall v \in H^3(\Omega). \end{aligned}$$

Indeed, for a test function  $v \in H^3(\Omega)^d$  we have:

$$\begin{aligned} &\sum_{i=1}^d \int_{\Omega_{th}} \sum_{j=0}^m b_j \binom{m}{j} \langle D^j\hat{w}_t^i, D^jv^i \rangle dx \\ &= \sum_{i=1}^d \int_{\Omega_{th}^-} \sum_{j=0}^m b_j \binom{m}{j} \langle D^j\hat{w}_t^i, D^jv^i \rangle dx + \int_{\Omega_{th}^+} \sum_{j=0}^m b_j \binom{m}{j} \langle D^j\hat{w}_t^i, D^jv^i \rangle dx. \end{aligned}$$

So integrating by parts:

$$\begin{aligned}
& \sum_{i=1}^d \int_{\Omega_{th}} \sum_{j=0}^m b_j \binom{m}{j} \langle D^j \hat{w}_t^i, D^j v^i \rangle dx \\
&= \sum_{i=1}^d \int_{\Omega_{th}^-} B \hat{w}_t^i \cdot v^i dx + \sum_{i=1}^d \int_{\Omega_{th}^+} B \hat{w}_t^i \cdot v^i dx \\
&+ \int_{q_{th} \circ S^1} b_2 \binom{m}{2} \langle \llbracket D^3 \hat{w}_t^i \rrbracket, v^i \rangle - b_3 \binom{m}{3} \langle \llbracket D^5 \hat{w}_t^i \rrbracket, v^i \rangle \\
&+ b_3 \binom{m}{3} \llbracket D^4 \hat{w}_t^i D v^i \rrbracket - b_3 \binom{m}{3} \llbracket D^3 \hat{w}_t^i D^2 v^i \rrbracket ds.
\end{aligned}$$

Since the integral over  $\Omega_{th}^-$  and  $\Omega_{th}^+$  of  $B \hat{w}_t^i \cdot v^i$  vanishes we can relate the momentum term to a jump condition on  $q_{th} \circ S^1$  via the following variational equality:

$$\begin{aligned}
\int_{S^1} p_{th} \cdot v \circ q_{th} d\theta &= \sum_{i=1}^d \sum_{e \in q_{th} \circ S^1} \int_e b_2 \binom{m}{2} \langle \llbracket D^3 \hat{w}_t^i \rrbracket, v^i \rangle - b_3 \binom{m}{3} \langle \llbracket D^5 \hat{w}_t^i \rrbracket, v^i \rangle \\
&+ b_3 \binom{m}{3} \langle \llbracket D^4 \hat{w}_t^i \rrbracket, D v^i \rangle \\
&- b_3 \binom{m}{3} \langle \llbracket D^3 \hat{w}_t^i \rrbracket, D^2 v^i \rangle ds, \quad \forall v \in H^3(\Omega).
\end{aligned} \tag{3.20}$$

Then:

$$\begin{aligned}
a_h(\hat{w}_t, y_h - \Pi_h^c y_h) &= \sum_{i=1}^2 \sum_{K \in \Omega_{th}} \underbrace{\int_K B \hat{w}_t^i \cdot (y_h^i - [\Pi_h^c y_h]^i) dx}_{\text{sum cancels by definition}} \\
&+ \underbrace{\int_{\partial K} b_1 \binom{m}{1} \langle D \hat{w}_t^i, \eta_K (y_h^i - [\Pi_h^c y_h]^i) \rangle ds}_{\text{sum cancels because integrand is } H^1} \\
&+ \int_{\partial K} b_2 \binom{m}{2} \langle \eta_K D^2 \hat{w}_t^i, D(y_h^i - [\Pi_h^c y_h]^i) \rangle ds \\
&- \int_{\partial K} b_2 \binom{m}{2} \langle D^3 \hat{w}_t^i, \eta_K (y_h^i - [\Pi_h^c y_h]^i) \rangle ds \\
&- \int_{\partial K} b_3 \binom{m}{3} \langle D^5 \hat{w}_t^i, \eta_K (y_h^i - [\Pi_h^c y_h]^i) \rangle ds \\
&+ \int_{\partial K} b_3 \binom{m}{3} \langle \eta_K D^4 \hat{w}_t^i, D(y_h^i - [\Pi_h^c y_h]^i) \rangle ds \\
&- \int_{\partial K} b_3 \binom{m}{3} \langle D^3 \hat{w}_t^i, \eta_K D^2 (y_h^i - [\Pi_h^c y_h]^i) \rangle ds.
\end{aligned}$$

We now write the boundary terms as jumps on the mesh skeleton. The expressions  $y_h^i - [\Pi_h^c y_h]^i$  and  $D^2 \hat{w}_t^i$  have unique traces everywhere, while the expressions  $D^j \hat{w}_t^i$ ,  $j = 3, 4, 5$  only have unique traces on  $\mathring{\mathcal{E}}_{th} \setminus q_{th} \circ S^1$ . Lastly, the expressions  $D^j(y_h^i - [\Pi_h^c y_h]^i)$ ,  $j = 1, 2$  are not uniquely defined anywhere on  $\mathring{\mathcal{E}}_{th}$ . With this we can write:

$$a_h(\hat{w}_t, y_h - \Pi_h^c y_h) = \sum_{i=1}^2 \int_{\mathring{\mathcal{E}}_{th}} b_2 \binom{m}{2} \langle D^2 \hat{w}_t^i, \llbracket D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle ds \quad (3.21a)$$

$$+ \int_{\mathring{\mathcal{E}}_{th} \setminus q_{th} \circ S^1} b_3 \binom{m}{3} \langle D^4 \hat{w}_t^i, \llbracket D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle \quad (3.21b)$$

$$- b_3 \binom{m}{3} \langle D^3 \hat{w}_t^i, \llbracket D^2(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle ds \quad (3.21c)$$

$$+ \int_{q_{th} \circ S^1} -b_2 \binom{m}{2} \langle \llbracket D^3 \hat{w}_t^i \rrbracket, y_h^i - [\Pi_h^c y_h]^i \rangle \quad (3.21d)$$

$$- b_3 \binom{m}{3} \langle \llbracket D^5 \hat{w}_t^i \rrbracket, y_h^i - [\Pi_h^c y_h]^i \rangle \quad (3.21e)$$

$$+ b_3 \binom{m}{3} \llbracket D^4 \hat{w}_t^i D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \quad (3.21f)$$

$$- b_3 \binom{m}{3} \llbracket D^3 \hat{w}_t^i D^2(y_h^i - [\Pi_h^c y_h]^i) \rrbracket ds, \quad (3.21g)$$

where we recall that  $\llbracket D^4 \hat{w}_t^i D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket$  and  $\llbracket D^3 \hat{w}_t^i D^2(y_h^i - [\Pi_h^c y_h]^i) \rrbracket$  on  $e = K_e^+ \cap K_e^-$  are defined as:

$$\begin{aligned} \llbracket D^4 \hat{w}_t^i D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket &= \langle \eta_{K_e^+} D^4 \hat{w}_t^i|_{K_e^+}, D(y_h^i - [\Pi_h^c y_h]^i)|_{K_e^+} \rangle \\ &\quad + \langle \eta_{K_e^-} D^4 \hat{w}_t^i|_{K_e^-}, D(y_h^i - [\Pi_h^c y_h]^i)|_{K_e^-} \rangle, \\ \llbracket D^3 \hat{w}_t^i D^2(y_h^i - [\Pi_h^c y_h]^i) \rrbracket &= \langle \eta_{K_e^+} \cdot D^3 \hat{w}_t^i|_{K_e^+}, D^2(y_h^i - [\Pi_h^c y_h]^i)|_{K_e^+} \rangle \\ &\quad + \langle \eta_{K_e^-} \cdot D^3 \hat{w}_t^i|_{K_e^-}, D^2(y_h^i - [\Pi_h^c y_h]^i)|_{K_e^-} \rangle. \end{aligned}$$

Many of the terms in (3.21) are estimated using the same technique and rely on the weak continuity result for  $\tilde{V}_{th}^d$  in lemma 28. First we highlight the main inequalities on which we rely. Let  $e$  be an edge of a arbitrary  $K \in \Omega_{th}$  with length  $|e|$ . Further, let  $\mathcal{K}_e$  denote the set of triangles in  $\Omega_{th}$  having  $e$  as an edge. We recall [42, Equations 10.3.8, 10.3.9] and the trace theorem:

$$|e|^{-1} \|f\|_{0,e}^2 \lesssim h_K^{-2} \|f\|_{0,K}^2 + \|f\|_{1,K}^2, \quad f \in H^1(K) \quad (3.22a)$$

$$|e| \|\llbracket w \rrbracket\|_{0,e}^2 \lesssim h_K^2 \sum_{K \in \mathcal{K}_e} \|w\|_{1,K}^2, \quad w|_K \in H^1(K), \quad (3.22b)$$

$$\|f|_K\|_{0,e}^2 \lesssim \|f\|_{1,K}^2, \quad f|_K \in H^1(K). \quad (3.22c)$$

**Remark 56.** *The inequalities in (3.22a) and (3.22b) are not proved in [42], so for completeness we prove these in appendix C. Readers can also refer to this appendix for an example of some of the techniques we use here albeit applied to a simple Poisson equation in the plane with a source term that is an integral of a function defined over a subset of the mesh skeleton. We believe this can help ease the exposition especially for readers with little exposure to nonconforming FEM. When discretising Poisson using nonconforming linear elements this problem has a similar structure to the one we consider here and may aid the reader in the exposition. What we show is that a linear convergence rate in the broken norm may be recovered even if the source term is, a priori,  $H^{-1}$  over the domain but with a mesh-aligned singularity.*

We now estimate each term in (3.21) starting with (3.21a). We now drop the binomial coefficients to ease the notation.

For any constant  $c_e$  and using Cauchy-Schwartz:

$$\begin{aligned} & \sum_{i=1}^2 \int_{\mathcal{E}_{th}} b_2 \langle D^2 \hat{w}_t^i, \llbracket D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle ds \\ &= \sum_{i=1}^2 \int_{\mathcal{E}_{th}} b_2 \langle D^2 \hat{w}_t^i - c_e, \llbracket D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle ds \\ &\leq b_2 \sum_{i=1}^2 \sum_{e \in \mathcal{E}_{th}} \|D^2 \hat{w}_t^i - c_e\|_{0,e} \|\llbracket D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket\|_{0,e}. \end{aligned}$$

Now we multiply by  $1 = |e|^{-1/2}|e|^{1/2}$ :

$$\begin{aligned} & \sum_{i=1}^2 \int_{\mathcal{E}_{th}} b_2 \langle D^2 \hat{w}_t^i, \llbracket D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle ds \\ &= b_2 \sum_{i=1}^2 \sum_{e \in \mathcal{E}_{th}} |e|^{-1/2} \|D^2 \hat{w}_t^i - c_e\|_{0,e} |e|^{1/2} \|\llbracket D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket\|_{0,e} \\ &\leq b_2 \left( \sum_{i=1}^2 \sum_{e \in \mathcal{E}_{th}} |e|^{-1} \|D^2 \hat{w}_t^i - c_e\|_{0,e}^2 \right)^{1/2} (|e| \|\llbracket D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket\|_{0,e}^2)^{1/2}. \end{aligned}$$

Now using (3.22):

$$\begin{aligned} & \sum_{i=1}^2 \int_{\mathcal{E}_{th}} b_2 \langle D^2 \hat{w}_t^i, \llbracket D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle ds \\ &\lesssim b_2 \left( \sum_{i=1}^2 \sum_{e \in \mathcal{E}_{th}} \min_{K \in \mathcal{K}_e} [h_K^{-2} \|D^2 \hat{w}_t^i - c_e\|_{0,K}^2 + |D^2 \hat{w}_t^i|_{1,K}^2] \right)^{1/2} \\ &\quad \times (h_K^2 |D(y_h^i - [\Pi_h^c y_h]^i)|_{1,K}^2)^{1/2}. \end{aligned}$$

Now since  $c_e$  was arbitrary we can take it to be the constant in  $\mathbb{R}$  that is the closest approximation of  $D^2\hat{w}_t^i$ . Projection of Sobolev functions onto constants is well-understood; by e.g. [89, Proposition 1.135] we know that:

$$\|D^2\hat{w}_t^i - |K|^{-1} \int_K D^2\hat{w}_t^i dx\|_{0,K} \lesssim h_K |D^2\hat{w}_t^i|_{1,K},$$

with a constant independent of  $h_K$ . Squaring on both sides cancels the dependence on  $h_K$  in the previous estimate so we get:

$$\begin{aligned} & \sum_{i=1}^2 \int_{\hat{\mathcal{E}}_{th}} b_2 \min_{K \in \mathcal{K}_e} \langle D^2\hat{w}_t^i, \llbracket D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle ds \\ & \lesssim b_2 \left( \sum_{i=1}^2 \sum_{e \in \hat{\mathcal{E}}_{th}} \min_{K \in \mathcal{K}_e} |D^2\hat{w}_t^i|_{1,K}^2 \right)^{1/2} (h_K^2 |D(y_h^i - [\Pi_h^c y_h]^i)|_{1,K}^2)^{1/2} \\ & \lesssim b_2 h \|D^3\hat{w}_t^i\|_{0,\Omega_{th}} \|y_h^i - [\Pi_h^c y_h]^i\|_{m,\Omega_{th}} \\ & \lesssim b_2 h \|D^3\hat{w}_t^i\|_{0,\Omega_{th}} \|y_h\|_{m,\Omega_{th}}, \end{aligned}$$

using (3.18) in the last step.

Next we estimate (3.21b). It is clear that the only difference from (3.21a) here is the operator  $D^4$  instead of  $D^2$  applied to  $\hat{w}_t^i$ , so by the same steps as previously:

$$\begin{aligned} & \sum_{i=1}^2 \int_{e \in \hat{\mathcal{E}}_{th} \setminus q_{th} \circ S^1} b_3 \langle D^4\hat{w}_t^i, \llbracket D(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle ds \\ & \lesssim b_3 h \|D^5\hat{w}_t^i\|_{0,\Omega_{th}} \|y_h\|_{m,\Omega_{th}}. \end{aligned}$$

We now treat (3.21c). For an arbitrary constant  $c_e$ :

$$\begin{aligned} & \sum_{i=1}^2 \int_{\hat{\mathcal{E}}_{th} \setminus q_{th} \circ S^1} -b_3 \langle D^3\hat{w}_t^i, \llbracket D^2(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle ds \\ & = \sum_{i=1}^2 \int_{\hat{\mathcal{E}}_{th} \setminus q_{th} \circ S^1} -b_3 \langle D^3\hat{w}_t^i - c_e, \llbracket D^2(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle ds \\ & \leq b_3 \left( \sum_{i=1}^2 \sum_{e \in \hat{\mathcal{E}}_{th} \setminus q_{th} \circ S^1} \|D^3\hat{w}_t^i - c_e\|_{0,e}^2 \right)^{1/2} (\|\llbracket D^2(y_h^i - [\Pi_h^c y_h]^i) \rrbracket\|_{0,e}^2)^{1/2}. \end{aligned}$$

Now using (3.22) again and Cauchy-Schwartz:

$$\begin{aligned}
 & \sum_{i=1}^2 \int_{\hat{\mathcal{E}}_{th} \setminus q_{th} \circ S^1} -b_3 \langle D^3 \hat{w}_t^i, \llbracket D^2(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle ds \\
 & \leq b_3 \left( \sum_{i=1}^2 \sum_{e \in \hat{\mathcal{E}}_{th} \setminus q_{th} \circ S^1} \min_{K \in \mathcal{K}_e} h_K^{-2} \|D^3 \hat{w}_t^i - c_e\|_{0,K}^2 + |D^3 \hat{w}_t^i|_{1,K}^2 \right)^{1/2} \\
 & \quad \times (h_K^2 |D^2(y_h^i - [\Pi_h^c y_h]^i)|_{1,K}^2)^{1/2} \\
 & \lesssim b_3 \left( \sum_{i=1}^2 \sum_{e \in \hat{\mathcal{E}}_{th} \setminus q_{th} \circ S^1} |D^3 \hat{w}_t^i|_{1,K}^2 \right)^{1/2} (h_K^2 |D^2(y_h^i - [\Pi_h^c y_h]^i)|_{1,K}^2)^{1/2}.
 \end{aligned}$$

Using the trace theorem we get:

$$\begin{aligned}
 & \sum_{i=1}^2 \int_{\hat{\mathcal{E}}_{th} \setminus q_{th} \circ S^1} -b_3 \langle D^3 \hat{w}_t^i, \llbracket D^2(y_h^i - [\Pi_h^c y_h]^i) \rrbracket \rangle ds \\
 & \lesssim b_3 h \left( \sum_{i=1}^2 \sum_{K \in \Omega_{th}} |D^3 \hat{w}_t^i|_{1,K}^2 \right)^{1/2} \|y_h^i - [\Pi_h^c y_h]^i\|_{m, \Omega_{th}} \\
 & = b_3 h \|D^4 \hat{w}_t^i\|_{0, \Omega_{th}} \|y_h^i - [\Pi_h^c y_h]^i\|_{m, \Omega_{th}} \\
 & \lesssim b_3 h \|D^4 \hat{w}_t^i\|_{0, \Omega_{th}} \|y_h\|_{m, \Omega_{th}}.
 \end{aligned}$$

Estimating (3.21d), (3.21e) and (3.21f) rely on the trace theorem and a property of the conforming relatives operator in (3.18). For (3.21d) we have:

$$\begin{aligned}
 & \sum_{i=1}^2 \int_{q_{th} \circ S^1} -b_2 \langle \llbracket D^3 \hat{w}_t^i \rrbracket, y_h^i - [\Pi_h^c y_h]^i \rangle ds \\
 & \leq b_2 \sum_{i=1}^2 \sum_{e \in q_{th} \circ S^1} \|\llbracket D^3 \hat{w}_t^i \rrbracket\|_{0,e} \|y_h^i - [\Pi_h^c y_h]^i\|_{0,e}.
 \end{aligned}$$

By the trace theorem,  $\|y_h^i - [\Pi_h^c y_h]^i\|_{0,e} \lesssim \sum_{K \in \mathcal{K}_e} \|y_h^i - [\Pi_h^c y_h]^i\|_{1,K}$  so by using this and (3.18) in the above estimate:

$$\begin{aligned}
 & \sum_{i=1}^2 \int_{q_{th} \circ S^1} -b_2 \langle \llbracket D^3 \hat{w}_t^i \rrbracket, y_h^i - [\Pi_h^c y_h]^i \rangle ds \\
 & \lesssim b_2 \sum_{i=1}^2 \sum_{e \in q_{th} \circ S^1} \|\llbracket D^3 \hat{w}_t^i \rrbracket\|_{0,e} \|y_h^i - [\Pi_h^c y_h]^i\|_{1, \Omega_{th}} \\
 & \lesssim b_2 h^2 \sum_{i=1}^2 \sum_{e \in q_{th} \circ S^1} \|\llbracket D^3 \hat{w}_t^i \rrbracket\|_{0,e} \|y_h\|_{m, \Omega_{th}}.
 \end{aligned}$$

By a similar argument for (3.21e):

$$\begin{aligned}
& \sum_{i=1}^2 \int_{q_{th} \circ S^1} -b_3 \langle [D^5 \hat{w}_t^i], y_h^i - [\Pi_h^c y_h]^i \rangle ds \\
& \lesssim b_3 \sum_{i=1}^2 \sum_{e \in q_{th} \circ S^1} \| [D^5 \hat{w}_t^i] \|_{0,e} \| y_h^i - [\Pi_h^c y_h]^i \|_{1, \Omega_{th}} \\
& \lesssim b_2 h^2 \sum_{i=1}^2 \sum_{e \in q_{th} \circ S^1} \| [D^5 \hat{w}_t^i] \|_{0,e} \| y_h \|_{m, \Omega_{th}}.
\end{aligned}$$

Term (3.21f) is estimated similarly:

$$\begin{aligned}
& \sum_{i=1}^2 \int_{q_{th} \circ S^1} b_3 [D^4 \hat{w}_t^i D(y_h^i - [\Pi_h^c y_h]^i)] ds \\
& = b_3 \sum_{i=1}^2 \int_{q_{th} \circ S^1} \langle \eta_{K_e^+} D^4 \hat{w}_t^i|_{K_e^+}, D(y_h^i - [\Pi_h^c y_h]^i)|_{K_e^+} \rangle \\
& \quad + \langle \eta_{K_e^-} D^4 \hat{w}_t^i|_{K_e^-}, D(y_h^i - [\Pi_h^c y_h]^i)|_{K_e^-} \rangle ds.
\end{aligned}$$

Using Cauchy-Schwartz twice, the trace theorem and (3.18) we proceed as previously:

$$\begin{aligned}
& \sum_{i=1}^2 \int_{q_{th} \circ S^1} b_3 [D^4 \hat{w}_t^i D(y_h^i - [\Pi_h^c y_h]^i)] ds \\
& \leq b_3 \sum_{i=1}^2 \sum_{e \in q_{th} \circ S^1} \| \eta_{K_e^+} D^4 \hat{w}_t^i|_{K_e^+} \|_{0,e} \| D(y_h^i - [\Pi_h^c y_h]^i)|_{K_e^+} \|_{0,e} \\
& \quad + \| \eta_{K_e^-} D^4 \hat{w}_t^i|_{K_e^-} \|_{0,e} \| D(y_h^i - [\Pi_h^c y_h]^i)|_{K_e^-} \|_{0,e} \\
& \lesssim b_3 \sum_{i=1}^2 \sum_{e \in q_{th} \circ S^1} (\| \eta_{K_e^+} D^4 \hat{w}_t^i|_{K_e^+} \|_{0,e} + \| \eta_{K_e^-} D^4 \hat{w}_t^i|_{K_e^-} \|_{0,e}) \| D(y_h^i - [\Pi_h^c y_h]^i) \|_{1, \Omega_{th}} \\
& \lesssim b_3 h \left( \sum_{i=1}^2 \sum_{e \in q_{th} \circ S^1} (\| \eta_{K_e^+} D^4 \hat{w}_t^i|_{K_e^+} \|_{0,e} + \| \eta_{K_e^-} D^4 \hat{w}_t^i|_{K_e^-} \|_{0,e}) \right) \| y_h \|_{m, \Omega_{th}} \\
& \leq b_3 h \| D^4 \hat{w}_t \|_{0, \Omega_{th}} \| y_h \|_{m, \Omega_{th}}.
\end{aligned}$$

Finally we estimate (3.21g):

$$\begin{aligned}
& \sum_{i=1}^2 \int_{q_{th} \circ S^1} -b_3 [D^3 \hat{w}_t^i D^2(y_h^i - [\Pi_h^c y_h]^i)] ds \\
& = -b_3 \sum_{i=1}^2 \int_{q_{th} \circ S^1} \{ D^3 \hat{w}_t^i \} \cdot [D^2(y_h^i - [\Pi_h^c y_h]^i)] + [D^3 \hat{w}_t^i] \{ D^2(y_h^i - [\Pi_h^c y_h]^i) \} ds \\
& = T_1 + T_2.
\end{aligned}$$

We estimate  $T_1$  using the same method as for the term (3.21c). The only difference is the presence of the averaging operator, but we can write  $\{D^3\hat{w}_t^i\} = \frac{1}{2}D^3\hat{w}_t^i|_{K_e^-} + \frac{1}{2}D^3\hat{w}_t^i|_{K_e^+}$ , and since the restrictions of  $D^3\hat{w}_t^i$  to  $K_e^+$  and  $K_e^-$  are smooth there are no changes to the procedure for the term (3.21c).

$$T_1 \lesssim b_3 h \|D^4\hat{w}_t\|_{0,\Omega_{th}} \|y_h\|_{m,\Omega_{th}}.$$

Obtaining a convergence rate for  $T_2$  is more difficult owing to the presence of the  $\{D^2(y_h^i - [\Pi_h^c y_h]^i)\}$  term. It is easy to see why, since for an arbitrary edge  $e \in \mathcal{E}_{th}$ ,  $\|D^2(y_h^i - [\Pi_h^c y_h]^i)\|_{0,e} \lesssim \|D^2(y_h^i - [\Pi_h^c y_h]^i)\|_{1,K}$  for some  $K \in \Omega_{th}$  with  $e$  as an edge. Then by (3.18) we know:

$$\left(\sum_{i=1}^2 \sum_{K \in \Omega_{th}} \|D^2(y_h^i - [\Pi_h^c y_h]^i)\|_{1,K}^2\right)^{1/2} \lesssim \|y_h\|_{m,\Omega_{th}},$$

but we do not recover a factor of  $h$  in this bound. We can however bound (3.21g) as follows using an averaging operator over a neighbourhood near an edge (see also [231, Equation 3.12] for an example). We define  $P_{\mathcal{K}_e}$  as the projection operator on  $L^2(\mathcal{K}_e)$  defined by:

$$P_{\mathcal{K}_e} v = \left(\sum_{K \in \mathcal{K}_e} |K|\right)^{-1} \sum_{K \in \mathcal{K}_e} \int_K v \, dx, \quad v \in L^2(\Omega).$$

Then,

$$\begin{aligned} T_2 &= -b_3 \sum_{i=1}^2 \int_{q_{th} \circ S^1} \llbracket D^3\hat{w}_t^i \rrbracket \{D^2(y_h^i - [\Pi_h^c y_h]^i)\} \, ds \\ &= -b_3 \sum_{i=1}^2 \int_{q_{th} \circ S^1} \llbracket D^3\hat{w}_t^i - P_{\mathcal{K}_e} [D^3\hat{w}_t^i] \rrbracket \{D^2(y_h^i - [\Pi_h^c y_h]^i)\} \, ds \\ &\leq b_3 \sum_{i=1}^2 \sum_{e \in q_{th} \circ S^1} \|\llbracket D^3\hat{w}_t^i - P_{\mathcal{K}_e} [D^3\hat{w}_t^i] \rrbracket\|_{0,e} \|\{D^2(y_h^i - [\Pi_h^c y_h]^i)\}\|_{0,e}. \end{aligned}$$

Now using the trace theorem and (3.18):

$$\begin{aligned} T_2 &\lesssim b_3 \sum_{i=1}^2 \sum_{e \in q_{th} \circ S^1} \|\llbracket D^3\hat{w}_t^i - P_{\mathcal{K}_e} [D^3\hat{w}_t^i] \rrbracket\|_{0,e} \|y_h^i - [\Pi_h^c y_h]^i\|_{m,\Omega_{th}} \\ &\lesssim b_3 \sum_{i=1}^2 \sum_{e \in q_{th} \circ S^1} \|\llbracket D^3\hat{w}_t^i - P_{\mathcal{K}_e} [D^3\hat{w}_t^i] \rrbracket\|_{0,e} \|y_h\|_{m,\Omega_{th}}. \end{aligned}$$

Collecting the bounds we have for (3.21) so far and dividing by  $\|y_h\|_{m,\Omega_{th}}$  yields:

$$\begin{aligned} \frac{a_h(\hat{w}_t - w_{th}, y_h)}{\|y_h\|_{m,\Omega_{th}}} &\lesssim h \left( \sum_{i=1}^2 \|D^3 \hat{w}_t^i\|_{0,\Omega_{th}} + \|D^5 \hat{w}_t^i\|_{0,\Omega_{th}} + \|D^4 \hat{w}_t^i\|_{0,\Omega_{th}} \right. \\ &\quad + h \sum_{e \in q_{th} \circ S^1} \| [D^3 \hat{w}_t^i] \|_{0,e} \\ &\quad + h \sum_{e \in q_{th} \circ S^1} \| [D^5 \hat{w}_t^i] \|_{0,e} + 2 \| D^4 \hat{w}_t^i \|_{0,\Omega_{th}} \Big) \\ &\quad + \sum_{e \in q_{th} \circ S^1} \| [D^3 \hat{w}_t^i - P_{\mathcal{K}_e} [D^3 \hat{w}_t^i]] \|_{0,e} \end{aligned}$$

Using Strang's lemma 41 and corollary 47 we can therefore say:

$$\begin{aligned} \|\hat{w}_t - w_{th}\|_{m,\Omega_{th}} &\lesssim h \left( \sum_{i=1}^2 \|D^3 \hat{w}_t^i\|_{0,\Omega_{th}} + \|D^5 \hat{w}_t^i\|_{0,\Omega_{th}} + \|D^4 \hat{w}_t^i\|_{0,\Omega_{th}} \right. \\ &\quad + h \sum_{e \in q_{th} \circ S^1} \| [D^3 \hat{w}_t^i] \|_{0,e} \\ &\quad + h \sum_{e \in q_{th} \circ S^1} \| [D^5 \hat{w}_t^i] \|_{0,e} + 2 \| D^4 \hat{w}_t^i \|_{0,\Omega_{th}} \Big) \\ &\quad + \sum_{e \in q_{th} \circ S^1} \| [D^3 \hat{w}_t^i - P_{\mathcal{K}_e} [D^3 \hat{w}_t^i]] \|_{0,e} \\ &\quad + h^s |\hat{w}_t|_{m+s,\Omega_{th}}, \end{aligned}$$

for any  $s \geq 0$ , and since  $\|\hat{u}_t - u_{th}\|_{m,\Omega_{th}} \leq \|\hat{w}_t - w_{th}\|_{m,\Omega_{th}}$  the result follows.  $\square$

**Remark 57** (Conforming reconstruction operator). *In the analysis above we make heavy use of the projection operator  $P_{\mathcal{K}_e}$  as it is constructed in such a way as to be continuous over the edge  $e$  (trivially so as the range of this operator is the space of constants). [104] proposes a  $C^1$  reconstruction operator with attractive properties (see e.g. equation (16) in the reference) that could be explored instead. These properties concern error estimation of the projection error on elements  $K$  of the triangulation. However, convergence in terms of  $h$  is not readily obtained using such operators as the preceding analysis deals with terms of like  $\sum_{e \in q_{th} \circ S^1} \| [D^3 \hat{w}_t^i - P_{\mathcal{K}_e} [D^3 \hat{w}_t^i]] \|_{0,e}$  concerning the jump over edges. By using a trace inequality to pull the estimate to  $K$  we lose an order of differentiation which is reflected in the estimation properties of the aforementioned operator (roughly speaking: the  $L^2$  norm on  $\partial K$  can be bounded by the  $H^1$  norm on  $K$ ). It is therefore not certain that such a conforming reconstruction operator yields estimates that scale well with  $h$ .*

**Lemma 58** (Curve convergence). *Suppose assumption 26, 33 and 40 hold and  $q_0, p_0 \in H^j(S^1)^d$  for  $j \geq 1$  and  $p_{0h} = \mathbb{I}_{S_h^p}[p_0]$  is the canonical interpolate of  $p_0$  so that  $\|p_0 - p_{0h}\|_{0,S^1} \leq h \|p_0\|_{1,S^1}$ ,  $0 < h < h^*$ , where  $h^*$  stems from*

assumption 40. Suppose further that corollary 51 holds so that for all  $t \in [0, 1]$ ,  $\lim_{h \rightarrow 0} \|\hat{u}_t - u_{th}\|_{m, \Omega_{th}} = 0$ . Then:

$$\lim_{h \rightarrow 0} \|q_t - q_{th}\|_{0, S^1} = 0,$$

and we have the following bounds for  $t \in [0, 1]$ :

$$\begin{aligned} \|q_t - q_{th}\|_{0, S^1} &\lesssim h \|q_0\|_{1, S^1} + \int_0^t (1 + h^{m-1}) (\|p_s - p_{sh}\|_{0, S^1} + \|\hat{u}_s - u_{sh}\|_{m, \Omega_{th}}) \\ &\quad + h^{m-1} \|u_s\|_{m, \Omega_{th}} \, ds, \end{aligned}$$

and

$$\begin{aligned} \|p_t - p_{th}\|_{0, S^1} &\lesssim h \|p_0\|_{1, S^1} + \int_0^t \|q_s - q_{sh}\|_{0, S^1} + h(1 + \|\sigma_s\|_{4, \Omega} + \|\sigma_s\|_{5, \Omega}) \\ &\quad + \|\hat{u}_s - u_{sh}\|_{m, \Omega_{th}} \, ds. \end{aligned}$$

**Remark 59** (Convergence rate). *The convergence rate of the integrands  $\|\hat{u}_s - u_{sh}\|_{m, \Omega_{th}}$  is a limitation to linear convergence of  $\|q_t - q_{th}\|_{0, S^1}$  as a function of  $h$ . This rate is discussed at length later and summarised in corollary 54.*

*Proof.* Using (3.6c), (3.11c) and Jensen's inequality:

$$\begin{aligned} \|q_t - q_{th}\|_{0, S^1} &\leq \|q_0 - q_{0h}\|_{0, S^1} + \int_0^t \|u_s \circ q_s - \mathbb{I}_\Omega^{1,d}[u_{sh}] \circ q_{sh}\|_{0, S^1} \, ds \\ &\leq \|q_0 - q_{0h}\|_{0, S^1} + \int_0^t \|u_s \circ q_s - u_s \circ q_{sh}\|_{0, S^1} \\ &\quad + \|u_s \circ q_{sh} - u_{sh} \circ q_{sh}\|_{0, S^1} \\ &\quad + \|u_{sh} \circ q_{sh} - \mathbb{I}_\Omega^{1,d}[u_{sh}] \circ q_{sh}\|_{0, S^1} \, ds. \quad (3.23) \end{aligned}$$

The first term in the time integral is bounded by  $\|u_s\|_{W^{1,\infty}(\Omega)} \|q_s - q_{sh}\|_{0, S^1}$ . For the second term above we reuse the notation from definition 32 and let  $\hat{I}$  denote a reference interval (e.g.  $[0, 1]$ ) of measure  $|\hat{I}|$ . Let  $d\theta$ ,  $ds$  and  $d\hat{s}$  denote the Lebesgue measures over  $S^1$ ,  $\mathcal{E}_{th}$ , and  $\hat{I}$ , respectively. Then we know  $|\frac{\partial\theta}{\partial\hat{s}}| = \frac{2\pi}{|\hat{I}|N_{\mathcal{J}}}$ . We denote by  $|\frac{\partial\hat{s}}{\partial s}|_i^{-1} = h_{q_{sh} \circ S_i^1}$  the length of the line segment  $q_{sh} \circ S_i^1$ . Using assumption 33 we see that:

$$\begin{aligned} \|u_s \circ q_{sh} - u_{sh} \circ q_{sh}\|_{0, S^1}^2 &= \sum_{i \in \mathcal{J}} \|u_s \circ q_{sh} - u_{sh} \circ q_{sh}\|_{0, S_i^1}^2 \\ &= \sum_{i \in \mathcal{J}} \frac{2\pi}{h_{q_{sh} \circ S_i^1} |\hat{I}| N_{\mathcal{J}}} \|u_s - u_{sh}\|_{0, q_{sh} \circ S^1}^2 \\ &\leq C_0 \frac{2\pi}{|\hat{I}|} \|u_s - u_{sh}\|_{0, q_{sh} \circ S^1}^2. \end{aligned}$$

Since  $u_s - u_{sh} \in H^1(\Omega_{th})$  we invoke the trace theorem to chain together the bounds:

$$\|u_s \circ q_{sh} - u_{sh} \circ q_{sh}\|_{0, S^1} \lesssim \|u_s - u_{sh}\|_{0, q_{sh} \circ S^1} \lesssim \|u_s - u_{sh}\|_{1, \Omega_{th}},$$

which in turn is controlled by the energy norm  $\|u_s - u_{sh}\|_{m, \Omega_{th}}$ . Before we proceed to show convergence for this we first treat the third term on the right-hand side of (3.23). Note that for some sufficiently smooth  $f$  assumption 33 provides the estimate:

$$\begin{aligned} \|f \circ q_{sh}\|_{0, S^1}^2 &= \sum_{i \in \mathcal{J}} \|f \circ q_{sh}\|_{0, S_i^1}^2 \\ &\leq C_0 \frac{2\pi}{|\hat{I}|} \sum_{i \in \mathcal{J}} \|f\|_{0, q_{sh} \circ S_i^1}^2. \end{aligned} \quad (3.24)$$

Using (3.24) where  $f = u_{sh} - \mathbb{I}_\Omega^{1,d}[u_{sh}]$  we get:

$$\begin{aligned} \|u_{sh} \circ q_{sh} - \mathbb{I}_\Omega^{1,d}[u_{sh}] \circ q_{sh}\|_{0, S^1} &\leq \sqrt{\frac{C_0 2\pi}{|\hat{I}|}} \|u_{sh} - \mathbb{I}_\Omega^{1,d}[u_{sh}]\|_{0, q_{sh} \circ S^1} \\ &\lesssim \|u_{sh} - \mathbb{I}_\Omega^{1,d}[u_{sh}]\|_{1, \Omega_{th}} \\ &\lesssim h^{m-1} |u_{sh}|_{m, \Omega_{th}} \\ &\leq h^{m-1} \|u_{sh}\|_{m, \Omega_{th}}, \end{aligned}$$

where we have used the change of variables to  $q_{sh} \circ S^1$  in the first step, and then trace theorem in the second. The third step above relies on standard local interpolation error estimate, [42, Theorem 4.4.4 and 4.4.20]. Based on this we can say:

$$\begin{aligned} \|q_t - q_{th}\|_{0, S^1} &\lesssim \|q_0 - q_{0h}\|_{0, S^1} + \int_0^t \|u_s\|_{W^{1,\infty}(\Omega)} \|q_s - q_{sh}\|_{0, S^1} \\ &\quad + \|u_s - u_{sh}\|_{m, \Omega_{th}} + h^{m-1} \|u_{sh}\|_{m, \Omega_{th}} \, ds \\ &\leq \|q_0 - q_{0h}\|_{0, S^1} + \int_0^t \|u_s\|_{W^{1,\infty}(\Omega)} \|q_s - q_{sh}\|_{0, S^1} \\ &\quad + (1 + h^{m-1}) \|u_s - u_{sh}\|_{m, \Omega_{th}} \\ &\quad + h^{m-1} \|u_s\|_{m, \Omega_{th}} \, ds. \end{aligned}$$

**Remark 60** (On introducing the auxiliary problem (3.13)). *Solving this auxiliary problem is a key step in the proof. To see why, first note that there is no obvious way of comparing  $u_s$  to  $u_{sh}$  in sense of convergence. It is important to understand why this is so in order to follow the analysis. Because the right-hand sides of (3.6b) and (3.11b) do not match, we cannot use readily apply convergence results from the nonconforming literature (see Strang's result later in lemma 41). This is not a limitation inherent to nonconforming finite elements but simply because the question of convergence no longer makes sense if the source terms differ. Indeed we introduce (3.13) because  $\hat{w}_t$  is a continuous quantity (and therefore we can use it as a test function in (3.6b)), but with a similar source term to (3.11b). We cannot dispense with this step.*

By the triangle inequality we have

$$\|u_s - u_{sh}\|_{m, \Omega_{th}} \leq \|u_s - \hat{u}_s\|_{m, \Omega_{th}} + \|\hat{u}_s - u_{sh}\|_{m, \Omega_{th}}. \quad (3.25)$$

First, we bound  $\|u_s - \hat{u}_s\|_{m, \Omega_{th}}$ . By using  $u_s - \hat{u}_s$  as a test function in (3.13a) and (3.6b) and subtracting these equations yields:

$$a(u_s - \hat{u}_s, u_s - \hat{u}_s) = \langle w_s - \hat{w}_s, u_s - \hat{u}_s \rangle_{0, \Omega},$$

so  $\|u_s - \hat{u}_s\|_{m, \Omega_{th}} \leq \|w_s - \hat{w}_s\|_{0, \Omega_{th}}$ . By a similar argument leading to (3.12) in theorem 38 and adding and subtracting  $p_s \cdot (w_s - \hat{w}_s) \circ q_{sh}$  we have:

$$\begin{aligned} \|w_s - \hat{w}_s\|_{m, \Omega}^2 &\leq \left| \int_{S^1} p_s \cdot (w_s - \hat{w}_s) \circ q_s \, d\theta - \int_{S^1} p_{sh} \cdot (w_s - \hat{w}_s) \circ q_{sh} \, d\theta \right| \\ &\leq \left| \int_{S^1} p_s \cdot (w_s - \hat{w}_s) \circ q_s + p_s \cdot (w_s - \hat{w}_s) \circ q_{sh} \right. \\ &\quad \left. - p_s \cdot (w_s - \hat{w}_s) \circ q_{sh} - p_{sh} \cdot (w_s - \hat{w}_s) \circ q_{sh} \, d\theta \right| \\ &\lesssim \|p_s\|_{1, S^1} \|w_s - \hat{w}_s\|_{m, \Omega_{th}} \|q_s - q_{sh}\|_{0, S^1} \\ &\quad + \|w_s \circ q_{sh} - \hat{w}_s \circ q_{sh}\|_{\infty, S^1} \|p_s - p_{sh}\|_{0, S^1}. \end{aligned}$$

Let  $K_i^+$  and  $K_i^-$  denote the two elements of  $\Omega_{th}$  such that  $K_i^+ \cap K_i^- = q_{sh} \circ S_i^1$  and  $\gamma_{K_i^+}, \gamma_{K_i^-}$  the associated trace operators on  $H^1(K_i^+)$  and  $H^1(K_i^-)$ . For the term  $\|w_s \circ q_{sh} - \hat{w}_s \circ q_{sh}\|_{\infty, S^1}$  we notice:

$$\begin{aligned} \|w_s \circ q_{sh} - \hat{w}_s \circ q_{sh}\|_{\infty, S^1} &= \|w_s - \hat{w}_s\|_{\infty, q_{sh} \circ S^1} \\ &\lesssim \|w_s - \hat{w}_s\|_{m, \Omega_{th}} \\ &= \|w_s - \hat{w}_s\|_{m, \Omega}, \end{aligned}$$

where we have used the Sobolev embedding theorem and the boundary conditions on the diffeomorphism to state that  $\partial\Omega_{th}$  is constant in time and therefore bounds the same domain as  $\partial\Omega$ . Now using theorem 22:

$$\begin{aligned} \|u_s - \hat{u}_s\|_{m, \Omega_{th}} &\leq \|w_s - \hat{w}_s\|_{m, \Omega} \\ &\lesssim \|p_s\|_{1, S^1} \|q_s - q_{sh}\|_{0, S^1} + \|p_s - p_{sh}\|_{0, S^1}, \end{aligned} \quad (3.26)$$

which takes care of the first term on the right-hand side in (3.25). Applying Grönwall's inequality [234, Theorem C.8] here yields:

$$\begin{aligned} \|q_t - q_{th}\|_{0, S^1} &\lesssim e^{\int_0^t \|u_s\|_{m, \Omega_{th}} + (1+h^{m-1})\|p_s\|_{1, S^1} \, ds} \times \left( \|q_0 - q_{0h}\|_{0, S^1} \right. \\ &\quad \left. + \int_0^t (1 + h^{m-1}) (\|p_s - p_{sh}\|_{0, S^1} + \|\hat{u}_s - u_{sh}\|_{m, \Omega_{th}}) \right. \\ &\quad \left. + h^{m-1} \|u_s\|_{m, \Omega_{th}} \, ds \right). \end{aligned} \quad (3.27)$$

By the hypothesis,  $\|q_0 - q_{0h}\|_{0, S^1} \leq h \|q_0\|_{0, S^1}$ . The exponential term is uniformly bounded in  $h$ , so we need only examine the momentum term  $\|p_s - p_{sh}\|_{0, S^1}$  using the evolution equations (3.6d) and (3.11d) as well as  $\|\hat{u}_s - u_{sh}\|_{m, \Omega_{th}}$  to establish convergence of  $\|q_t - q_{th}\|_{0, S^1}$ . First we treat the momentum term:

$$\begin{aligned} \|p_t - p_{th}\|_{0, S^1} &\leq \|p_0 - p_{0h}\|_{0, S^1} + \int_0^t \|\sigma_s \circ q_s p_s - \sigma_{sh} \circ q_{sh} p_{sh}\|_{0, S^1} \\ &\quad + \|\sigma_{sh} \circ q_{sh} p_{sh} - \mathbb{P}_{S_h^p}[\sigma_{sh} \circ q_{sh} p_{sh}]\|_{0, S^1} \, ds. \end{aligned}$$

We estimate the second term on the right-hand side of the above using a standard interpolation result [42, Theorem 4.4.4]. Let  $\mathbb{I}_{q_{sh}}^0[\sigma_{sh}]$  denote the matrix whose entries are comprised of the piecewise average of  $\sigma_{sh}$  on  $q_{sh}$  defined by the relation:

$$\mathbb{I}_{q_{sh}}^0[\sigma_{sh}]^{k,l}|_{q_{sh} \circ S_i^1} = \frac{1}{q_{sh} \circ S_i^1} \int_{q_{sh} \circ S_i^1} \sigma_{sh}^{k,l} ds, \quad i \in \mathcal{J}, \quad k, l = 1, d.$$

This should not be confused with  $\mathbb{I}_{S_h^p}$ . Then using the previously mentioned interpolation properties and rearranging:

$$\begin{aligned} & \|\sigma_{sh} \circ q_{sh} p_{sh} - \mathbb{P}_{S_h^p}[\sigma_{sh} \circ q_{sh} p_{sh}]\|_{0,S^1} \\ & \leq \|\sigma_{sh} \circ q_{sh} p_{sh} - \mathbb{I}_{q_{sh}}^0[\sigma_{sh}] \circ q_{sh} p_{sh}\|_{0,S^1} \\ & = \|(\sigma_{sh} \circ q_{sh} - \mathbb{I}_{q_{sh}}^0[\sigma_{sh}] \circ q_{sh}) p_{sh}\|_{0,S^1} \end{aligned}$$

Now taking the supremum of the term in parentheses:

$$\begin{aligned} & \|\sigma_{sh} \circ q_{sh} p_{sh} - \mathbb{P}_{S_h^p}[\sigma_{sh} \circ q_{sh} p_{sh}]\|_{0,S^1} \\ & \leq \|\sigma_{sh} \circ q_{sh} - \mathbb{I}_{q_{sh}}^0[\sigma_{sh}] \circ q_{sh}\|_{\infty,S^1} \|p_{sh}\|_{0,S^1} \\ & = \|\sigma_{sh} - \mathbb{I}_{q_{sh}}^0[\sigma_{sh}]\|_{\infty,q_{sh} \circ S^1} \|p_{sh}\|_{0,S^1} \end{aligned}$$

Since the quantity is  $\|\sigma_{sh} - \mathbb{I}_{q_{sh}}^0[\sigma_{sh}]\|_{\infty,q_{sh} \circ S^1}$  is just a number, we can use the  $\{\cdot\}$  operator to write it as the sum of the restrictions from either side of the edges that form  $q_{sh} \circ S^1$ :

$$\begin{aligned} & \|\sigma_{sh} \circ q_{sh} p_{sh} - \mathbb{P}_{S_h^p}[\sigma_{sh} \circ q_{sh} p_{sh}]\|_{0,S^1} \\ & \leq \|\sigma_{sh} - \mathbb{I}_{q_{sh}}^0[\sigma_{sh}]\|_{\infty,q_{sh} \circ S^1} \|p_{sh}\|_{0,S^1} \\ & = \|\{\sigma_{sh} - \mathbb{I}_{q_{sh}}^0[\sigma_{sh}]\}\|_{\infty,q_{sh} \circ S^1} \|p_{sh}\|_{0,S^1} \end{aligned}$$

Using Cauchy-Schwartz, the properties of the operator  $\mathbb{I}_{q_{sh}}^0$  and the trace theorem we finally obtain:

$$\begin{aligned} & \|\sigma_{sh} \circ q_{sh} p_{sh} - \mathbb{P}_{S_h^p}[\sigma_{sh} \circ q_{sh} p_{sh}]\|_{0,S^1} \\ & \lesssim \sum_{i \in \mathcal{J}} (\|\sigma_{sh}|_{K_i^+} - \mathbb{I}_{q_{sh}}^0[\sigma_{sh}]\|_{\infty,q_{sh} \circ S_i^1} + \|\sigma_{sh}|_{K_i^-} - \mathbb{I}_{q_{sh}}^0[\sigma_{sh}]\|_{\infty,q_{sh} \circ S_i^1}) \|p_{sh}\|_{0,S^1} \\ & \lesssim h \sum_{i \in \mathcal{J}} (|\sigma_{sh}|_{K_i^+}|_{1,q_{sh} \circ S_i^1} + |\sigma_{sh}|_{K_i^-}|_{1,q_{sh} \circ S_i^1}) \|p_{sh}\|_{0,S^1} \\ & \lesssim h \|\sigma_{sh}\|_{m,\Omega_{th}} \|p_{sh}\|_{0,S^1}. \end{aligned}$$

Using the triangle inequality we obtain the bound:

$$\|\sigma_{sh} \circ q_{sh} p_{sh} - \mathbb{P}_{S_h^p}[\sigma_{sh} \circ q_{sh} p_{sh}]\|_{0,S^1} \leq h \|\sigma_{sh}\|_{m,\Omega_{th}} (\|p_s\|_{0,S^1} + \|p_s - p_{sh}\|_{0,S^1}).$$

Using this estimate and continuing:

$$\begin{aligned} \|p_t - p_{th}\|_{0,S^1} &\leq \|p_0 - p_{0h}\|_{0,S^1} + \int_0^t \|\sigma_s \circ q_s p_s - \sigma_{sh} \circ q_{sh} p_{sh}\|_{0,S^1} \\ &\quad + h \|\sigma_{sh}\|_{m,\Omega_{th}} (\|p_s\|_{0,S^1} + \|p_s - p_{sh}\|_{0,S^1}) \, ds. \end{aligned}$$

By adding and subtracting  $\sigma_s \circ q_{sh} p_s$  next and then adding and subtracting  $\sigma_{sh} \circ q_{sh} p_s$  as well as using the Lipschitz continuity of  $\sigma_s$ :

$$\begin{aligned} \|p_t - p_{th}\|_{0,S^1} &\leq \|p_0 - p_{0h}\|_{0,S^1} + \int_0^t \|\sigma_s \circ q_s p_s - \sigma_s \circ q_{sh} p_s\|_{0,S^1} \\ &\quad + \|\sigma_s \circ q_{sh} p_s - \sigma_{sh} \circ q_{sh} p_{sh}\|_{0,S^1} \\ &\quad + h \|\sigma_{sh}\|_{m,\Omega_{th}} (\|p_s\|_{0,S^1} + \|p_s - p_{sh}\|_{0,S^1}) \, ds \\ &\leq \|p_0 - p_{0h}\|_{0,S^1} + \int_0^t \|p_s\|_{1,S^1} \|\sigma_s\|_{W^{1,\infty}(\Omega)} \|q_s - q_{sh}\|_{0,S^1} \\ &\quad + \|\sigma_s \circ q_{sh} p_s - \sigma_{sh} \circ q_{sh} p_{sh}\|_{0,S^1} \\ &\quad + \|\sigma_{sh} \circ q_{sh} p_s - \sigma_{sh} \circ q_{sh} p_{sh}\|_{0,S^1} \\ &\quad + h \|\sigma_{sh}\|_{m,\Omega_{th}} (\|p_s\|_{0,S^1} + \|p_s - p_{sh}\|_{0,S^1}) \, ds. \end{aligned}$$

Now we use the regularity of  $p_s$  to take it out of the norm in the second term on the integral on the right-hand side, and use Lipschitz continuity of  $\sigma_{sh} \circ q_{sh}$ :

$$\begin{aligned} \|p_t - p_{th}\|_{0,S^1} &\leq \|p_0 - p_{0h}\|_{0,S^1} + \int_0^t \|p_s\|_{1,S^1} (\|\sigma_s\|_{W^{1,\infty}(\Omega)} \|q_s - q_{sh}\|_{0,S^1} \\ &\quad + \|\sigma_s \circ q_{sh} - \sigma_{sh} \circ q_{sh}\|_{0,S^1} + h \|\sigma_{sh}\|_{m,\Omega_{th}}) \\ &\quad + (h \|\sigma_{sh}\|_{m,\Omega_{th}} + \|\sigma_{sh} \circ q_{sh}\|_{\infty,S^1}) \|p_s - p_{sh}\|_{0,S^1} \, ds. \end{aligned}$$

We remind ourselves that  $\|\sigma_s\|_{W^{1,\infty}(\Omega)} \leq \|\nabla u_s\|_{W^{1,\infty}(\Omega)}$  so this quantity is uniformly bounded by some constant. Rearranging and summarising the momentum bound:

$$\begin{aligned} \|p_t - p_{th}\|_{0,S^1} &\leq \|p_0 - p_{0h}\|_{0,S^1} + \int_0^t \|p_s\|_{1,S^1} (\|\sigma_s\|_{W^{1,\infty}(\Omega)} \|q_s - q_{sh}\|_{0,S^1} \\ &\quad + \|\sigma_s - \sigma_{sh}\|_{m-1,\Omega_{th}} + h \|\sigma_{sh}\|_{m,\Omega_{th}}) \\ &\quad + (1 + h) \|\sigma_{sh}\|_{m,\Omega_{th}} \|p_s - p_{sh}\|_{0,S^1} \, ds. \end{aligned}$$

For convergence of the term  $\|\sigma_s - \sigma_{sh}\|_{m-1,\Omega_{th}}$  we introduce the auxiliary problem (cf. remark 60) for  $\tilde{\sigma}_{th} \in \tilde{V}_{th}^{d \times d}$ ,  $t \in [0, 1]$ :

$$\langle \tilde{\sigma}_{th} - \nabla u_t^\top, v_h \rangle_{m-1,\Omega_{th}} = 0, \quad \forall v_h \in \tilde{V}_{th}^{d \times d},$$

and proceed by bounding:

$$\|\sigma_s - \sigma_{sh}\|_{m-1,\Omega_{th}} \leq \|\sigma_s - \tilde{\sigma}_{sh}\|_{m-1,\Omega_{th}} + \|\tilde{\sigma}_{sh} - \sigma_{sh}\|_{m-1,\Omega_{th}}. \quad (3.28)$$

Since  $\nabla u_s \in H^5(\Omega)^{d \times d}$ ,  $\sigma_s \in H^{m+4}(\Omega)^{d \times d}$  by inspection of (3.6e), so by [231, Theorem 4.4] we obtain the bound:

$$\|\sigma_s - \tilde{\sigma}_{sh}\|_{m-1, \Omega_{th}} \lesssim h(|\sigma_s|_{4, \Omega} + |\sigma_s|_{5, \Omega}).$$

The second term in (3.28) is clearly bounded by  $\|u_s - u_{sh}\|_{m, \Omega_{th}}$ . Using this we get:

$$\begin{aligned} \|p_t - p_{th}\|_{0, S^1} &\leq \|p_0 - p_{0h}\|_{0, S^1} + \int_0^t \|p_s\|_{1, S^1} \left( \|\sigma_s\|_{W^{1, \infty}(\Omega)} \|q_s - q_{sh}\|_{0, S^1} \right. \\ &\quad \left. + h(|\sigma_s|_{4, \Omega} + |\sigma_s|_{5, \Omega}) \right. \\ &\quad \left. + \|u_s - u_{sh}\|_{m, \Omega_{th}} + h\|\sigma_{sh}\|_{m, \Omega_{th}} \right) \\ &\quad + (1 + h)\|\sigma_{sh}\|_{m, \Omega_{th}} \|p_s - p_{sh}\|_{0, S^1} ds. \end{aligned}$$

Now using the triangle inequality on  $\|u_s - u_{sh}\|_{m, \Omega_{th}}$  and (3.26) we get:

$$\begin{aligned} \|p_t - p_{th}\|_{0, S^1} &\leq \|p_0 - p_{0h}\|_{0, S^1} + \int_0^t \|p_s\|_{1, S^1} \left( \|\sigma_s\|_{W^{1, \infty}(\Omega)} \|q_s - q_{sh}\|_{0, S^1} \right. \\ &\quad \left. + h(|\sigma_s|_{4, \Omega} + |\sigma_s|_{5, \Omega}) \right. \\ &\quad \left. + \|p_s\|_{1, S^1} \|q_s - q_{sh}\|_{0, S^1} + \|p_s - p_{sh}\|_{0, S^1} \right. \\ &\quad \left. + \|\hat{u}_s - u_{sh}\|_{m, \Omega_{th}} + h\|\sigma_{sh}\|_{m, \Omega_{th}} \right) \\ &\quad + (1 + h)\|\sigma_{sh}\|_{m, \Omega_{th}} \|p_s - p_{sh}\|_{0, S^1} ds. \end{aligned}$$

Yielding the Grönwall estimate again:

$$\begin{aligned} \|p_t - p_{th}\|_{0, S^1} &\leq \left( \|p_0 - p_{0h}\|_{0, S^1} + \int_0^t \|p_s\|_{1, S^1} \left( \|\sigma_s\|_{W^{1, \infty}(\Omega)} \|q_s - q_{sh}\|_{0, S^1} \right. \right. \\ &\quad \left. \left. + h(|\sigma_s|_{4, \Omega} + |\sigma_s|_{5, \Omega}) \right. \right. \\ &\quad \left. \left. + \|p_s\|_{1, S^1} \|q_s - q_{sh}\|_{0, S^1} \right. \right. \\ &\quad \left. \left. + \|\hat{u}_s - u_{sh}\|_{m, \Omega_{th}} + h\|\sigma_{sh}\|_{m, \Omega_{th}} \right) ds \right) \\ &\quad \times e^{\int_0^t \|p_s\|_{1, S^1} + (1+h)\|\sigma_{sh}\|_{m, \Omega_{th}} ds}. \end{aligned}$$

By assumption 40,  $\|\sigma_{sh}\|_{m, \Omega_{th}} \leq M_h$  taking  $h$  is sufficiently small. Now taking  $C > 0$  sufficiently large we can write:

$$\begin{aligned} \|p_t - p_{th}\|_{0, S^1} &\leq CM_h e^{(1+h)M_h} \left( \|p_0 - p_{0h}\|_{0, S^1} + \int_0^t \|q_s - q_{sh}\|_{0, S^1} \right. \\ &\quad \left. + h(1 + \|\sigma_s\|_{4, \Omega} + \|\sigma_s\|_{5, \Omega}) + \|\hat{u}_s - u_{sh}\|_{m, \Omega_{th}} ds \right). \end{aligned}$$

Again,  $\|p_0 - p_{0h}\|_{0, S^1}$  vanishes as a function of  $h$ . We assume that the limit  $h \rightarrow 0$  is taken for all previous times  $0 \leq s$  so the term  $\|q_s - q_{sh}\|_{0, S^1}$  also vanishes, at which point we simply bootstrap to  $t > s$ . Finally the term  $\|\hat{u}_s - u_{sh}\|_{m, \Omega_{th}}$  is dealt with by assumption.  $\square$

We briefly summarise our results. First, theorem 22 showed that a  $H^3$  non-conforming but globally continuous finite elements leads to Lipschitz stable spaces. We were then able to confirm an existence result in theorem 38 for the discretised geodesics under very mild assumptions. The assumption on which we rely is also easily checked computationally at essentially no additional cost during simulations. Our main result, lemma 58 showed under which conditions the discretised geodesics approaches the continuous solutions. We saw that the presence of the singular curve integral as a source term prohibits global regularity, and that despite recovering almost everywhere  $C^\infty$  regularity we cannot attain linear convergence as a function of  $h$  using the conforming relatives used in corollary 54.

In summary, the finite element discretisation in (3.11) is convergent for any initial scalar momentum in  $p_{0h} \in H^1(S^1)^d$  and initial curve  $q_{0h}$ . Moreover, the results in this section indicate that the initial momentum  $p_0$  contains all the information necessary to compute the position of the curve  $q$  at time  $t = 1$ , and that we have found a Lipschitz-stable discretisation allowing us to do so. We note, however, that owing to the symmetry-breaking of the discrete system we cannot expect that the discrete kinetic energy is conserved.

### 3.3.1 ALTERNATIVE APPROACHES & PERSPECTIVE

Over the course of this project several discretisations of (3.6) were considered. This section discusses some of these alternative approaches as well as their drawbacks to motivate the choices we have made in section 3.3. We also remark on a few bibliographic details and how our problem is situated in the wider literature.

- The *medius* error analysis approach introduced in [116] may be applicable to our setting. At the core, this is an error analysis technique for elliptic problems when higher regularity of the continuous solution cannot be obtained. An example of this is the  $C^0$ -conforming penalty method in [117] which could be considered for our sixth order problem. Here, Galerkin orthogonality is not preserved either at the price of more involved *a posteriori* error analysis due to presence of penalty term. We have not pursued this direction as it is not clear how to treat certain oscillation terms in the authors' analysis, nor how to arrive at certain technical results (see Lemma 4.1 in the reference) for our space  $\tilde{V}_{th}^d$ .
- We mention that [100] presents a stable mixed finite element method for polyharmonic problems where the stability is independent of the regularity of the solution. This amounts to decomposing e.g. an  $2m^{\text{th}}$  order problem into  $m$  second-order problems. While finite element *convergence* is possible in this context we do not have control over the Lipschitz norm of the discrete velocity. Using the example from remark 46 we see that a second-order problem does not prevent the Lipschitz norm from becoming

arbitrarily large. It is therefore completely essential that the norm we use for the velocity bounds the Lipschitz norm, it is not sufficient to recover higher-order regularity from standard elliptic regularity (again by remark 46). This demonstrates the need for the finite element space introduced in [231] as it provides subspace of Lipschitz functions.

- Theorem 25 gives us an explicit form of the momentum and thus potentially alleviates the need to integrate forward the momentum equation. Using analytical solution  $p_t$  (resp.  $p_{th}$ ) in the velocity equation forces us to treat the remaining system for the velocity and the curve as nonlinear because the Jacobian of the flow map is not linear in the velocity. We conjecture that Schaefer's fixed point theorem can potentially be applied to provide existence of a solution at the continuous level for  $H^6$  velocities (as this, in turn, provides a highly regular Jacobian determinant). It is out of scope of this text to treat the nonlinear case and its discretisation, and we favour results from linear functional analysis to solve the discrete system (3.11).
- An alternative to the moving mesh method is to apply, for all  $t > 0$  a pullback to the system (3.11a),  $\varphi_{th}^{-1}$ , to  $t = 0$ , so the domain never moves. This fixes the source term as an integral over  $S^1$  where the test functions are now always evaluated on  $q_{0h}$  so that, by a simple change of variables, the integration involved in the source term can be done on  $q_{0h}$  for all times and allows us to define the curves using higher-order polynomials (defined over  $q_{0h}$ ). This approach comes at the cost of modifying the bilinear forms  $a(\cdot, \cdot)$  (resp.  $a_h(\cdot, \cdot)$ ) to take into account the Jacobians from the pullback. At the discrete level, these Jacobians are polynomials so continuity and coercivity of the bilinear form is not difficult to prove provided a boundedness assumption on the Jacobians (otherwise the proof of corollary 51 does not translate given the way we construct the sequences  $\{\hat{w}_{th_k}\}_{k \geq 0}$  and the mesh-dependence of our discrete bilinear forms). For the same reason it is not trivial to show that the discrete bilinear form converges to the continuous version since the second and third derivatives of the flow map are not uniformly bounded in  $h$ . While our proposed moving mesh method does not preserve the conservation property of  $u_{th}$  (cf. assumption 40), practitioners may favour a pullback method provided the aforementioned flow map derivatives do not, at least numerically, go to infinity as the mesh is refined.
- We comment on the limitation that  $S_h^q$  is restricted to an affine space meaning the discrete curves are comprised of piecewise affine segments of  $\mathbb{R}^d$ . If the embedded curve  $q_{th}$  were a higher-order polynomial a moving mesh method would move with a higher-order polynomial diffeomorphism  $\varphi_{th}$  and no longer preserve convexity of triangles  $K \in \Omega_{th}$ . We therefore lose key local approximation and regularity results which cannot be dispensed with. While an isoparametric [42, Section 10.4] approach may

be viable we find the restriction of  $q_{th}$  to an affine space sufficient at high enough mesh resolution and has the critical advantage that we can easily write an explicit ODE for its evolution cf. (3.11c).

- Owing to the homogeneity of the operator  $B$  over the domain it may be useful to seek solutions to (3.11) via boundary element methods provided an explicit Green's function can be constructed. We leave this as future work.
- We could cast (3.11b) as a type of transmission problem in sense of [13] (see for instance equation (1.8) in this reference), where the solution  $u_{th}$  on the inside and outside of the curve could be computed separately until a condition similar to (3.20) is verified for the joint flux across the curve  $q_{th} \circ S^1$ . We do, however, foresee the exact same issues viz a viz regularity across the boundary and simply mention this approach for completeness.
- [20] describes an unfitted finite element method which allows domain perturbations to a boundary integral. This is useful when e.g. a higher-order polynomial representation of our curve is approximated using a piece-wise linear representation. This geometric unfit could be useful in the convergence analysis of a higher-order moving mesh method although it is not obvious to construct dynamics for the perturbed curve. Recalling that such explicit dynamics are a precondition for our applications of Grönwall's inequality we have not pursued this area of research further.
- We make a final remark about the term (3.21g) and regularity of  $\hat{w}_t$ . A natural way to bound this is to use higher-order regularity and refine in a tubular neighbourhood around  $q_{th}$  so as to recover some factor of  $h$ . Using a naïve method we could be tempted to use the trace theorem to say:

$$\begin{aligned}
 \sum_{e \in q_{th} \circ S^1} \| [D^3 \hat{w}_t^i] \|_{0,S^1} &\lesssim \sum_{e \in q_{th} \circ S^1} \sum_{K \in \mathcal{K}_e} \| D^4 \hat{w}_t^i |_K \|_{1,K} \\
 &\lesssim (|D^4 \hat{w}_t^i|_{\infty, \Omega_{th}} + |D^5 \hat{w}_t^i|_{\infty, \Omega_{th}}) \sum_{e \in q_{th} \circ S^1} \sum_{K \in \mathcal{K}_e} |K|^{1/2} \\
 &\lesssim (|D^4 \hat{w}_t^i|_{\infty, \Omega_{th}} + |D^5 \hat{w}_t^i|_{\infty, \Omega_{th}}) N_{\mathcal{J}_h} 2h,
 \end{aligned}$$

here we mark  $N_{\mathcal{J}_h}$  with explicit dependence on  $h$  to show that, as the mesh is refined, the amount of edges that form the discrete curve  $q_{th}$  still increases (although many of them will be collinear). By shape-regularity,  $N_{\mathcal{J}_h} = \mathcal{O}(h^{-1})$  ("is on the order of"), which cancels the factor of  $h$  from the fact that  $|K| \rightarrow 0$ .

### 3.3.2 COMPUTING THE FORWARD MAP

In this section we describe a practical implementation of the discretisation, namely how to carry out the integration over  $S^1$  present in (3.11b) and (3.11d)

as well as the time discretisation of the ODEs in (3.11).

Time is discretised into  $T$  discrete time steps  $t \in [0, 1 \dots, T - 1]$  of size  $\Delta t = T^{-1}$ . We therefore approximate the time derivative in (3.11d) by forward Euler,  $\frac{p_{(t+1),h} - p_{th}}{\Delta t}$ , so the discrete-time momentum update is written as

$$\int_{S^1} p_{(t+1),h} \cdot \delta p \, d\theta = \int_{S^1} (p_{th} - \sigma_{th} \circ q_{th} p_{th} \Delta t) \cdot \delta p \, d\theta. \quad (3.29)$$

where we test against all  $\delta p \in \mathbf{S}_h^p$ . In general, finite element software can only integrate expressions over simplices defined by the mesh. As a result we perform a change of variables so the integration over  $S^1$  in this equation (and in the right-hand side of (3.11b)) is carried out over  $q_{th} \circ S^1$ . By definition 30 this curve is presented by a subset of the mesh skeleton for all  $t \in [0, 1]$ . Using the same notation for the reference interval  $\hat{I}$  in lemma 58 we use an affine change of variables from  $S^1$  to  $q_{sh}$  (linear segment by linear segment), so we can write (3.29) as

$$\begin{aligned} & \sum_{i \in \mathcal{J}} \int_{q_{th} \circ S_i^1} p_{(t+1),h} \circ q_{th}^{-1} \cdot \delta p \circ q_{th}^{-1} \, dx \\ &= \sum_{i \in \mathcal{J}} \int_{q_{th} \circ S_i^1} (p_{th} \circ q_{th}^{-1} - \sigma_{th} p_{th} \circ q_{th}^{-1} \Delta t) \cdot \delta q \circ q_{th}^{-1} \, dx, \end{aligned} \quad (3.30)$$

where the Jacobian term  $\frac{2\pi h_{q_{th} \circ S_i^1}}{|\hat{I}|N_{\mathcal{J}}}$  cancels on both sides. By affine invariance of  $\mathbf{S}_h^q$  this update equation is easily computed. In practice,  $p_{(t+1),h} \circ q_{(t+1),h}^{-1}$  is obtained from  $p_{(t+1),h} \circ q_{th}^{-1}$  simply by moving the mesh at each time step. What this shows is that in practice we can compute with a “momentum” function  $p_{th} \circ q_{th}^{-1}$  defined as polynomial over the mesh skeleton which is widely supported in finite element software as well as alleviating the need for mapping between  $S^1$  and  $\Omega_{th}$ . The same applies to the equation for  $q_{th}$  since  $q_{th} = \varphi_{th} \circ q_{0h}$  by (3.8). We also perform a change of variables in the right-hand side of (3.11b):

$$\int_{S^1} p_{th} \cdot v_h \circ q_{th} \, d\theta = \sum_{i \in \mathcal{J}} \frac{2\pi h_{q_{th} \circ S_i^1}}{|\hat{I}|N_{\mathcal{J}}} \int_{q_{th} \circ S_i^1} p_{th} \circ q_{th}^{-1} \cdot v_h \, d\theta.$$

**Remark 61.** Suppose  $q_{th} \in \tilde{V}_{th}^d$ , then the equation:

$$a_h(w_{th}, v_h) = \int_{S^1} p_{th} \cdot v_h \circ q_{th} \, d\theta, \quad \forall v_h \in \tilde{V}_{th}^d, \quad (3.31)$$

is of course still solvable on paper. Numerically, however, this system is very difficult to approximate when  $q_{th}$  is not affine. When this is the case we cannot evaluate the right-hand side term when this using finite element software. This is not the only issue with (3.31). Even if  $q_{th}$  is affine it may not aligned with the mesh, so we have so-called cut cells, i.e. elements  $K$  where the curve

$q_{th} \circ S^1$  intersects the interior of  $K$ . While such cases can be handled, see e.g. [50], the fact that we are using nonconforming elements prohibits any Galerkin orthogonality to be recovered which is key in dealing with such cut cells. Because of this fact we cannot apply the techniques from the convergence result in lemma 58.

This thesis takes the first step towards a finite element discretisation of Hamilton's equations for curves. As demonstrated in the previous section, the discretisation of the geodesics is highly nontrivial and highly technical. We require very strong assumptions to begin to approach the problem. A gap therefore remains between what we can prove, and what we can compute. We remind the reader that the continuous time, discrete space system (3.11) relies on the fact that at every time instance  $t \in [0, 1]$  there exists a shape-regular, quasi-uniform triangulation around the curve  $q_{th}$ . It is extremely important to note that while the discrete curve is moved by the linear diffeomorphism, the theory above does not suggest a moving mesh method because this would mean the entire triangulation moves with the flow. Such a method would, a priori, violate the shape-regularity and quasi-uniformity assumptions on which we rely. A practical discrete time method compliant with the theory above would, at each time step, generate a mesh, say  $\Omega_{th}$ , that is part of a shape-regular, quasi-uniform family given a discrete curve  $q_{th}$ ,  $t \in [0, 1]$ , such that  $q_{th}$  is embedded in the mesh skeleton of  $\Omega_{th}$ . While we have access to mesh generators in the form of Gmsh [105] that produce output such as that shown in figure 3.2, we cannot arbitrarily map discrete curves  $q_{th} : S^1 \rightarrow \mathbb{R}^2$  to triangulations of the ambient space such that the segments of  $q_{th}(S^1)$  is embedded in the mesh skeleton. While our theory above is sound, the lack of such a practical tool forms the gap between theory and practice. In spite of this issue we can present an algorithm that, while dispensing with some rigour, shows promise of our approach to motivate further research into filling the gaps outlined above. We now present algorithm 1 for the discrete-time forward map, after which we remark further on extensions and limitations.

---

**Algorithm 1** Discrete-time forward map

---

```

procedure DISCRETEFORWARDMAP( $p_{0h}, q_{0h}, T, \Delta t$ )
   $t \leftarrow 0$ 
   $\varphi_{th} \leftarrow \text{id}$ 
  while  $t < T$  do
    Compute  $u_{th}$  by solving (3.11a) and (3.11b) with  $p_{th}$  and  $q_{th}$  fixed
     $\varphi_{(t+1)h} \leftarrow \varphi_{th} + \mathbb{I}_{\Omega}^{1,d}[u_{th}]\Delta t$ 
    Update the momentum  $p_{(t+1)h}$  by solving (3.11e), (3.30)
    Update the mesh  $\Omega_{(t+1)h} \leftarrow \varphi_{(t+1)h} \circ \Omega_{0h}$ 
     $t \leftarrow t + 1$ 
  return  $\varphi_{th} \circ q_{0h}$ 

```

---

We use the Firedrake software package with the PETSc back-end [191, 19, 18,

70]. The Zenodo entry [237] for the specific versions of the software components used. The simulations in this chapter require a very bespoke setup, see appendix B for the details. The implementation of the finite element spaces from [231] and their integration in Firedrake was done by Robert Kirby, but code used to generate the results here are the sole work of the author of this thesis.

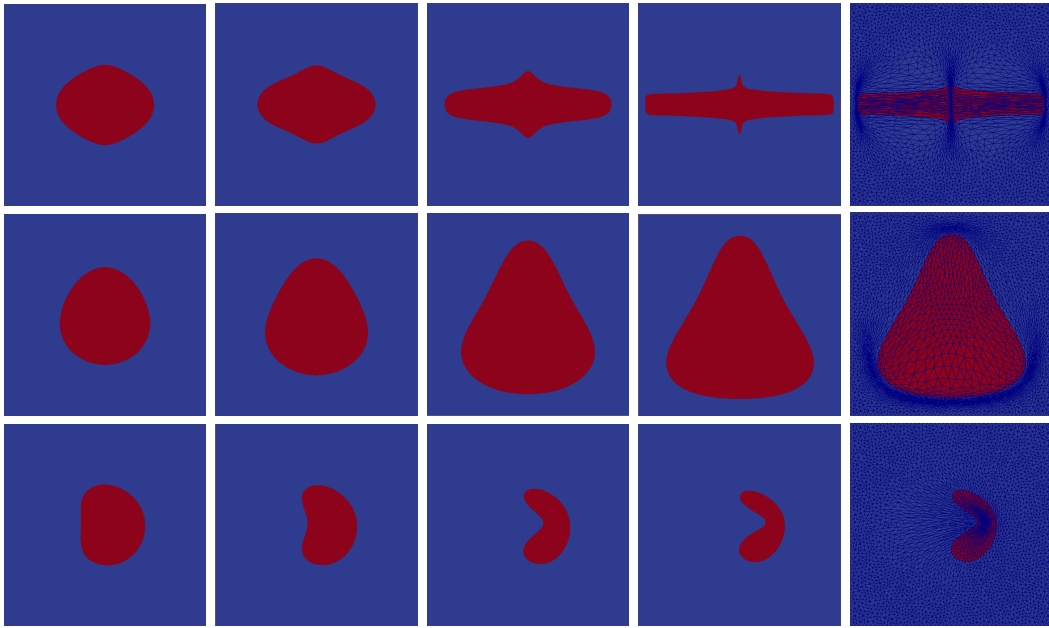


FIGURE 3.3: Evolution of the circle from figure 3.2 obtained by forward integration of the system (3.11) for different initial momenta  $p_{0h}$ . The figures are generated by plotting indicator functions of the interior of the curve  $q_{th} \circ S^1$ . The right-most shows the mesh distortion as well.

Figure (3.3) shows one such result of algorithm 1 for different initial signals  $p_{0h}$  with  $q_{0h}$ , depicted in figure 3.2, being the same for all examples. The evolution of the curve  $q_{th}$  is highlighted. The three initial momenta generating the geodesics in the above figure are given as follows, where  $x$  and  $y$  denote the spatial coordinates of the mesh and  $\eta_{q_{0h}}$  the outward normal to  $q_{0h}$ :

- Top:  $p_{0h} = \tilde{P}\eta_{q_{0h}}$ , where  $\tilde{P} = 87 \cos(\frac{2\pi x}{5})$ ,
- Middle:  $p_{0h} = \tilde{P}\eta_{q_{0h}}$ , where  $\tilde{P} = \begin{cases} 70, & \text{if } y < 0, \\ 90 \exp(-\frac{x^2}{5}), & \text{otherwise,} \end{cases}$
- Bottom:  $p_{0h} = (\tilde{P}^1, \tilde{P}^2)^\top$ , where:

$$\tilde{P}^1 = \begin{cases} 2.0 \times 10^3 \exp(-\frac{y^2}{5}), & \text{if } x < -1 \\ 0, & \text{otherwise,} \end{cases}$$

$$\tilde{P}^2 = -40 \sin(\frac{x}{5})|y|.$$

Here we have chosen  $T = 5$ ,  $b = 1$  (the length scale) and a direct solver was used to invert the discrete system. Here we use lower-upper (LU) decomposition which scales on the order of  $n^3$ , where  $n$  is the number of rows in the square finite element matrix corresponding to the bilinear form in 4 (see [2] for details). At each time step, four linear systems must be computed corresponding to (3.11a), (3.11b), (3.11d) and (3.11e) with the boundary conditions (3.11f). These simulations take three minutes on a 2014 MacBook Pro with a 2.5GHz Intel Core i7 processor and 16GB of 1600MHz DDR3 RAM with no parallelisation or preconditioning.

**Remark 62** (Towards a moving mesh method). *This algorithm assumes that at each time step  $\Omega_{(t+1)h}$  is a member of a shape-regular, quasi-uniform family of triangulations. While  $\varphi_{(t+1)h}$  is Lipschitz it may occur that elements  $K \in \Omega_{(t+1)h}$  become too flat to satisfy condition (2.13). This problem can easily be solved by adding a procedure to algorithm 1 that, after the mesh update, loops over the mesh and flags triangles in violation of (2.13) for refinement until this condition is satisfied. This can be done by, at each time step treating the domain inside and outside of the curve as separate domains, and triangulating these separately to the same refinement level. While this may be a practical solution for the stability of numerical experiments, the moving mesh method presented in algorithm 1 is not guaranteed to be adequately consistent with the theory. A practical solution here would be to choose shape-regularity and quasi-uniformity constants a priori, and numerically validate these a posteriori: if the generated meshed do not show to satisfy the necessary assumptions the numerical experiment is invalidated and new constants must be chosen. A sufficient condition for closing this gap is to prove that, at each time step, the domain  $\Omega_{(t+1)h}$  possesses the necessary shape-regularity and quasi-uniformity requirements invoked in the preceding theory. We conjecture that a shape-regular, quasi-uniform family retains its properties under the group action (of the ambient space) of a sufficiently smooth diffeomorphism. We believe the constants appearing in shape-regularity and quasi-uniformity estimates (i.e. the degree of anisotropy seen in figure (3.3)) would depend on this diffeomorphism and its Jacobian. It is beyond the scope of this thesis to investigate this claim, and leave it as future research.*

We remark that it is not trivial to numerically validate the convergence results obtain in this section. The convergence results that we have shown here is for continuous time. For the reasons given above this obviously cannot be realised numerically, so it is a difficult task to design an experiment providing numerical evidence for curve convergence as a function of spatial mesh resolution. There is a priori no guarantee that limits between time and space commute.

### 3.4 INVERSE PROBLEM

As we see in figure 3.3, different shapes are produced depending on the choice of initial momentum. In this section we discuss the inverse problem of identify-

ing the correct initial momentum bringing into alignment a template with a prescribed target. First, we must describe a way for such an algorithm to systematically *explore* the function space  $H^1(S^1)^d$  and a way to *evaluate* how suitable a given point in this space is. We comment on both of these aspects and describe future work.

As mentioned in the previous chapter, linearity of  $H^1(S^1)^d$  makes shooting an attractive method for solving shape matching problems. Starting with some initial guess defined as a function over the initial curve  $q_{0h}$ , a solution to the discrete geodesic provides a candidate curve,  $q_{1h}$  approximating the target. If  $q_{1h}$  adequately approximates the target (which we comment on later), the procedure terminates. If not, the adjoint equations associated with the problem (3.11) can be solved backwards in time to provide an update to the initial guess. The solution of these equations is also sometimes referred to as *backpropagation*, in particular in the machine learning community. In practice, the adjoint equations are seldom explicit in user-level code due to *automatic differentiation*. This is a software abstraction allowing for the numerical evaluation of the Jacobian of expressions that are given as functions in a programming language [182]. These essentially articulate chain and product rules for computer programmes. The implementation of such an abstraction permits a user to evaluate gradients of for instance complex composite functions for use in an optimisation loop. While Firedrake supports an automatic differentiation engine [99, 177] some additional work is necessary to implement a shooting method for (3.11). We conclude this section by explaining these steps.

To leverage the linearity of  $H^1(S^1)^d$  in a shooting method we also need to equip an inverse problem solver with what is commonly referred to as a *loss*, or *mismatch* function. This is a means of measuring how far  $q_{1h}$  is from a specified target. If the target, say  $q_1$ , is infinite-dimensional, we do not have a practical way of numerically evaluating  $\|q_{1h} - q_1\|_{0,S^1}$ , so we restrict our discussion to the case where the target is only partially observed and specified as a finite-dimensional set of e.g.  $M$  points in  $\mathbb{R}^d$ . In this case, a loss function can be specified as a Euclidean distance or Karcher mean [151]. The parameterisation of the template becomes relevant because we do not a priori know which points on the infinite-dimensional template correspond to each target point. Figure 3.4 depicts this difficulty. In addition to addressing this issue, Firedrake does not currently support differentiation of  $q_{th}$  (in particular when  $t = 1$ ) with respect to the  $M$  point on  $S^1$  that would be necessary in a gradient descent method.

### 3.5 SUMMARY

In this chapter we have shown new theoretical and numerical results for solving curve matching problems. We have proved convergence for a Lipschitz-conforming finite element method and verified the implementation of the finite element space  $\tilde{V}_{th}^d$  of [231] for use as a forward operator. The forward operator

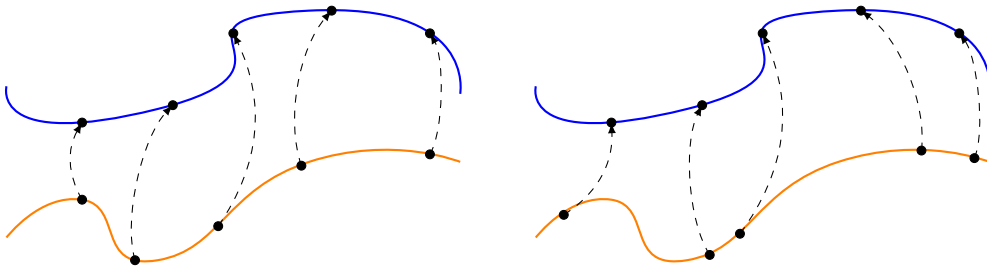


FIGURE 3.4: Matching of curves for different parameterisations of the template (bottom) with the target (top). For simplicity we have only illustrated a subset of the curves.

can easily be integrated into other black box optimisation loops or for use with third party software. Our work here highlights the interplay between Hamiltonian systems (requiring Lipschitz in time), diffeomorphisms (requiring Lipschitz in space) and minimal order regularity for elliptic PDEs. It is particularly striking how the variational crime of nonconforming methods propagate throughout our computations and proofs, in particular forcing the use of a 12<sup>th</sup> order operator in the continuous system. A conforming method would naturally improve all aspects of our work. We have also found a characterisation of the momentum in theorem 25. As mentioned before, this characterisation did not explicitly enter our convergence analysis. The source terms considered for the velocity solutions depend strongly on the velocity field via (3.1). Establishing a convergence proof for the discrete problem with the explicit characterisation of the momentum is an open problem.

We have chosen an outer metric to perform matching with discretised curves. Another option would have been to restrict the motion to a finite set of landmarks as in e.g. [37], where a kernel method was used in the time integration. Such methods are incredibly popular and are amenable to efficient computation via GPU acceleration see e.g. [92, 54]. Forthcoming work also includes applying these methods to real data e.g. in the medical sciences. We also wish to explore parameterisation-independent curve matching using the shape space  $\text{Emb}(S^1, \mathbb{R}^d)/\text{Diff}(S^1)$  which leads to a simplification of the momentum [64]. We also believe this may be a way to solve the problem described by figure 3.4.

# CHAPTER 4

---

## SPACE-TIME METAMORPHOSIS

---

This chapter is dedicated to the metamorphosis problem for image matching introduced in (2.9). The aim is to derive geodesic equations for images in the form of space-time Eulerian frame PDEs and develop finite element discretisations amenable to efficient computation. Defining solution spaces over space-time allow us to leverage computational platforms featuring parallelism in both space and time at the expense of dealing with a larger system. This then defines a forward operator taking as input a time-dependent family of velocity fields and provides the associated geodesics for the images, motivating a gradient descent method to solve the inverse problem for the space-time velocity.

We let  $\Omega = [0, 1] \times [0, 1]^d$ , where  $d$  denotes the dimension of the images we are matching and the first coordinate denotes time. We let  $\Gamma_0$  and  $\Gamma_1$  denote the  $d$ -dimensional submanifolds of  $\Omega$  given by the restriction of  $\Omega$  in the first coordinate to the boundaries i.e.  $t = 0$  (resp.  $t = 1$ ) specifying where we impose the images  $I_j$ ,  $j = 0, 1$  that we aim to match. When  $d = 1$ ,  $\Omega$  is a simple unit square and  $\Gamma_j$ ,  $j = 0, 1$  are simply opposite boundaries; for  $d = 2$  these are opposing facets of the unit cube. Further,  $\Gamma = \Gamma_0 \cup \Gamma_1$ . The subject of this chapter is the study the following abstract problem:

$$\inf_{u, I \in V \times G} \frac{1}{2} \int_0^1 \frac{1}{2} \|u\|_V^2 + \sigma^{-2} \left\| \frac{\partial I}{\partial t} + u \cdot \nabla I \right\|_{0, \Omega}^2 dt \quad (4.1a)$$

$$\text{subject to } I|_{\Gamma_j} = I_j, \quad j = 0, 1. \quad (4.1b)$$

where  $I_j$ ,  $j = 0, 1$  are the template and target images respectively and  $u$  and  $I$  depend both on space and time and  $\sigma^{-2} > 0$  is a parameter. Note that we generally drop explicit time-dependence as we are working in space-time. We aim to solve this problem and determine suitable function spaces  $V$  and  $G$  defined over *space-time* so that  $\frac{\partial}{\partial t} + u \cdot \nabla : G \rightarrow L^2(\Omega)$ . Deferring for the moment to define the spaces  $V$  and  $G$  we first sketch our approach to solving (4.1). We aim to solve a type of relaxation of this problem and implicitly define  $I$  by the velocity via a forward operator in the form of a PDE. Indeed for fixed  $u$ , the second term in (4.1) is a *least-squares advection problem* for  $I$ . With the

previous idea in mind we write (4.1) as:

$$\inf_{(u,z) \in V \times L^2(\Omega)} \frac{1}{2} \int_0^1 \|u\|_V^2 + \sigma^{-2} \|z\|_{0,\Omega}^2 dt \quad (4.2a)$$

$$\text{subject to } z = \frac{\partial I^*}{\partial t}[u] + u \cdot \nabla I^*[u] \quad \text{in } L^2(\Omega), \quad (4.2b)$$

$$I^*[u] \triangleq \arg \inf_{\substack{I \in G, \\ I|_{\Gamma_0} = I_0, \\ I|_{\Gamma_1} = I_1}} \left\| \frac{\partial I}{\partial t} + u \cdot \nabla I \right\|_{L^2(\Omega)}^2. \quad (4.2c)$$

Since the optimisation problem in the constraint is convex in  $I$  we can differentiate the functional in (4.2c) with respect to a variation in  $I$  and find necessary and sufficient optimality conditions for an "optimal" image. To summarise the main idea: we wish to show that  $I$  satisfies a system of linear equations *implicitly defined* for any candidate  $u \in V$  to cast (4.2) as an inverse problem in the  $u$  variable to which a gradient method can be applied.

We highlight some differences between the setting we propose and the literature. Neither the space-time nor the finite element approaches are novel here. Generally speaking, space-time methods are advantageous when a *one-shot* (i.e. solving a PDE for all time and space simultaneously) solution approach is preferred to a time-marching scheme. Early references applying space-time finite elements include [137, 138] in the context of hyperbolic problems and elastodynamics, motivated by the desire to resolve discontinuities in the solution via adaptivity. This approach permits a so-called *unstructured* mesh which means that the elements form an irregular pattern as opposed to e.g. a uniform quadrilateral discretisation in space and time. We are therefore adopting this framework to not only develop a more computationally expedient method but also to pave the way towards an adaptive strategy, where local mesh resolutions in space-time can help accurately resolve sharp image gradients. See the recent textbook [159] for more general space-time methods for PDEs. Finite element methods enjoy a rich literature in the computational anatomy community, see [40, 118]. Much of this work uses anatomical models i.e. applying finite element methods to model tissues from volumetric data. To the best of our knowledge using finite elements to directly model both the image and velocity in an Eulerian setting has not yet been investigated and this forms our contribution to the literature. Related is the work [194] wherein a shooting method for image metamorphosis is developed using reproducing kernel Hilbert spaces. Although this is not a space-time method we believe this is a candidate for the state of the art given that kernel methods are extremely amenable to GPU acceleration as we discuss in a later chapter. We would like to compare our framework with both the firm theoretical setting (of optimal control theory) in this paper as well as the results they produce. We return to this setting also in chapter 5. Finally, very recently, [87] develops a shooting method for metamorphosis of weakly differentiable images via a time-discrete forward map. The regularity

of deformation is ensured through cubic splines on a coarse mesh, while the image is discretised using bilinear elements on a fine mesh.

**Remark 63** (Riemannian formulation). *The work of Trouné [220] attached differentiable structure to an abstract image space  $\mathcal{I}$  to build a Riemannian distance on this space. We highlight some interesting analytical aspects as a result of the space-time construction for metamorphosis that we favour in this chapter. In the aforementioned reference, an image solution solving the metamorphosis problem is found as a curve in image space  $\mathcal{I}$  i.e. a continuous time-indexed map occupying e.g.  $H^1([0, 1]; \mathcal{I})$ .  $H^1([0, 1]; \mathcal{I})$  is a natural choice of space when opting for a Riemannian construction since infinitesimal changes are specified in  $\mathcal{I}$  and then integrated in time. In space-time we desire partial differential equations over one space-time field so our function spaces ( $V$  and  $G$  above) are defined over the joint field  $[0, 1] \times \Omega$  allowing us to use the advection operator  $\frac{\partial}{\partial t} + u \cdot \nabla$ . This operator is not entirely trivial to interpret when the space  $G$  is replaced by  $H^1([0, 1]; \mathcal{I})$  since  $\frac{\partial}{\partial t}$  and  $u \cdot \nabla$  are understood with respect to different spaces<sup>1</sup>. The space  $H^1([0, 1]; \mathcal{I})$  is used in the work [220] because it the metric over the space of images is Riemannian: it is generated from integration of local infinitesimal time-independent changes in image space  $\mathcal{I}$ . Viewing metamorphosis in the space-time frame does not allow for the same characterisation. Therefore the distance measured by (4.1) is, in a sense, a "global" and not Riemannian metric.*

The rest of this chapter is structured as follows. Section 4.1 studies the subproblem given by (4.2c) which provides the forward operator  $u \mapsto I$  as a variational problem. We refer to it as the *inner problem*. We explore several formulations of this issue and contribute with many theoretical results. Some numerical experiments based on a finite element discretisation of these equations also validate this theory. Next, section 4.2 studies the so-called *outer problem*, which is the inverse problem of finding a suitable minimiser  $u \in V$  of the problem (4.2) where we use the inner problem to eliminate the  $I$  variable. Section 4.3 concludes this chapter and highlights our accomplishments as well as the main challenges going forward.

## 4.1 INNER PROBLEM

In this section we solve the least-squared advection problem in (4.2c) for a fixed  $u$ . A minimiser of this convex problem can be shown to satisfy a (necessary and sufficient) coupled system between advection and continuity. We first solve this problem analytically in section 4.1.1 and establish some theoretical results. Next, section 4.1.2 develops a completely novel variational setting in which well-posedness is obtained and from which finite element methods

<sup>1</sup> In fact, the advection operator needs to be interpreted in sense of distributions, as the dual operator to  $i_t \mapsto -\operatorname{div}(ui_t)$ , for  $i_t \in \mathcal{I}$  cf. [220, Definition 9] via an appropriate Green's theorem.

can be derived and contains the main results of this chapter. Section 4.1.3 extends this theory and develops the machinery to describe image matching for discontinuous image matching.

Throughout this chapter,  $u \in V$  will be a fixed but possibly non-constant velocity satisfying the following assumptions:

**Assumption 64.**

1. Let  $V$  be a function space over both space and time such that for  $t \in [0, 1]$ ,  $x \mapsto u_t(x)$  is a Lipschitz mapping in  $x$ ,  $u_t|_{\partial[0,1]^d} = 0$  and  $t \mapsto u_t$  is  $L^1([0, 1])$ , see remark 65 below.
2. Denote by  $\varphi$  the flow map associated with each  $u \in V$  such that for any point  $\hat{x} \in \Gamma_0$ ,  $\varphi$  solves the following system uniquely [234]:

$$\begin{cases} \dot{\varphi}_t(\hat{x}) = u(\varphi_t(\hat{x}), t), & t \in [0, 1], \\ \varphi_0 = \text{id}. \end{cases} \quad (4.3)$$

3. For  $\mathbf{b} = (u, 1)$  there exists constants  $\beta_0$  and  $\beta_1$  such that  $0 < \beta_0 \leq |\mathbf{b}| \leq \beta_1 < \infty$ .

For ease of notation we define  $\frac{\partial I}{\partial t} + u \cdot \nabla I := \mathbf{b} \cdot \nabla_t I$ . Here we use the  $d$ -dimensional row vector  $\mathbf{b} = (u, 1)^\top$ ,  $\nabla_t = (\nabla, \frac{\partial}{\partial t})$  and  $-\text{div}_t$  its adjoint.

**Remark 65.** The spatial regularity for admissible velocity fields prescribed by assumption 64.1 is deserving of a remark. While the Lipschitz requirement is obvious from classical theory we take the intersection with  $H_0^1([0, 1]^d)$  which is kernel of the trace map on  $\partial[0, 1]^d$  in  $H^1([0, 1]^d)$ . To clarify what is mean we give an example when  $d = 1$ :  $\Omega$  is the unit square, and imposed on the "north" and "south" boundaries are the images. A candidate velocity field with the regularity above is Lipschitz in space and vanishes on the "east" and "west" boundaries of  $\Omega$ . This ensures that the induced  $\varphi_t$  does not move the boundary of  $\Omega$  which in the Eulerian setting eliminates the presence of in-flow and out-flow boundaries. The assumption above also establishes a bijection between  $\Gamma_0$  and  $\Gamma_1$  in the sense that for any  $\hat{x}_0 \in \Gamma_0$  there exists a unique  $\hat{x}_1 \in \Gamma_1$  such that  $\hat{x}_0 = \varphi_1^{-1}(\hat{x}_1)$ .

For a fixed velocity field  $u$  we aim to construct the image  $I$  which minimises advection while satisfying the boundary conditions as described by problem 3 below.

**Problem 3 (Inner Problem)**

$$\text{Find } I^* = \arg \inf_{I \in G} \|\mathbf{b} \cdot \nabla_t I\|_{0,\Omega} \quad (4.4a)$$

$$\text{subject to the fixed end-point conditions } I_0 \text{ and } I_1. \quad (4.4b)$$

The rest of this section is structured as follows. First, section 4.1.1 looks at the strong form equations resulting from applying a variational principle to problem 3. We solve this equation analytically and formulate a regularity theorem for the solution. Next, we solve the weak equations associated to the inner problem in section 4.1.2, and investigate the properties of the resulting bilinear operator and the properties of  $G$  such as trace and Poincaré theorems. Section 4.1.3 studies an even weaker formulation amenable for discontinuous images.

#### 4.1.1 ANALYTICAL SOLUTION

To derive the analytical solution to the convex problem described by problem 3 we assume for the moment that everything is smooth and that a unique minimiser to this problem exists then retrace our steps to impose the minimum required regularity of our solutions. Substituting  $\mathbf{b} \cdot \nabla_t I = z$  in problem 3 and formally applying a variational principle to problem results in the strong form equations:

$$\operatorname{div}_t(\mathbf{b}z) = 0, \quad (4.5a)$$

$$\mathbf{b} \cdot \nabla_t I = z, \quad (4.5b)$$

$$I_0 \text{ and } I_1 \text{ fixed}, \quad (4.5c)$$

since continuity is the adjoint of advection. To solve this we need a property of the Jacobian determinant of the flow map  $\varphi_t$  generated by  $u$ . Let  $J_t = \frac{\partial \varphi_t}{\partial \hat{x}}$  denote the Jacobian of  $\varphi_t$  at  $\hat{x}$  and the Jacobian determinant by:

$$\Phi_t = \det J_t. \quad (4.6)$$

Since  $\varphi_t \circ \varphi_t^{-1}(x) = x$ ,  $\Phi_t$  is not degenerate and so the inverse function theorem holds:

$$\begin{aligned} \Phi_t(\hat{x}) &= (\Phi_t^{-1} \circ \varphi_t(\hat{x}))^{-1}, \\ \Phi_t^{-1}(x) &= (\Phi_t \circ \varphi_t^{-1}(x))^{-1}. \end{aligned}$$

This quantity can be seen to satisfy the continuity equation since for  $t \in [0, 1]$  a time slice of the space-time domain  $\Omega_t = \{(x, t) \mid x \in [0, 1]^d\}$  we have:

$$\int_{\Omega_t} \Phi_t^{-1} dx = \int_{\Gamma_0} \Phi_t^{-1} \circ \varphi_t \Phi_t d\hat{x} = \int_{\Gamma_0} d\hat{x} = \text{constant}, \quad (4.7)$$

implying:

$$\frac{d}{dt} \int_{\Omega_t} \Phi_t^{-1} dx = 0,$$

so the integral of  $\Phi_t^{-1}$  is conserved. By Reynold's transport theorem [58] (differentiation under the integral sign and using the divergence theorem) we obtain:

$$\operatorname{div}_t(\mathbf{b}\Phi_t^{-1}) = 0, \quad (4.8a)$$

$$\Phi_0^{-1} = 1. \quad (4.8b)$$

Therefore, since the governing equation (4.5a) for  $z$  is the same it must be of the form  $z = \lambda \Phi_t^{-1}$ , where  $\lambda$  solves  $\frac{\partial \lambda}{\partial t} + u \cdot \nabla \lambda = 0$  as shown by the following computation:

$$\begin{aligned} \frac{\partial z}{\partial t} &= \frac{\partial \lambda}{\partial t} \Phi_t^{-1} + \lambda \frac{\partial \Phi_t^{-1}}{\partial t} \\ &= -u \cdot \nabla \lambda \Phi_t^{-1} - \lambda \operatorname{div}(u \Phi_t^{-1}) \\ &= -\operatorname{div}(u \lambda \Phi_t^{-1}) \\ &= -\operatorname{div}(uz). \end{aligned}$$

so clearly  $\operatorname{div}_t(\mathbf{b}z) = \frac{\partial z}{\partial t} + \operatorname{div}(uz) = 0$ .  $\lambda$  is constant along streamlines which we denote by  $\lambda_{\hat{x}}$  for each selection of initial  $\hat{x} \in \Gamma_0$ . As a result:

$$\frac{\partial I}{\partial t} + u \cdot \nabla I = \lambda_{\hat{x}} \Phi_t^{-1}. \quad (4.9)$$

Letting  $\gamma_t(\hat{x}) = \int_0^t \Phi_s^{-1} \circ \varphi_s(\hat{x}) \, ds$  we define the ansatz solution as a kind of interpolation between the boundary conditions along the streamlines of  $u$ :

$$I(\varphi_t(\hat{x}), t) := \tilde{I}(\hat{x}, t) := \frac{\gamma_t(\hat{x})}{\gamma_1(\hat{x})} [I_1 \circ \varphi_1(\hat{x}) - I_0(\hat{x})] + I_0(\hat{x}). \quad (4.10)$$

For convenience later  $\tilde{I}$  is defined above as simply the image solution in streamline coordinates. Now we differentiate (4.10) with respect to time yields. On the left-hand side we get:

$$\begin{aligned} \frac{d}{dt} I(\varphi_t(\hat{x}), t) &= \nabla I(\varphi_t(\hat{x}), t) \cdot \dot{\varphi}_t \circ \varphi_t(\hat{x}) + \frac{\partial I}{\partial t}(\varphi_t(\hat{x}), t) \\ &= \nabla I(x, t) \cdot u(x, t) + \frac{\partial I}{\partial t}(x, t), \end{aligned}$$

and on the right:

$$\frac{d}{dt} \frac{\gamma_t(\hat{x})}{\gamma_1(\hat{x})} [I_1 \circ \varphi_1(\hat{x}) - I_0(\hat{x})] + I_0(\hat{x}) = \lambda_{\hat{x}} \frac{\partial \gamma_t}{\partial t}(\hat{x}) = \lambda_{\hat{x}} \Phi_t^{-1} \circ \varphi_t(\hat{x}),$$

where  $\lambda_{\hat{x}} = \frac{1}{\gamma_1(\hat{x})} [I_1 \circ \varphi_1(\hat{x}) - I_0(\hat{x})]$ . In summary we see that (4.10) satisfies (4.9), so the solution to (4.5) is given in terms of  $x$  and  $t$  by:

$$I(x, t) = \tilde{I}(\varphi_t^{-1}(x), t) \quad (4.11a)$$

$$z(x, t) = \lambda_{\varphi_t^{-1}(x)} \Phi_t^{-1}(x). \quad (4.11b)$$

(4.11) is valid for  $I_j \in L^2(\Gamma_j)$ ,  $j = 0, 1$ . Next, we state conditions that imply higher-order regularity of the solutions obtained above. Prescribing regularity of Jacobian determinants is in and of itself a field of study, and is particularly non-trivial when restricted to functions  $W^{1,\infty}([0, 1]^d)$ , see [82, 204]. This poses a challenge for the study of regularity of the solutions obtained here. We provide one example below of what we can expect when restricting  $V$  to be a sufficiently smooth subspace.

**Proposition 66** (Streamline regularity). *Suppose  $u$  is such that  $\varphi_t \in \mathbf{C}^2([0, 1]^d)^d$  and assume  $I_j \in H^1(\Gamma_j)$ ,  $j = 0, 1$ . Then for any  $\hat{x} \in \Gamma_0$  the map  $t \mapsto \tilde{I}(\hat{x}, t)$  has a continuous second derivative.*

*Proof.* By using  $\frac{\partial \tilde{I}}{\partial t} = \lambda_{\varphi_t^{-1}(x)} \Phi_t^{-1} \circ \varphi_t$  we simply apply direct computation to arrive at the following expression for the second derivative:

$$\begin{aligned} \frac{\partial^2 \tilde{I}}{\partial t^2} &= \lambda_{\varphi_t^{-1}(x)} \frac{d}{dt} \Phi_t^{-1} \circ \varphi_t \\ &= \lambda_{\varphi_t^{-1}(x)} (\nabla \Phi_t^{-1}) \circ \varphi_t \cdot \frac{\partial \varphi_t}{\partial t} \\ &= \lambda_{\varphi_t^{-1}(x)} (\nabla \Phi_t^{-1}) \circ \varphi_t \cdot u. \end{aligned}$$

Now the  $i^{\text{th}}$  component of  $\nabla \Phi_t^{-1}$ ,  $i = 1, d$ , can be expressed using Jacobi's formula (see e.g. [58]) as:

$$[\nabla \Phi_t^{-1}]^i = \Phi_t^{-1} \text{trace} \left[ J_t \frac{\partial J_t^{-1}}{\partial x_i} \right].$$

The right-hand side is well-defined if the second derivative of the flow map is continuous i.e.  $\frac{\partial^2 \varphi_t^{k,j}}{\partial x_j \partial x_k} \in \mathbf{C}^0([0, 1]^d)$ , for all  $k, j = 1, d$ . Since it is a diffeomorphism the same regularity holds for its inverse and we are done.  $\square$

#### 4.1.2 PRIMAL FORMULATION

This section studies the least-square advection problem described in problem 3. We examine the stability of the resulting least-squares boundary value problem for the image by differentiating the functional in this problem with respect to a variation in the image. What we present here is not the first study of hyperbolic equations in the (least-squares) FE literature, so we first provide a brief historical review. Indeed, the least-squares framework has enjoyed success for application to hyperbolic systems. The early work of [183] presents a least-squares space-time method for the advection-diffusion equation. The direction we take here is mostly similar to the work of Perrochet in 1995 [189]. See the textbook reference [35] for an overview of least-square FEM.

Motivated by the analysis in the preceding section we now determine a candidate definition for the image space induced by the space-time vector field  $\mathbf{b}$ . We define an *energy semi-norm*:

$$\|f\|_{E(\mathbf{b}; \Omega)} = \|\mathbf{b} \cdot \nabla_t f\|_{0, \Omega}, \quad (4.12)$$

and the *graph norm*:

$$\|f\|_{G(\mathbf{b}; \Omega)}^2 = \|f\|_{0, \Omega}^2 + \|f\|_{E(\mathbf{b}; \Omega)}^2. \quad (4.13)$$

The solution space for the images is then given by:

$$G(\mathbf{b}; \Omega) = \{f \in L^2 \mid \|f\|_{G(\mathbf{b}; \Omega)} < \infty\}. \quad (4.14)$$

It is clear that  $H^1(\Omega) \subset G(\mathbf{b}; \Omega)$ . We define the boundary norm:

$$\|f\|_{2, \Gamma}^2 = \sum_{i=0,1} \|f\|_{2, \Gamma_i}^2, \quad \|f\|_{2, \Gamma_i}^2 = \int_{\Gamma_i} f^2 \, d\hat{x}, \quad i = 0, 1. \quad (4.15)$$

We start by stating the following proposition which is central to the analysis carried out in this section.

**Proposition 67.**  *$G(\mathbf{b}; \Omega)$  is a Hilbert space in which  $H^1(\Omega)$  is dense.*

*Proof.* This result is stated for general graph norms in [35, Section 10.2.1].  $\square$

We now state and prove results for these spaces that are considered standard for Sobolev spaces: a trace inequality similar to that shown for linear first-order PDEs in [133], Poincaré-type inequality, a trace theorem and existence and uniqueness for the inner problem. We highlight the reference [73] as a foundation on which our work relies as we can adapt to the setting described therein with relative ease, itself using similar techniques to [165] where a neutron transport problem was interpreted as a least-squares problem.

**Theorem 68** (Trace inequality). *Given assumption 64,  $\exists C > 0$  depending only on  $\Omega$  and  $u$  such that for any  $I \in G(\mathbf{b}; \Omega)$ ,*

$$\|I\|_{2, \Gamma_i}^2 \leq C \|I\|_{G(\mathbf{b}; \Omega)}^2, \quad i = 0, 1.$$

*Proof.* Let  $\hat{x} \in \Gamma_0$  and let  $f \in C^\infty(\bar{\Omega})$  such that the following integration by parts formula holds:

$$f(x, t) = f(\hat{x}, 0) + \int_0^t \frac{\partial}{\partial s} f(\varphi_s(\hat{x}), s) + u_s \circ \varphi_s(\hat{x}) \cdot \nabla f(\varphi_s(\hat{x}), s) \, ds.$$

In particular, for  $f = p^2$ , where  $p$  is some smooth function we obtain:

$$p^2(\hat{x}, 0) \leq p^2(x, t) + \int_0^t \left| \frac{\partial}{\partial s} p^2(\varphi_s(\hat{x}), s) + u_s \circ \varphi_s(\hat{x}) \cdot \nabla(p^2)(\varphi_s(\hat{x}), s) \right| \, ds.$$

Now integrating in time we get:

$$1 \cdot p^2(\hat{x}, 0) \leq \int_0^1 p^2(x, t) + \int_0^1 \left| \frac{\partial}{\partial t} p^2(\varphi_t(\hat{x}), t) + u_t \circ \varphi_t(\hat{x}) \cdot \nabla(p^2)(\varphi_t(\hat{x}), t) \right| \, dt. \quad (4.16)$$

We integrate over  $\Gamma_0$  on both sides of (4.16) and use  $p^2(x, t) = p^2(\varphi_t(\hat{x}), t)$ :

$$\begin{aligned} \|p\|_{2, \Gamma_0}^2 &= \int_{\Gamma_0} \int_0^1 p^2(\varphi_t(\hat{x}), t) \, dt \, d\hat{x} \\ &\quad + \int_{\Gamma_0} \int_0^1 \left| \frac{\partial}{\partial t} p^2(\varphi_t(\hat{x}), t) + u_t \circ \varphi_t(\hat{x}) \cdot \nabla(p^2)(\varphi_t(\hat{x}), t) \right| \, dt \, d\hat{x}. \end{aligned}$$

Examining the first term on the right-hand side:

$$\int_{\Gamma_0} \int_0^1 p^2(\varphi_t(\hat{x}), t) dt d\hat{x} = \int_{\varphi_t \circ \Gamma_0} \int_0^1 p^2(x, t) \left| \frac{\partial \varphi_t^{-1}}{\partial x} \right| dt dx \leq C_u \|p\|_{0,\Omega}^2,$$

where we have used the fact that  $\varphi_t \circ \Gamma_0 \times [0, 1] = \Omega$ , for all  $t \in [0, 1]$  due to the spatial boundary conditions (cf. remark 65) for  $u$  and where  $C_u = \sup_{(x,t) \in \Omega} \left| \frac{\partial \varphi_t^{-1}}{\partial x} \right|$ , which by assumption 64 holds. Now the second term can also be bounded by the same reasoning and an application of Hölder's inequality:

$$\begin{aligned} & \int_{\Gamma_0} \int_0^1 \left| \frac{\partial}{\partial t} p^2(\varphi_t(\hat{x}), t) + u_t \circ \varphi_t(\hat{x}) \cdot \nabla(p^2)(\varphi_t(\hat{x}), t) \right| dt d\hat{x} \\ & \leq \int_{\varphi_t \circ \Gamma_0} \int_0^1 \left| \frac{\partial}{\partial t} p^2(x, t) + u_t(x) \cdot \nabla(p^2)(x, t) \right| \left| \frac{\partial \varphi_t^{-1}}{\partial x} \right| dt dx \\ & \leq C_u \|\mathbf{b} \cdot \nabla_t p^2\|_{0,\Omega}^2 \\ & \leq 2C_u \|p\|_{0,\Omega} \|\mathbf{b} \cdot \nabla_t p\|_{0,\Omega}. \end{aligned}$$

Now using Young's  $\epsilon$ -inequality on this and collecting the terms above we get

$$\begin{aligned} \|p\|_{2,\Gamma_0}^2 & \leq \|p\|_{0,\Omega}^2 + \frac{C_u^2}{2} (\|p\|_{0,\Omega}^2 + \|\mathbf{b} \cdot \nabla_t p\|_{0,\Omega}^2) \\ & \leq \left(1 + \frac{C_u^2}{2}\right) \|p\|_{G(\mathbf{b};\Omega)}^2 \end{aligned}$$

showing the result for  $\Gamma_0$ . Since the ODE (4.3) is time-reversible (compose  $\varphi_t$  with  $g(t) = -t$  in time and take the derivative: the same equations are obtained), the same procedure can be carried out if we follow the streamlines backwards in time from  $\Gamma_1$  so adding these proves the claim for smooth functions. Now we simply extend this from the dense set  $C^\infty(\bar{\Omega})$  to  $G(\mathbf{b};\Omega)$  by a limiting procedure (see appendix E).  $\square$

**Remark 69.** *The streamline integration used in theorem 68 emanate from points on  $\Gamma_0$ . This choice is completely arbitrary and we could have integrated the domain from  $\Gamma_1$  and followed the streamlines backwards. Because there is no important distinction between  $\Gamma_0$  and  $\Gamma_1$  in the space-time formulations here we shall not make any further mention of this choice.*

**Theorem 70** (Poincaré inequality). *Given assumption 64,  $\exists C > 0$  depending only on  $\Omega$  and  $u$  such that for any  $I \in G(\mathbf{b};\Omega)$ ,*

$$\|I\|_{0,\Omega}^2 \leq C \left( \|I\|_{2,\Gamma}^2 + \|\mathbf{b} \cdot \nabla_t I\|_{0,\Omega}^2 \right)$$

*Proof.* As before, let  $\hat{x} \in \Gamma_0$  and let  $f \in C^\infty(\bar{\Omega})$  be such that the streamline integration by parts formula from theorem 68 holds:

$$f(x, t) = f(\hat{x}, 0) + \int_0^1 \frac{\partial}{\partial t} f(\varphi_s(\hat{x}), s) + u_s \circ \varphi_s(\hat{x}) \cdot \nabla f(\varphi_s(\hat{x}), s) ds.$$

Square both sides, using Jensen's inequality:

$$f^2(x, t) \leq 2f^2(\hat{x}, 0) + \int_0^t \left| \frac{\partial}{\partial t} f(\varphi_s(\hat{x}), s) + u_s \circ \varphi_s(\hat{x}) \cdot \nabla f(\varphi_s(\hat{x}), s) \right|^2 ds.$$

Now we multiply both sides of the equation by  $\Phi_t^{-1}$  defined in (4.6) and integrate over time and  $\Gamma_0$ :

$$\begin{aligned} \int_{\Gamma_0} \int_0^1 f^2(x, t) \Phi_t^{-1} dt d\hat{x} &\leq \int_{\Gamma_0} f^2(\hat{x}, 0) \Phi_t^{-1} d\hat{x} \\ &+ \int_{\Gamma_0} \int_0^1 \left| \frac{\partial}{\partial t} f(\varphi_t(\hat{x}), t) \right. \\ &\quad \left. + u_t \circ \varphi_t(\hat{x}) \cdot \nabla f(\varphi_t(\hat{x}), t) \right|^2 \Phi_t^{-1} dt d\hat{x}. \end{aligned}$$

Using the results from the theorem 68 this inequality simplifies to:

$$\|f\|_{0,\Omega}^2 \leq \int_{\Gamma_0} f^2(\hat{x}, 0) \Phi_t^{-1} d\hat{x} + \|f\|_{G(\mathbf{b};\Omega)}^2,$$

so can state:

$$\|f\|_{0,\Omega}^2 \leq (1 + C_u)(\|f\|_{2,\Gamma_0}^2 + \|f\|_{G(\mathbf{b};\Omega)}^2)$$

as required, where  $C_u$  is defined as in theorem 68. The proof follows by density.  $\square$

We also have a trace theorem in the setting presented thus far.

**Corollary 71** (Trace theorem). *There exists a bounded surjective operator:  $\gamma_{G(\mathbf{b};\Omega)} : G(\mathbf{b};\Omega) \rightarrow L^2(\Gamma)$ .*

*Proof.* Proposition 67 states that  $H^1(\Omega)$  is dense in  $G(\mathbf{b};\Omega)$  so the standard trace operator  $\gamma_{H^1(\Omega)} : H^1(\Omega) \rightarrow L^2(\Gamma)$  admits a continuous extension which we denote by  $\gamma_{G(\mathbf{b};\Omega)}$ . Boundedness follows by theorem 68.  $\square$

Note there that it is essential that  $V$  is such that the Jacobian is well-defined, see e.g. [35, Theorem 10.3] where  $u_t \in C^1(\bar{\Omega})$  is assumed.

This theorem does *not* follow from the analytical solution (4.10). Equation (4.10) tells us that for any data  $g_\Gamma \in L^2(\Gamma)$ , the analytical solution (4.10) provides a function  $g \in G(\mathbf{b};\Omega)$  such that  $g|_\Gamma = g_\Gamma$ . This does not on its own guarantee that we can evaluate the trace of an arbitrary function in  $G(\mathbf{b};\Omega)$ .

Equipped with theorem 71 we can define the space  $\mathring{G}(\mathbf{b};\Omega)$  of functions that vanish on  $\Gamma$  as follows:

$$\mathring{G}(\mathbf{b};\Omega) = \{f \in G(\mathbf{b};\Omega) \mid \gamma_{G(\mathbf{b};\Omega)} f = 0\}. \quad (4.17)$$

**Remark 72** (Historical comments). *Theorem 70 was first called a curved Poincaré result in the early work [189] but was not proved there. We also mention that the theory of Friedrich systems [133] forms some of the basis for the analysis of least-squared advection systems, or can at the very least be applied in certain circumstances e.g. in the construction of trace theorems.*

We return to show well-posedness of problem 3.

**Theorem 73** (Existence of a minimiser). *There exists a unique minimiser to problem 3 and the associated necessary first-order optimality conditions are necessary and sufficient.*

*Proof.* The functional (4.4a) is clearly bounded from below and admits a minimizing sequence  $\mathcal{I} = \{I_k\}_{k \geq 0}$ . To match the boundary conditions, for any function  $\iota \in G(\mathbf{b}; \Omega)$  extending the data in (4.4b) to the domain define the sequence  $\mathring{\mathcal{I}} = \mathcal{I} - \iota$ . Then by the Poincaré estimate in theorem 70,  $\mathring{\mathcal{I}}$  is bounded in  $G(\mathbf{b}; \Omega)$  which implies  $\mathcal{I}$  is bounded in  $G(\mathbf{b}; \Omega)$ . Because this space is reflexive we can therefore find a subsequence of  $\mathcal{I}$  that converges weakly to a minimiser of (4.4a). Convexity implies lower semi-continuity of the functional, and the unique minimiser is obtained.  $\square$

We now assume  $V$  is such that the Jacobian of the flow map is well-defined. With the results above we can differentiate the functional in problem 3 with respect to an arbitrary variation  $\delta I \in G(\mathbf{b}; \Omega)$  in the image  $I$  and obtain what we call the *primal inner problem*:

**Problem 4 (Primal Inner Problem)**

$$\text{Find } I \in G(\mathbf{b}; \Omega) \text{ such that} \quad (4.18a)$$

$$\int_{\Omega} (\mathbf{b} \cdot \nabla_t I)(\mathbf{b} \cdot \nabla_t \delta I) \, d(x, t) = 0, \quad \forall \delta I \in G(\mathbf{b}; \Omega) \quad (4.18b)$$

$$I|_{\Gamma_i} = I_i, \quad i = 0, 1 \quad (4.18c)$$

**Theorem 74** (Existence, uniqueness and stability for the inner problem). *For any boundary data  $I_i \in L^2(\Gamma_i)$ ,  $i = 0, 1$ , there exists a unique  $I \in G(\mathbf{b}; \Omega)$  solving problem 4 such that*

$$\|I\|_{G(\mathbf{b}; \Omega)} \lesssim \|I\|_{2, \Gamma}.$$

*Proof.* Let  $g$  be any extension of the boundary data  $I_i$ ,  $i = 0, 1$  to the domain via corollary 71. Thanks to theorem 70 the inner product associated to  $\|\cdot\|_{E(\mathbf{b}; \Omega)}$

is equivalent to  $\|\cdot\|_{G(\mathbf{b};\Omega)}$  on the subspace  $\mathring{G}(\mathbf{b};\Omega)$ . By the Riesz representation theorem we know that we can find a unique  $\xi \in \mathring{G}(\mathbf{b};\Omega)$  such that:

$$\langle \xi, v \rangle_{E(\mathbf{b};\Omega)} = -\langle g, v \rangle_{E(\mathbf{b};\Omega)}, \quad \forall v \in \mathring{G}(\mathbf{b};\Omega).$$

By construction we have  $I = \xi + g$ . The bound is provided by the trace operator cf. theorem 71. □

We now turn to discretising problem 4 using a conforming finite element space. Based on the results above the rest of the analysis here is trivial. The Galerkin [42] finite element space  $G_h \subset G(\mathbf{b};\Omega)$  denotes a finite-dimensional  $C^0(\Omega)$ -conforming subset of  $G(\mathbf{b};\Omega)$  spanned by the standard  $\mathcal{P}^1(K)$  shape functions associated with the triangulation  $\Omega_h$  of  $\Omega$ . The bilinear form in problem 4 remains coercive when restricted to  $G_h$  by conformity and hence the results above remain valid and this finite-dimensional problem is well-posed.

#### 4.1.2.1 NUMERICAL EXAMPLES

We manufacture solutions to problem 4 and show convergence using a conforming piecewise linear finite element space. Table 4.1 shows four such manufactured solutions with exact solution  $I$  and velocity field  $u$  (where  $\chi_M$  is the indicator function i.e. equal to 1 on  $M$  and 0 otherwise). Note that forcing terms have been added to the right-hand side of the bilinear form in problem 4 so the computed solutions match those in table 4.1. As mentioned, we use the Firedrake software package [191] which in turns uses PETSc [19] as a numerical back-end, see also [18, 70]. The Zenodo entry [237] for the specific versions of the software components used. The LU decomposition from section 3.3.2 was also used to invert the discrete systems above. Figure 4.1 depicts the computed matches for each of these manufactured solutions using Firedrake. The intensities across the four solutions are normalised to the same colour scale as their values are not important here and are only presented to develop an intuition about the convergence rates we can observe in figure 4.2. These simulations take less than four minutes on a 2014 MacBook Pro with a 2.5GHz Intel Core i7 processor and 16GB of 1600MHz DDR3 RAM. Here we show  $L^2$  and energy convergence rates. The latter measures the semi-norm specified in (4.12). These errors and the convergence rates for the four examples in table 4.1 are shown in tables 4.2 through 4.5, respectively. We observe some interesting phenomena. For example 0 the boundary data is discontinuous along  $\Gamma$ , so even with a simple constant velocity field the method cannot recover a positive convergence rate in the energy norm. Note that example 3 exhibits some oscillations in the energy norm.

**Remark 75.** *The results for example 0 are not contradictory to our theoretical findings but rather indicating that a linear  $C^0(\Omega)$  finite element space poorly approximates the analytical solution in this specific case.*

Some convergence in the  $L^2$  norm is recovered as this norm is somehow agnostic to the discontinuity in the derivative. Upon inspection of example 0 in figure 4.1 we see that the continuous approximation space fails to resolve the fine discontinuity along the streamline of  $u = -0.2$  and the smoothing effect is clearly seen on the interior of the domain. Examples 1, 2 and 3 provide a manufactured solution with smooth data and the finite element space is suitable for approximating the global smoothness of the solutions. The rates are spurious before entering an asymptotic regime. These rates also indicate that some higher-order convergence may be recovered with more in-depth error analysis. We highlight that example 2 is essentially identical to example 0 albeit with smooth data, and we clearly see that this is reflected in the convergence rate of the method. [36, 132] look at *a posteriori analysis* for scalar hyperbolic and transport problems (which are particularly difficult for  $L^2$  estimates) and it may be of interest to investigate these further in the context above. Such an analysis is carried out in [107] for systems of hyperbolic conservation laws. In addition, see [186] for a review of posteriori error analysis and adaptivity for conservation laws. We also note that for general discontinuous solutions of hyperbolic equations (in sense of functions of bounded variation), the theory of Krushkov [154] may be applicable to the setting here and is subject to future investigation (see e.g. [74]).

**Remark 76.** *The condition number of a matrix is, roughly speaking, an indicator of how accurately we can solve a linear system of equations involving said matrix [217]. As the methods in this chapter deal with matrices at a discrete level we make a comment about their conditioning. For certain elliptic problems it is known that the condition number of linear systems arising from finite element discretisations depend the Poincaré constant [88, Section 3.2]. This constant enters Céa’s result in lemma 16. Although these results are not readily available for problem 4 we conjecture the following: since the Poincaré constant in theorem 70 depends on the determinant of the Jacobian of the flow map generated by the velocity, the error estimate implied by a curved Céa’s lemma would have an explicit dependence on this constant. An understanding of this interplay could be useful in the design of preconditioners for problem 4.*

TABLE 4.1: Manufactured solutions to problem 4 where  $d = 1$ .

| No. | $I(x, t)$                             | $u(x, t)$                 |
|-----|---------------------------------------|---------------------------|
| 0   | $\chi_{[0.6-0.2t < x < 0.8-0.2t]}$    | $-0.2$                    |
| 1   | $e^{-100(x-t(1-t)-0.5)^2}$            | $1 - 2t$                  |
| 2   | $e^{\frac{-(x-0.4(1-t)-0.3)^2}{125}}$ | $-0.4$                    |
| 3   | $e^{-25(-4.3x+5.16x^3+0.5+t)^2}$      | $-\frac{5}{77.4x^2-21.5}$ |

For completeness appendix A shows similar results for  $d = 2$ . Given this numerical evidence that the inner problems are well-posed given a velocity

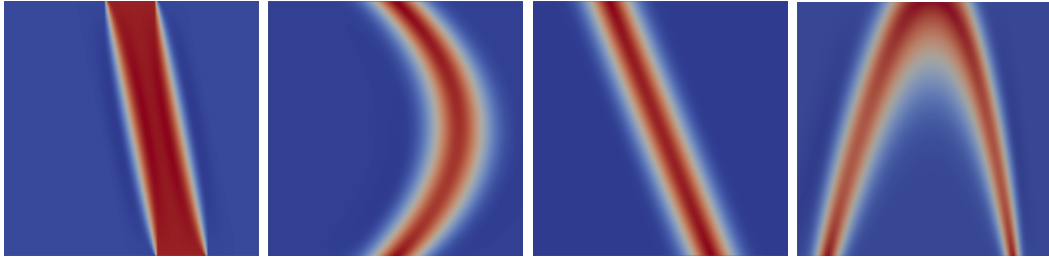


FIGURE 4.1: Density matches computed from the manufactured solutions to problem 4 in table 4.1. Left to right: example 0 through 3.

TABLE 4.2: Convergence rates for example 0 from table 4.1.

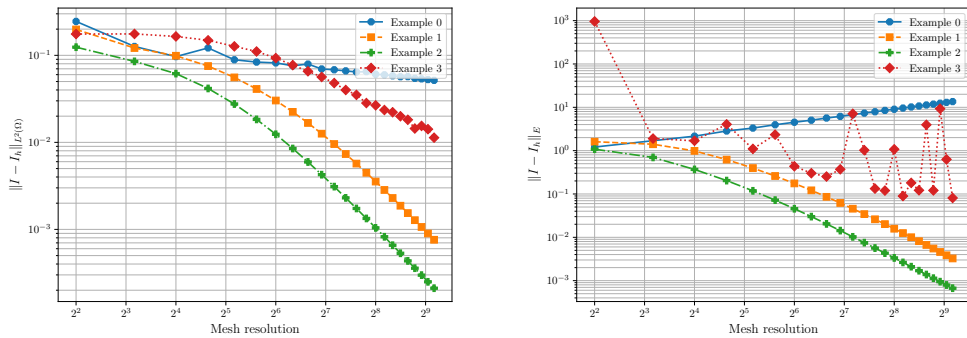
| $h^{-1}$ | $L^2$ error | Order | Energy error | Order | CPU time (s) |
|----------|-------------|-------|--------------|-------|--------------|
| 4        | 2.4544e-01  | -     | 1.2110e+00   | -     | 0.043        |
| 9        | 1.2650e-01  | 0.82  | 1.6965e+00   | -0.42 | 0.073        |
| 16       | 9.6662e-02  | 0.47  | 2.1562e+00   | -0.42 | 0.077        |
| 25       | 1.2163e-01  | -0.51 | 2.8403e+00   | -0.62 | 0.070        |
| 36       | 8.8937e-02  | 0.86  | 3.3204e+00   | -0.43 | 0.074        |
| 49       | 8.3665e-02  | 0.20  | 3.9564e+00   | -0.57 | 0.079        |
| 64       | 8.1851e-02  | 0.08  | 4.5218e+00   | -0.50 | 0.087        |
| 81       | 7.5917e-02  | 0.32  | 5.0401e+00   | -0.46 | 0.100        |
| 100      | 7.9340e-02  | -0.21 | 5.6601e+00   | -0.55 | 0.123        |
| 121      | 6.9729e-02  | 0.68  | 6.1802e+00   | -0.46 | 0.162        |
| 144      | 6.8222e-02  | 0.13  | 6.7847e+00   | -0.54 | 0.247        |
| 169      | 6.6289e-02  | 0.18  | 7.3505e+00   | -0.50 | 0.282        |
| 196      | 6.4335e-02  | 0.20  | 7.8859e+00   | -0.47 | 0.393        |
| 225      | 6.4961e-02  | -0.07 | 8.4868e+00   | -0.53 | 0.508        |
| 256      | 6.0614e-02  | 0.54  | 9.0213e+00   | -0.47 | 0.682        |
| 289      | 5.9747e-02  | 0.12  | 9.6137e+00   | -0.52 | 1.422        |
| 324      | 5.8212e-02  | 0.23  | 1.0179e+01   | -0.50 | 1.239        |
| 361      | 5.6753e-02  | 0.23  | 1.0723e+01   | -0.48 | 1.608        |
| 400      | 5.6874e-02  | -0.02 | 1.1315e+01   | -0.52 | 2.354        |
| 441      | 5.4450e-02  | 0.45  | 1.1857e+01   | -0.48 | 2.960        |
| 484      | 5.3689e-02  | 0.15  | 1.2443e+01   | -0.52 | 4.515        |
| 529      | 5.2667e-02  | 0.22  | 1.3008e+01   | -0.50 | 4.703        |
| 576      | 5.1613e-02  | 0.24  | 1.3556e+01   | -0.48 | 5.797        |

satisfying assumption 64.

The next section presents a DPG version of problem 3 but the literature on numerical schemes for hyperbolic equations is vast, so we briefly mention other possible intermediate formulations between DPG and the primal system studied above. The reader should also consult the references in section 2.2.2.1. *Stabilised* methods could be explored for problem 3. These are, roughly speaking, methods

TABLE 4.3: Convergence rates for example 1 from table 4.1.

| $h^{-1}$ | $L^2$ error | Order | Energy error | Order | CPU time (s) |
|----------|-------------|-------|--------------|-------|--------------|
| 4        | 1.9828e-01  | -     | 1.6074e+00   | -     | 0.041        |
| 9        | 1.2141e-01  | 0.60  | 1.4055e+00   | 0.17  | 0.072        |
| 16       | 9.8889e-02  | 0.36  | 9.8769e-01   | 0.61  | 0.078        |
| 25       | 7.5253e-02  | 0.61  | 6.2023e-01   | 1.04  | 0.158        |
| 36       | 5.5783e-02  | 0.82  | 3.9579e-01   | 1.23  | 0.174        |
| 49       | 4.1067e-02  | 0.99  | 2.5985e-01   | 1.36  | 0.129        |
| 64       | 3.0265e-02  | 1.14  | 1.7539e-01   | 1.47  | 0.128        |
| 81       | 2.2414e-02  | 1.27  | 1.2138e-01   | 1.56  | 0.123        |
| 100      | 1.6720e-02  | 1.39  | 8.5924e-02   | 1.64  | 0.135        |
| 121      | 1.2584e-02  | 1.49  | 6.2090e-02   | 1.70  | 0.165        |
| 144      | 9.5646e-03  | 1.58  | 4.5721e-02   | 1.76  | 0.280        |
| 169      | 7.3456e-03  | 1.65  | 3.4258e-02   | 1.80  | 0.313        |
| 196      | 5.7016e-03  | 1.71  | 2.6083e-02   | 1.84  | 0.435        |
| 225      | 4.4727e-03  | 1.76  | 2.0154e-02   | 1.87  | 0.649        |
| 256      | 3.5451e-03  | 1.80  | 1.5786e-02   | 1.89  | 0.787        |
| 289      | 2.8381e-03  | 1.83  | 1.2520e-02   | 1.91  | 1.056        |
| 324      | 2.2939e-03  | 1.86  | 1.0044e-02   | 1.93  | 1.409        |
| 361      | 1.8709e-03  | 1.88  | 8.1435e-03   | 1.94  | 1.987        |
| 400      | 1.5390e-03  | 1.90  | 6.6672e-03   | 1.95  | 2.389        |
| 441      | 1.2762e-03  | 1.92  | 5.5076e-03   | 1.96  | 3.080        |
| 484      | 1.0663e-03  | 1.93  | 4.5875e-03   | 1.96  | 3.816        |
| 529      | 8.9729e-04  | 1.94  | 3.8505e-03   | 1.97  | 4.588        |
| 576      | 7.6005e-04  | 1.95  | 3.2548e-03   | 1.97  | 6.270        |

FIGURE 4.2: Convergence as a function of mesh resolution  $h^{-1}$  for the examples in table 4.1. Left:  $L^2$  errors. Right:  $E(\mathbf{b}; \Omega)$  errors.

that use penalty terms in the bilinear forms of finite element methods to enforce some interelement continuity. For a DG method with stabilisation applied to a scalar hyperbolic equation see the early reference [44]. This also contains a comparison of this method with a streamline upwind DG method [136, 47].

TABLE 4.4: Convergence rates for example 2 from table 4.1.

| $h^{-1}$ | $L^2$ error | Order | Energy error | Order | CPU time (s) |
|----------|-------------|-------|--------------|-------|--------------|
| 4        | 1.2408e-01  | -     | 1.0790e+00   | -     | 0.0356       |
| 9        | 8.5228e-02  | 0.46  | 6.9771e-01   | 0.54  | 0.079        |
| 16       | 6.1681e-02  | 0.56  | 3.6876e-01   | 1.11  | 0.079        |
| 25       | 4.1645e-02  | 0.88  | 2.0371e-01   | 1.33  | 0.095        |
| 36       | 2.7604e-02  | 1.13  | 1.1835e-01   | 1.49  | 0.088        |
| 49       | 1.8364e-02  | 1.32  | 7.2029e-02   | 1.61  | 0.092        |
| 64       | 1.2381e-02  | 1.48  | 4.5682e-02   | 1.71  | 0.108        |
| 81       | 8.4991e-03  | 1.60  | 3.0052e-02   | 1.78  | 0.126        |
| 100      | 5.9517e-03  | 1.69  | 2.0425e-02   | 1.83  | 0.163        |
| 121      | 4.2532e-03  | 1.76  | 1.4290e-02   | 1.87  | 0.165        |
| 144      | 3.1000e-03  | 1.82  | 1.0259e-02   | 1.90  | 0.253        |
| 169      | 2.3022e-03  | 1.86  | 7.5355e-03   | 1.93  | 0.327        |
| 196      | 1.7398e-03  | 1.89  | 5.6493e-03   | 1.94  | 0.415        |
| 225      | 1.3360e-03  | 1.91  | 4.3129e-03   | 1.96  | 0.507        |
| 256      | 1.0412e-03  | 1.93  | 3.3464e-03   | 1.97  | 0.793        |
| 289      | 8.2242e-04  | 1.95  | 2.6346e-03   | 1.97  | 0.996        |
| 324      | 6.5764e-04  | 1.96  | 2.1014e-03   | 1.98  | 1.385        |
| 361      | 5.3180e-04  | 1.96  | 1.6960e-03   | 1.98  | 1.765        |
| 400      | 4.3446e-04  | 1.97  | 1.3834e-03   | 1.99  | 2.260        |
| 441      | 3.5828e-04  | 1.98  | 1.1395e-03   | 1.99  | 3.329        |
| 484      | 2.9800e-04  | 1.98  | 9.4688e-04   | 1.99  | 3.581        |
| 529      | 2.4984e-04  | 1.98  | 7.9323e-04   | 1.99  | 4.647        |
| 576      | 2.1098e-04  | 1.99  | 6.6946e-04   | 1.99  | 5.932        |

[51] presents a unified setting for both continuous and discontinuous Galerkin approximations of hyperbolic equations and also contains an excellent review of the literature on this topic. In addition, the setting for the inner problem in this chapter is either globally continuous or globally discontinuous. An intermediate setting could also be explored where the finite element space is only continuous in parts of the domain. For images that are discontinuous at a finite number of points we mention this as a possible extension to the above. In the next section we will see that the mixed advection/continuity structure of the inner problem lends itself well to a DPG formulation.

#### 4.1.3 DUAL FORMULATION

The previous section presented a variational setting leading to a stable continuous Galerkin finite element method for the inner problem. In this section we weaken the regularity of the solution space  $G(\mathbf{b}; \Omega)$  with the aim of describing a *discontinuous* Galerkin method as a mixed system for  $I$  and  $z$ . The motivation here is to provide a set equations describing geodesic motion for *non-smooth images* i.e. images that are represented by function spaces

TABLE 4.5: Convergence rates for example 3 from table 4.1.

| $h^{-1}$ | $L^2$ error | Order | Energy error | Order  | CPU time (s) |
|----------|-------------|-------|--------------|--------|--------------|
| 4        | 1.7576e-01  | -     | 9.6060e+02   | -      | 0.057        |
| 9        | 1.7620e-01  | -0.00 | 1.9000e+00   | 7.68   | 0.103        |
| 16       | 1.6483e-01  | 0.12  | 1.6866e+00   | 0.21   | 0.093        |
| 25       | 1.4885e-01  | 0.23  | 4.0564e+00   | -1.97  | 0.103        |
| 36       | 1.2681e-01  | 0.44  | 1.0990e+00   | 3.58   | 0.100        |
| 49       | 1.1072e-01  | 0.44  | 2.3390e+00   | -2.45  | 0.113        |
| 64       | 9.2946e-02  | 0.66  | 4.3551e-01   | 6.29   | 0.132        |
| 81       | 7.7202e-02  | 0.79  | 3.0148e-01   | 1.56   | 0.135        |
| 100      | 6.5567e-02  | 0.78  | 2.5252e-01   | 0.84   | 0.176        |
| 121      | 5.6455e-02  | 0.78  | 3.7138e-01   | -2.02  | 0.249        |
| 144      | 4.8037e-02  | 0.93  | 7.1032e+00   | -16.96 | 0.296        |
| 169      | 3.9981e-02  | 1.15  | 1.0250e+00   | 12.09  | 0.368        |
| 196      | 3.5126e-02  | 0.87  | 1.3289e-01   | 13.78  | 0.586        |
| 225      | 2.8296e-02  | 1.57  | 1.1974e-01   | 0.76   | 0.653        |
| 256      | 2.6659e-02  | 0.46  | 1.0756e+00   | -17.01 | 0.841        |
| 289      | 2.3535e-02  | 1.03  | 8.9558e-02   | 20.50  | 1.127        |
| 324      | 2.2196e-02  | 0.51  | 1.8030e-01   | -6.12  | 1.626        |
| 361      | 1.9955e-02  | 0.98  | 1.2159e-01   | 3.64   | 2.210        |
| 400      | 1.8179e-02  | 0.91  | 3.9522e+00   | -33.94 | 2.564        |
| 441      | 1.4424e-02  | 2.37  | 1.2082e-01   | 35.74  | 3.051        |
| 484      | 1.5454e-02  | -0.74 | 9.2763e+00   | -46.66 | 3.405        |
| 529      | 1.4173e-02  | 0.97  | 6.2777e-01   | 30.29  | 4.710        |
| 576      | 1.1319e-02  | 2.64  | 8.0891e-02   | 24.07  | 5.889        |

that are discontinuous on  $\Omega$ . As for the primal formulation we begin with a review of the literature for discontinuous methods for advection-type problems.

Advection-type problems often deal with discontinuous solutions with shocks or wave-fronts. Therefore significant resources have been dedicated to applying globally discontinuous approximation spaces for such problems with the aim of developing stable methods that more accurately represent advection phenomena. In addition to the sources cited in chapter 2 we highlight the work of Broersen et al. [46] which analyses the stability of a DPG formulation of the transport equation and describes how optimal test functions can be computed for certain flows. We rely on this work for the structure of the proofs of lemma 81, corollary 82 and theorem 83. Our aim is to develop this theory further and apply it to the mixed system for  $I$  and  $z$ . The technical work [69] analysis an adaptive Petrov-Galerkin method for such transport equations, albeit not for purely advective operators, see also the sequel [68] on anisotropic methods.

To pose a discontinuous method for the inner problem we begin with the

following simple reformulation of problem 3:

**Problem 5 (Dual Inner Problem)**

$$\begin{aligned} \inf_{z, I} \quad & F = \frac{1}{2} \|z\|_{0, \Omega}^2 \\ \text{subject to} \quad & \mathbf{b} \cdot \nabla_t I = z \\ & I|_{\Gamma_i} = I_i, \quad i = 0, 1 \end{aligned}$$

We derive a weak dual version of the variational version in problem 4 formally and later make precise which function spaces our variables of interest occupy. Here, we will treat a system of two linear first order hyperbolic equations in space-time. The space-time aspect has been investigated before, see [178] wherein a space-time Trefftz-DG method is presented. We can relax the constraints in problem 5 with multipliers  $\phi$  and  $\mu$  dual to the constraints and integrate by parts in  $I$ :

$$\begin{aligned} F = \frac{1}{2} \|z\|_{0, \Omega}^2 - \langle \operatorname{div}_t (\mathbf{b}\phi), I \rangle_{0, \Omega} - \langle \phi, z \rangle_{0, \Omega} \\ + \langle \eta \cdot \mathbf{b}\phi, I \rangle_{0, \Gamma} + \langle \eta \cdot \mathbf{b}\mu, I^* - I \rangle_{0, \Gamma}, \end{aligned} \quad (4.19)$$

where  $I^*$  represents the boundary data in the sense that:

$$\langle \eta \cdot \mathbf{b}\mu, I^* - I \rangle_{0, \Gamma} = \sum_{i=0,1} \langle \eta \cdot \mathbf{b}\mu, I_i - I \rangle_{0, \Gamma_i}.$$

We take variations in  $F$  in the four variables i.e.  $\delta F = 0$  and inspecting the resulting equations for  $z \in G(\mathbf{b}; \Omega)$ ,  $\phi \in G(\mathbf{b}; \Omega)$ ,  $I \in L^2(\Omega)$  and  $\mu \in L^2(\Gamma)$ :

$$\langle z, \delta z \rangle_{0, \Omega} = \langle \phi, \delta z \rangle_{0, \Omega}, \quad \forall \delta z \in G(\mathbf{b}; \Omega), \quad (4.20a)$$

$$\begin{aligned} - \langle z, \delta \phi \rangle_{0, \Omega} - \langle \operatorname{div}_t (\mathbf{b}\delta \phi), I \rangle_{0, \Omega} \\ + \langle \mathbf{b} \cdot \eta \delta \phi, I^* \rangle_{0, \Gamma} = 0, \quad \forall \delta \phi \in G(\mathbf{b}; \Omega), \end{aligned} \quad (4.20b)$$

$$\begin{aligned} - \langle \operatorname{div}_t (\mathbf{b}\phi), \delta I \rangle_{0, \Omega} + \langle \mathbf{b} \cdot \eta \phi, \delta I \rangle_{0, \Gamma} \\ - \langle \mathbf{b} \cdot \eta \mu, \delta I \rangle_{0, \Gamma} = 0, \quad \forall \delta I \in L^2(\Omega), \end{aligned} \quad (4.20c)$$

$$\langle \mathbf{b} \cdot \eta \delta \mu, I^* - I \rangle_{0, \Gamma} = 0, \quad \forall \delta \mu \in L^2(\Gamma). \quad (4.20d)$$

The first equation states an equivalence between  $z$  and  $\phi$  so we can reduce the system by simple substitution. Further, choosing  $z|_{\Gamma} = \mu$  as ansatz we obtain two equations instead of four:

$$\langle z, \delta z \rangle + \langle \operatorname{div}_t (\mathbf{b}\delta z), I \rangle_{0, \Omega} = \langle \mathbf{b} \cdot \eta \delta z, I^* \rangle_{0, \Gamma}, \quad \forall \delta z \in G(\mathbf{b}; \Omega) \quad (4.21a)$$

$$- \langle \operatorname{div}_t (\mathbf{b}z), \delta I \rangle_{0, \Omega} = 0, \quad \forall \delta I \in L^2(\Omega) \quad (4.21b)$$

Now any solution to (4.21) is also a solution to (4.20). Indeed, by equivalence of problem 5 to the primal problem for which existence and uniqueness was

established in theorem 73, these Euler-Lagrange equations are sufficient conditions. Therefore, an ansatz solution to (4.21) is in fact the unique solution we seek. To summarise we write (4.21) on the following form.

**Problem 6 (Conforming Ultra Weak Formulation)**

Find  $(z, I) \in G(\mathbf{b}; \Omega) \times L^2(\Omega)$  such that:

$$a(z, \delta z) - b(\delta z, I) = f_{I^*}(\delta z) \quad \forall \delta z \in G(\mathbf{b}; \Omega) \quad (4.22a)$$

$$b(z, \delta I) = 0 \quad \forall \delta I \in L^2(\Omega) \quad (4.22b)$$

where

$$a(z, \delta z) = \langle z, \delta z \rangle_{0, \Omega}$$

$$b(f, g) = -\langle \operatorname{div}_t(\mathbf{b}f), g \rangle_{0, \Omega}$$

$$f_{I^*}(\delta z) = \langle \mathbf{b} \cdot \eta \delta z, I^* \rangle_{0, \Gamma}$$

We say this is an *ultra weak* formulation because the operator (and therefore regularity requirement) is shifted from the trial function  $I$  to the test function. This idea was first presented for the Helmholtz equation in 1998 by Cessenat and Després [53], see also [49, 126]. We recommend [98, Section 2] for a more recent exposition of DPG with an ultra weak formulation.

**Remark 77** (Comparison with the Riemannian framework). *We believe that the formulations in this section are quite close, in some sense, to the Riemannian construction of [220], see also remark 63. According to this seminal work the natural Riemannian way of interpreting " $u \cdot \nabla I$ " is indeed via partial integration onto a test function. In our space-time framework this is similar to what we do by shifting regularity onto  $z$ .*

We now recall similar notation to section 2.2, the goal being to solve the problem 6 by going along the lines of roadmap 20.

Let the operator  $\mathcal{B} : G(\mathbf{b}; \Omega) \rightarrow L^2(\Omega)^* = L^2(\Omega)$  and its transpose  $\mathcal{B}^* : L^2(\Omega) \rightarrow G(\mathbf{b}; \Omega)^*$  be defined by:

$$\langle \mathcal{B} \delta z, I \rangle_{L^2(\Omega), L^2(\Omega)} = \langle \mathcal{B}^* I, \delta z \rangle_{G(\mathbf{b}; \Omega)^*, G(\mathbf{b}; \Omega)} = b(\delta z, I), \quad (4.23)$$

$\forall I \in L^2(\Omega), \forall \delta z \in G(\mathbf{b}; \Omega)$ .

It is not always the case that the image of an arbitrary infinite-dimensional operator is closed in its target space. To quote from [39], "*This fact (that the image might not be closed) puts pains thousandfold upon the mathematicians*" and is truly what separates analysis of equations in infinite dimensions from their finite dimensional relatives. An essential ingredient for the well-posedness of problem 6 is surjectivity of  $\mathcal{B}$  which is sufficient for the aforementioned closedness. This condition is shown to hold for certain transport problems,

see [69] wherein two conditions are assumed to hold. In the setting here these conditions are: (i) injectivity of  $\mathcal{B}$  and  $\text{Im}(\mathcal{B})$  is dense in  $L^2(\Omega)$ . These conditions are sufficient to determine that  $\mathcal{B}$  and its formal adjoint are isometries [69, Proposition 2.1] and thereby providing the surjectivity condition from before. In [69] these are easier to obtain because of the presence of a *reactive term* i.e. for an operator of the form  $v \mapsto cv - \text{div}_t(\mathbf{b}v)$  for some  $c > 0$ . Moreover, additional assumptions are made about the interplay between  $c$  and  $\mathbf{b}$ . We cannot directly apply this theory so we make the following assumption and provide a roadmap towards a rigorous proof.

**Assumption 78** (Surjectivity of  $\mathcal{B}$ ). *We assume  $\text{Im}(\mathcal{B}) = L^2(\Omega)$ .*

**Roadmap 79** (Towards a proof of assumption 78). *We show that a regularity result is sufficient to validate assumption 78. Indeed, thanks to theorem 70, given arbitrary  $f \in L^2(\Omega)$  we can solve following problem for a unique  $\phi \in \mathring{G}(\mathbf{b}; \Omega)$ :*

$$\int_{\Omega} (\mathbf{b} \cdot \nabla_t \phi)(\mathbf{b} \cdot \nabla_t \psi) \, dx = -\langle f, \mathbf{b} \cdot \nabla_t \psi \rangle_{0, \Omega}, \quad \forall \psi \in \mathring{G}(\mathbf{b}; \Omega).$$

*If we can appeal to a regularity result and say that the solution  $\phi$  is such that  $\mathbf{b} \cdot \nabla_t \phi \in G(\mathbf{b}; \Omega)$ , then we can let  $\delta z = \mathbf{b} \cdot \nabla_t \phi$  and the problem above is equivalent to:*

$$\int_{\Omega} \text{div}_t(\mathbf{b} \delta z) \psi \, dx = \langle f, \psi \rangle_{0, \Omega}, \quad \forall \psi \in L^2(\Omega),$$

*which is exactly what we need to say that for every  $f \in L^2(\Omega)$ , there is a function  $\delta z \in G(\mathbf{b}; \Omega)$  that is mapped to this  $f$  by  $\mathcal{B}$ .*

*Interestingly, proposition 66 showed that the analytical solution (4.11) to (4.5) has some higher-order regularity depending on the smoothness of the initial images as well as the velocity field. This is not, however, enough to state a general result for functions in  $G(\mathbf{b}; \Omega)$ . Note also that the regularity stated in the conditions of proposition 66 need also apply for this proof strategy.*

**Theorem 80** (Well-posedness of problem 6). *Provided assumption 78 holds there exists a unique pair  $(z, I) \in G(\mathbf{b}; \Omega) \times L^2(\Omega)$  solving (6) when  $I_i \in H^1(\Gamma_i)$ ,  $i = 0, 1$ .*

*Proof.* First note that  $b$  is clearly continuous:

$$\begin{aligned} |b(f, g)| &= |(-\text{div}_t(\mathbf{b}f), g)_{0, \Omega}| \\ &\leq \|\mathbf{b} \cdot \nabla_t f + f \text{div}_t(\mathbf{b})\|_{0, \Omega} \|g\|_{0, \Omega} \\ &\leq (1 + \|\text{div}_t(\mathbf{b})\|_{\infty, \Omega}) \|f\|_{G(\mathbf{b}; \Omega)} \|g\|_{0, \Omega} \\ &\leq (1 + \|\text{div}_t(\mathbf{b})\|_{\infty, \Omega}) \|f\|_{G(\mathbf{b}; \Omega)} \|g\|_{G(\mathbf{b}; \Omega)}. \end{aligned}$$

Here we have used that  $\operatorname{div}_t(\mathbf{b}) \in L^\infty(\Omega)$  by assumption 64. We follow a standard strategy in the literature see e.g. [39, Theorem 4.2.1] beginning with existence. Define the kernel of  $\mathcal{B}$  by:

$$\ker(\mathcal{B}) = \{\delta z \in G(\mathbf{b}; \Omega) \mid \langle \mathcal{B}\delta z, I \rangle_{L^2(\Omega), L^2(\Omega)} = 0, \forall I \in L^2(\Omega)\}.$$

Selecting  $\delta z \in \ker(\mathcal{B})$  we see that (4.22a) reduces to:

$$a(z, \delta z) = f_{I^*}(\delta z). \quad (4.24)$$

$a$  is naturally coercive on  $\ker(\mathcal{B})$  since the derivative terms are mapped to zero on this kernel, hence (4.24) has a (unique) solution  $z \in \ker(\mathcal{B})$ . Note that as a consequence (4.22b) is therefore naturally satisfied. Next we show that there exists an  $I$  such that (4.22a) is satisfied. Now that  $z \in \ker(\mathcal{B})$  is fixed we can write this equation as:

$$b(\delta z, I) = f_{I^*}(\delta z) - a(z, \delta z) = l(\delta z), \quad \forall \delta z \in G(\mathbf{b}; \Omega). \quad (4.25)$$

If we define  $\mathcal{L} \in G(\mathbf{b}; \Omega)^*$  by:

$$\langle \mathcal{L}, \delta z \rangle_{G(\mathbf{b}; \Omega)^*, G(\mathbf{b}; \Omega)} = l(\delta z),$$

then we can write (4.25) using (4.23):

$$\mathcal{B}^* I = \mathcal{L} \quad \text{in } G(\mathbf{b}; \Omega)^*. \quad (4.26)$$

By assumption 78 we have surjectivity of the operator  $\mathcal{B}$ , so since  $l$  actually vanishes on  $\ker(\mathcal{B})$  i.e.  $\mathcal{L}$  annihilates the kernel:  $\mathcal{L} \in \ker(\mathcal{B})^\circ$ . We can now appeal to the closed range theorem for surjective operators. Indeed, from assumption 78 we know from [39, Theorem 4.1.4] that:

$$\ker(\mathcal{B})^\circ = \operatorname{Im}(\mathcal{B}^*).$$

That is  $f = l$  is in the image of  $\mathcal{B}^*$ . We can therefore solve (4.26): there exists an  $I$  such that:

$$\langle \mathcal{B}^* I, \delta z \rangle_{G(\mathbf{b}; \Omega)^*, G(\mathbf{b}; \Omega)} = \langle \mathcal{L}, \delta z \rangle_{G(\mathbf{b}; \Omega)^*, G(\mathbf{b}; \Omega)}, \quad \forall \delta z \in G(\mathbf{b}; \Omega).$$

Now we show uniqueness. We have already proved uniqueness of  $z$  so we sketch the argument. Coercivity of  $a(\cdot, \cdot)$  on  $\ker(\mathcal{B})$  forces uniqueness for  $z$  since by taking  $z$  as a test function means  $\|z\|_{G(\mathbf{b}; \Omega)} \lesssim a(z, z)$  which shows that the bilinear form is bounding and therefore injective. For uniqueness of  $I$  we recall that  $\mathcal{B}$  is surjective, so  $\mathcal{B}^*$  must be bounding by the closed range theorem (see e.g. [39, Corollary 4.1.1]) and therefore injective.  $\square$

The advantage of problem 6 over the primal inner formulation in problem 4 is that we may request  $z \in G(\mathbf{b}; \Omega)$  and  $I \in L^2(\Omega)$  allowing for fully discontinuous images at the cost of computing a source function  $z$ . This is sometimes called an *ultra weak* formulation when the regularity requirement is shifted

onto an auxiliary function. Although it may be interesting to discretise at this level purely for mathematical novelty, we endeavour to break the global streamline continuity requirement of  $z$  in problem 6 in the next (and final) formulation so we pursue this avenue of research no further. Our aim is to use this problem as a stepping stone to a DPG formulation where we can leverage potential computational speed-ups available in this framework. These come in the form of hybridization techniques recently added to the Firedrake package [106]. Roughly speaking, the goal here is to compute, in parallel, solutions to smaller local *elemental subproblems* that are combined to yield a global solution. This effectively allows a space-time solution to be computed in parallel.

The final version of the inner problem transgresses the continuity requirement (in line with the DPG framework introduced in chapter 2) for an even weaker notion of a solution to the inner problem. The rest of this section will be increasingly technical as we steer our analysis towards building a suitable variational setting for the dual inner problem. It is therefore necessary for us to construct these various formulations of the inner problem, the final of which requires some additional notation.

Given  $\Omega$ , define a triangulation  $\Omega_h$  of  $\Omega$ . For a prescribed velocity field  $u$  we can decompose  $\partial K$ ,  $K \in \Omega_h$  as follows:

$$\partial K = \partial K_{\text{in}} \cup \partial K_{\text{out}} \cup \partial K_0,$$

where:

$$\partial K_{\text{in}} = \{(x, t) \in \partial K \mid \mathbf{b} \cdot \eta_K < 0\}, \quad (4.27a)$$

$$\partial K_{\text{out}} = \{(x, t) \in \partial K \mid \mathbf{b} \cdot \eta_K > 0\}, \quad (4.27b)$$

$$\partial K_0 = \{(x, t) \in \partial K \mid \mathbf{b} \cdot \eta_K = 0\}, \quad (4.27c)$$

where  $\eta_K$  is the outward normal to  $K$ . Recall definition 11 in section 2.2.2 which defines the mesh skeleton  $\mathcal{E}_h(\Omega_h)$  associated with  $\Omega_h$ . In view of the decomposition above it will be convenient to define the part of the mesh skeleton where  $\mathbf{b} \cdot \eta$  does not vanish. Indeed let:

$$\mathcal{E}_h^\neq = \mathcal{E}_h(\Omega_h) \setminus \left[ \bigcup_{K \in \Omega_h} \partial K_0 \right].$$

Further we define the interior mesh skeleton  $\mathcal{E}_{th}^{\neq, \text{int}} = \mathcal{E}_h^\neq \setminus \Gamma$ . Next introduce the function space :

$$G(\mathbf{b}; \Omega_h) = \{f \in L^2(\Omega) \mid \mathbf{b} \cdot \nabla_t f \in L^2(K), \forall K \in \Omega_h\}, \quad (4.28)$$

We also write the second condition on the right-hand side of (4.28) as i.e.  $\mathbf{b} \cdot \nabla_{th} f \in L^2(\Omega)$  using the *broken* operator  $\mathbf{b} \cdot \nabla_{th}$ , where the restriction of  $\mathbf{b} \cdot \nabla_{th} f$  to  $K \in \Omega_h$  is equal to  $\mathbf{b} \cdot \nabla_t f|_K$ . The broken operator  $\text{div}_{th}(\mathbf{b}f)$  is defined analogously. We have an associated norm given by:

$$\|f\|_{G(\mathbf{b}; \Omega_h)}^2 = \|f\|_{0, \Omega}^2 + \|\mathbf{b} \cdot \nabla_{th} f\|_{0, \Omega_h}^2.$$

In view of theorem 70 we can now make sense of local trace spaces on  $K$  via the space  $G(\mathbf{b}; \Omega_h)$ . Indeed, we define:

$$G(\mathbf{b}; \mathcal{E}_h^\neq) = \{w|_{\mathcal{E}_h^\neq} \mid w \in \mathring{G}(\mathbf{b}; \Omega)\}, \quad (4.29)$$

equipped with the norm:

$$\|\theta\|_{G(\mathbf{b}; \mathcal{E}_h^\neq)} = \inf_{w \in \mathring{G}(\mathbf{b}; \Omega)} \{\|w\|_{G(\mathbf{b}; \Omega_h)} \mid w|_{\mathcal{E}_h^\neq} = \theta\}.$$

Breaking global continuity of  $z$  in problem 6 means supplying the bilinear form  $b(\cdot, \cdot)$  with jump terms from partial integration on  $K \in \Omega_h$ . This necessitates an evaluation of  $I$  on  $\mathcal{E}_{th}^{\neq, \text{int}}$ , which is a priori not defined for  $I \in L^2(\Omega)$ . Using the new local trace spaces designed above we may, as a result, add a "flux variable"  $\theta$  and pose a *broken ultra weak formulation* of the inner problem as follows, suppressing obvious dependencies of our variables on  $h$ :

**Problem 7 (Broken Ultra Weak Formulation)**

Find  $(z, I, \theta) \in G(\mathbf{b}; \Omega_h) \times L^2(\Omega) \times \mathring{G}(\mathbf{b}; \mathcal{E}_h^\neq)$  such that:

$$a(z, \delta z) - b_h(\delta z, I, \theta) = f_{I^*}(\delta z), \quad \forall \delta z \in G(\mathbf{b}; \Omega_h), \quad (4.30a)$$

$$b_h(z, \delta I, \delta \theta) = 0, \quad \forall (\delta \theta, \delta I) \in L^2(\Omega) \times G(\mathbf{b}; \mathcal{E}_h^\neq), \quad (4.30b)$$

where:

$$a(z, \delta z) = \langle z, \delta z \rangle_{0, \Omega},$$

$$b_h(f, g, l) = -\langle \text{div}_{th}(\mathbf{b}f), g \rangle_{0, \Omega_h} + \langle \llbracket \mathbf{b}f \rrbracket, l \rangle_{0, \mathcal{E}_{th}^{\neq, \text{int}}},$$

$$f_{I^*}(\delta z) = \langle \mathbf{b} \cdot \eta \delta z, I^* \rangle_{0, \Gamma}.$$

The following technical lemma is needed to show existence of a  $\theta$  satisfying the broken formulation of the inner problem. We define the dual norm:

$$\|f\|_{\mathring{G}(\mathbf{b}; \mathcal{E}_h^\neq)^*} = \sup_{v \in \mathring{G}(\mathbf{b}; \mathcal{E}_h^\neq)} \frac{|\int_{\mathcal{E}_h} f v \, ds|}{\|v\|_{\mathring{G}(\mathbf{b}; \mathcal{E}_h^\neq)}}, \quad f \in G(\mathbf{b}; \mathcal{E}_h^\neq).$$

**Lemma 81.** For  $\delta z \in G(\mathbf{b}; \Omega_h)$  we have  $\llbracket \mathbf{b} \delta z \rrbracket \in \mathring{G}(\mathbf{b}; \mathcal{E}_h^\neq)^*$  and the following bounds hold:

$$(\|\mathcal{B}^{-*}\|(2 + \|\text{div}_t(\mathbf{b})\|_{\infty, \Omega}))^{-1} \leq \frac{\|\llbracket \mathbf{b} \delta z \rrbracket\|_{\mathring{G}(\mathbf{b}; \mathcal{E}_h^\neq)^*}}{\inf_{f \in G(\mathbf{b}; \Omega)} \|\delta z - f\|_{G(\mathbf{b}; \Omega_h)}} \leq (2 + \|\text{div}_t(\mathbf{b})\|_{\infty, \Omega}),$$

when  $\delta z \in G(\mathbf{b}; \Omega_h) \setminus \mathring{G}(\mathbf{b}; \Omega)$  and  $\mathcal{B}^{-*}$  denotes the inverse of the transpose of  $\mathcal{B}$ .

*Proof.* Let  $\delta z \in G(\mathbf{b}; \Omega_h)$  and  $w \in \mathring{G}(\mathbf{b}; \Omega)$  and note:

$$\int_{\mathcal{E}_h} \llbracket \delta z \mathbf{b} \rrbracket w \, ds = \sum_{K \in \Omega_h} \int_K w \operatorname{div}_t(\mathbf{b} \delta z) + \delta z \mathbf{b} \cdot \nabla_t w \, dx \quad (4.31)$$

$$\leq (2 + \|\operatorname{div}_t(\mathbf{b})\|_{\infty, \Omega}) \|\delta z\|_{G(\mathbf{b}; \Omega_h)} \|w\|_{G(\mathbf{b}; \Omega)}. \quad (4.32)$$

Hence:

$$\|\llbracket \mathbf{b} \delta z \rrbracket\|_{\mathring{G}(\mathbf{b}; \mathcal{E}_h^*)} \leq (2 + \|\operatorname{div}_t(\mathbf{b})\|_{\infty, \Omega}) \|\delta z\|_{G(\mathbf{b}; \Omega_h)}. \quad (4.33)$$

For  $f \in G(\mathbf{b}; \Omega)$  and  $g \in \mathring{G}(\mathbf{b}; \Omega)$  we have through (4.31) that  $\int_{\mathcal{E}_h} \llbracket \mathbf{b} f \rrbracket g \, ds = 0$  which implies  $\|\llbracket \mathbf{b} f \rrbracket\|_{\mathring{G}(\mathbf{b}; \mathcal{E}_h^*)} = 0$ . To see why this is so, note that  $f$  and  $g$  are smooth along streamlines, so have no jumps across the interior facets. Since  $g$  is zero on the boundary  $\Gamma$  we see that the statement is true. Consequently we can quotient out the smooth subspace  $G(\mathbf{b}; \Omega)$  and make a stronger statement than (4.33):

$$\|\llbracket \mathbf{b} \delta z \rrbracket\|_{\mathring{G}(\mathbf{b}; \mathcal{E}_h^*)} \leq (2 + \|\operatorname{div}_t(\mathbf{b})\|_{\infty, \Omega}) \inf_{q \in G(\mathbf{b}; \Omega)} \|\delta z - q\|_{G(\mathbf{b}; \Omega_h)}, \quad (4.34)$$

showing the second estimate.

For the first inequality, observe that  $\operatorname{div}_t(\mathbf{b} \cdot)$  is invertible, so for a given  $\delta z \in G(\mathbf{b}; \Omega_h)$ , let  $v \in G(\mathbf{b}; \Omega)$  solve  $\mathcal{B}v = -\operatorname{div}_t(\mathbf{b}v) = -\operatorname{div}_{th}(\mathbf{b} \delta z)$ . We let  $\phi$  solve:

$$\int_{\Omega} (\mathbf{b} \cdot \nabla_t \phi)(\mathbf{b} \cdot \nabla_t \psi) \, dx = - \int_{\Omega_h} \operatorname{div}_{th}(\mathbf{b} \delta z) \psi \, dx, \quad \forall \psi \in \mathring{G}(\mathbf{b}; \Omega),$$

and letting  $v = \mathbf{b} \cdot \nabla_t \phi$ . This means:

$$\operatorname{div}_{th}(\mathbf{b}(\delta z - v)) = \operatorname{div}_t(\mathbf{b})(\delta z - v) + \mathbf{b} \cdot \nabla_{th}(\delta z - v) = 0, \quad (4.35)$$

which implies:

$$\|\mathbf{b} \cdot \nabla_{th}(\delta z - v)\|_{0, \Omega} \leq \|\operatorname{div}_t(\mathbf{b})\|_{\infty, \Omega} \|\delta z - v\|_{0, \Omega}. \quad (4.36)$$

This is a highly nontrivial statement because it means we are able to bound the derivative of a function by its  $L^2$  norm (up to the divergence term). Now we construct a  $w \in \mathring{G}(\mathbf{b}; \Omega)$  such that  $\mathcal{B}^*w = \mathbf{b} \cdot \nabla_t w = \delta z - v$  by again using namely theorem 74 and assumption 78:

$$\langle \mathbf{b} \cdot \nabla_t w, \mathbf{b} \cdot \nabla_t \psi \rangle_{0, \Omega} = \langle f, \mathbf{b} \cdot \nabla_t \psi \rangle_{0, \Omega}, \quad \forall \psi \in \mathring{G}(\mathbf{b}; \Omega), \quad (4.37)$$

giving the desired equality in the sense of  $L^2(\Omega)$ . As a result:

$$\begin{aligned}
\|\delta z - v\|_{0,\Omega}^2 &= \int_{\Omega} (\delta z - v) \mathbf{b} \cdot \nabla_t w \, dx \\
&= \sum_K \int_K (\delta z - v) \mathbf{b} \cdot \nabla_t w \, dx \\
&= \sum_K \int_K \underbrace{\operatorname{div}_t (\mathbf{b}(v - \delta z))}_{=0 \text{ by (4.35)}} w \, dx + \int_{\partial K} \mathbf{b} \cdot \eta_K w (\delta z - v) \, ds \\
&= \int_{\mathcal{E}_{th}^{\neq, \text{int}}} \llbracket \mathbf{b} \delta z \rrbracket w \, ds + \int_{\Gamma} \mathbf{b} \cdot \eta \delta z w \, ds \\
&= \sum_{K \in \Omega_h} \int_{\partial K} \mathbf{b} \cdot \eta_K \delta z w \, ds.
\end{aligned}$$

The last term does not include  $v$  because it is not from a broken space (both  $w$  and  $v$  are smooth along streamlines). Then:

$$\begin{aligned}
\|\delta z - v\|_{0,\Omega}^2 &\leq \|\llbracket \mathbf{b} \delta z \rrbracket\|_{\mathring{G}(\mathbf{b}; \mathcal{E}_h^{\neq})^*} \|w\|_{G(\mathbf{b}; \Omega)} \\
&\leq \|\llbracket \mathbf{b} \delta z \rrbracket\|_{\mathring{G}(\mathbf{b}; \mathcal{E}_h^{\neq})^*} \|\mathcal{B}^{-*}\| \|\delta z - v\|_{0,\Omega},
\end{aligned}$$

giving:

$$\|\delta z - v\|_{0,\Omega} \leq \|\mathcal{B}^{-*}\| \|\llbracket \mathbf{b} \delta z \rrbracket\|_{\mathring{G}(\mathbf{b}; \mathcal{E}_h^{\neq})^*}.$$

Using (4.36) and  $\|\delta z - v\|_{G(\mathbf{b}; \Omega_h)} \leq \|\delta z - v\|_{0,\Omega} + \|\mathbf{b} \cdot \nabla_{th}(\delta z - v)\|_{0,\Omega}$  we get:

$$\inf_{w \in G(\mathbf{b}; \Omega)} \|\delta z - w\|_{G(\mathbf{b}; \Omega_h)} \leq \|\mathcal{B}^{-*}\| (2 + \|\operatorname{div}_t (\mathbf{b})\|_{\infty, \Omega}) \|\llbracket \mathbf{b} \delta z \rrbracket\|_{\mathring{G}(\mathbf{b}; \mathcal{E}_h^{\neq})^*},$$

and we are done.  $\square$

As a consequence we can establish an isomorphism:

**Corollary 82.**  $\mathring{G}(\mathbf{b}; \mathcal{E}_h^{\neq})^* \simeq G(\mathbf{b}; \Omega_h) / \mathring{G}(\mathbf{b}; \Omega)$ . The norm on  $\mathring{G}(\mathbf{b}; \mathcal{E}_h^{\neq})^*$  is equivalent to the norm on  $G(\mathbf{b}; \Omega_h)$  for functions in  $G(\mathbf{b}; \Omega_h) / \mathring{G}(\mathbf{b}; \Omega)$ .

Intuitively, this isomorphism makes sense: taking the boundary values out of the equation, the norm of a broken function is related to the sup-norm of its jumps (in the direction of  $\mathbf{b}$ ).

**Theorem 83** (Well-posedness of problem 7). *Let  $I_i \in H^1(\Gamma_i)$ ,  $i = 0, 1$ . There exists a unique triple  $(z, I, \theta) \in G(\mathbf{b}; \Omega_h) \times L^2(\Omega) \times \mathring{G}(\mathbf{b}; \mathcal{E}_h^{\neq})$  such that (7) is satisfied.*

*Proof.* First we verify that  $\mathcal{B}$  has closed range by conditions **i-iii** of theorem 7.

**i.** This is obvious from the definition of  $\mathcal{B}$  in (4.23).

**ii.** Next we show uniqueness. Suppose  $(I, \theta) \in L^2(\Omega) \times \mathring{G}(\mathbf{b}; \mathcal{E}_h^\neq)$  are such that:

$$b_h(\delta z, I, \theta) = 0 \quad \forall \delta z \in G(\mathbf{b}; \Omega_h). \quad (4.38)$$

Note that if  $\delta z \in G(\mathbf{b}; \Omega)$  then  $I \equiv 0$  because on this subspace, the operator applied to  $I$  is already an isomorphism by theorem 80, specifically we know by (4.26) that  $\mathcal{B}^*$  is injective on this subspace. Next we consider  $\delta z$  such that  $\text{supp}(\delta z) = K$  for an arbitrary  $K \in \Omega_h$ . Since  $I = 0$  was shown on smooth subspaces for the test function  $\delta z$ , it must also hold for this choice. Then, (4.38) implies  $\int_{\partial K} \mathbf{b} \cdot \eta_K \delta z \theta \, ds = 0 \Rightarrow \theta \equiv 0$  and uniqueness is shown.

**iii.** We verify an inf-sup condition for  $b_h$  directly in the form of (2.12).

Let  $\delta z \in G(\mathbf{b}; \Omega_h)$ . Here we use the bracket notation  $\delta z_1[\delta z]$  when necessary to make it clear how  $\delta z_1$  is computed: by the previous results exists  $\delta z_1[\delta z] \in G(\mathbf{b}; \Omega)$  such that:

$$-\text{div}_t(\mathbf{b} \delta z_1[\delta z]) = -\text{div}_{th}(\mathbf{b} \delta z), \quad (4.39)$$

with:

$$\begin{aligned} \|\delta z_1\|_{G(\mathbf{b}; \Omega)} &\leq \|\mathcal{B}^{-1}\| \|\text{div}_{th}(\mathbf{b} \delta z)\|_{0, \Omega} \\ &\leq \|\mathcal{B}^{-1}\| (2 + \|\text{div}_t(\mathbf{b})\|_{\infty, \Omega}) \|\delta z\|_{G(\mathbf{b}; \Omega_h)}. \end{aligned}$$

This in turn implies:

$$\|\delta z_1 - \delta z\|_{G(\mathbf{b}; \Omega)} \leq \underbrace{(1 + \|\mathcal{B}^{-1}\| (2 + \|\text{div}_t(\mathbf{b})\|_{\infty, \Omega}))}_{=\tilde{C}} \|\delta z\|_{G(\mathbf{b}; \Omega_h)}. \quad (4.40)$$

We can make this bound tighter through the following observation: when it is the case that  $\delta z$  occupies the smooth subspace  $G(\mathbf{b}; \Omega)$  i.e.  $\delta z \in G(\mathbf{b}; \Omega)$ ,  $\delta z \equiv \delta z_1$  which means that for any  $v \in G(\mathbf{b}; \Omega)$ :

$$\delta z_1[\delta z - v] - \delta z_1[\delta z] = -v. \quad (4.41)$$

This means that if we change the right-hand side of the equation  $\mathcal{B}x = y$  by say  $y \mapsto y + w$  then we get a change in  $x$  exactly equal to  $w$ :

$$\begin{aligned} \text{div}_{th}(\mathbf{b}(\delta z - v)) &= \text{div}_t(\mathbf{b} \delta z_1(\delta z)) - \text{div}_t(\mathbf{b} v), \\ \text{div}_{th}(\mathbf{b}(\delta z - v)) &= \text{div}_t(\mathbf{b} \delta z_1(\delta z - v)), \end{aligned}$$

so that:

$$\text{div}_t(\mathbf{b}(\delta z_1[\delta z - v] - \delta z_1[\delta z])) = \text{div}_t(\mathbf{b} v),$$

which gives us (4.41). As a result (4.40) becomes:

$$\|\delta z_1 - \delta z\|_{G(\mathbf{b}; \Omega)} \leq \tilde{C} \inf_{v \in G(\mathbf{b}; \Omega)} \|\delta z - v\|_{G(\mathbf{b}; \Omega_h)}. \quad (4.42)$$

**Remark 84.** *Another way of seeing this is that if we write a test function  $\delta z$  as a sum of  $\delta z_{G(\mathbf{b};\Omega)} \in G(\mathbf{b};\Omega)$  and  $\delta z_{G(\mathbf{b};\Omega_h) \setminus G(\mathbf{b};\Omega)} \in G(\mathbf{b};\Omega_h) \setminus G(\mathbf{b};\Omega)$ , then  $\delta z_{G(\mathbf{b};\Omega)} - \delta z_1$  will vanish due to the equivalence in (4.39).*

Applying lemma 81 to the right-hand side of (4.42) yields:

$$\begin{aligned} \|\delta z_1 - \delta z\|_{G(\mathbf{b};\Omega)} &\leq \tilde{C} \inf_{v \in G(\mathbf{b};\Omega)} \|\delta z - v\|_{G(\mathbf{b};\Omega_h)} \\ &\leq \hat{C} \|[\![\mathbf{b}\delta z]\!]\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)^*}, \end{aligned} \quad (4.43)$$

where  $\hat{C} = \|\mathcal{B}^{-*}\|(2 + \|\operatorname{div}_t(\mathbf{b})\|_{\infty,\Omega})\tilde{C}$ .

Since the space  $\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)$  is a Hilbert space it is reflexive, so by James' theorem [139] we know that the operator norm  $\|\cdot\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)^*}$  attains its supremum. As a result we can guarantee the existence of a  $\bar{\theta} \in \mathring{G}(\mathbf{b};\mathcal{E}_h^\#)$  such that:

$$\|[\![\mathbf{b}\delta z]\!]\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)^*} = \frac{\int_{\mathcal{E}_h^\#} [\![\mathbf{b}\delta z]\!] \bar{\theta} \, ds}{\|\bar{\theta}\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)}}.$$

We see that  $\bar{\theta}$  is only defined up to a multiplicative constant by the nature of the expression above, so we let  $\theta$  be defined by:

$$\theta = \frac{\bar{\theta}}{\|\bar{\theta}\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)}} \hat{C}^{-1} \|\delta z_1 - \delta z\|_{G(\mathbf{b};\Omega_h)}, \quad (4.44)$$

so that, for any  $\bar{\theta}$  we can select:

$$\|[\![\mathbf{b}\delta z]\!]\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)^*} = \frac{\int_{\mathcal{E}_h^\#} [\![\mathbf{b}\delta z]\!] \theta \, ds}{\|\theta\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)}},$$

with  $\|\theta\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)} = \hat{C}^{-1} \|\delta z_1 - \delta z\|_{G(\mathbf{b};\Omega_h)}$  and observe that since:

$$\|\theta\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)} = \frac{\int_{\mathcal{E}_h^\#} [\![\mathbf{b}\delta z]\!] \theta \, ds}{\|[\![\mathbf{b}\delta z]\!]\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)^*}},$$

and that the denominator is greater than or equal to  $\hat{C}^{-1} \|\delta z_1 - \delta z\|_{G(\mathbf{b};\Omega_h)}$  by (4.43) then we can conclude that:

$$\|\theta\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)} \leq \frac{\int_{\mathcal{E}_h^\#} [\![\mathbf{b}\delta z]\!] \theta \, ds}{\|\theta\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)}},$$

implying:

$$\|\theta\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\#)}^2 \leq \int_{\mathcal{E}_h^\#} [\![\mathbf{b}\delta z]\!] \theta \, ds, \quad (4.45)$$

for any choice of  $\delta z \in G(\mathbf{b}; \Omega_h)$ .

Next, there exists an  $I \in L^2(\Omega)$  for each  $\delta z_1 \in G(\mathbf{b}; \Omega)$  such that :

$$\|\mathcal{B}\delta z_1\|_{0,\Omega} = \frac{\langle -\operatorname{div}_t(\mathbf{b}\delta z_1), I \rangle_{0,\Omega}}{\|I\|_{0,\Omega}},$$

which holds since  $I \mapsto \langle \mathcal{B}\delta z_1 I \rangle_{0,\Omega}$  attains its supremum by James' theorem because  $L^2(\Omega)$  is reflexive. Note that we have, for  $\|\mathcal{B}^{-1}\| = \sup_{i \in L^2(\Omega)} \frac{\|\mathcal{B}^{-1}i\|_{G(\mathbf{b};\Omega)}}{\|i\|_{0,\Omega}}$ :

$$\begin{aligned} \|\mathcal{B}^{-1}\mathcal{B}\delta z_1\|_{G(\mathbf{b};\Omega)} &= \|\delta z_1\|_{G(\mathbf{b};\Omega)}, \\ \|\mathcal{B}^{-1}\|\|\mathcal{B}\delta z_1\|_{0,\Omega} &\geq \|\delta z_1\|_{G(\mathbf{b};\Omega)}, \\ \|\mathcal{B}\delta z_1\|_{0,\Omega} &\geq \|\mathcal{B}^{-1}\|^{-1}\|\delta z_1\|_{G(\mathbf{b};\Omega)}, \\ \|\mathcal{B}\delta z_1\|_{0,\Omega}^{-1} &\leq \frac{1}{\|\mathcal{B}^{-1}\|^{-1}\|\delta z_1\|_{G(\mathbf{b};\Omega)}}, \end{aligned}$$

so we can say:

$$\begin{aligned} \|I\|_{0,\Omega} &= \frac{\langle -\operatorname{div}_t(\mathbf{b}\delta z_1), I \rangle_{0,\Omega}}{\|\mathcal{B}\delta z_1\|_{0,\Omega}} \\ &\leq \frac{\langle -\operatorname{div}_t(\mathbf{b}\delta z_1), I \rangle_{0,\Omega}}{\|\mathcal{B}^{-1}\|^{-1}\|\delta z_1\|_{G(\mathbf{b};\Omega)}}. \end{aligned}$$

Select now  $I$  such that  $\|I\|_{0,\Omega} = \|\mathcal{B}^{-1}\|^{-1}\|\delta z_1\|_{G(\mathbf{b};\Omega)}$  (by scaling similar to (4.44)). In turn, this implies:

$$\|I\|_{0,\Omega}^2 \leq \langle -\operatorname{div}_t(\mathbf{b}\delta z_1), I \rangle_{0,\Omega}.$$

Using the fact that  $\exists \delta z_1 \in G(\mathbf{b}; \Omega)$  such that  $\operatorname{div}_t(\mathbf{b}\delta z_1) = \operatorname{div}_{th}(\mathbf{b}\delta z)$  for any choice of  $\delta z \in G(\mathbf{b}; \Omega_h)$  we derive:

$$\|I\|_{0,\Omega}^2 \leq \langle -\operatorname{div}_{th}(\mathbf{b}\delta z), I \rangle_{0,\Omega}, \quad \forall \delta z \in G(\mathbf{b}; \Omega_h). \quad (4.46)$$

Next, we use (4.46) and (4.45) to show that for  $b_h$  on  $G(\mathbf{b}; \Omega_h) \times (L^2(\Omega) \times G(\mathbf{b}; \mathcal{E}_h^\neq))$ :

$$\begin{aligned} (\|\mathcal{B}^{-1}\|^2 + \hat{C}^2)^{-0.5} \|\delta z\|_{G(\mathbf{b}; \Omega_h)} &\leq \frac{\|\delta z\|_{G(\mathbf{b}; \Omega_h)} + \|\delta z_1 - \delta z\|_{G(\mathbf{b}; \Omega_h)}}{(\|\mathcal{B}^{-1}\|^2 + \hat{C}^2)^{0.5}} \\ &\leq \sqrt{\|\mathcal{B}^{-1}\|^2 \|\delta z\|_{G(\mathbf{b}; \Omega_h)}^2 + \hat{C}^2 \|\delta z_1 - \delta z\|_{G(\mathbf{b}; \Omega_h)}^2} \\ &\leq \frac{\|\mathcal{B}^{-1}\|^2 \|\delta z\|_{G(\mathbf{b}; \Omega_h)}^2 + \hat{C}^2 \|\delta z_1 - \delta z\|_{G(\mathbf{b}; \Omega_h)}^2}{\sqrt{\|\mathcal{B}^{-1}\|^2 \|\delta z\|_{G(\mathbf{b}; \Omega_h)}^2 + \hat{C}^2 \|\delta z_1 - \delta z\|_{G(\mathbf{b}; \Omega_h)}^2}} \\ &= \frac{\|\mathcal{B}^{-1}\|^2 \|\delta z\|_{G(\mathbf{b}; \Omega_h)}^2 + \hat{C}^2 \|\delta z_1 - \delta z\|_{G(\mathbf{b}; \Omega_h)}^2}{\sqrt{\|I\|_{0,\Omega}^2 + \|\theta\|_{G(\mathbf{b}; \mathcal{E}_h^\neq)}^2}} \\ &\leq \frac{b_h(\delta z, I, \theta)}{\sqrt{\|I\|_{0,\Omega}^2 + \|\theta\|_{G(\mathbf{b}; \mathcal{E}_h^\neq)}^2}}. \end{aligned}$$

Here,  $\sqrt{\|I\|_{0,\Omega}^2 + \|\theta\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\neq)}^2}$  is a norm on  $L^2(\Omega) \times \mathring{G}(\mathbf{b};\mathcal{E}_h^\neq)$ . We therefore have:

$$\frac{b_h(\delta z, I, \theta)}{\sqrt{\|I\|_{0,\Omega}^2 + \|\theta\|_{\mathring{G}(\mathbf{b};\mathcal{E}_h^\neq)}^2}} \geq \gamma \|\delta z\|_{G(\mathbf{b};\Omega_h)}, \quad \forall \delta z \in G(\mathbf{b};\Omega_h),$$

where  $\gamma = (\|\mathcal{B}^{-1}\|^2 + \hat{C}^2)^{-0.5}$ . This implies the existence result in theorem 7 so we verify the inf-sup condition. The uniqueness argument is analogous to that in theorem 80.

Let  $\gamma = (\|\mathcal{B}^{-1}\|^2 + \hat{C}^2)^{-0.5}$ . We now proceed to examine the equations in (7) together. Define the operator  $\mathcal{B}_h(I, \theta) = b_h(\cdot, I, \theta) \in G(\mathbf{b};\Omega_h)^*$ , together with its kernel  $\ker(\mathcal{B}_h) = \{\delta z \in G(\mathbf{b};\Omega_h) \mid \mathcal{B}_h(I, \theta)(\delta z) = 0, \forall (I, \theta) \in L^2(\Omega) \times \mathring{G}(\mathbf{b};\mathcal{E}_h^\neq)\}$ .

Select now  $\delta z \in \ker(\mathcal{B}_h)$  in (7) so that (4.30a) reduces to  $a(z, \delta z) = f_{I^*}(\delta z)$ . For  $z \in \ker(\mathcal{B}_h)$ ,  $a(\cdot, \cdot)$  is naturally coercive since  $a(z, z) = \|z\|_2^2 = \|z\|_{G(\mathbf{b};\Omega_h)}^2$ . This equation is uniquely solvable by the Lax-Milgram theorem.

It remains to show that there exists a pair  $(I, \theta) \in L^2(\Omega) \times \mathring{G}(\mathbf{b};\mathcal{E}_h^\neq)$  such that:

$$b_h(\delta z, I, \theta) = f_{I^*}(\delta) - a(z, \delta z).$$

This follows immediately from the inf-sup condition we just verified.  $\square$

#### 4.1.3.1 DISCRETE FORMULATION

In this section we consider the problem of finding suitable finite-dimensional subspaces for our trial and test functions to recover a stable discretisation. That is, we want to find subspaces:

$$\begin{aligned} \mathbb{U}_h &= \mathbb{U}_h^I \times \mathbb{U}_h^\theta \subset L^2(\Omega) \times G(\mathbf{b};\mathcal{E}_h^\neq), \\ \mathbb{V}_h &\subset G(\mathbf{b};\Omega_h), \end{aligned}$$

parameterised by the mesh size  $h$ . These subspace will be written in product form e.g.:

$$\mathbb{V}_h = \Pi_{K \in \Omega_h} \mathbb{V}_K, \quad (4.47a)$$

$$\mathbb{U}_h^I = \Pi_{K \in \Omega_h} \mathbb{U}_K, \quad (4.47b)$$

$$\mathbb{U}_h^\theta = \Pi_{F \in \mathcal{E}_h^\neq} \mathbb{U}_F, \quad (4.47c)$$

where  $\mathbb{V}_K$  and  $\mathbb{U}_K$  denote a polynomial restricted to  $K$  and  $\mathbb{U}_F$  is a polynomial space restricted to the  $d$ -dimensional facets of  $\mathcal{E}_h^\neq$ . The main task in this section is to determine the polynomial orders of these spaces as the well-posedness of the following the discrete version of problem 7 will depend directly on this choice.

**Problem 8 (Discrete Broken Ultra Weak Formulation)**Find  $(z, I, \theta) \in \mathbb{V}_h \times \mathbb{U}_h$  such that

$$a(z, \delta z) - b_h(\delta z, I, \theta) = f_{I^*}(\delta z) \quad \forall \delta z \in \mathbb{V}_h \quad (4.48a)$$

$$b_h(z, \delta I, \delta \theta) = 0 \quad \forall (\delta \theta, \delta I) \in \mathbb{U}_h \quad (4.48b)$$

Since  $\mathbb{U}_h$  and  $\mathbb{V}_h$  are exterior approximations we cannot directly apply the result from theorem 83 to show well-posedness - in particular existence. In order to do so we verify what forms the subject of this section, namely the following *discrete* inf-sup condition:

$$0 < \gamma \leq \inf_{(I, \theta) \in \mathbb{U}_h \setminus \{0\}} \sup_{\delta z \in \mathbb{V}_h \setminus \{0\}} \frac{b_h(\delta z, I, \theta)}{\|(I, \theta)\|_{\mathbb{U}_h} \|\delta z\|_{\mathbb{V}_h}}, \quad (4.49)$$

where  $\gamma$  is *independent of*  $h$ . To this end we use the following result.

**Theorem 85** (Fortin operator [97]). *Suppose there exists a linear operator  $\mathcal{F} : \mathbb{V} \rightarrow \mathbb{V}_h$  satisfying the following conditions:*

1.  $b_h(\delta z - \mathcal{F}\delta z, I, \theta) = 0, \quad \forall (I, \theta) \in \mathbb{U}_h, \quad \forall \delta z \in \mathbb{V}$
2.  $\|\mathcal{F}\delta z\|_{\mathbb{V}} \leq C\|\delta z\|_{\mathbb{V}}, \quad \forall \delta z \in \mathbb{V}$

*with a constant  $C$  that does not degenerate as  $h \rightarrow 0$ . Then the inequality (4.49) holds.*

*Proof.* Let  $(I, \theta) \in \mathbb{U}_h$  and  $\delta z \in \mathbb{V}$ . Then,

$$\begin{aligned} \gamma \|(I, \theta)\|_{\mathbb{U}_h} &\leq \sup_{\delta z \in \mathbb{V}} \frac{b(\delta z, I, \theta)}{\|\delta z\|_{\mathbb{V}}} \\ &= \sup_{\delta z \in \mathbb{V}} \frac{b(\mathcal{F}\delta z, I, \theta)}{\|\delta z\|_{\mathbb{V}}} \\ &\leq \sup_{\delta z \in \mathbb{V}} \frac{b(\mathcal{F}\delta z, I, \theta)}{\|\mathcal{F}\delta z\|_{\mathbb{V}}}. \end{aligned}$$

□

The advantage of the element-wise construction of the broken spaces in (4.47) is that the Fortin operator can be constructed locally. It also happens that we can do this by constructing two operators,  $\Pi_h^1$  and  $\Pi_h^2$  from  $\mathbb{V}$  to  $\mathbb{V}_h$ . We request that these operators satisfy:

$$\begin{aligned} \|\Pi_h^1 \delta z\|_{\mathbb{V}} &\lesssim \|\delta z\|_{\mathbb{V}}, & \forall \delta z \in \mathbb{V}, \\ \|\Pi_h^2 [\mathbf{id} - \Pi_h^1] \delta z\|_{\mathbb{V}} &\lesssim \|\delta z\|_{\mathbb{V}}, & \forall \delta z \in \mathbb{V}, \\ b_h(\delta z - \Pi_h^2 \delta z, I, \theta) &= 0, & \forall (I, \theta) \in \mathbb{U}_h, \quad \forall \delta z \in \mathbb{V}. \end{aligned}$$

Then we can construct the operator  $\mathcal{F} = \Pi_h^1 + \Pi_h^2(\mathbf{id} - \Pi_h^1)$  which verifies theorem 85. Before we construct these operators we shall make the following assumption about the velocity:

**Assumption 86.** *The velocity field  $u$  is a piece-wise linear polynomial in space-time.*

$u$  remains in  $W^{1,\infty}(\Omega)$  given this assumption.

**Remark 87.** *Note that if  $\mathbf{b}$  were given as an element-wise constant function then a Fortin operator for our problem would be similar to the one produced in [91] for the stationary Stokes equation. However, a piece-wise constant  $\mathbf{b}$  does not generate unique streamlines as the spatial continuity is violated.*

First we construct  $\Pi_h^1$ . Interpolation of functions in  $\mathbb{V}$  is not completely trivial since point evaluation is not supported for this equivalence class. As such we have the usual choice of selecting either the Clément [62] (its usage for Fortin construction was first suggested by Verfürth in [223]) or Scott-Zhang [200] operators. The latter is more convenient for matching Dirichlet boundary conditions, but since we do not impose these for the test functions in  $\mathbb{V}$  we shall consider this option no further.

Next we construct  $\Pi_h^2$  such that:

$$\|\Pi_h^2[\mathbf{id} - \Pi_h^1]\delta z\|_{\mathbb{V}} \leq C\|\delta z\|_{\mathbb{V}}, \quad (4.50)$$

and such that:

$$b_h(\delta z - \Pi_h^2\delta z, I, \theta) = -\langle \operatorname{div}_{th}(\mathbf{b}(\delta z - \Pi_h^2\delta z)), I \rangle + \langle \llbracket \mathbf{b}(\delta z - \Pi_h^2\delta z) \rrbracket, \theta \rangle_{\mathcal{E}_{th}^{\neq, \text{int}}} = 0,$$

for any choice of  $(I, \theta) \in \mathbb{U}_h$ . We want to set up a square system to determine the polynomial orders of  $\mathbb{U}_h$  and  $\mathbb{V}_h$  so this condition above holds. Recall that  $\llbracket \mathbf{b}(\delta z - \Pi_h^2\delta z) \rrbracket = (\mathbf{b}\delta z)^+ \cdot \eta^+ + (\mathbf{b}\delta z)^- \cdot \eta^-$ . So writing  $\eta = |\eta_1 \ \eta_2|^\top$  the jump term can be written as

$$\begin{aligned} \langle \llbracket \mathbf{b}(\delta z - \Pi_h^2\delta z) \rrbracket, \theta \rangle_{\mathcal{E}_{th}^{\neq, \text{int}}} &= \sum_F \int_F [(\mathbf{b}\delta z)^+ \cdot \eta^+ + (\mathbf{b}\delta z)^- \cdot \eta^-] \theta \\ &= \sum_F \int_F [(\eta_1^+ + u\eta_2^+)\delta z^+ + (\eta_1^- + u\eta_2^-)\delta z^-] \theta. \end{aligned}$$

Let us define the integers  $p_u$ ,  $p_z$ ,  $p_I$  and  $p_\theta$  as the polynomial orders of the various variables on  $K$ , noting that by assumption 86,  $p_u = 1$ . Note the term  $(\eta_1^+ + u\eta_2^+)\theta$  is of order  $p_u + p_\theta$  (resp. for  $-$ ). We therefore define the augmented condition:

$$\int_{\partial K} (\delta z - \Pi_h^2\delta z)\phi \, ds = 0, \forall \phi \in \mathcal{P}^{p_u + p_\theta}. \quad (4.51)$$

Next, we look at the first term of the orthogonality condition and examine the polynomial orders here. It is sufficient to request:

$$\langle \operatorname{div}_{th}(\mathbf{b}(\delta z - \Pi_h^2\delta z)), I \rangle = 0.$$

Integrating by parts locally this is equivalent to:

$$\langle \operatorname{div}_{th}(\mathbf{b}(\delta z - \Pi_h^2 \delta z)), I \rangle_K = \langle -(\delta z - \Pi_h^2 \delta z) \mathbf{b} \cdot \nabla_t I \rangle_K + \int_{\partial K} (\delta z - \Pi_h^2 \delta z) \mathbf{b} \cdot \eta I.$$

The boundary term vanishes provided  $p_I \leq p_\theta$  by the results we just derived for  $p_\theta$ . The interior integral condition then reads:

$$\int_K (\delta z - \Pi_h^2 \delta z) \psi \, ds = 0, \forall \psi \in \mathcal{P}^{p_u+p_I-1}. \quad (4.52)$$

because the order of  $\mathbf{b} \cdot \nabla_t I$  is  $p_u + p_I - 1$  (because of the gradient). We fix  $p_u$ , so we wish to find the degree  $p_z$  such that:

$$\int_K (\delta z - \Pi_h^2 \delta z) \psi \, dx = 0, \forall \psi \in \mathcal{P}^{p_I}, \quad (4.53a)$$

$$\int_{\partial K} (\delta z - \Pi_h^2 \delta z) \phi \, ds = 0, \forall \phi \in \mathcal{P}^{p_\theta+1}. \quad (4.53b)$$

It is the subject of future work to determine the Fortin operator above - or suitable degrees  $p_I$ ,  $p_\theta$  and  $p_z$  such that the discrete inf-sup condition is met. The goal is to leverage a technology called *static condensation* - or *hybridisation* - where the discrete system (8) is decomposed into local subproblems on  $K$  in terms of the trace variables  $\theta$  (which can be done in parallel), and then "glued" together to reconstruct the image  $I$  and source  $z$ . We do not attempt to explain the details of this parallel optimisation method and simply refer to [106] for a Firedrake abstraction of this method that we aim to apply to the system (8).

Broersen et al. [46] construct *near-optimal* test functions for a bilinear form similar to  $b_h(\cdot, \cdot)$  above. These are derived from streamline integrals of the test functions on each  $K \in \Omega_h$  using a piecewise constant approximation of the velocity field. It was deemed out of scope of this project to investigate whether such test functions could prove useful in proving a discrete inf-sup condition. In the seminal papers [80, 77] optimal test functions for a transport problem are also computed, although in the latter it is assumed that the field  $u$  is incompressible which is in general not the case for the problem we treat here. The difficulties that remain for the DPG formulation are analytical, namely the proof of assumption 78 and those related to the discrete optimal test functions. We are however able to compute with the DPG formulation using the Firedrake package in certain cases. The main challenge numerically is to treat  $\theta$  on those edges where  $\mathbf{b} \cdot \eta = 0$ , where  $\eta$  is an outward normal of the mesh skeleton. Recall that the function space  $G(\mathbf{b}; \mathcal{E}_h^\neq)$  for  $\theta$  in (4.29) passes via the decomposition in (4.27), so in fact  $\theta$  is not defined on subsets of  $\mathcal{E}_h^\neq$  positive measure where  $\mathbf{b} \cdot \eta = 0$ . Numerically,  $\theta$  is a piece-wise polynomial defined on the edges so a practical way of enforcing this condition is to zero the degrees of freedom associated with  $\theta$  on an edge if  $\int_e \mathbf{b} \, ds = 0$ . It is easy to see that this method may not be granular enough depending on the velocity field, so this approach may constitute a variational crime that is the subject of further analysis.

## 4.2 OUTER PROBLEM

The various forms of the inner problem equip us with a set of *forward operators* that map the one-parameter family of velocity fields  $u$  to a solution to the inner problem. These forward operators motivate a gradient descent method to solve the inverse problem of finding the optimal velocity  $u$  minimising the metamorphosis function in (4.1).

### 4.2.1 GRADIENT DESCENT

Central to the preceding analysis was the space  $G(\mathbf{b}; \Omega)$  which depends explicitly on the velocity  $u$  via  $\mathbf{b}$ . If we wish to explore the linear space  $V$  by means of a gradient descent method perturbing a velocity  $u \mapsto u + \delta u$  (where  $\delta u$  is some appropriate search direction) implies a change in the Hilbert space  $G(\mathbf{b}; \Omega)$ . Since a variation  $\delta u$  in the velocity is arbitrary, it is not straight-forward to characterise the derivative of image  $I \in G(\mathbf{b}; \Omega)$  since it surely will not occupy  $G(\mathbf{b}; \Omega)$  unless the variation  $\delta u$  vanishes. We do not, to the best of the author's knowledge, have at our disposal analytical tools to treat such a problem for the full space  $G(\mathbf{b}; \Omega)$ . We briefly remark on a possible way of taking this research further.

**Remark 88** (Analytical derivative). *It may not be necessary to specify differentiability of all functions in  $G(\mathbf{b}; \Omega)$  with respect to the velocity, but only for images that are "optimal" for a given velocity field. For sufficiently smooth data the image is given by (4.10). Despite not having a convenient computational way of evaluating this formula, it is possible to state an explicit form of the derivative of  $I$  with respect to a variation in the velocity field. We refer to the work of François-Xavier Vialard for a semi-differentiation lemma for discontinuous image matching with (special) functions of bounded variation [224, 225]. With more work we believe that some of the analytical tools used here may be adapted to the space-time setting.*

A pragmatic solution to this problem is therefore simply to restrict the space of images to be continuous everywhere on  $\Omega$ , and importantly finite-dimensional by using the conforming discretisation of (4). The approach we take here is called a *discretise-then-optimize* strategy, where we replace the function spaces with finite-dimensional surrogates and *then* apply a gradient descent method. This is in contrast with an *optimize-then-discretise* scheme, where an optimality system (sometimes referred to as the Karush-Kuhn-Tucker, or KKT, system) is derived for an infinite-dimensional optimisation problem after which the resulting equations are discretised together. We return to this briefly later on, see also [125, Chapter 3] for details.

For some finite-dimensional spaces  $V_h$  and  $G_h$  we use the inner problem formu-

lations above to pose the inverse problem as follows:

$$\inf_{(u,I) \in V_h \times G_h} H = \frac{1}{2} \int_0^1 \|u\|_V^2 + \sigma^{-2} \left\| \frac{\partial I^*[u]}{\partial t} + u \cdot \nabla I^*[u] \right\|_{0,\Omega}^2 dt \quad (4.54a)$$

$$\text{subject to } I^*[u] \triangleq \arg \inf_{\substack{I \in G_h, \\ I|_{\Gamma_0} = I_0, \\ I|_{\Gamma_1} = I_1}} \left\| \frac{\partial I}{\partial t} + u \cdot \nabla I \right\|_{0,\Omega}^2. \quad (4.54b)$$

We write this as a *reduced* problem for the control variable  $u$ :

$$\inf_{u \in V_h} H_R = \frac{1}{2} \int_0^1 \|u\|_V^2 + \sigma^{-2} \left\| \frac{\partial I^*[u]}{\partial t} + u \cdot \nabla I^*[u] \right\|_{0,\Omega}^2 dt \quad (4.55a)$$

$$\text{subject to } I^*[u] \text{ solves problem 4.} \quad (4.55b)$$

$I^*[u]$  is the *optimal* source term given  $u$  in sense of the inner problems in the previous section<sup>2</sup>. The space  $V_h$  is described later. Remark that the inverse problem is for  $u$  in the whole time, not just initial because we are using space-time. In this section we treat the inverse problem directly as an optimisation problem using a gradient descent approach. Alternative ways of solving this inverse problem are discussed in section 4.3.

We note that (4.54) and (4.55) lead to the same stationary points, so the reformulation is idempotent. Setting  $\delta H = 0$  yields:

$$\begin{aligned} \langle u, \delta u \rangle_V &= -\sigma^{-2} \langle \mathbf{b} \cdot \nabla_t I, \delta u \cdot \nabla I \rangle_2, & \forall \delta u \in V_h \\ \langle \mathbf{b} \cdot \nabla I, \mathbf{b} \cdot \nabla \delta I \rangle_2 &= 0, & \forall \delta I \in G_h \end{aligned}$$

for all  $t = 0, \dots, 1$ , with the image at time  $t = 0, 1$  fixed. Since  $G_h$  is a globally continuous piecewise finite-dimensional space, the spatial derivative is well-defined. Doing the same for  $H_R$  we get:

$$\langle u, \delta u \rangle_V = -\sigma^{-2} \langle \mathbf{b} \cdot \nabla_t I^*[u], \mathbf{b} \cdot \nabla_t \frac{\delta I^*}{\delta u}[u] \delta u + \delta u \cdot \nabla I^*[u] \rangle_2, \quad \forall \delta u \in V_h,$$

where  $\frac{\delta I^*}{\delta u}[u] \delta u$  is the Fréchet derivative of the solution  $I^*[u]$  at  $u$  with respect to  $u$ . In the PDE-constrained community this is commonly referred to as the *adjoint*. We notice that the reduced nature of the reformulation (4.55) introduces a new term in the equation for  $u$ . However, since  $\frac{\delta I^*}{\delta u}[u] \delta u \in G_h$  (since the derivative is taken with respect to the inner product on  $G_h$  in sense of Fréchet), then:

$$\langle \mathbf{b} \cdot \nabla_t I^*[u], \mathbf{b} \cdot \nabla_t \frac{\delta I^*}{\delta u}[u] \delta u \rangle_2 = 0,$$

since  $I^*[u]$  solves problem 4, which is a convex problem, so the equivalence follows.

---

<sup>2</sup> Here we have used problem 4, but future work aims at replacing this with the DPG formulations.

We comment on possible discretisations of the space  $V$ . To this end we recall that for  $u$  to generate diffeomorphisms it is sufficient that the map  $t \mapsto u_t$  is integrable, while (for any  $t$ ),  $x \mapsto u_t(x)$  is at least Lipschitz [234, Chapter 8]. Firedrake supports a variety of meshes e.g. triangular meshes and quadrilaterals [130] on which function space can be built leading to different choices for the discretisation of  $V$ . This section briefly discusses how these can be realised numerically using Firedrake [191] as well as how knowledge of the existence of the gradient from the previous section motivates an adjoint-based automatic differentiation algorithm. The focus here is simply to motivate the usefulness of the inner problem formulations.

$V$  can be discretised with simple piecewise linear  $C^0$ -conforming finite element space leading to Lipschitz velocity fields (in space-time). This was used in the numerical evidence shown in the previous section. Firedrake also allows the user to define extruded quadrilateral meshes [168, 34]. These make it possible to separate the regularity between space and time by a tensor-product structure, therefore allowing e.g. the finite element of [231] used in the previous chapter to be applied in the spatial direction, while time can be either a simple  $C^0$ -conforming space or possibly discontinuous. In the rest of this section we explore the use of the piecewise linears in space-time for simplicity and leave the investigation of tensor-product elements to future work. This section concludes with a discussion and some numerical examples of solving the inverse problem (4.55) for space-time  $u$ .

*Automatic differentiation* is a software abstraction allowing for the numerical evaluation of the Jacobian of expressions that are given as functions in a programming language [182]. These essentially articulate chain and product rules for computer programmes. The implementation of such an abstraction permits a user to evaluate gradients of for instance complex composite functions for use in an optimisation loop. For instance, formally speaking, the derivative of the functional in (4.55) with respect to the velocity requires some sense to be made about the expression  $\frac{\delta z^*[u]}{\delta u}$ . In the PDE-constrained optimisation community this expression is called the *adjoint* and it is precisely the machinery of automatic differentiation that (upon suitable discretisation) allows for its numerical evaluation. This machinery is indeed necessary as no explicit gradient can be evaluated from  $z^*[u]$ . As we have mentioned before, Firedrake supports an automatic differentiation engine [177, 99] (we use version 2018.1). With only a few lines of code we can express the optimisation problem (4.55) which can then be solved by e.g. a Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [184]. We used the Rapid Optimization Library (ROL) (version 0.0.9), see [trilinos.github.io](https://trilinos.github.io) (Python interface version 0.0.16). For details see the previous references. Our setup to run the simulations is identical to that described in section 4.1.2.1 Here we provide but a few numerical examples to support our work in section 4.1.2.

**Remark 89.** *The DPG system for the inner problem naturally involves globally non-smooth functions. Numerical experiments suggest standard optimisation tools available via Firedrake are not ideal in this setting. We observe a degradation in the speed and convergence of the optimiser and we therefore suggest that looking further into the non-smooth optimisation literature can address some of these issues. For this reason we only provide numerical examples for the outer problem using the conforming discretisation of the inner problem.*

Figure 4.3 and 4.4 show the image (or density)  $I$  results of a gradient descent applied to the problem in (4.55) using Firedrake’s automatic differentiation abstraction and the conforming discretisation from section 4. In these figures,  $h = 0.025$  in space and time. It took 15 minutes to compute all of these solutions. Both the image and the velocity field is discretised using a globally continuous piecewise linear finite element space on a spatially periodic domain. These figures show that our method is able to compute meaningful matches between different densities. In figure 4.3.A and E we see that transport solutions are recovered, even in the presence of a template with a sharp spatial gradient. For the other cases we have chosen densities which can be matched without pure transport. In some instances, i.e. F and D, the gradient descent finds intuitive matches between the data, while B, C and G struggle to recover the same visual effect. For these three configurations we find that advection is recovered for some parts of the image but we do not see a *merging* effect as before. This may be due to the gradient method settling in local minima. For instance, in figure 4.3.C it is more economical to seek an approximately constant velocity field to match the left-most mode of the template while causing the right-most mode to vanish.

We also comment on the role of  $\sigma$  which we recall acts as a penalisation parameter for advection term. In figure 4.3 the two terms in (4.55) are weighted equally. Figure 4.4 uses the same data as in figure 4.3, albeit with  $\sigma^{-2} = 10^5$ . We observe similar results as before with the exception of B and C. Adjusting the value of the penalty parameter allows the gradient method to recover, with moderate success, some of the density merging behaviour observed previously. With a more complete theoretical understanding of the space-time gradient descent method, and more expertise in numerical optimisation, we can begin to carry out more careful parameter studies for  $\sigma$ . This is deferred to future work. We also present some results for  $d = 2$  in figure 4.5 and 4.6 for different values of  $\sigma$ . These took only 9 minutes to compute as we quickly observe some error stagnation here earlier on than for the  $d = 1$  case and therefore simply exit the optimisation procedure more quickly. In the spatial dimension,  $h = 0.025$  and extruded in time with 15 subdivisions. The observations are similar to those in the one-dimensional case. We comment on a few of these. For transport-type problems (the first and third rows of figure 4.5 and 4.6), the parameter  $\sigma$  does not change the nature of the solutions, albeit smearing out the smooth data in the first row when the penalty parameter is low. This behaviour is not present for discontinuous data which we attribute the flow being almost grid-aligned and constant in time.  $\sigma$  does not appear to alter the *merging* effect in the

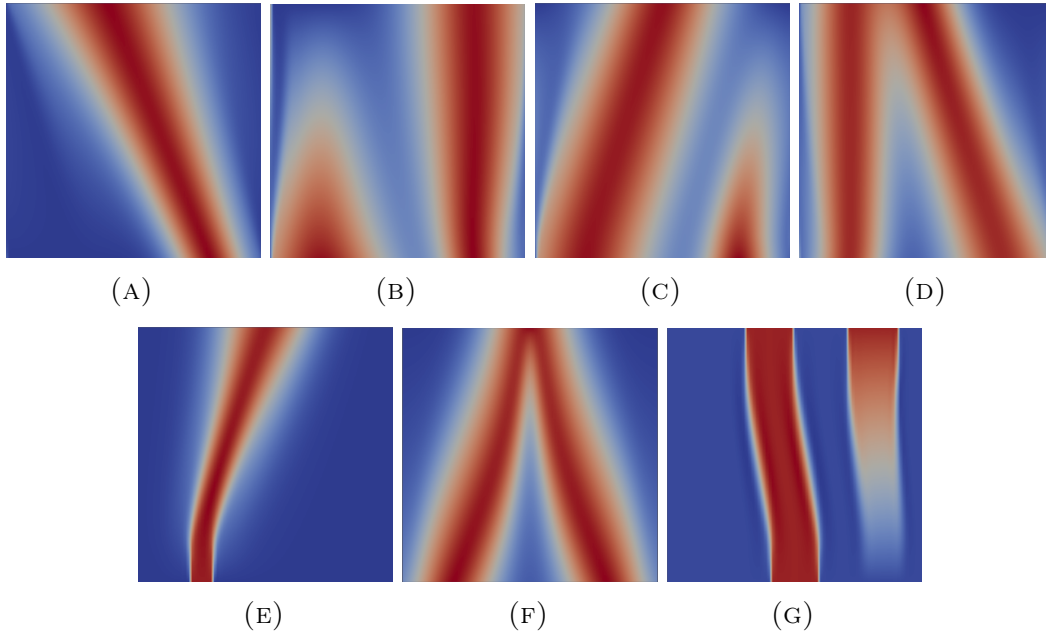


FIGURE 4.3: Results for the outer problem for various template and target pairs in one dimension. The vertical dimension is time and the horizontal dimension is space, so the *north* and *south* faces of the domain represent the target and template, respectively. Here we have taken  $\sigma^{-2} = 1$ .

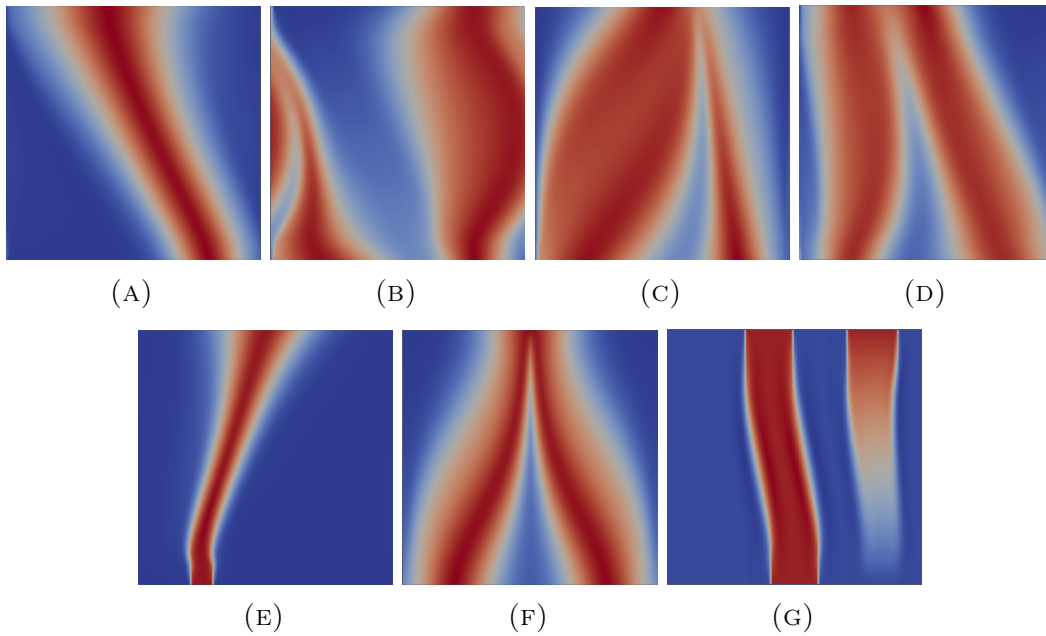


FIGURE 4.4: Results for the outer problem for various template and target pairs in one dimension with the same template and target densities as in figure 4.3. Here we have taken  $\sigma^{-2} = 10^5$  thereby forcing the algorithm to seek more diffeomorphic geodesics.

example provided in the second row of the figures, though overall the intensity values are lower whenever  $\sigma = 1$ . Again this can be explained by advection being enforced to a lesser extent, but not changing the direction of the velocity field. Lastly are the fourth and fifth rows where we truly see the influence of  $\sigma$ . When  $\sigma = 1$  in figure 4.5, very little advection is observed and the motion between the template and the target can be viewed as a simple "fading in/out" effect. In figure 4.6, however, we see some interesting results showing more meaningful (qualitatively speaking), geodesics between the images. Although these results are still of a toy problem nature they show clear promise for our approach. These results can be generated in less than 24 hours on standard consumer hardware (without using any preconditioning whatsoever). One may therefore conjecture that, with the built-in parallelism and high-performance support in Firedrake [191] we may be able to take this method further and generate geodesic between high resolution images.

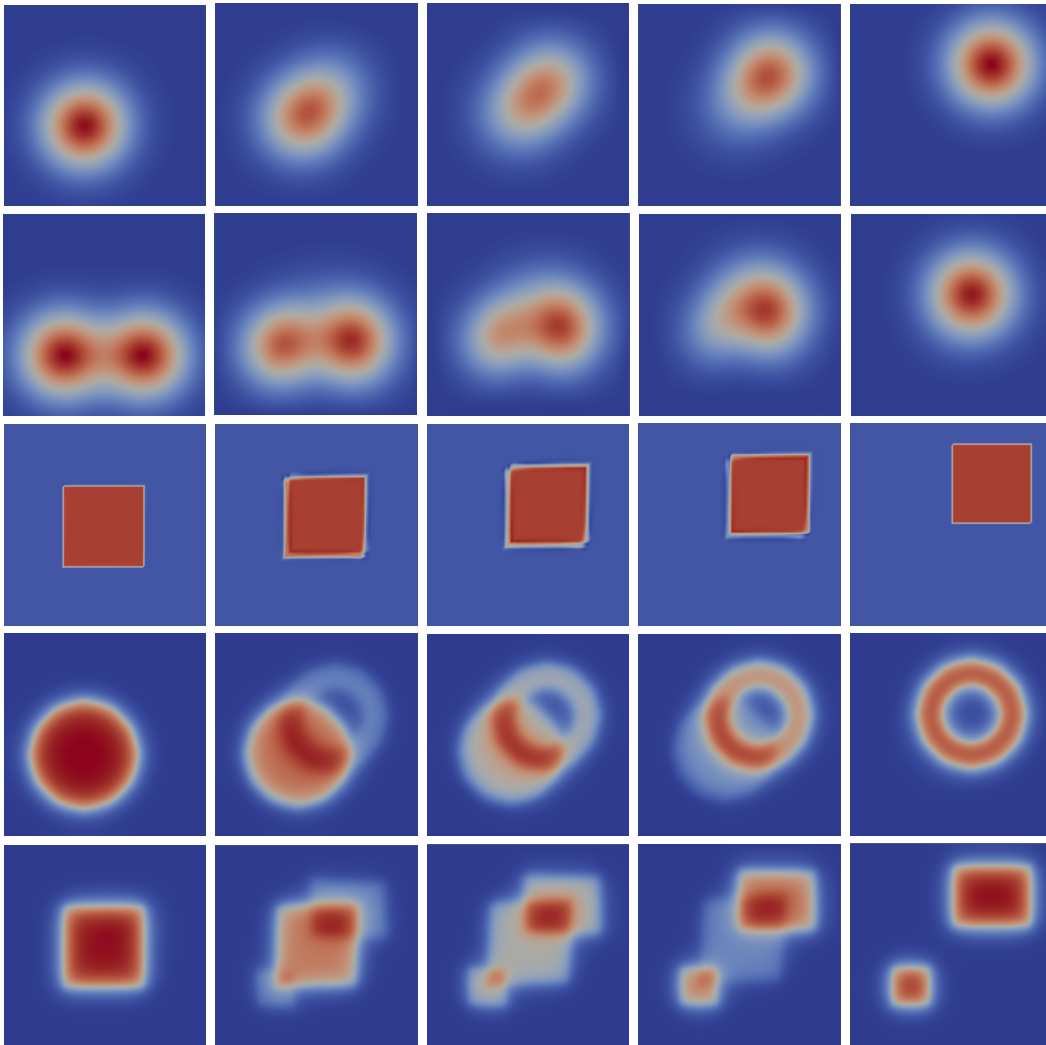


FIGURE 4.5: Results for the outer problem for various template and target pairs in two dimensions for  $\sigma^{-2} = 1$ . Each row shows the evolution at time  $t = 0, 0.25, 0.5, 0.75, 1$ .

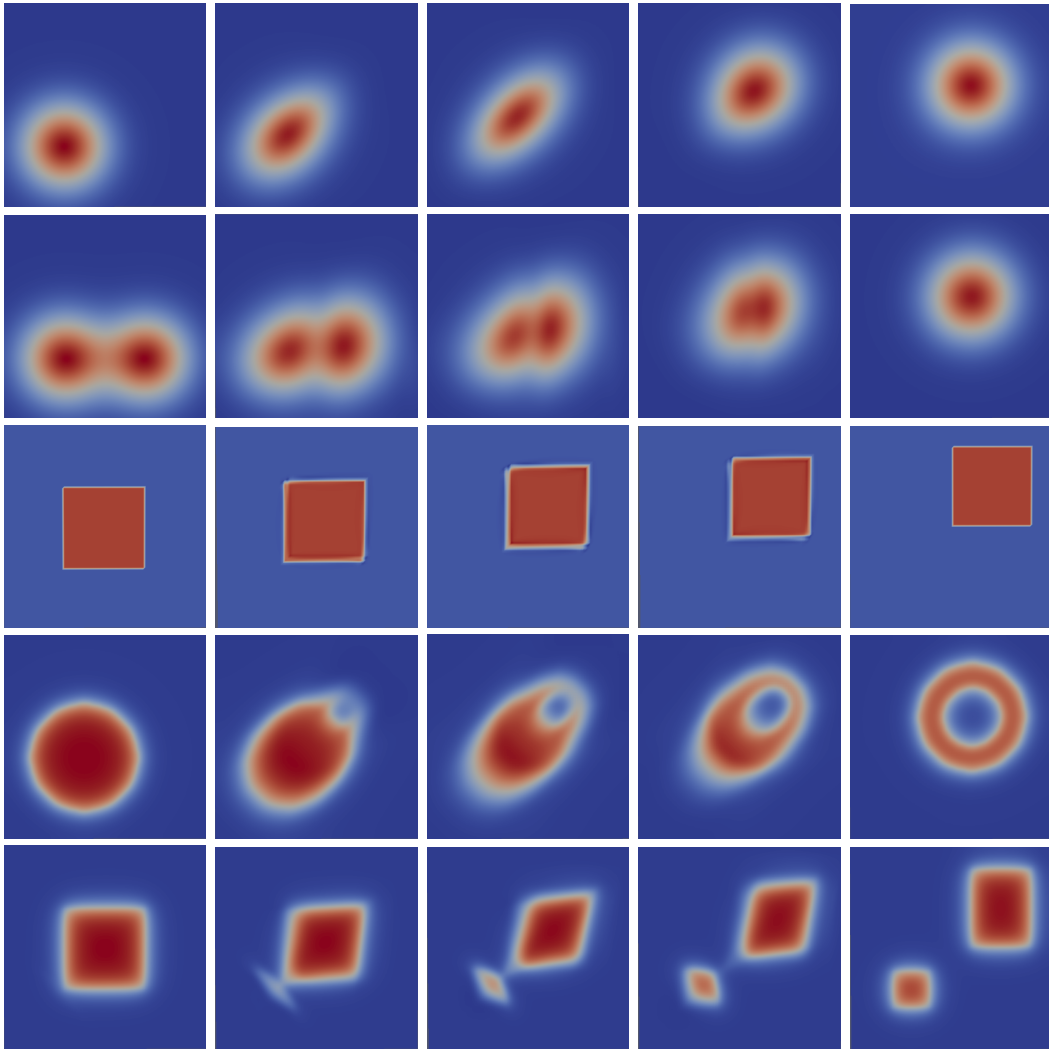


FIGURE 4.6: Results for the outer problem for various template and target pairs in two dimensions for  $\sigma^{-2} = 10^5$  to seek more diffeomorphic geodesics. Each row shows the evolution at time  $t = 0, 0.25, 0.5, 0.75, 1$ .

### 4.3 SUMMARY

This chapter successfully casts the metamorphosis in a variational Hilbertian space-time setting. The variational setting introduced in section 4.1 provides a novel framework to study weakly space-time differentiable geodesics for metamorphosis. We identified, in various ways, three formulations of the so-called inner problem with different interpretations of the advective operator  $\mathbf{b} \cdot \nabla_t$ . A *primal* formulation in section 4.1.2 and two *dual* formulations in section 4.1.3 for  $L^2$  images. From the vantage points of a least-squared variational setting and theory of Friedrich systems we derived deep theoretical results for the inner problem in section 4.1 which essentially reinterprets what is assumed to hold for most elliptic problems (continuity, coercivity, trace theorems) in the least-squared advection setting. Based on this we presented a simple conforming

finite element method. The dual formulations cast the geodesic equations in the discontinuous Petrov-Galerkin framework. We also paved the way towards the design of a Fortin operator for our mixed method. Finally we showed a way to solve the inverse problem to find a suitable velocity field  $u$ .

There are many open problems and options yet to be explored for space-time metamorphosis. The main theoretical challenges for the inner problem is showing a regularity result akin to the usual elliptic regularity for  $H^1$  variational problems. Second to show the existence of a Fortin operator for the discrete DPG formulation. Ideally, the order of the polynomials determined in (4.53) would open up the possibility to use static condensation for the system and greatly reduce the computational load of the inner problem see e.g. [193] for an example of this. The DPG formulation also highlights the need for a non-smooth optimiser as a result of the discontinuous functions inherent in the system. In addition, preconditioning of any of the discrete formulations presented above would also help in accomplishing this task. For instance, [5] offers some insights into preconditioning for a space-time wave equation and offers excellent numerical evidence that might be useful in this endeavour. Space-time adaptivity could also be explored in order to resolve details of images represented as finite element functions, in which case one would dispense with the quasi-uniformity assumption that we presented in chapter 2. Based on our numerical experiments in 2D the combination of preconditioning, static condensation techniques and more bespoke numerical optimisers can lead to a powerful method on par with contemporary methods.

The outer problem also presents some possibilities for future work. In this project we have chosen to treat the outer problem with a *discretise-then-optimize* approach because of the availability of the seamless adjoint framework in Firedrake. Another approach that was attempted was to derive the Euler-Lagrange equations for the "full" metamorphosis problem (also called EPDiff in the literature) to solve for  $u$ ,  $I$  and  $z$  *simultaneously*, and then discretise this nonlinear system of equations. A Newton method could then be applied for this system, some work has been done previously in this area for LDDMM [123]. Also, the work [238] establishes a fast algorithm for computing finite-dimensional velocity fields which could be used in conjunction with the inner problem formulations above. This may be morally similar to discretising the velocity field using finite elements but we simply mention this as a possible avenue owing to the numerical evidence in support of the authors' work. Finally, a Bayesian inverse formulation is also yet to be explored. Formally one could define a distribution  $\rho$  over the space  $V$  of velocity fields by:

$$\rho(u) = e^{\int_0^1 \langle u_\varphi, u_\varphi \rangle_V + \sigma^{-2} \|\mathbf{b} \cdot \nabla_t i\|_{0,\Omega}^2 dt}, \quad (4.56)$$

subject to the boundary conditions for the image  $i$ . Letting  $\rho$  denote the *likelihood* of a candidate  $u$  and define a prior on the function space  $V$  we can - formally speaking - reason about a *maximum likelihood* strategy for (4.56) see

e.g. [65, 67, 66] for an illustration of this idea and [211, 71, 72] for the theory behind these.

# CHAPTER 5

---

## SELECTIVE METAMORPHOSIS

---

In this chapter we present the results of a project initially conducted in collaboration with Dr. Alexis Arnaudon. This generated a conference paper [37], presented at *Geometric Science of Information 2019* in Toulouse which forms the basis of this chapter.

**Remark 90** (Attribution). *The new ideas presented in section 5.1.2 were generated by both authors, the technical results done by the author. The implementation of the forward problems in section 5.1 are the work of Dr. Arnaudon, whereas the implementation of those presented in section 5.1.2 was carried out in collaboration. The algorithm and implementation thereof in section 5.2 were done by the author. The two ideas represented by (5.19) in section 5.3 were a collaborate effort, although the computations in this section were carried out by the author.*

Chapter 3 presented a framework for curve matching where the target is approximated in the orbit of a subgroup of diffeomorphisms. Next, chapter 4 presented the metamorphosis metric on image space which allowed for transgression of the diffeomorphic constraint by allowing for intensity changes. The choice of metric clearly depends on the application and the data. Purely diffeomorphic matching is ill-suited for modelling certain growth phenomena. This chapter considers the question: what if the application calls for a more granular approach in the sense that we request a diffeomorphic evolution on certain parts of the shape while allowing for growth in other parts? This chapter presents a framework which, in some sense, interpolates between the frameworks used in chapter 3 and 4 which we call *selective metamorphosis*. Here we focus mainly on modelling aspects using landmark data and how making changes to the classical framework of metamorphosis lead to new insights.

### 5.1 LANDMARKS

We recall that a set of landmarks is a finite point cloud in  $\mathbb{R}^d$  typically representing locations of particular interest sampled from some higher-dimensional shape space. To set up the notation we present some preliminaries here and describe the classic LDDMM and metamorphosis frameworks in section 5.1.1

before selective metamorphosis in section 5.1.2.

We let  $V = L^2([0, 1], V)$  denote the space for our time-dependent velocities  $u$  such that  $u_t \in V \hookrightarrow C_0^k(\Omega, \mathbb{R}^d)$ , where  $\Omega \subset \mathbb{R}^d$ ,  $k \geq 2$ ,  $d = 2$  is a connected set. In this chapter,  $\|\cdot\|_{\mathbb{R}^d}$  denotes the standard Euclidean norm.

**Definition 91** (Reproducing Kernel Hilbert Space [234, Chapter 9]). *Let  $X \hookrightarrow L^2(\Omega, \mathbb{R}^n)$ ,  $\Omega \subset \mathbb{R}^d$  be a Hilbert space where with the isometry map denoted  $K_X : X^* \rightarrow X$ .  $X$  is a reproducing kernel Hilbert space if the point evaluation Dirac  $\delta_x$  on  $X$  defined by  $\delta_x v = v(x)$  is a bounded linear map i.e. for all  $v \in X$  and  $x \in \Omega$ :*

$$(\delta_x | v)_{X^*, X} \lesssim \|v\|_X.$$

As a result there exists, for every  $x \in \Omega$ , an element  $K_x \in X$  such that

$$(\delta_x | v)_{X^*, X} = \langle K_x, v \rangle_X. \quad (5.1)$$

Now letting  $v = K_y$  we define the function  $K_X : \Omega \times \Omega \rightarrow \mathbb{R}^n$  called the kernel of  $X$  by  $K_y(x) = (\delta_x | K_y)_{X^*, X} = \langle K_x, K_y \rangle_X = K_X(x, y)$  and is related to the isometry map in the following way when  $a \in \mathbb{R}^n$ :

$$K_X(x, y)a = K_X(a\delta_y)(x). \quad (5.2)$$

This is easily seen by using (5.1):

$$\begin{aligned} (a\delta_y | K_x)_{X^*, X} &= \langle K_X(a\delta_y), K_x \rangle_X \\ &= \langle aK_X(\delta_y), K_x \rangle_X \\ &= \langle aK_y, K_x \rangle_X \\ &= K_X(x, y)a. \end{aligned}$$

Equipped with this definition we denote by  $L_V : V \rightarrow L^2(\Omega, \mathbb{R}^d)$  the symmetric operator generating  $V$  which for all  $u \in V$  satisfies the following:

$$\begin{aligned} \|u\|_V^2 &= \langle L_V u, u \rangle_{2^d, \Omega}, \\ \langle u, u \rangle_{2^d, \Omega} &\lesssim \langle L_V u, u \rangle_{2^d, \Omega}, \end{aligned}$$

with an associated inverse described by the kernel  $K_V$ . It will be useful to define an inverse  $\hat{L}_V = K_V^{-1}$  of the isometry map  $\hat{L}_V : V \rightarrow V^*$  taking values in  $V^*$  via  $L_V$  and the representation theorem:

$$(\hat{L}_V u | v)_{X^*, X} = \langle L_V u, v \rangle_{2^d, \Omega}.$$

This is also known as the Friedrich's extension [234, Theorem 9.8] of  $L_V$ .

As usual the one-parameter curve of velocities  $t \mapsto u_t$  induces a curve  $t \mapsto \varphi_t$  on a subgroup  $\text{Diff}_V(\Omega)$  of diffeomorphisms [11] via the equation:

$$\dot{\varphi}_t = u_t \circ \varphi_t, \quad \varphi_0 = \text{id}. \quad (5.3)$$

In this section we are concerned with landmarks so we shall use bold variables to ease notation. That is, we use  $\mathbf{q}_t$  to denote a collection of positions of  $M$  landmarks  $\{q_t^i\}_{i=1}^M$  at time  $t$ , with  $q_t^i \in \mathbb{R}^d$ ,  $d = 2$ . When the time index is omitted  $\mathbf{q}$  refers to the one-parameter family of landmark positions. Further, when  $\varphi_t$  is a diffeomorphism we understand the action of  $\varphi_t$  on  $\mathbf{g} = \{g^i\}_{i=1}^M$ ,  $g^i \in \mathbb{R}^d$  as:

$$\varphi_t \cdot \mathbf{g} = \{\varphi_t \cdot g^i\}_{i=1}^M,$$

where  $\varphi_t \cdot g^i = \varphi_t \circ g^i$ . We shall also use the notation:

$$\dot{\mathbf{q}}_t = u_t \circ \mathbf{q}_t$$

to describe the evolution of the collection of landmarks whereby each landmark  $i = 1, \dots, M$  is governed by

$$\dot{q}_t^i = u_t \circ q_t^i.$$

We also use this notation for operations such as addition and multiplication, where the operations are understood as component-wise along the *landmark* dimension  $i = 1, \dots, M$ . With this we shall throughout this chapter assume that the kernel  $K_V$  is Gaussian and given by the following expression:

$$K_V(x, y) = e^{-\frac{\|x-y\|_{\mathbb{R}^d}^2}{2\tau^2}}, \quad (5.4)$$

where  $\tau > 0$  is a kernel parameter that determines the interaction of the landmarks. This is easily seen: as  $\tau \rightarrow 0$ ,  $x$  and  $y$  may be closer in the plane before  $K_V(x, y)$  takes values away from zero. Because the kernel is smooth and induces a diffeomorphism the parameter  $\tau$  determines, in a sense, the “size” of the landmarks.

### 5.1.1 PREVIOUS WORK

Here we describe the classic LDDMM and metamorphosis frameworks on landmark manifolds. These finite-dimensional objects are convenient for matching experiments and can offer deep insight into new models (in particular in stochastic landmark and bridge dynamics [205, 7, 8, 6]) since there is no need to discretise the shape space *per se*. We refer to [129] for theoretical background, and [155] for an introduction to implementation of such systems.

For completeness we derive the equation for LDDMM and metamorphosis to set up the notation. Finally we compare some numerical experiments for these two frameworks to highlight the goal of the proposed model in the next section.

The LDDMM matching problem between two configurations of landmarks  $\mathbf{q}_0$  and  $\mathbf{q}_1$  seeks to minimise the following functional:

$$\frac{1}{2} \int_0^1 \|u_t\|_V^2 dt, \quad (5.5)$$

subject to the evolution equation  $\dot{\mathbf{q}}_t = u_t \circ \mathbf{q}_t$  and the boundary conditions  $\mathbf{q}_0$  and  $\mathbf{q}_1$ . Letting  $\mathbf{p}_t$  denote the conjugate momentum we can write the *Hamiltonian* [224] corresponding to this constrained optimisation problem:

$$H_{\text{LDDMM}} = \int_0^1 \langle \mathbf{p}_t, u_t \circ \mathbf{q}_t \rangle_{\mathbb{R}^{d \times M}} - \frac{1}{2} \|u_t\|_V^2 dt,$$

where  $\langle h, g \rangle_{\mathbb{R}^{d \times M}} = \sum_{i=1}^M \langle h^i, g^i \rangle_{\mathbb{R}^d}$ . Differentiating  $H_{\text{LDDMM}}$  with respect to an arbitrary variation in the velocity field gives us the familiar equation:

$$\langle L_V u_t, v \rangle_{0^d, \Omega} = \langle \mathbf{p}_t, v \circ \mathbf{q}_t \rangle_{\mathbb{R}^{d \times M}}, \quad \forall v \in V.$$

Using the properties of the RKHS this variational equation has an explicit solution in terms of  $\mathbf{p}_t$  and  $\mathbf{q}_t$ . These equations above are actually those for *measure* metamorphosis [129, Section 6.1] as landmarks can be lifted to  $V^*$  by the delta function and we write the right-hand  $\langle \mathbf{p}_t, v \circ \mathbf{q}_t \rangle_{\mathbb{R}^{d \times M}}$  as a function:

$$x \mapsto \sum_{i=1}^M p_t^i \cdot v(x) \delta_{q_t^i}(x),$$

where  $\delta_y(x)$  equals 1 when  $x = y$  and is otherwise 0. Using  $L_V^{-1} = K_V$  we can write Hamilton's equations as follows where we seek  $u \in V$  and  $\mathbf{p}_t = \{p_t^i\}_{i=1}^M$ ,  $\mathbf{q}_t = \{q_t^i\}_{i=1}^M$ ,  $q_t^i, p_t^i \in \mathbb{R}^d$  for  $t \in [0, 1]$ :

$$u_t(q_t^j) = \sum_{i=1}^M K_V(q_t^i, q_t^j) p_t^i, \quad j = 1, \dots, M, \quad (5.6a)$$

$$\dot{\mathbf{p}}_t = -\nabla u_t^\top \circ \mathbf{q}_t \mathbf{p}_t, \quad (5.6b)$$

$$\dot{\mathbf{q}}_t = u_t \circ \mathbf{q}_t. \quad (5.6c)$$

subject the boundary conditions  $\mathbf{q}_0$  and  $\mathbf{q}_1$ . The motion of  $\mathbf{p}_t$  and  $\mathbf{q}_t$  are computed from functional derivatives of  $H_{\text{LDDMM}}$  with respect to variations in  $\mathbf{q}_t$  and  $\mathbf{p}_t$ , respectively. The kinetic energy is therefore also given by the kernel:

$$\|u_t\|_V^2 = \sum_{i=1}^M \sum_{j=1}^M K_V(q_t^i, q_t^j) p_t^i \cdot p_t^j.$$

Since  $K_V$  is usually a parameter as given by (5.4) the motion of the landmarks is fully described by  $\mathbf{q}_t$  and  $\mathbf{p}_t$ .

The metamorphosis version of landmark motion introduces a so-called *template variable*  $\boldsymbol{\eta}_t = \{\eta_t^i\}_{i=1}^M$  such that the deformation of a set of landmarks is written as the composition of the landmark position  $\mathbf{q}_t$  and a template velocity  $\mathbf{z}_t$ :

$$\mathbf{q}_t = \varphi_t \cdot \boldsymbol{\eta}_t, \quad (5.7a)$$

$$\mathbf{z}_t = \varphi_t \cdot \dot{\boldsymbol{\eta}}_t, \quad (5.7b)$$

so the governing equation for  $\mathbf{q}_t$  becomes:

$$\dot{\mathbf{q}}_t = u_t(\mathbf{q}_t) + \mathbf{z}_t. \quad (5.8)$$

The template velocity measures the intensity changes in the coordinate  $\mathbf{q}_t$  that is not brought on by the infinitesimal change  $u_t$  in the diffeomorphism  $\varphi_t$ . Given the same boundary conditions as before we solve the landmark metamorphosis problem by minimisation of the following functional in  $u$  and  $\mathbf{z}$ :

$$\frac{1}{2} \int_0^1 \|u_t\|_V^2 + \sum_{i=1}^M \sigma^{-2} \|z_t^i\|_{\mathbb{R}^d}^2 dt, \quad (5.9)$$

where  $\sigma$  is a scalar acting as a penalisation parameter for the amount of metamorphic action  $\mathbf{z}$  on the motion of  $\mathbf{q}$ . The associated Hamiltonian is then:

$$H_{\text{MM}} = \int_0^1 \langle \mathbf{p}_t, u_t \circ \mathbf{q}_t + \mathbf{z}_t \rangle_{\mathbb{R}^d \times M} - \frac{1}{2} \|u_t\|_V^2 - \frac{1}{2} \sum_{i=1}^M \sigma^{-2} \|z_t^i\|_{\mathbb{R}^d}^2 dt.$$

Differentiation of this with respect to a variation in the intensity field  $\mathbf{z}_t$  yields:

$$\mathbf{z}_t = \sigma^2 \mathbf{p}_t,$$

resulting in the Hamiltonian system where we seek  $u \in V$  and  $\mathbf{p}_t = \{p_t^i\}_{i=1}^M$ ,  $\mathbf{q}_t = \{q_t^i\}_{i=1}^M$ ,  $q_t^i, p_t^i \in \mathbb{R}^d$  for  $t \in [0, 1]$ :

$$u_t(q_t^j) = \sum_{i=1}^M K_V(q_t^i, q_t^j) p_t^i, \quad j = 1, \dots, M, \quad (5.10a)$$

$$\dot{\mathbf{p}}_t = -\nabla u_t^\top \circ \mathbf{q}_t \mathbf{p}_t, \quad (5.10b)$$

$$\dot{\mathbf{q}}_t = u_t \circ \mathbf{q}_t + \sigma^2 \mathbf{p}_t. \quad (5.10c)$$

Note that the velocities  $u_t$  above are defined everywhere on  $\Omega$  but need only be evaluated on the landmarks to compute the dynamics. We recall that the LDDMM landmark equations can therefore be thought of as a formal limit of metamorphosis when  $\sigma$  vanishes so the penalty for violating a diffeomorphic match is infinite.

For simplicity we discretise the time derivatives in (5.6) and (5.10) with a forward Euler scheme. Time is discretised into  $T$  time steps  $t \in [0, 1 \dots, T-1]$  of size  $\Delta t = T^{-1}$ . We therefore approximate the time derivatives of  $\mathbf{p}$  and  $\mathbf{q}$  as follows:

$$\begin{aligned} \dot{\mathbf{p}}_t &\approx \frac{\mathbf{p}_{t+1} - \mathbf{p}_t}{\Delta t}, \\ \dot{\mathbf{q}}_t &\approx \frac{\mathbf{q}_{t+1} - \mathbf{q}_t}{\Delta t}. \end{aligned}$$

Since the kernel  $K_V$  is given explicitly by (5.4) we do not need to discretise  $L_V$  as in chapter 3. This approximation will be used throughout this chapter. To

approximate the time-discrete Hamiltonian systems above we apply a *shooting method*, where we replace the end-point condition  $\mathbf{q}_1$  with a guess for  $\mathbf{p}_0$ , and iteratively update  $\mathbf{p}_0$  until the mismatch between the landmarks at time  $t = 1$  and the prescribed locations of  $\mathbf{q}_1$  below a certain tolerance measured in the Euclidean norm. We use  $T := 100$  and carry out a maximum of 5000 shooting iterations allowing us to compute geodesics in roughly 5 minutes depending on the nature of the template and target. As in the previous chapters we use a 2014 MacBook Pro with a 2.5GHz Intel Core i7 processor and 16GB of 1600MHz DDR3 RAM to run our simulations. When replacing the endpoint conditions for the systems (5.6) and (5.10) with an initial condition we refer to these as *forward systems*. In practice, the updates to the initial momentum are computed using automatically generated adjoint equations using the automatic differentiation package Theano [215] (we use version 1.0.2) and they are therefore omitted here. The landmark equations are solved extremely efficiently as they are simple ODEs in space and only require the gradient of the velocity field and no other derivatives (time is approximated by forward Euler). The Theano back-end carries out execution speed optimisations by compiling parts of the symbolic code into executable code, allowing Theano programmes to be potentially faster than pure Python. We can simply write down symbolic expressions of the forward systems and minimise the mismatch as a function  $\mathbf{p}_0$ , see e.g. [156, 155] for more examples and a discussion of deep learning libraries for differential geometry. The authors also argue the case for GPU acceleration and scalability of this approach for solving shape matching problems.

Although diffeomorphic (LDDMM) paths always exist for landmark problems [119] it can be challenging to realise certain matchings numerically depending on the template and target. The conditioning of certain problems can be degenerate and it is evidenced that problems where geodesics approach intersection yield large kinetic energies and spurious geodesics that do not contribute to an intuitive matching. A comparison of the LDDMM and metamorphosis frameworks can be seen in figure 5.1 where we carry out matchings between sets of landmarks in  $\mathbb{R}^d$ . A template set of landmarks (marked with dots) are matched to their targets (crosses). The geodesic paths are marked with lines between. In the top two panels the template and the target is chosen to be aligned with the  $y$ -axis, and we impose a correspondence so the geodesics must "invert" the order of the landmarks. The metamorphosis model clearly recovers the symmetry implied by the template and target, while the LDDMM framework struggles to align with the target. The same applies to two bottom panels where we match a template on the unit circle to a target in the shape of the number 8. The metamorphosis matching is able to overlap the right and left-most points to the origin, while keeping the other points stationary. The overlap at the origin is almost attained for LDDMM at the cost of spurious geodesics that cause even the points that should be stationary to move. To summarise, the diffeomorphic nature makes it impossible for the landmarks to cross at the origin. For metamorphosis, equation (5.10c) explains

how the momentum  $\mathbf{p}_t$  (or equivalently, the intensity variable  $\mathbf{z}_t$ ) can account for violations of the diffeomorphic orbit implied by the  $\varphi_t$  implied by the velocity.

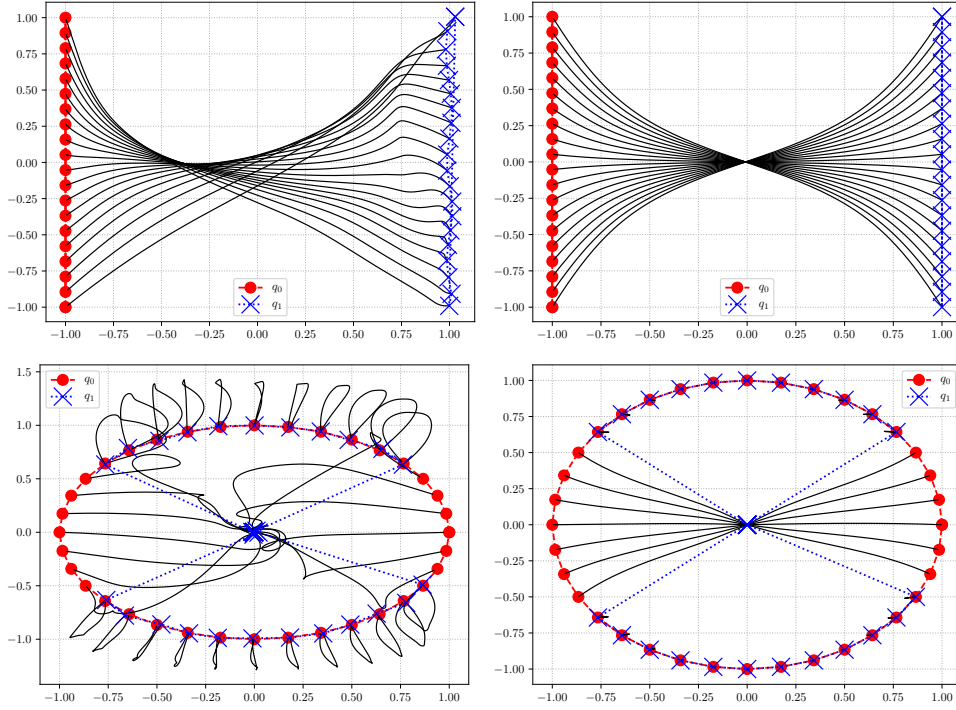


FIGURE 5.1: Landmark matching in the plane with diffeomorphic and metamorphic models. Left column: inexact purely diffeomorphic matches. Right column: the same matches where we have applied metamorphosis. These simulations were done for landmarks with a kernel size  $\tau = 0.7$  over 100 timesteps.

The aim of the next section is to investigate a meaningful "interpolation" between these two approaches. The  $\sigma$  parameter in the metamorphosis model controls, at least formally, how far the geodesics can deviate from those of LDDMM. In this chapter, we modify metamorphosis to include a spatially dependent control parameter  $x \mapsto \nu(x)$  instead of  $\sigma^{-2}$  in order to selectively allow non-diffeomorphic (*metamorphic*) matching in parts of the domain. For  $\nu = \sigma^2$  our theory recovers the standard metamorphosis model. However, with a localised control penalty, we can selectively introduce metamorphosis and model local topological effects such as growth phenomena. The key observation here is to observe that metamorphosis does not subsume LDDMM. Indeed, since  $\mathbf{z} = \sigma^2 \mathbf{p}_i$  means metamorphosis cannot result in purely diffeomorphic matching - if  $\mathbf{z}$  vanishes then so does the momentum in which case the template does not move. By allowing *spatial* control over  $\sigma^2$  we may combine these two notions. The difficulty of this problem is now to infer the function  $\nu$  without prior knowledge of the location of the topological effects. This problem is similar to the one treated in [6], where such functions were parameterising the randomness in LDDMM matching of shapes. We will use a Markov chain

Monte Carlo (MCMC) [67] approach to infer appropriate functions  $\nu$ , such that the topological effects are well described and a large part of the matching remains diffeomorphic.

### 5.1.2 SELECTIVE LANDMARK METAMORPHOSIS

We can now extend the metamorphosis setting to be able to locally control the amount of non-diffeomorphic evolution. For this, we introduce a function  $\nu : \mathbb{R}^d \rightarrow \mathbb{R}$  replacing the parameter  $\sigma^2$  such that  $\nu(x) = \sigma^2$  corresponds to the classic landmark metamorphosis. We now seek minimisers of the following energy:

$$\frac{1}{2} \int_0^1 \|u_t\|_V^2 + \sum_{i=1}^M \frac{1}{\nu(q_t^i)} \|z_t^i\|_{\mathbb{R}^d}^2 dt, \quad (5.11)$$

subject to the reconstruction equation (5.8) and the boundary conditions  $\mathbf{q}_0$  and  $\mathbf{q}_1$ . As before this leads to the relation:

$$z_t^i = \nu(q_t^i) p_t^i, \quad i = 1, \dots, M. \quad (5.12)$$

So even as  $\nu \rightarrow 0$  on, say, an open set  $\omega \in \Omega$  containing a landmark, we may describe diffeomorphic deformation (i.e. non-zero momentum) while curtailing metamorphic deformation on this set. Knowing this we can eliminate the template variable  $\mathbf{z}_t$  and write an equivalent energy without the singularity as  $\nu$  vanishes:

$$\frac{1}{2} \int_0^1 \|u_t\|_V^2 + \sum_{i=1}^M \nu(q_t^i) \|p_t^i\|_{\mathbb{R}^d}^2 dt. \quad (5.13)$$

The problem defined by (5.13) yields the following Hamiltonian equations for selective metamorphosis for landmarks where again we seek  $u \in V$  and  $\mathbf{p}_t = \{p_t^i\}_{i=1}^M$ ,  $\mathbf{q}_t = \{q_t^i\}_{i=1}^M$ ,  $q_t^i, p_t^i \in \mathbb{R}^d$  for  $t \in [0, 1]$ :

$$u_t(q_t^j) = \sum_{i=1}^M K_V(q_t^j, q_t^i) p_t^i, \quad j = 1, \dots, M \quad (5.14a)$$

$$\dot{p}_t^j = -\nabla u_t^\top \circ q_t^j p_t^j - \frac{1}{2} \nabla \nu(q_t^j) \|p_t^j\|_{\mathbb{R}^d}^2, \quad j = 1, \dots, M \quad (5.14b)$$

$$\dot{\mathbf{q}}_t = u_t \circ \mathbf{q}_t + \nu(\mathbf{q}_t) \mathbf{p}_t, \quad (5.14c)$$

with  $\mathbf{q}_0, \mathbf{q}_1$  fixed. The landmark dynamics follow standard LDDMM trajectories as  $\nu(x)$  vanishes in parts of the domain.

Figure 5.2 shows the geodesics (5.14) for the previous problem instances in figure 5.1. Here we have taken  $\nu$  to be a Gaussian with mode at the origin with isotropic covariance  $0.2I$ , where  $I$  is the identity matrix. Selective metamorphosis has the advantage of only violating the diffeomorphic property

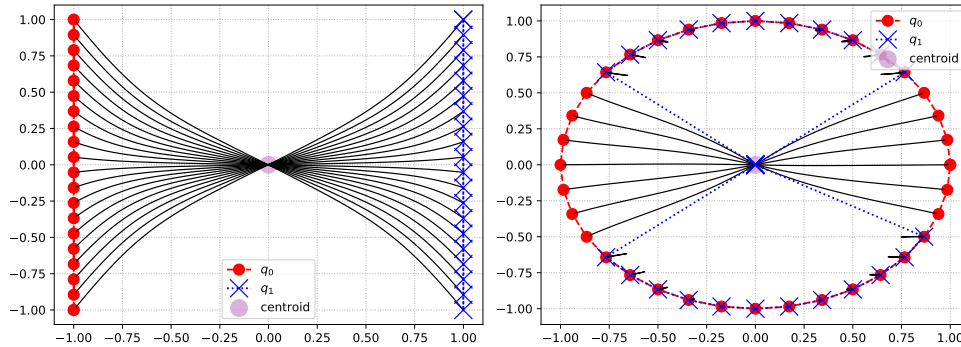


FIGURE 5.2: This figure shows how a carefully chosen spatially dependent parameter  $\nu$  can recover the metamorphosis geodesics in figure 5.1. The dot at the origin highlights the mode of the Gaussian function  $\nu$  (with a covariance of  $0.2I$ ).

where needed in along the matching which, for these problems, is near the origin.

The aim now is to pose a Bayesian inverse problem to determine  $\nu$  since we usually have no a priori knowledge of where in the domain we require metamorphic evolution e.g. due to intersecting geodesics. This is useful also as a diagnostic tool since inspection of the kinetic energy as a function of, say, a family of candidate functions  $\nu$ . As we provide numerical evidence later on, this provides, at least formally, an estimate of whether one such candidate is accurately representing where in the domain we are likely to encounter non-diffeomorphic paths. We return to this in section 5.2. This section concludes with a theoretical result before we extend the landmark equations above to image data.

**Theorem 92.** *Let  $\nu$  be such that for all  $x \in \Omega$ ,  $0 < \nu_{\inf} \leq \nu(x) \leq \sigma^2$  for some  $0 < \nu_{\inf}$  and  $0 < \sigma^2$ . Then there exists a minimiser  $(u, \mathbf{q}, \mathbf{w}) \in V \times H^1([0, 1], \mathbb{R}^{d \times M}) \times L^2([0, 1], \mathbb{R}^{d \times M})$  of (5.13) admissible to (5.14c).*

*Proof.* The functional in (5.13) is not convex so we work with a reformulation to ensure lower semi-continuity. Define a variable  $w_t^i = \sqrt{\nu(q_t^i)} p_t^i$  in the problem:

$$\begin{aligned} \inf_{\substack{u \in V \\ \mathbf{q} \in H^1([0, 1], \mathbb{R}^{d \times M}) \\ \mathbf{w} \in L^2([0, 1], \mathbb{R}^{d \times M})}} & \frac{1}{2} \int_0^1 \|u_t\|_V^2 + \sum_{i=1}^M \|w_t^i\|_{\mathbb{R}^d}^2 dt \\ & \dot{\mathbf{q}}_t = u_t(\mathbf{q}_t) + \sqrt{\nu(\mathbf{q}_t)} \mathbf{w}_t \\ & \mathbf{q}_0, \mathbf{q}_1 \text{ fixed} \end{aligned}$$

First, note that owing to the constraint equation we cannot always find a  $\mathbf{q}$  for arbitrary pairs of  $(u, \mathbf{w})$ . For any continuous curve (e.g. linear landmark-wise interpolation)  $\mathbf{q}$  between  $\mathbf{q}_0$  and  $\mathbf{q}_1$  any  $u$  we define  $\mathbf{w}$  implicitly by the following

operator  $(q_t^i, u_t) \mapsto \frac{\dot{q}_t^i - u_t(q_t^i)}{\sqrt{\nu(q_t^i)}} \triangleq w_t^i$ ,  $i = 1, \dots, M$ . Given our assumptions this operator is clearly bounded:

$$\|w_t^i\|_{\mathbb{R}^d}^2 = \left\| \frac{\dot{q}_t^i - u_t \circ q_t^i}{\sqrt{\nu \circ q_t^i}} \right\|_{\mathbb{R}^d}^2 \leq \nu_{\inf}^{-1} \left( \|\dot{q}_t^i\|_{\mathbb{R}^d}^2 + \|u_t\|_V^2 \right).$$

From this fact we can say that any minimising sequence  $(u^k, \mathbf{q}^k, \mathbf{w}^k)_{k \geq 0}$  to the functional in the reformulation above is bounded and is admissible to (5.15). We show that the reformulated constraint equation is continuous with respect to the weak topology on  $V \times H^1([0, 1], V) \times L^2([0, 1], \mathbb{R}^{d \times M})$  i.e.  $e(\mathbf{q}_t^k, \mathbf{w}_t^k, u_t^k) \rightharpoonup e(\mathbf{q}_t, \mathbf{w}_t, u_t)$  where  $e(q, w, u) \triangleq \dot{q} - u(q) - \sqrt{\nu(q)}w$ . Then,

$$\langle \sqrt{\nu(\mathbf{q}_t)} \mathbf{w}_t - \sqrt{\nu(\mathbf{q}_t^k)} \mathbf{w}_t^k, \phi \rangle \lesssim \nu_{\inf} \langle \mathbf{w}_t - \mathbf{w}_t^k, \phi \rangle \rightarrow 0, \quad \forall \phi \in L^2([0, 1], \mathbb{R}^{d \times M}).$$

Further, for  $\phi \in L^2([0, 1], V)$ ,

$$\langle u_t(\mathbf{q}_t) - u_t^k(\mathbf{q}_t^k), \phi \rangle = \langle u_t(\mathbf{q}_t) - u_t^k(\mathbf{q}_t), \phi \rangle + \langle u_t^k(\mathbf{q}_t) - u_t^k(\mathbf{q}_t^k), \phi \rangle.$$

The first term vanishes trivially, while for the second we see

$$\langle u_t^k(\mathbf{q}_t) - u_t^k(\mathbf{q}_t^k), \phi \rangle \leq \text{Lip}(u_t^k) \langle \mathbf{q}_t - \mathbf{q}_t^k, \phi \rangle \rightarrow 0$$

Linear operators such as the time derivative are naturally compatible with the weak topology, so  $\mathbf{q} \mapsto \dot{\mathbf{q}}$  is weakly continuous. Passing to subsequences where necessary we can by classic results extract bounded subsequences converging to weak limits where necessary to obtain a minimiser. Convexity of the functional above implies weak lower semi-continuity concluding the proof.  $\square$

**Theorem 93.** Assume  $\nu \in W^{2,\infty}(\Omega)$  and  $V$  is embedded in  $\mathcal{C}_0^k(\Omega)$ ,  $k \geq 2$  (continuous functions with continuous derivatives to order  $k$  vanishing at infinity). Given  $\mathbf{p}_0, \mathbf{q}_0 \in \mathbb{R}^{d \times M}$ , (5.14) with (5.4) there exists a  $t_0 > 0$  such that these are integrable on  $[0, t_0]$ .

*Proof.* Establishing appropriate Lipschitz conditions implies integrability of the system akin to [65, Theorem 5]. We note that the kernel in (5.4) is Lipschitz in  $(p, q)$  by assumption, so the composition  $(p, q) \mapsto u \circ q$  is also Lipschitz.  $u(q) \mapsto \nabla u(q)^T$  consider  $v, w \in V$  and  $x, y \in \mathbb{R}^d$ :

$$\|\nabla v(x) - \nabla w(y)\|_{0,\Omega} \lesssim \|v\|_V \|x - y\|_{0,\Omega} + \|v - w\|_V \|y\|_{0,\Omega} \quad (5.15)$$

so the mapping is Lipschitz in both the position and velocity. Given the conditions on  $\nu$  the mappings

$$\begin{aligned} (q, p) &\mapsto \nu(q)p \\ (q, p) &\mapsto \nabla \nu(q) \|p\|_{\mathbb{R}^d}^2 \end{aligned} \quad (5.16)$$

are globally Lipschitz in  $q$  by assumption and  $\|\cdot\|_{\mathbb{R}^d}^2$  is locally Lipschitz. Consequently we verify that for any  $(\mathbf{p}_0, \mathbf{q}_0) \in \mathbb{R}^{d \times M} \times \mathbb{R}^{d \times M}$ , the system (5.14) is locally Lipschitz. This implies that there exists a  $t_0 > 0$  for which we can integrate the system (5.14).  $\square$

Since we have no a priori control of the growth on  $\|p\|_{\mathbb{R}^d}^2$  in (5.16) we cannot state a global Lipschitz result. Appealing to conservation of the Hamiltonian does not imply boundedness of  $\|p\|_{\mathbb{R}^d}^2$  since this condition only bounds the velocity field.

Finally, the numerical results for selective metamorphosis depend on the nature of  $\nu$ . It is subject of future research to analyse how the complexity and numerical conditioning of the system (5.14) is affected by the choice of  $\nu$  and the landmark configurations.

## 5.2 BAYESIAN FRAMEWORK

We now place a stochastic model on  $\nu$  and use the functional (5.11) as a measure of the *fitness*, or *likelihood* of such candidate controls. The aim is to solve the Bayesian inverse problem [211] of inferring the support of  $\nu$  i.e. where in the domain growth phenomena is likely to occur. The shooting system (5.14) therefore affords a convenient way of computing the geodesics to assess the magnitude of (5.11). See also [65, 199, 3, 4] for similar Bayesian approaches in computational anatomy. We remark that we could have considered the function  $\nu$  as part of the variational problem in section 5.1.2 and enlarged the state in the shooting procedure. However, because the functional (5.11) is not a priori convex for arbitrary  $\nu$  we may find several extrema corresponding to different values of the control. Because of the implied non-uniqueness we consider it a better option to take  $\nu$  as a fixed quantity in the shooting and solve a sequence of shooting problems given by an iterative Bayesian model such as MCMC. Apart from the practical aspects of having a smaller system for numerical shooting the non-uniqueness also implies a level of qualitative consideration for what is a fitting control variable  $\nu$ . As such, finding a distribution of candidates  $\nu$  is more appropriate, see. e.g. [71, Example 4.1].

Formally speaking Bayes' rule states that the posterior distribution is proportional to the likelihood times the prior and provides the foundation for a Metropolis-Hastings algorithm [121, 169]. In the context of Gaussian measures over function space the likelihood is appropriately described the Radon-Nikodym derivative of the posterior with respect to the prior. The aim is not to provide a fully-fledged analysis of function space MCMC applied to the inverse problem above and instead refer the interested reader to [71, 67]. To avoid probability measures defined over function space we assume a finite-dimensional parametrisation of  $\nu$  in that we assume that all candidate functions  $\nu$  are the sum of  $N_\nu$  Gaussian functions:

$$\nu(x) = \sum_{j=1}^{N_\nu} e^{-a_j^{-2}\|c_j-x\|_{\mathbb{R}^d}^2}. \quad (5.17)$$

centred at  $c_j \in \mathbb{R}^d$ ,  $j = 1, \dots, N_\nu$  and with a priori fixed length-scales

$\mathbf{a} = \{a_j\}_{j=1}^{N_\nu}$ . Indeed, the results in figure 5.2 correspond to taking  $N_\nu = 1$  and  $a_1 = 0.2$  in (5.17) with  $c_1$  at the origin. Functions satisfying this parameterisation adhere to the conditions of theorem 92 and 93.

Returning to Bayes' rule we need some additional notation to define a prior on the space of controls. For any  $\mathbf{a}$  let:

$$\Xi_{\mathbf{a}} = \{\nu \mid \nu(x) = \sum_{j=1}^{N_\nu} e^{-a_j^{-2} \|c_j - x\|_{\mathbb{R}^d}^2}, c_j \in \mathbb{R}^d\}$$

define the space of functions on the form (5.17). It then makes sense to talk about points in  $\Xi_{\mathbf{a}}$  identified by  $\mathbf{c} = \{c_j\}_{j=1}^{N_\nu} \in \mathbb{R}^{d \times N_\nu}$  and we denote by  $\nu_{\mathbf{c}}$  the point in  $\Xi_{\mathbf{a}}$  whose coefficients in (5.17) are  $\mathbf{c}$ . The prior can therefore be specified as a probability distribution over  $\mathbb{R}^{d \times N_\nu}$ . Before making this precise we define the likelihood associated with a point in  $\Xi_{\mathbf{a}}$ .

For any point  $\nu_{\mathbf{c}} \in \Xi_{\mathbf{a}}$  we can compute solutions  $\mathbf{q}(\mathbf{c})$ ,  $\mathbf{p}(\mathbf{c})$ ,  $u(\mathbf{c})$  to (5.14) via (5.17) and theorem 93. These solutions define a density  $\rho(\mathbf{c})$  on the space  $\Xi_{\mathbf{a}}$  proportional to the energy (5.13) associated with its geodesics i.e.

$$\rho(\mathbf{c}) \propto e^{-\Phi_{\mathbf{c}}} = e^{-\frac{1}{2} \int_0^1 \|u_t(\mathbf{c})\|_V^2 + \sum_{i=1}^M \nu_{\mathbf{c}}(q_t^i(\mathbf{c})) \|p_t^i(\mathbf{c})\|_{\mathbb{R}^d}^2 dt}. \quad (5.18)$$

We call this the *likelihood* given  $\mathbf{c}$ . It is easy to see that the log-likelihood (sometimes called the *potential* in the literature) is precisely given, up to a sign, by (5.18). We do not attempt to prove existence of maximum likelihood estimators and readers are advised to consult [211, Theorem 5.4] and [72] for the relevant theory.

In light of the density (5.18) and the characterisation (5.17) we now wish to define a suitable prior measure over  $\mathbb{R}^{d \times N_\nu}$ . Our intuition from the previous sections tells us that the mode of the Gaussians defining  $\nu$  should be in the neighbourhood of singularities and we would like our prior to reflect this. To validate this intuition we first take  $N_\nu = 1$  and  $a_1 = 0.2$  in the definition of (5.17) and then compute the forward equations (5.14) to plot the value of the energy functional (5.11) as a function of the centroid  $c_1$ . Figure 5.3 shows this where  $c_1$  takes values on a grid in the neighbourhood of the template and target providing a heatmap indicating where we should place the mode of  $\nu$  to obtain the lowest value of (5.11). An interesting feature is observed for both these examples that were not anticipated. A bimodality appears (i.e. two ‘wells’) on these heatmaps. This means that it is cheaper in sense of (5.11) to place a single centroid slightly away from the origin rather than on top of it as our intuition alluded to in the previous section. We can explain this feature by noting that the mode of  $\nu$  in (5.17) may be such that placing the support of the mode at the origin is too costly (second term in (5.11)) when compared to the purely diffeomorphic energy (first term in (5.11)). It may therefore make

sense to misalign one such centroid away from the origin.

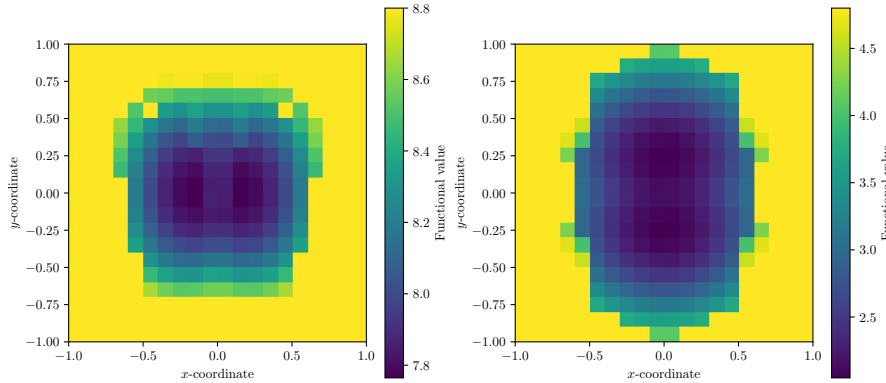


FIGURE 5.3: Heatmap of (5.11) as a function of the position of a single centroid  $c_1$  for  $a_1 = 0.2$  for the template and target pairs seen in figure 5.2. We see that the functional is minimised near the origin. Here we have used Gaussian kernel of variance  $\tau = 0.5$  cf. (5.4), 100 timesteps from  $t = 0$  to  $t = 1$ , and a metamorphosis kernel of variance 0.2.

Our prior distribution should therefore lead to an exploration of the domain  $\Omega$  leading to minimisers of the energy (5.11) starting with some initial guess  $\mathbf{c}_0 \in \mathbb{R}^{d \times N_\nu}$  and will be a simple Gaussian random walk in  $\Omega$  described later on.

We are now equipped with the ingredients to describe an MCMC algorithm of the *preconditioned Crank-Nicholson* (pCN) variety [120, Algorithm 1] which is an improvement on the standard *random walk Metropolis-Hastings algorithm* (RWMH). The aim is not to prove a law of large numbers or a limit theorem similar to e.g. [120, Section 4], but to describe a simple algorithm to generate a sequence of posterior densities that formally tend towards a minimiser of (5.11). When  $\mathbf{c}'$  is a new proposal and  $\mathbf{c}$  the current state we can define an acceptance probability by:

$$\min\left(1, \frac{e^{-\Phi_{\mathbf{c}'}}}{e^{-\Phi_{\mathbf{c}}}}\right),$$

where  $\Phi$  is defined in (5.18). This has the usual advantage that the normalising constant for the posterior need not be computed. Summarising these results we describe a pCN in algorithm 2 using the following notation:

- $N_{\text{iter}}$  : number of iterations of the algorithm,
- $\mathbf{c}_{\text{init}}$  : a collection of  $N_\nu$  points in  $\mathbb{R}^d$  representing an initial guess for the centroids,
- $\beta \in (0, 1]$ : a parameter (on which we elaborate later) that is tuned such that the acceptance rate in the range 0.5 – 0.8,
- $\Phi_{\mathbf{c}}$ : measures the potential of  $\mathbf{c}$  by (5.18),

- $\iota$ : measures the maximum distance the proposal can move away from the template and target.

The difference between pCN and RWMH lies in the coefficient  $\beta$  which scales between the previous sample and the new step  $\xi$  and chosen according to [120]. This needs to be calibrated as a too low value may increase the acceptance rate by taking shorter steps at the cost of slow exploration of the state space. Conversely, a too high value of  $\beta$  results in lower acceptance and thus convergence.

---

**Algorithm 2** pCN for selective metamorphosis

---

```

procedure PCNSM( $N_{\text{iter}}$ ,  $\mathbf{c}_{\text{init}}$ ,  $\mathbf{q}_0$ ,  $\mathbf{q}_1$ ,  $\beta$ ,  $\iota$ )
   $j \leftarrow 0$ 
   $\mathbf{c}^j \leftarrow \mathbf{c}_{\text{init}}$ 
  Solve (5.14) with  $\nu_{\mathbf{c}^j}$  and  $\mathbf{q}_0$ ,  $\mathbf{q}_1$  to obtain  $\Phi_{\mathbf{c}^j}$ 
  while  $j < N_{\text{iter}}$  do
    Sample a point  $\xi \in \mathcal{N}(0, I)$ 
     $\mathbf{c} \leftarrow \sqrt{1 - \beta^2} \mathbf{c}^j + \beta \xi$ 
    Solve (5.14) with  $\nu_{\mathbf{c}}$  and  $\mathbf{q}_0$ ,  $\mathbf{q}_1$  to obtain  $\Phi_{\mathbf{c}}$ 
     $\lambda \leftarrow$  random number in  $[0, 1]$ 
    if  $\lambda < \min(1, e^{\Phi_{\mathbf{c}^j} - \Phi_{\mathbf{c}}})$  and  $\max\{\text{dist}(\mathbf{c}, \mathbf{q}_0), \text{dist}(\mathbf{c}, \mathbf{q}_1)\} \leq \iota$  then
       $\mathbf{c}^{j+1} \leftarrow \mathbf{c}$     // accept
    else
       $\mathbf{c}^{j+1} \leftarrow \mathbf{c}^j$   // reject
     $j \leftarrow j + 1$ 
  return  $\{\mathbf{c}^j\}_{j=1}^N$ 

```

---

We have applied algorithm 2 to the previous template and target landmark configurations from figure 5.2. The pCN results for the inverted landmarks example are depicted in figure 5.4. First we notice that the results in the top panels are clearly aligned with our intuition from figure 5.3, so the prior that we have chosen (a random walk in the plane) explores the state space and provide a sample distribution for the centroid placement. We can also say that the autocorrelation quickly decreases as our sample size increases, although we do not experience convergence. These preliminary results demonstrate that for even a small number of samples the density of accepted samples corresponds at least heuristically to the analytical density histogram obtained by computing the value of the metamorphosis functional in (5.13).

We arrive at the same conclusion for the second example (right panel in figure 5.2) which was conducted to investigate how the presence of multiple centroids would interact. The results are shown in figure (5.5). From the top panels we see that both centroids are centered around the middle origin, and in combination with the histogram shown in the middle panel we can conclude that the random walk is able to find minimisers of the energy. The bottom

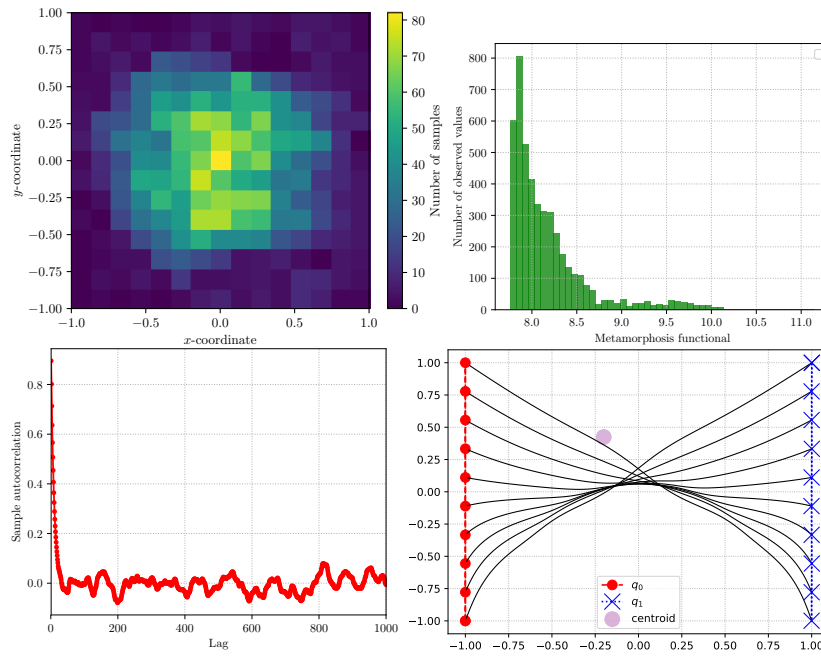


FIGURE 5.4: pCN results for the configuration in the left panel of figure 5.1. Top, left to right: a heatmap of the exploration by algorithm 2 of the space  $\Omega$  i.e. sampled positions of a centroid, a histogram indicating the number of evaluations of the selective metamorphosis functional at different values. Bottom, left to right: a sample autocorrelation plot as a function of the distance (lag) between the samples. An example of the geodesics for  $\mathbf{q}$  resulting from a centroid being placed near the origin. Here,  $\tau = 0.7$ ,  $N_\nu = 1$ ,  $\beta = 0.2$ ,  $a_1 = 0.2$ ,  $M = 10$ ,  $N_{\text{iter}} = 5000$ .

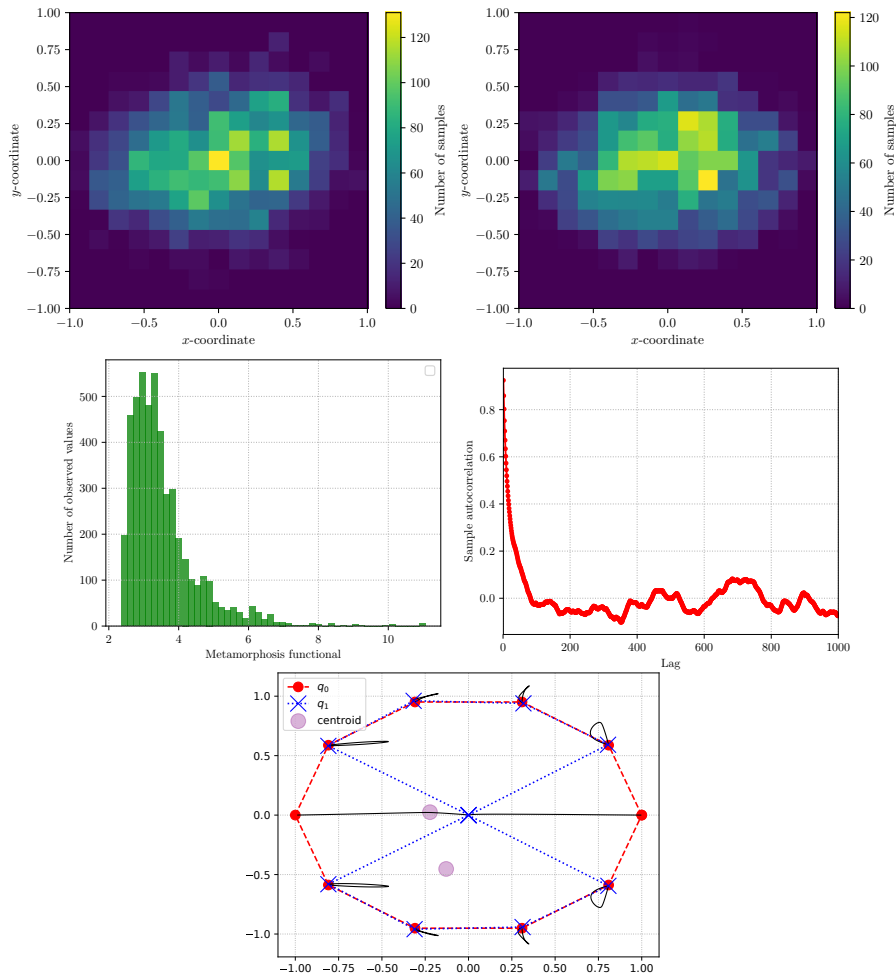


FIGURE 5.5: pCN results for the configuration in the right panel of figure 5.1. Top: heatmaps of the exploration by algorithm 2 of the space  $\Omega$  for two centroids. Middle, left to right: a histogram indicating the number of evaluations of the selective metamorphosis functional at different values, a sample autocorrelation plot as a function of the distance (lag) between the samples. Bottom: an example of the geodesics for  $\mathbf{q}$  resulting from two centroids being placed near the origin attaining the lowest energy. Here,  $\tau = 0.7$ ,  $N_\nu = 2$ ,  $\beta = 0.2$ ,  $a_1 = a_2 = 0.2$ ,  $M = 10$ ,  $N_{\text{iter}} = 5000$ .

panel shows the geodesics for one such minimiser, where the centroids are situated near the origin. We also note that the geodesic equations for  $p$  and  $q$  are time-reversible meaning that the configuration in figure 5.5 corresponds to both particle collapse as well as hole creation. Finally, we do not believe that we have enough samples to recognise the bimodality in figure 5.3.

### 5.3 IMAGES

We now formulate the equations for selective metamorphosis for images. These are given as functions that reside in a RKHS  $H$  with norm  $\|\cdot\|_H$ , also defined

over  $\Omega \subset \mathbb{R}^d$  with similar notation to that presented in section 5.1. The images are assumed to grey-scale so  $H$  will be a scalar-valued space. We assume that  $H$  is equivalent to a Sobolev space  $H^r(\Omega)$ ,  $r > 1$  so by the Sobolev embedding theorem  $H^r(\Omega) \hookrightarrow C^0(\Omega)$  and we know that  $H$  forms a Banach algebra. The full space-time image will occupy  $\mathbf{H} = L^2([0, 1], H)$ . The equations presented here are a generalisation of those derived by Richardson and Younes in their work [194]. The aim is to show how the choice of data (i.e. landmarks, images) affect the nature of the geodesic equations and, in particular, how these are much more complicated in the image case. We refer the interested reader to the reference above as we shall not pursue any rigorous results for the equations shown here.

In the previous section we introduced control of the degree of metamorphic deformation given by a function  $\nu$  acting on the landmark positions that are, in effect, measures. The setting is different for image data, as these are assumed to be supported everywhere on  $\Omega$  which leaves us with a modelling choice. Either we let  $\nu$  act on the underlying field  $\Omega$ , allowing metamorphosis in only certain parts of  $\Omega$ , resulting in a control similar to that of landmarks. Alternatively,  $\nu$  may act on the intensity values of the image, allowing metamorphosis only in a chosen band of intensities. With a localised control, we may selectively introduce or curtail metamorphosis in an image and model local topological effects such as growth phenomena. *Horizontal* metamorphosis shall refer to the case when  $\nu$  acts on  $\Omega$ , and *vertical* metamorphosis shall refer to the case when  $\nu$  acts on the image intensities. Mathematically, this leads to two functionals  $S^h$  and  $S^v$ , defined as follows:

$$S^h = \frac{1}{2} \int_0^1 \|u_t\|_V^2 + \left\| \frac{z_t}{\nu} \right\|_H^2 dt, \quad (5.19a)$$

$$S^v = \frac{1}{2} \int_0^1 \|u_t\|_V^2 + \left\| \frac{z_t}{\nu \circ I_t} \right\|_H^2 dt, \quad (5.19b)$$

such that the following reconstruction equation is satisfied:

$$\dot{I}_t + u_t \cdot \nabla I_t = z_t, \quad (5.20a)$$

$$I_l = j_l, \quad (5.20b)$$

with the image data  $H \ni j_l$ ,  $l = 1, 2$ . The division by (5.19) is well-defined by  $H$  forming an algebra when  $\nu$  is not identically zero. We shall focus on horizontal metamorphosis in this paper as we want to model control of the metamorphosis in spatial locations, although vertical metamorphosis could provide a useful approach if an application calls for granular control of the image intensities.

Given images  $H \ni j_l$ ,  $l = 1, 2$  and  $\nu \in H$  for which there exist  $\nu_{\inf} > 0$  such that  $\nu_{\inf} \leq \nu(x)$ ,  $\forall x \in \Omega$ , we now seek a minimiser  $(u, z, I) \in \mathbf{V} \times \mathbf{H} \times \mathbf{H}$  of the

following energy:

$$\frac{1}{2} \int_0^1 \|u_t\|_V^2 + \left\| \frac{z_t}{\nu} \right\|_H^2 dt, \quad (5.21)$$

such that the reconstruction equations (5.20) are satisfied. To avoid having to treat spatial derivatives we propose a similar change of variables to [194] to write PDE in (5.20) as a simple ODE amenable to a shooting procedure. We therefore define  $m_t = I_t \circ \varphi_t$  so that:

$$\dot{m}_t = z_t \circ \varphi_t.$$

As a result, we can now seek a minimiser  $(u, z, m) \in \mathbf{V} \times \mathbf{H} \times \mathbf{H}$  of the action:

$$\int_0^1 \frac{1}{2} \|u_t\|_V^2 + \left\| \frac{z_t}{\nu} \right\|_H^2 dt,$$

subject to:

$$\dot{m}_t = z_t \circ \varphi_t, \quad (5.22a)$$

$$m_l = j_l \circ \varphi_l, \quad l = 0, 1, \quad (5.22b)$$

and as usual  $\dot{\varphi}_t = u_t \circ \varphi_t$ . The reformulation eliminates the spatial derivative by explicitly solving for the diffeomorphism  $\varphi_t$  thereby allowing for a fully Lagrangian method.

We now use Lagrange curves  $\gamma \in \mathbf{H}^*$  and  $\lambda \in \mathbf{V}^*$  for the constraints, respectively, and look at critical points of the following quantity:

$$\int_0^1 \frac{1}{2} \|u_t\|_V^2 + \left\| \frac{z_t}{\nu} \right\|_H^2 + (\gamma_t | \dot{m}_t - z_t \circ \varphi_t)_{H^*, H} + (\lambda_t | \dot{\varphi}_t - u_t \circ \varphi_t)_{V^*, V} dt. \quad (5.23)$$

Define the operator  $\mathbf{D}_{\varphi_t} : g \mapsto g \circ \varphi_t$  and  $\mathbf{D}_{\varphi_t}^*$  its adjoint and  $\delta_w u$  denote differentiation of  $u$  with respect to a variation in  $w$ . Let  $\partial_\varphi$  denote a formal derivative with respect to the diffeomorphism  $\varphi$ , see [234] for more details. From (5.23) we obtain the following equations that hold for  $t \in [0, 1]$  where we seek  $u \in \mathbf{V}$ ,  $z \in \mathbf{H}$ ,  $\gamma \in \mathbf{H}^*$ ,  $\lambda \in \mathbf{V}^*$ ,  $m \in \mathbf{H}$  and such that for all  $t \in [0, 1]$ ,  $\varphi_t \in \text{Diff}(\Omega)$ :

$$\langle L_V u, \delta u \rangle_{2^d, \Omega} - (\lambda_t | \delta u \circ \varphi_t)_{V^*, V} = 0 \quad \forall \delta u \in V, \quad (5.24a)$$

$$\left\langle \frac{z_t}{\nu}, \frac{\delta z}{\nu} \right\rangle_H = (\gamma_t | \delta z \circ \varphi_t)_{H^*, H} \quad \forall \delta z \in H, \quad (5.24b)$$

$$(\delta \dot{m}_t | \gamma)_{H^*, H} = 0 \quad \forall \delta m \in H, \quad (5.24c)$$

$$\dot{\lambda}_t = -\partial_\varphi(\gamma | z_t \circ \varphi_t)_{H^*, H} - \partial_\varphi(\lambda | u_t \circ \varphi_t)_{V^*, V}, \quad (5.24d)$$

$$\dot{\varphi}_t = u \circ \varphi_t, \quad (5.24e)$$

$$\dot{m}_t = z_t \circ \varphi_t. \quad (5.24f)$$

$\gamma$  is invariant over time by a simple integration by parts, so from now on we omit its temporal subscript. We can write (5.24b) in a more convenient form:

$$(\mathbf{K}_H^{-1} \left[ \frac{z_t}{\nu} \right] \mid \frac{\delta z}{\nu})_{H^*, H} = (\mathbf{D}_{\varphi_t}^* \gamma_t \mid \delta z)_{H^*, H}.$$

Now assuming that we can write this expression as

$$(\mathbf{K}_H^{-1} \left[ \frac{z_t}{\nu} \right] \mid \delta z)_{H^*, H} = (\nu \mathbf{D}_{\varphi_t}^* \gamma_t \mid \delta z)_{H^*, H},$$

so that (with a slight abuse of notation)  $\nu \mathbf{D}_{\varphi_t}^* \gamma_t$  is considered an element of  $H^*$  we get the explicit formula for  $z_t$  given by:

$$z_t = \nu \left[ \mathbf{K}_H \left[ \nu \mathbf{D}_{\varphi_t}^* \gamma_t \right] \right] \quad \text{in sense of } H.$$

We can in fact consider the first term of  $\nu \mathbf{D}_{\varphi_t}^* \gamma_t$  as the lifting of  $\nu$  to  $H^*$  by a delta function e.g.  $v \mapsto v \delta_x$  so that the multiplication of this linear form with  $\mathbf{D}_{\varphi_t}^* \gamma_t$  still occupies  $H^*$ . The same reasoning yields an explicit formula for  $u_t$  and (5.24) can be written:

$$u_t = \mathbf{K}_V \mathbf{D}_{\varphi_t}^* \lambda_t, \tag{5.25a}$$

$$z_t = \nu \left[ \mathbf{K}_H \left[ \nu \mathbf{D}_{\varphi_t}^* \gamma_t \right] \right], \tag{5.25b}$$

$$\dot{\gamma}_t = 0, \tag{5.25c}$$

$$\dot{\lambda}_t = -\partial_\varphi(\gamma \mid z_t \circ \varphi_t)_{H^*, H} - \partial_\varphi(\lambda \mid u_t \circ \varphi_t)_{V^*, V}, \tag{5.25d}$$

$$\dot{\varphi}_t = u \circ \varphi_t, \tag{5.25e}$$

$$\dot{m}_t = z_t \circ \varphi_t. \tag{5.25f}$$

In the section we showed how the order of operators in equation (5.25b) results in an interesting interaction with the isometry map (and therefore the kernel). For a discussion of boundary conditions for this system we refer to [194, Section 2].

### 5.3.1 SINGULAR SOLUTIONS

Similar to the paper [194], we now look at singular solutions of (5.25) in order to derive an ODE system for images. We begin by developing some intuition to better understand the RKHS framework for image data. We select  $M$  control points  $\{x_k(t)\}_{k=1}^M$  defined by  $x_k(t) = \varphi_t(x_k(0))$ . These can be thought of as points in the domain that “store” values of the image which via the kernel representation of  $H$  defines the image everywhere similarly to how the velocity field  $u_t$  is defined everywhere. It is important to distinguish this from landmark data, as the latter is in fact measure-valued. To understand the RKHS framework in the context of images it is useful to consider a purely diffeomorphic matching framework. In this case the image intensities do not

change, so we are only looking for control points  $\{x_k(1)\}_{k=1}^M$  that move with a diffeomorphism  $\varphi_1$  at time  $t = 1$  such that the expression

$$I_0 - I_1 \circ \varphi_1,$$

is minimised in some sense e.g. point-wise or  $L^2$ . As such, the kernel approach provides a convenient way of representing a diffeomorphism on the whole space by a finite number of control points. This is precisely expressed by the dynamics (5.22) whenever  $z_t = 0, \forall t$ . The metamorphic version now also allows for intensity changes and therefore considers an evolution of a scalar-valued object  $m_t$  which is associated with each control point to be thought of as the grey-scale value at a certain point in the domain  $\Omega$ .

We therefore select  $\lambda$  and  $\gamma$  of the form:

$$\begin{aligned}\lambda_t &= \sum_{k=1}^M w_k^\lambda(t) \delta_{x_k(0)}, \\ \gamma &= \sum_{k=1}^M w_k^\gamma \delta_{x_k(0)},\end{aligned}$$

with:

$$\begin{aligned}w^\lambda(t) &= \{w_k^\lambda(t)\}_{k=1}^M, \quad w_k^\lambda(t) \in \mathbb{R}^d, \\ w^\gamma &= \{w_k^\gamma\}_{k=1}^M, \quad w_k^\gamma \in \mathbb{R}.\end{aligned}$$

Notice that  $w_k^\gamma$  is time-independent as  $\dot{\lambda} = 0$ . By direct computation,  $(\lambda | \delta u \circ \varphi_t) = \sum_{k=1}^M w_k^\lambda(t) \cdot \delta u \circ x_k(t)$ , so:

$$u_t(x) = \sum_{k=1}^M K_V(x, x_k(t)) \cdot w_k^\lambda(t). \quad (5.26)$$

Similarly,  $(\gamma | \delta z \circ \varphi_t) = \sum_{k=1}^M w_k^\gamma \delta z \circ x_k(t)$  which we use in (5.25b) as follows:

$$z_t(x) = \nu(x) \sum_{k=1}^M \nu(x_k(t)) w_k^\gamma K_H(x, x_k(t)), \quad (5.27)$$

To see this we use the expression  $\mathbf{D}_{\varphi_t}^* \gamma_t = \sum_{k=1}^M w_k^\gamma \delta_{x_k(t)}$ :

$$\begin{aligned}z_t(x) &= \nu \left[ \mathbf{K}_H \left[ \nu \mathbf{D}_{\varphi_t}^* \gamma_t \right] \right] \\ &= \nu \left[ \mathbf{K}_H \left[ \nu \sum_{k=1}^M w_k^\gamma \delta_{x_k(t)}(x) \right] \right] \\ &= \nu(x) \sum_{k=1}^M \nu(x_k(t)) w_k^\gamma K_H(x, x_k(t)).\end{aligned}$$

Inserting the singular solutions in (5.25d):

$$\begin{aligned} \sum_{k=1}^M \dot{w}_k^\lambda(t) \cdot \delta\varphi \circ x_k(0) &= - \sum_{k=1}^M w_k^\lambda(t) \cdot \nabla u_t \circ \varphi_t \circ x_k(0) \delta\varphi \circ x_k(0) \\ &\quad - \sum_{k=1}^M w_k^\gamma \nabla z_t \circ \varphi_t \circ x_k(0) \delta\varphi \circ x_k(0), \end{aligned}$$

so the singular equation for  $\lambda$  becomes:

$$\dot{w}_k^\lambda(t) = -w_k^\lambda(t) \cdot \nabla u_t \circ x_k(t) - w_k^\gamma \nabla z_t \circ x_k(t)$$

and using (5.26) and (5.27) we arrive at the *forward equations* for selective image metamorphosis:

$$\begin{aligned} \dot{w}_k^\lambda(t) &= - \sum_{j=1}^M \nabla_{x_k} K_V(x_k(t), x_j(t)) w_k^\lambda(t) \cdot w_j^\lambda(t) \\ &\quad - \sum_{j=1}^M \nu(x_j(t)) \nabla_{x_k} \left[ K_H(x_k(t), x_j(t)) \nu(x_k(t)) \right] w_k^\gamma w_j^\gamma, \end{aligned} \quad (5.28a)$$

$$\dot{x}_k(t) = \sum_{j=1}^M K_V(x_k(t), x_j(t)) \cdot w_j^\lambda(t), \quad (5.28b)$$

$$\dot{m}_k(t) = \nu(x_k(t)) \sum_{j=1}^M \nu(x_j(t)) K_H(x_k(t), x_j(t)) w_j^\gamma. \quad (5.28c)$$

We take the analysis of these equations no further in this project and instead comment on future research plans using these equations in a Bayesian setting akin to that presented for the landmark case in section 5.2.

## 5.4 SUMMARY

We presented an approach allowing for control over the intensity variation for metamorphic matching. We analysed the selective metamorphosis problem, the associated geodesic equations and demonstrated a pCN algorithm inferring a simple parameterisation of  $\nu$ . This generalises LDDMM and metamorphosis and could provide a first-order exploratory tool for physicians to see if the development of a biological feature stems from a few violations of diffeomorphic evolution. While we do not attempt to provide any convergence results (whether of time discretisations or relating to MCMC), we believe our work on the landmark case and the preliminary calculations in section 5.3 motivates future work for image data. In particular we believe that the kernel operation library KeOps<sup>1</sup> [54] is perfectly suited for the shooting problems in this paper.

---

<sup>1</sup> [www.kernel-operations.io](http://www.kernel-operations.io)

See [92] for an application of this framework. KeOps not only provides a convenient Python interface for specifying kernel products for the relevant RKHS we shall use (such as those ubiquitous in the equations in section 5.3), it also provides seamless graphical processing unit (GPU) acceleration known to be well-suited to particle methods. Moreover, KeOps also includes an automatic differentiation engine via PyTorch [187] that automatically formulates and integrates the adjoint equations under the bonnet thereby making it ideal for shooting. The main advantage here is therefore that the user need only specify the forward equations and a loss function i.e. (5.11). This would seamlessly allow for deeper experiments with a stochastic deformation framework such as presented in [8].

The stochastic model developed in section 5.2 also requires a rigorous treatment. As mentioned the inverse problem introduced here should be interpreted through the appropriate theoretical lens of e.g. [211]. Instead of a simple parameterisation like (5.17) treating the control as a function could also open the possibility for discretisation of such a function via a truncated Fourier series to produce efficient numerical methods. In the same vein we could also investigate using a Metropolis-adjusted Langevin algorithm [195] instead of the pCN. This essentially includes gradient information of the log-likelihood into a stochastic differential equation exploring the state space (which here would be a suitable function space for  $\nu$ ).

Finally, it is our hope that we can extend the probabilistic approach developed here to encompass classic metamorphosis as well; that is to say, to develop the necessary theory in order to place a stochastic model on the state space consisting of velocities and source functions and sample from function space. This provides a derivative-free method of solving classic metamorphosis (or other problems in shape analysis) at the expense of interpreting the results probabilistically.

# CHAPTER 6

---

## CLOSURE

---

In this thesis we studied the equations of motion associated with shape matching problems using curve, image and landmark data. Our contributions pertain mainly to theoretical or modelling aspects of these equations, in particular the study of convergence of finite element methods. This final chapter summarises our work and highlights several open problems subject to future research.

### 6.1 SUMMARY

Chapter 2 contained a comprehensive introduction to diffeomorphometry and how distances on nonlinear shape can be computed via integration of the norm of tangent linearisations of a group action. These linearisations are vector fields, and we outlined the main conditions for well-posedness of the geodesics associated with the distance. Shape matching problems were also shown to be posed as inverse problems with the forward motion (in time) of the shape geodesics acting as an observation operator. We compiled several key results from abstract variational and functional analysis and showed how some of these can be leveraged in the finite-dimensional setting of finite element methods.

Chapter 3 contributed with a novel discretisation of the geodesic equations for curve matching. We proved that a nonconforming  $H^3$  discretisation was sufficient to prove a Lipschitz property for the existence of a diffeomorphism of  $S^1$ . We carefully show that under mild and, importantly, *numerically verifiable* conditions we are able to integrate forward the discrete geodesic equations. We also showed that these converge to their continuous counterparts under mesh refinement. We investigated the convergence rate using the nonconforming finite element space with a pathological a priori  $H^{-1}$  source term in the form of the  $S^1$  integral of the momentum variable times a test function evaluated on the curve. The limitations to a linear convergence result were traced to a single term involving the jump of the third derivative of the continuous velocity variable  $\hat{w}_t$ . We provided several numerical examples of the resulting forward operator easily implemented with most standard finite element software.

Next, chapter 4 contained a new space-time formulation for potential parallel-in-time metamorphosis of images. We show that it is natural to cast the equations of motion in a classical Hilbertian *least-square finite element method* framework. Several key theoretical results were provided in this context such as trace, Poincaré and well-posedness theorems. Convergence of a conforming discretisation to the least-squared system was evidenced and we demonstrated that a gradient descent method can provide solutions to the space-time metamorphosis problem. We also laid the theoretical foundation for a mixed DPG method and highlighted the challenges that still remain for a stability proof of a discrete system.

Finally chapter 5 explored a formal interpolation between LDDMM and metamorphosis primarily using landmark data. Some theoretical results were obtained here and an extension to image data was sketched. We presented an MCMC method that explores the support of the data as a way of characterising the most areas in which metamorphosis is most likely to occur.

## 6.2 OPEN PROBLEMS

We suggest several questions to be addressed by future research.

The methods used in chapters 3, 4 and 5 make use of Lipschitz velocities to generate diffeomorphisms. This common theme therefore deserves a comment for completeness. At first glance this motivates research into direct discretisation of the Banach space  $W^{1,\infty}(\Omega)$ . There is, to the best of the author's ability, no obvious way of making sense of a discrete analogue of this space or how to compute its norm. As a consequence, we (and the literature in general) rely on the Sobolev embedding theorem to select Hilbert spaces  $V$  with the sought-after Lipschitz properties i.e.  $V \hookrightarrow W^{1,\infty}(\Omega)$ . In most cases the inclusion is strict, so morally we are not modelling the full space of Lipschitz velocities that generate diffeomorphism. In other words we are "leaving out" certain diffeomorphisms that could adequately be applied in shape matching, so it may make sense to develop methods that allow us to represent the space  $W^{1,\infty}(\Omega)$  without having to rely on embeddings. As mentioned in section 3.2 incorporating [157] into the variational setting here may present a path forward.

Our work on curve matching offers several paths for future research. The  $H^3$  nonconforming finite element space  $\tilde{V}_{th}^d$  that we used guaranteed a key Lipschitz property of its elements while possessing enough inter-elemental regularity in the form of a weak continuity property. This made it possible to discretise the 6<sup>th</sup> order operator generating the velocity field. Since the system we analyse requires the velocity  $u_t$  *as well as its gradient* to be Lipschitz, we resorted to a 12<sup>th</sup> order system. An improvement would be to truly find a  $C^2$  conforming finite element space which would correctly provide the Lipschitz property of both  $u_t$  and  $\nabla u_t$ . While it is numerically feasible to invert the 12<sup>th</sup> order

operator with the element from [231], several proofs will simplify and the communication of the discretisation will be much clearer. A conforming finite element space also provides an orthogonality estimate which means the error analysis conducted in corollary 54 greatly simplifies as well, see e.g. [50] for a  $H^1$  example. Naturally, we also seek to improve the error estimate provided by corollary 54, although we conjecture that the loss of regularity across the curve is too prohibitive. In [50], a PDE is *explicitly* solved on the curve to recover some regularity in this area and a similar approach could be explored for curves given a suitable conforming finite element space. Aside from the theoretical details, the next natural step for our research is to solve the inverse problem for the momentum and use our forward operator to solve curve matching problems. As we alluded to, a Bayesian method similar in spirit to [65] can be applied so the goal is to find a distribution of momentum that, on average, maps the specified template to a target shape. Another interesting path to investigate is a more local method using gradient information. Automatic differentiation software (cf. section 4.2.1) already exists in the Firedrake framework that we used to generate the curve evolution in figure 3.3, so if we can provide *loss function* on the manifold of discrete curves this would constitute an inverse problem solver. From a practical point of view, the target will be specified as a finite-dimensional set of detector points and a loss given as a Euclidean distance. There is currently no support for automatic code (adjoint) generation for the derivative of an arbitrary functional with respect to point evaluation on the curve, but this is apparently in the Firedrake pipeline.

Parameterisation-independent matching can also be explored. Since the curves considered here are elements of  $\text{Emb}(S^1, \mathbb{R}^d)$ , the shapes are invariant to diffeomorphisms of  $S^1$ . We could therefore look at choosing  $\text{Emb}(S^1, \mathbb{R}^d)/\text{Diff}(S^1)$  as shape space as in [64, 63] and look at the consequences for this choice both theoretically and numerically. Finally, we wish to extend our work to surface matching problems.

While the conforming space-time methods of chapter 4 showed excellent promise, some challenges remain for the discontinuous methods in section 4.1.3. Roadmap 79 showed that a regularity theorem is shown to be sufficient for surjectivity of the operator  $\mathcal{B}$  to obviate assumption 78. We also recall that the design of optimal test functions or the existence of a Fortin operator would instantly verify the conditions of theorem 85 and show that a inf-sup condition for the discrete system is verified. Such a proof would likely be constructive and therefore provide the local polynomial orders necessary for numerical experiments. We also wish to explore the nascent analytical tools developed in [224] in the space-time setting to improve the gradient descent algorithm and our numerical results. The bilinear form in problem 4 is symmetric positive definite on  $\mathring{G}(\mathbf{b}; \Omega)$  so a discrete solution to this problem can be sought via a Krylov-type method such as the conjugate gradient (CG) method [198]. As conjectured the conditioning of this system depends heavily on the nature of

the field  $\mathbf{b}$ . It may therefore be natural to consider a preconditioning method taking this into account in a CG loop.

In chapter 5 we discussed the combination of the KeOps library with the PyTorch autodiff engine which represents a huge step forward in making (selective) metamorphosis available to users. With the seamless GPU acceleration advertised by KeOps, high resolution image matching problems could be solved by users such as domain experts with very little exposure to high performance computing. An exercise in software engineering is therefore to build a suitable abstraction of these libraries to facilitate this functionality.

Finally we mention a more philosophical question on metamorphosis that arose as a consequence of the exploratory modelling in chapter 5. Many resources have been devoted to constructing smooth diffeomorphisms for shape matching. In the context of curves such a group is extremely valuable as it preserves the topology of the shape. Metamorphosis is also a diffeomorphometric framework although the presence of the  $z$  term in e.g. the equation  $\mathbf{b} \cdot \nabla_t I = z$  makes the transformation between shapes *homeomorphic*. One may therefore ask whether it makes sense to model a smooth group of diffeomorphisms if the model seeks to violate its smoothness by the additive term  $z$ . One can answer in the affirmative if the metamorphosis truly subsumes the geodesics of LDDMM. What we mean by this is that if the metamorphosis model is able to *fully* diffeomorphic i.e.  $z = 0$ , then it is in fact a *generalisation* of LDDMM. However we see that this is false upon inspection of e.g. the geodesic equations for  $z = 0$ : for selective metamorphosis we recall (5.12) read:

$$z_t^i = \nu(q_t^i)p_t^i,$$

for each landmark  $i = 1, \dots, M$ . As such,  $p_t^i$  is forced to zero whenever  $\nu$  is some positive constant. A zero momentum leads to a zero velocity *so the shape cannot move*. To summarise: in the classic metamorphosis framework the only purely diffeomorphic geodesic is the trivial case when the shape does not move i.e.  $\varphi_t = \text{id}$ ,  $t \in [0, 1]$  and the target always equals the template. It therefore seems much more natural and mathematically elegant to seek homeomorphic geodesics and *then* seek to recover diffeomorphic regularity *a posteriori*. We conclude by mentioning that the field of optimal transport (OTT) [227] may provide at least a partial answer. Indeed, several papers address this exact problem using so-called *unbalanced optimal transport* where the template and target are measure-valued objects with possibly different  $L^1$  norms [57, 164, 161, 25, 93]. See also the PhD thesis of Lénaïc Chizat [56]. We believe that the strong theoretical foundation of OTT as well as the numerical evidence make this framework the primary candidate in taking forward the metamorphosis model in favour of the PDE-constrained setting introduced in chapter 4.

## BIBLIOGRAPHY

- [1] R. A. Adams. “Sobolev Spaces, volume 65 of”. In: *Pure and applied mathematics* (1975).
- [2] G. Allaire and S. M. Kaber. *Numerical linear algebra*. Vol. 55. Springer, 2008.
- [3] S. Allasonnière, Y. Amit, and A. Trouvé. “Towards a coherent statistical framework for dense deformable template estimation”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.1 (2007), pp. 3–29.
- [4] S. Allasonnière, E. Kuhn, and A. Trouvé. “Map estimation of statistical deformable templates via nonlinear mixed effects models: Deterministic and stochastic approaches”. In: *2nd MICCAI Workshop on Mathematical Foundations of Computational Anatomy*. 2008, pp. 80–91.
- [5] M. Anderson and J. H. Kimn. “A numerical approach to space-time finite elements for the wave equation”. In: *Journal of Computational Physics* 226.1 (2007), pp. 466–476.
- [6] A. Arnaudon, D. D. Holm, and S. Sommer. “A geometric framework for stochastic shape analysis”. In: *Foundations of Computational Mathematics* 19.3 (2019), pp. 653–701.
- [7] A. Arnaudon, D. Holm, and S. Sommer. “String methods for stochastic image and shape matching”. In: *Journal of Mathematical Imaging and Vision* 60.6 (2018), pp. 953–967.
- [8] A. Arnaudon et al. “A stochastic large deformation model for computational anatomy”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2017, pp. 571–582.
- [9] D. N. Arnold, F. Brezzi, and M. Fortin. “A stable finite element for the Stokes equations”. In: *Calcolo* 21.4 (1984), pp. 337–344.
- [10] D. N. Arnold et al. “Unified analysis of discontinuous Galerkin methods for elliptic problems”. In: *SIAM journal on numerical analysis* 39.5 (2002), pp. 1749–1779.
- [11] V. I. Arnold. “Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l’hydrodynamique des fluides parfaits”. In: *Ann. Inst. Fourier* 16.1 (1966), pp. 319–361.

- [12] W. Arthur. “D’Arcy Thompson and the theory of transformations”. In: *Nature Reviews Genetics* 7.5 (2006), p. 401.
- [13] C. Athanasiadis and I. G. Stratis. “On some elliptic transmission problems”. In: *Annales Polonici Mathematici*. Vol. 63. 2. 1996, pp. 137–154.
- [14] I. Babuška, U. Banerjee, and J. E. Osborn. “On principles for the selection of shape functions for the generalized finite element method”. In: *Computer methods in applied mechanics and engineering* 191.49-50 (2002), pp. 5595–5629.
- [15] R. Bajcsy and S. Kovačič. “Multiresolution elastic matching”. In: *Computer vision, graphics, and image processing* 46.1 (1989), pp. 1–21.
- [16] R. Bajcsy, R. Lieberman, and M. Reivich. *A computerized system for the elastic matching of deformed radiographic images to idealized atlas images*. 1983.
- [17] G. A. Baker. “Finite element methods for elliptic equations using non-conforming elements”. In: *Mathematics of Computation* 31.137 (1977), pp. 45–59.
- [18] S. Balay et al. “Efficient Management of Parallelism in Object Oriented Numerical Software Libraries”. In: *Modern Software Tools in Scientific Computing*. Ed. by E. Arge, A. M. Bruaset, and H. P. Langtangen. Birkhäuser Press, 1997, pp. 163–202.
- [19] S. Balay et al. *PETSc Users Manual*. Tech. rep. ANL-95/11 - Revision 3.11. Argonne National Laboratory, 2019.
- [20] J. W. Barrett and C. M. Elliott. “A finite-element method for solving elliptic equations with Neumann data on a curved boundary using unfitted meshes”. In: *IMA Journal of Numerical Analysis* 4.3 (1984), pp. 309–325.
- [21] E. Barron, L. Evans, and R. Jensen. “The infinity Laplacian, Aronsson’s equation and their generalizations”. In: *Transactions of the American Mathematical Society* 360.1 (2008), pp. 77–101.
- [22] F. Bassi and S. Rebay. “A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations”. In: *Journal of computational physics* 131.2 (1997), pp. 267–279.
- [23] M. Bauer and M. Bruveris. “A new Riemannian setting for surface registration”. In: *arXiv preprint arXiv:1106.0620* (2011).
- [24] M. Bauer, M. Bruveris, and P. W. Michor. “Overview of the geometries of shape spaces and diffeomorphism groups”. In: *Journal of Mathematical Imaging and Vision* 50.1-2 (2014), pp. 60–97.

- [25] M. Bauer, S. Joshi, and K. Modin. “Diffeomorphic density matching by optimal information transport”. In: *SIAM Journal on Imaging Sciences* 8.3 (2015), pp. 1718–1751.
- [26] M. Bauer et al. “A numerical framework for Sobolev metrics on the space of curves”. In: *SIAM Journal on Imaging Sciences* 10.1 (2017), pp. 47–73.
- [27] M. Bauer et al. “Curve matching with applications in medical imaging”. In: *arXiv preprint arXiv:1506.08840* (2015).
- [28] M. Bauer et al. “Varifold-based matching of curves via Sobolev-type Riemannian metrics”. In: *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*. Springer, 2017, pp. 152–163.
- [29] G. Bazeley. “Triangular elements in bending-conforming and non-conforming solutions”. In: *Proc. Conf. Matrix Methods in Struct. Mech., Air Force Inst. of Tech., Wright Patterson AF Base, Ohio*. 1965.
- [30] M. F. Beg et al. “Computing large deformation metric mappings via geodesic flows of diffeomorphisms”. In: *International Journal of Computer Vision* 61.2 (2005), pp. 139–157.
- [31] M. F. Beg and A. Khan. “Computing an average anatomical atlas using LDDMM and geodesic shooting”. In: *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006*. IEEE. 2006, pp. 1116–1119.
- [32] T. Beier and S. Neely. “Feature-based image metamorphosis”. In: *ACM SIGGRAPH computer graphics* 26.2 (1992), pp. 35–42.
- [33] T. Belytschko, R. Gracie, and G. Ventura. “A review of extended/generalized finite element methods for material modeling”. In: *Modelling and Simulation in Materials Science and Engineering* 17.4 (2009), p. 043001.
- [34] G. T. Bercea et al. “A structure-exploiting numbering algorithm for finite elements on extruded meshes, and its performance evaluation in Firedrake”. In: *arXiv preprint arXiv:1604.05937* (2016).
- [35] P. B. Bochev and M. D. Gunzburger. *Least-squares finite element methods*. Vol. 166. Springer Science & Business Media, 2009.
- [36] P. Bochev and J. Choi. “Improved least-squares error estimates for scalar hyperbolic problems”. In: *Computational Methods in Applied Mathematics Comput. Methods Appl. Math.* 1.2 (2001), pp. 115–124.
- [37] A. Bock, A. Arnaudon, and C. Cotter. “Selective metamorphosis for growth modelling with applications to landmarks”. In: *International Conference on Geometric Science of Information*. Springer. 2019, pp. 39–48.

- [38] D. Boffi, F. Brezzi, and M. Fortin. “Finite elements for the Stokes problem”. In: *Mixed finite elements, compatibility conditions, and applications*. Springer, 2008, pp. 45–100.
- [39] D. Boffi, F. Brezzi, M. Fortin, et al. *Mixed finite element methods and applications*. Vol. 44. Springer, 2013.
- [40] S. K. Boyd. “Image-based finite element analysis”. In: *Advanced Imaging in Biology and Medicine*. Springer, 2009, pp. 301–318.
- [41] S. C. Brenner. “A two-level additive Schwarz preconditioner for non-conforming plate elements”. In: *Numerische Mathematik* 72.4 (1996), pp. 419–447.
- [42] S. Brenner and R. Scott. *The Mathematical Theory of Finite Element Methods*. Vol. 15. Springer Science & Business Media, 2007.
- [43] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.
- [44] F. Brezzi, L. D. Marini, and E. Süli. “Discontinuous Galerkin methods for first-order hyperbolic problems”. In: *Mathematical models and methods in applied sciences* 14.12 (2004), pp. 1893–1903.
- [45] F. Brezzi et al. “Discontinuous finite elements for diffusion problems”. In: *Atti Convegno in onore di F. Brioschi (Milano 1997), Istituto Lombardo, Accademia di Scienze e Lettere* (1999), pp. 197–217.
- [46] D. Broersen, W. Dahmen, and R. Stevenson. “On the stability of DPG formulations of transport equations”. In: *Mathematics of Computation* 87.311 (2018), pp. 1051–1082.
- [47] A. N. Brooks and T. J. Hughes. “Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations”. In: *Computer Methods in Applied Mechanics and Engineering* 32.1 (1982), pp. 199–259.
- [48] M. Bruveris. “Riemannian geometry for shape analysis and computational anatomy”. In: *arXiv preprint arXiv:1807.11290* (2018).
- [49] A. Buffa and P. Monk. “Error estimates for the ultra weak variational formulation of the Helmholtz equation”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 42.6 (2008), pp. 925–940.
- [50] E. Burman, P. Hansbo, and M. G. Larson. “A simple finite element method for elliptic bulk problems with embedded surfaces”. In: *Computational Geosciences* 23.1 (2019), pp. 189–199.
- [51] E. Burman, A. Quarteroni, and B. Stamm. “Interior penalty continuous and discontinuous finite element approximations of hyperbolic equations”. In: *Journal of Scientific Computing* 43.3 (2010), pp. 293–312.

- [52] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan. “Breaking spaces and forms for the DPG method and applications including Maxwell equations”. In: *Computers & Mathematics with Applications* 72.3 (2016), pp. 494–522.
- [53] O. Cessenat and B. Després. “Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problem”. In: *SIAM journal on numerical analysis* 35.1 (1998), pp. 255–299.
- [54] B. Charlier, J. Feydy, and J. Glaunes. “KeOps: Calcul rapide sur GPU dans les espaces à noyaux”. In: *Proceedings of Journées de Statistique de la SFdS. Paris, France* (2018).
- [55] N. Charon and A. Trounev. “The varifold representation of nonoriented shapes for diffeomorphic registration”. In: *SIAM Journal on Imaging Sciences* 6.4 (2013), pp. 2547–2580.
- [56] L. Chizat. “Unbalanced optimal transport: models, numerical methods, applications”. PhD thesis. 2017.
- [57] L. Chizat et al. “An interpolating distance between optimal transport and Fisher–Rao metrics”. In: *Foundations of Computational Mathematics* 18.1 (2018), pp. 1–44.
- [58] A. J. Chorin, J. E. Marsden, and J. E. Marsden. *A mathematical introduction to fluid mechanics*. Vol. 3. Springer, 1990.
- [59] G. E. Christensen, R. D. Rabbitt, and M. I. Miller. “Deformable templates using large deformation kinematics”. In: *IEEE transactions on image processing* 5.10 (1996), pp. 1435–1447.
- [60] P. G. Ciarlet. *The finite element method for elliptic problems*. Vol. 40. Siam, 2002.
- [61] C. Clason. “Finite Element Methods”. In: *arXiv preprint arXiv:1709.08618* (2017).
- [62] P. Clément. “Approximation by finite element functions using local regularization”. In: *Revue française d’automatique, informatique, recherche opérationnelle. Analyse numérique* 9.R2 (1975), pp. 77–84.
- [63] C. J. Cotter, A. Clark, and J. Peiró. “A reparameterisation based approach to geodesic constrained solvers for curve matching”. In: *International Journal of Computer Vision* 99.1 (2012), pp. 103–121.
- [64] C. J. Cotter and D. D. Holm. “Geodesic boundary value problems with symmetry”. In: *arXiv preprint arXiv:0911.2205* (2009).
- [65] C. J. Cotter, S. L. Cotter, and F. X. Vialard. “Bayesian data assimilation in shape registration”. In: *Inverse Problems* 29.4 (2013), p. 45011.
- [66] S. L. Cotter, M. Dashti, and A. M. Stuart. “Approximation of Bayesian inverse problems for PDEs”. In: *SIAM Journal on Numerical Analysis* 48.1 (2010), pp. 322–345.

- [67] S. L. Cotter et al. “MCMC methods for functions: modifying old algorithms to make them faster”. In: *Statistical Science* (2013), pp. 424–446.
- [68] W. Dahmen et al. “Adaptive anisotropic Petrov–Galerkin methods for first order transport equations”. In: *Journal of Computational and Applied Mathematics* 340 (2018), pp. 191–220.
- [69] W. Dahmen et al. “Adaptive Petrov–Galerkin methods for first order transport equations”. In: *SIAM journal on numerical analysis* 50.5 (2012), pp. 2420–2445.
- [70] L. D. Dalcin et al. “Parallel distributed computing using Python”. In: *Advances in Water Resources* 34.9 (2011). New Computational Methods and Software Tools, pp. 1124–1139. DOI: <http://dx.doi.org/10.1016/j.advwatres.2011.04.013>.
- [71] M. Dashti and A. M. Stuart. “The Bayesian approach to inverse problems”. In: *Handbook of Uncertainty Quantification* (2017), pp. 311–428.
- [72] M. Dashti et al. “MAP estimators and their consistency in Bayesian nonparametric inverse problems”. In: *Inverse Problems* 29.9 (2013), p. 95017.
- [73] H. De Sterck et al. “Least-squares finite element methods and algebraic multigrid solvers for linear hyperbolic PDEs”. In: *SIAM Journal on Scientific Computing* 26.1 (2004), pp. 31–54.
- [74] A. Dedner, C. Makridakis, and M. Ohlberger. “Error control for a class of Runge–Kutta discontinuous Galerkin methods for nonlinear conservation laws”. In: *SIAM journal on numerical analysis* 45.2 (2007), pp. 514–538.
- [75] L. Demkowicz. “Babuška  $\Leftrightarrow$  Brezzi”. In: *ICES Report* (2006), pp. 06–08.
- [76] L. F. Demkowicz and J. Gopalakrishnan. “An overview of the discontinuous Petrov Galerkin method”. In: *Recent developments in discontinuous Galerkin finite element methods for partial differential equations*. Springer, 2014, pp. 149–180.
- [77] L. Demkowicz and J. Gopalakrishnan. “A class of discontinuous Petrov–Galerkin methods. II. Optimal test functions”. In: *Numerical Methods for Partial Differential Equations* 27.1 (2011), pp. 70–105.
- [78] L. Demkowicz and J. Gopalakrishnan. “Discontinuous Petrov–Galerkin (DPG) Method”. In: *Encyclopedia of Computational Mechanics Second Edition* (2017), pp. 1–15.
- [79] L. Demkowicz, J. Gopalakrishnan, and A. H. Niemi. “A class of discontinuous Petrov–Galerkin methods. Part III: Adaptivity”. In: *Applied numerical mathematics* 62.4 (2012), pp. 396–427.

- [80] L. Demkowicz and J. Gopalakrishnan. “A class of discontinuous Petrov–Galerkin methods. Part I: The transport equation”. In: *Computer Methods in Applied Mechanics and Engineering* 199.23–24 (2010), pp. 1558–1572.
- [81] L. Demkowicz and J. Gopalakrishnan. “Analysis of the DPG method for the Poisson equation”. In: *SIAM Journal on Numerical Analysis* 49.5 (2011), pp. 1788–1809.
- [82] Y. Dong. “Prescribing the Jacobian determinant in Sobolev spaces”. In: *Annales de l’Institut Henri Poincaré (C) Non Linear Analysis*. Vol. 11. 3. Elsevier. 1994, pp. 275–296.
- [83] J. Douglas and T. Dupont. “Interior penalty procedures for elliptic and parabolic Galerkin methods”. In: *Computing methods in applied sciences*. Springer, 1976, pp. 207–216.
- [84] J. Du, L. Younes, and A. Qiu. “Whole brain diffeomorphic metric mapping via integration of sulcal and gyral curves, cortical surfaces, and images”. In: *NeuroImage* 56.1 (2011), pp. 162–173.
- [85] T. Dupont and R. Scott. “Polynomial approximation of functions in Sobolev spaces”. In: *Mathematics of Computation* 34.150 (1980), pp. 441–463.
- [86] P. Dupuis, U. Grenander, and M. I. Miller. “Variational problems on flows of diffeomorphisms for image matching”. In: *Quarterly of Applied Mathematics* (1998), pp. 587–600.
- [87] A. Effland, M. Rumpf, and F. Schäfer. “Image Extrapolation for the Time Discrete Metamorphosis Model - Existence and Applications”. In: *arXiv preprint arXiv:1705.04490* (2017).
- [88] A. Ern and J. L. Guermond. “Evaluation of the condition number in linear systems arising in finite element approximations”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 40.1 (2006), pp. 29–48.
- [89] A. Ern and J. L. Guermond. *Theory and practice of finite elements*. Vol. 159. Springer Science & Business Media, 2013.
- [90] L. C. Evans. *Partial differential equations*. Vol. 19. American Mathematical Soc., 2010.
- [91] R. S. Falk. “A Fortin operator for two-dimensional Taylor-Hood elements”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 42.3 (2008), pp. 411–424.
- [92] J. Feydy et al. “Interpolating between optimal transport and MMD using Sinkhorn divergences”. In: *arXiv preprint arXiv:1810.08278* (2018).
- [93] J. Feydy et al. “Optimal transport for diffeomorphic registration”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 291–299.

- [94] B. Fischer and J. Modersitzki. “Ill-posed medicine - an introduction to image registration”. In: *Inverse Problems* 24.3 (2008), p. 34008.
- [95] P. T. Fletcher et al. “Principal geodesic analysis for the study of nonlinear statistics of shape”. In: *IEEE transactions on medical imaging* 23.8 (2004), pp. 995–1005.
- [96] G. B. Folland. *Introduction to partial differential equations*. Princeton university press, 1995.
- [97] M. Fortin. “An analysis of the convergence of mixed finite element methods”. In: *RAIRO. Analyse numérique* 11.4 (1977), pp. 341–354.
- [98] T. Führer. “Superconvergence in a DPG method for an ultra-weak formulation”. In: *Computers & Mathematics with Applications* 75.5 (2018), pp. 1705–1718.
- [99] S. W. Funke and P. E. Farrell. “A framework for automated PDE-constrained optimisation”. In: *arXiv preprint arXiv:1302.3894* (2013).
- [100] D. Gallistl. “Stable splitting of polyharmonic operators by generalized Stokes systems”. In: *Mathematics of Computation* 86.308 (2017), pp. 2555–2577.
- [101] G. N. Gatica. “A simple introduction to the mixed finite element method”. In: *Theory and Applications. Springer Briefs in Mathematics. Springer, London* (2014).
- [102] F. Gay-Balmaz, D. D. Holm, and T. S. Ratiu. “Geometric dynamics of optimization”. In: *arXiv preprint arXiv:0912.2989* (2009).
- [103] J. C. Gee. “On matching brain volumes”. In: *Pattern Recognition* 32.1 (1999), pp. 99–111.
- [104] E. H. Georgoulis and T. Pryer. “Analysis of discontinuous Galerkin methods using mesh-dependent norms and applications to problems with rough data”. In: *Calcolo* 54.4 (2017), pp. 1533–1551.
- [105] C. Geuzaine and J. F. Remacle. “Gmsh: A 3-D finite element mesh generator with built-in pre-and post-processing facilities”. In: *International journal for numerical methods in engineering* 79.11 (2009), pp. 1309–1331.
- [106] T. H. Gibson et al. “Slate: extending Firedrake’s domain-specific abstraction to hybridized solvers for geoscience and beyond.” In: *Geoscientific model development*. 13.2 (2020), pp. 735–761.
- [107] J. Giesselmann, C. Makridakis, and T. Pryer. “A posteriori analysis of discontinuous Galerkin schemes for systems of hyperbolic conservation laws”. In: *SIAM Journal on Numerical Analysis* 53.3 (2015), pp. 1280–1303.
- [108] J. Glaunès et al. “Large deformation diffeomorphic metric curve mapping”. In: *International journal of computer vision* 80.3 (2008), p. 317.

- [109] J. Gopalakrishnan and W. Qiu. “An analysis of the practical DPG method”. In: *Mathematics of Computation* 83.286 (2014), pp. 537–552.
- [110] U. Grenander. *General Pattern Theory*. 1993.
- [111] U. Grenander and M. I. Miller. “Computational anatomy: An emerging discipline”. In: *Quarterly of applied mathematics* 56.4 (1998), pp. 617–694.
- [112] U. Grenander and M. I. Miller. *Pattern theory: from representation to inference*. Oxford University Press, 2007.
- [113] U. Grenander and M. I. Miller. “Representations of knowledge in complex systems”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1994), pp. 549–603.
- [114] P. Grisvard. *Elliptic problems in nonsmooth domains*. SIAM, 2011.
- [115] P. Grisvard. *Singularities in boundary value problems*. Vol. 22. Springer, 1992.
- [116] T. Gudi. “A new error analysis for discontinuous finite element methods for linear elliptic problems”. In: *Mathematics of computation* 79.272 (2010), pp. 2169–2189.
- [117] T. Gudi and M. Neilan. “An interior penalty method for a sixth-order elliptic equation”. In: *IMA journal of numerical analysis* 31.4 (2011), pp. 1734–1753.
- [118] A. Günther, H. Lamecker, and M. Weiser. “Direct LDDMM of discrete currents with adaptive finite elements”. In: 2011.
- [119] H. Guo, A. Rangarajan, and S. Joshi. “Diffeomorphic point matching”. In: *Handbook of Mathematical Models in Computer Vision*. Springer, 2006, pp. 205–219.
- [120] M. Hairer, A. M. Stuart, S. J. Vollmer, et al. “Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions”. In: *The Annals of Applied Probability* 24.6 (2014), pp. 2455–2490.
- [121] W. K. Hastings. “Monte Carlo sampling methods using Markov chains and their applications”. In: (1970).
- [122] J. Heinonen. *Lectures on Lipschitz analysis*. 100. University of Jyväskylä, 2005.
- [123] M. Hernandez and S. Olmos. “Gauss-Newton optimization in diffeomorphic registration”. In: *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2008, pp. 1083–1086.
- [124] I. Herrera. “Trefftz method: a general theory”. In: *Numerical Methods for Partial Differential Equations: An International Journal* 16.6 (2000), pp. 561–580.
- [125] M. Hinze et al. *Optimization with PDE constraints*. Vol. 23. Springer Science & Business Media, 2008.

- [126] R. Hiptmair, A. Moiola, and I. Perugia. “Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the p-version”. In: *SIAM Journal on Numerical Analysis* 49.1 (2011), pp. 264–284.
- [127] D. D. Holm. *Geometric mechanics*. Imperial College Press, 2008.
- [128] D. D. Holm, J. E. Marsden, and T. S. Ratiu. “The Euler–Poincaré equations and semidirect products with applications to continuum theories”. In: *Advances in Mathematics* 137.1 (1998), pp. 1–81.
- [129] D. D. Holm, A. Trounev, and L. Younes. “The Euler-Poincaré theory of metamorphosis”. In: *Quarterly of Applied Mathematics* (2009), pp. 661–685.
- [130] M. Homolya and D. A. Ham. “A parallel edge orientation algorithm for quadrilateral meshes”. In: *SIAM Journal on Scientific Computing* 38.5 (2016), S48–S61.
- [131] Z. Hong-qing and W. Ming. “On the compactness of quasi-conforming element spaces and the convergence of quasi-conforming element method”. In: *Applied Mathematics and Mechanics* 7.5 (1986), pp. 443–459.
- [132] P. Houston, R. Rannacher, and E. Süli. “A posteriori error analysis for stabilised finite element approximations of transport problems”. In: *Computer methods in applied mechanics and engineering* 190.11-12 (2000), pp. 1483–1508.
- [133] P. Houston et al. “A posteriori error analysis for numerical approximations of Friedrichs systems”. In: *Numerische Mathematik* 82.3 (1999), pp. 433–470.
- [134] J. Hu, R. Ma, and Z. Shi. “A new a priori error estimate of nonconforming finite element methods”. In: *Science China Mathematics* 57.5 (2014), pp. 887–902.
- [135] J. Hu and S. Zhang. “A canonical construction of  $H^m$ -nonconforming triangular finite elements”. In: *Ann. Appl. Math* 33.33 (2017), pp. 266–288.
- [136] T. J. Hughes and A. Brooks. “A multi-dimensioal upwind scheme with no crosswind diffusion.” In: *FINITE ELEMENT METHODS FOR CONVECTION DOMINATED FLOWS, ASME WINTER ANNUAL MEETING, TJR HUGHES (ED.), NEW YORK, USA*, vol. 34. 2017, pp. 19–35.
- [137] T. J. Hughes and G. M. Hulbert. “Space-time finite element methods for elastodynamics: formulations and error estimates”. In: *Computer methods in applied mechanics and engineering* 66.3 (1988), pp. 339–363.
- [138] G. M. Hulbert and T. J. Hughes. “Space-time finite element methods for second-order hyperbolic equations”. In: (1990).

- [139] R. C. James. “Reflexivity and the supremum of linear functionals”. In: *Annals of Mathematics* (1957), pp. 159–169.
- [140] C. Johnson and J. Pitkäranta. “An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation”. In: *Mathematics of computation* 46.173 (1986), pp. 1–26.
- [141] S. C. Joshi and M. I. Miller. “Landmark matching via large deformation diffeomorphisms”. In: *IEEE transactions on image processing* 9.8 (2000), pp. 1357–1370.
- [142] S. H. Joshi et al. “A novel representation for Riemannian analysis of elastic curves in  $\mathbb{R}^n$ ”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2007, pp. 1–7.
- [143] M. Juntunen and R. Stenberg. “Nitsche’s method for general boundary conditions”. In: *Mathematics of computation* 78.267 (2009), pp. 1353–1374.
- [144] I. Kaltenmark. “Geometrical Growth Models for Computational Anatomy”. PhD thesis. Université Paris-Saclay, 2016.
- [145] I. Kaltenmark and A. Trouvé. “Estimation of a Growth Development with Partial Diffeomorphic Mappings”. In: (2017).
- [146] I. Kaltenmark and A. Trouvé. “Existence and Continuity of Minimizers for the Estimation of Growth Mapped Evolutions for Current Data Term and Counterexamples for Varifold Data Term”. In: (2017).
- [147] I. Kaltenmark and A. Trouvé. “Partial Matchings and Growth Mapped Evolutions in Shape Spaces”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 28–36.
- [148] B. Khesin and R. Wendt. *The geometry of infinite-dimensional groups*. Vol. 51. Springer Science & Business Media, 2008.
- [149] R. C. Kirby. “A general approach to transforming finite elements”. In: *The SMAI journal of computational mathematics* 4 (2018), pp. 197–224.
- [150] R. C. Kirby and L. Mitchell. “Code Generation for Generally Mapped Finite Elements”. In: *ACM Transactions on Mathematical Software (TOMS)* 45.4 (2019), pp. 1–23.
- [151] E. Klassen et al. “Analysis of planar shapes using geodesic paths on shape spaces”. In: *IEEE transactions on pattern analysis and machine intelligence* 26.3 (2004), pp. 372–383.
- [152] A. Knauf. *Mathematical physics: classical mechanics*. Vol. 109. Springer, 2018.
- [153] V. A. Kondrat’ev. “Boundary value problems for elliptic equations in domains with conical or angular points”. In: *Trudy Moskovskogo Matematicheskogo Obshchestva* 16 (1967), pp. 209–292.

- [154] S. Krushkov. “First-order quasilinear equations in several independent variables”. In: *Math. Sb.* 123 (1970), pp. 228–255.
- [155] L. Kühnel, A. Arnaudon, and S. Sommer. “Differential geometry and stochastic dynamics with deep learning numerics”. In: *arXiv preprint arXiv:1712.08364* (2017).
- [156] L. Kühnel and S. Sommer. “Computational anatomy in Theano”. In: *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*. Springer, 2017, pp. 164–176.
- [157] O. Lakkis and T. Pryer. “An adaptive finite element method for the infinity Laplacian”. In: *Numerical Mathematics and Advanced Applications-ENUMATH 2013*. Springer, 2015, pp. 283–291.
- [158] O. Lakkis and T. Pryer. “Gradient recovery in adaptive finite-element methods for parabolic problems”. In: *IMA Journal of Numerical Analysis* 32.1 (2011), pp. 246–278.
- [159] U. ( Langer and O. S. (Ed.) *Space-Time Methods. Applications to Partial Differential Equations*. De Gruyter, 2019.
- [160] P. Lesaint and P. A. Raviart. “On a finite element method for solving the neutron transport equation”. In: *Publications mathématiques et informatique de Rennes S4* (1974), pp. 1–40.
- [161] D. Lombardi and E. Maitre. “Eulerian models and algorithms for unbalanced optimal transport”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 49.6 (2015), pp. 1717–1744.
- [162] J. Ma, M. I. Miller, and L. Younes. “A Bayesian generative model for surface template estimation”. In: *International journal of biomedical imaging* 2010 (2010).
- [163] J. Ma et al. “Bayesian template estimation in computational anatomy”. In: *NeuroImage* 42.1 (2008), pp. 252–261.
- [164] J. Maas et al. “A generalized model for optimal transport of images including dissipation and density modulation”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 49.6 (2015), pp. 1745–1769.
- [165] T. A. Manteuffel, K. J. Ressel, and G. Starke. “A boundary functional for the least-squares finite-element solution of neutron transport problems”. In: *SIAM Journal on Numerical Analysis* 37.2 (1999), pp. 556–586.
- [166] K. A. Mardal, J. Schöberl, and R. Winther. “A uniformly stable Fortin operator for the Taylor–Hood element”. In: *Numerische Mathematik* 123.3 (2013), pp. 537–551.
- [167] C. R. Maurer and J. M. Fitzpatrick. “A review of medical image registration”. In: *Interactive Image-guided Neurosurgery* 17 (1993).
- [168] A. T. McRae et al. “Automated generation and symbolic manipulation of tensor product finite elements”. In: *SIAM Journal on Scientific Computing* 38.5 (2016), S25–S47.

- [169] N. Metropolis et al. “Equation of state calculations by fast computing machines”. In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092.
- [170] P. W. Michor and D. Mumford. “An overview of the Riemannian metrics on spaces of curves using the Hamiltonian approach”. In: *Applied and Computational Harmonic Analysis* 23.1 (2007), pp. 74–113.
- [171] M. I. Miller, A. Trouné, and L. Younes. “Geodesic shooting for Computational Anatomy”. In: *Journal of Mathematical Imaging and Vision* 24.2 (2006), pp. 209–228.
- [172] M. I. Miller, A. Trouné, and L. Younes. “Hamiltonian systems and optimal control in computational anatomy: 100 years since D’Arcy Thompson”. In: *Annual review of biomedical engineering* 17 (2015), pp. 447–509.
- [173] M. I. Miller, A. Trouné, and L. Younes. “On the metrics and Euler-Lagrange equations of computational anatomy”. In: *Annual Review of Biomedical Engineering* 4.1 (2002), pp. 375–405.
- [174] M. I. Miller and L. Younes. “Group actions, homeomorphisms, and matching: A general framework”. In: *International Journal of Computer Vision* 41.1-2 (2001), pp. 61–84.
- [175] M. Miller et al. “Statistical methods in computational anatomy”. In: *Statistical methods in medical research* 6.3 (1997), pp. 267–299.
- [176] W. Mio, A. Srivastava, and S. Joshi. “On shape of plane elastic curves”. In: *International Journal of Computer Vision* 73.3 (2007), pp. 307–324.
- [177] S. Mitusch, S. Funke, and J. Dokken. “dolfin-adjoint 2018.1: automated adjoints for FEniCS and Firedrake”. In: *Journal of Open Source Software* 4.38 (2019), p. 1292.
- [178] A. Moiola and I. Perugia. “A space-time Trefftz discontinuous Galerkin method for the acoustic wave equation in first-order formulation”. In: *Numerische Mathematik* 138.2 (2018), pp. 389–435.
- [179] D. Mumford. “Pattern theory: the mathematics of perception”. In: *arXiv preprint math/0212400* (2002).
- [180] D. Mumford and P. W. Michor. “On Euler’s equation and ‘EPDiff’”. In: *arXiv preprint arXiv:1209.6576* (2012).
- [181] S. Nagaraj, S. Petrides, and L. F. Demkowicz. “Construction of DPG Fortin operators for second order problems”. In: *Computers & Mathematics with Applications* 74.8 (2017), pp. 1964–1980.
- [182] U. Naumann. *The art of differentiating computer programs: an introduction to algorithmic differentiation*. Vol. 24. Siam, 2012.
- [183] H. Nguyen and J. Reynen. “A space-time least-square finite element scheme for advection-diffusion equations”. In: *Computer Methods in Applied Mechanics and Engineering* 42.3 (Mar. 1984), pp. 331–342.

- [184] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
- [185] A. Oberman. “A convergent difference scheme for the infinity Laplacian: construction of absolutely minimizing Lipschitz extensions”. In: *Mathematics of computation* 74.251 (2005), pp. 1217–1230.
- [186] M. Ohlberger. “A review of a posteriori error control and adaptivity for approximations of non-linear conservation laws”. In: *International journal for numerical methods in fluids* 59.3 (2009), pp. 333–354.
- [187] A. Paszke et al. “Automatic differentiation in PyTorch”. In: (2017).
- [188] X. Pennec, S. Sommer, and T. Fletcher. *Riemannian Geometric Statistics in Medical Image Analysis*. Academic Press, 2019.
- [189] P. Perrochet and P. Azérad. “Space-time integrated least-squares: Solving a pure advection equation with a pure diffusion operator”. In: *Journal of computational physics* 117.2 (1995), pp. 183–193.
- [190] A. Qiu, L. Younes, and M. I. Miller. “Principal component based diffeomorphic surface mapping”. In: *IEEE transactions on medical imaging* 31.2 (2011), pp. 302–311.
- [191] F. Rathgeber et al. “Firedrake: automating the finite element method by composing abstractions”. In: *ACM Transactions on Mathematical Software (TOMS)* 43.3 (2016), pp. 1–27.
- [192] W. H. Reed and T. Hill. *Triangular mesh methods for the neutron transport equation*. Tech. rep. Los Alamos Scientific Lab., N. Mex.(USA), 1973.
- [193] S. Rhebergen and B. Cockburn. “Space-time hybridizable discontinuous galerkin method for the advection–diffusion equation on moving and deforming meshes”. In: *The Courant–Friedrichs–Lewy (CFL) Condition*. Springer, 2013, pp. 45–63.
- [194] C. L. Richardson and L. Younes. “Metamorphosis of images in reproducing kernel Hilbert spaces”. In: *Advances in Computational Mathematics* 42.3 (2016), pp. 573–603.
- [195] G. O. Roberts and O. Stramer. “Langevin diffusions and Metropolis-Hastings algorithms”. In: *Methodology and computing in applied probability* 4.4 (2002), pp. 337–357.
- [196] N. V. Roberts, T. Bui-Thanh, and L. Demkowicz. “The DPG method for the Stokes problem”. In: *Computers & Mathematics with Applications* 67.4 (2014), pp. 966–995.
- [197] D. Rueckert et al. “Diffeomorphic registration using B-splines”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2006, pp. 702–709.
- [198] Y. Saad. *Iterative methods for sparse linear systems*. Vol. 82. siam, 2003.

- [199] J. B. Schiratti et al. “A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations”. In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 4840–4872.
- [200] L. R. Scott and S. Zhang. “Finite element interpolation of nonsmooth functions satisfying boundary conditions”. In: *Mathematics of Computation* 54.190 (1990), pp. 483–493.
- [201] S. M. Selby. *Standard mathematical tables*. Tech. rep. 1973.
- [202] Z. C. Shi. “Nonconforming finite element methods”. In: *Journal of computational and applied mathematics* 149.1 (2002), pp. 221–225.
- [203] Z. C. Shi. “The FEM test for convergence of nonconforming finite elements”. In: *Mathematics of computation* 49.180 (1987), pp. 391–405.
- [204] W. Sickel and A. Youssfi. “The characterisation of the regularity of the Jacobian determinant in the framework of potential spaces”. In: *Journal of the London Mathematical Society* 59.1 (1999), pp. 287–310.
- [205] S. Sommer et al. “Bridge simulation and metric estimation on landmark manifolds”. In: *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*. Springer, 2017, pp. 79–91.
- [206] A. Srivastava et al. “Shape analysis of elastic curves in euclidean spaces”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.7 (2010), pp. 1415–1428.
- [207] E. M. Stein. *Singular integrals and differentiability properties of functions*. Vol. 2. Princeton university press, 1970.
- [208] G. Strang. “Variational crimes in the finite element method”. In: *The mathematical foundations of the finite element method with applications to partial differential equations*. Elsevier, 1972, pp. 689–710.
- [209] T. Strouboulis, K. Copps, and I. Babuška. “The generalized finite element method: an example of its implementation and illustration of its performance”. In: *International Journal for Numerical Methods in Engineering* 47.8 (2000), pp. 1401–1417.
- [210] T. Strouboulis, I. Babuška, and K. Copps. “The design and analysis of the generalized finite element method”. In: *Computer methods in applied mechanics and engineering* 181.1-3 (2000), pp. 43–69.
- [211] A. M. Stuart. “Inverse problems: a Bayesian perspective”. In: *Acta numerica* 19 (2010), pp. 451–559.
- [212] F. Stummel. “The limitations of the patch test”. In: *International journal for numerical methods in engineering* 15.2 (1980), pp. 177–188.
- [213] F. Stummel. “The generalized patch test”. In: *SIAM Journal on Numerical Analysis* 16.3 (1979), pp. 449–471.
- [214] E. Süli. “A brief excursion into the mathematical theory of mixed finite element methods”. In: *Lecture Notes, University of Oxford* (2013), pp. 24–29.

- [215] T. T. D. Team et al. “Theano: A Python framework for fast computation of mathematical expressions”. In: *arXiv preprint arXiv:1605.02688* (2016).
- [216] D. W. Thompson et al. “On growth and form.” In: *On growth and form*. (1942).
- [217] L. N. Trefethen and D. Bau III. *Numerical linear algebra*. Vol. 50. Siam, 1997.
- [218] A. Trouvé. “An infinite dimensional group approach for physics based models in pattern recognition”. In: *Preprint* (1995).
- [219] A. Trouvé. “Diffeomorphisms groups and pattern matching in image analysis”. In: *International Journal of Computer Vision* 28.3 (1998), pp. 213–221.
- [220] A. Trouvé and L. Younes. “Local geometry of deformable templates”. In: *SIAM Journal on Mathematical Analysis* 37.1 (2005), pp. 17–59.
- [221] A. Trouvé and L. Younes. “Metamorphoses through Lie group action”. In: *Foundations of Computational Mathematics* 5.2 (2005), pp. 173–198.
- [222] M. Vaillant et al. “Statistics on diffeomorphisms via tangent space representations”. In: *NeuroImage* 23 (2004), S161–S169.
- [223] R. Verfürth. “Error estimates for a mixed finite element approximation of the Stokes equations”. In: *RAIRO. Analyse numérique* 18.2 (1984), pp. 175–182.
- [224] F. X. Vialard. “Hamiltonian approach to shape spaces in a diffeomorphic framework: from the discontinuous image matching problem to a stochastic growth model”. PhD thesis. École Normale Supérieure de Cachan-ENS Cachan, 2009.
- [225] F. X. Vialard and F. Santambrogio. “Extension to BV functions of the large deformation diffeomorphisms matching approach”. In: *Comptes Rendus Mathématique* 347.1 (2009), pp. 27–32.
- [226] F. X. Vialard et al. “Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation”. In: *International Journal of Computer Vision* 97.2 (2012), pp. 229–241.
- [227] C. Villani. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media, 2008.
- [228] L. Wang et al. “Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type”. In: *IEEE transactions on medical imaging* 26.4 (2007), pp. 462–470.
- [229] M. Wang. “On the necessity and sufficiency of the patch test for convergence of nonconforming finite elements”. In: *SIAM journal on numerical analysis* 39.2 (2001), pp. 363–384.

- [230] M. Wang and J. Xu. “Minimal finite element spaces for  $2m$ -th-order partial differential equations in  $R^n$ ”. In: *Mathematics of Computation* 82.281 (2013), pp. 25–43.
- [231] S. Wu and J. Xu. “Nonconforming finite element spaces for  $2m$ th order partial differential equations on  $\mathbb{R}^n$  simplicial grids when  $m = n + 1$ ”. In: *Mathematics of Computation* 88.316 (2019), pp. 531–551.
- [232] L. Younes. “Computable elastic distances between shapes”. In: *SIAM Journal on Applied Mathematics* 58.2 (1998), pp. 565–586.
- [233] L. Younes. “Jacobi fields in groups of diffeomorphisms and applications”. In: *Quarterly of applied mathematics* (2007), pp. 113–134.
- [234] L. Younes. *Shapes and diffeomorphisms*. Vol. 171. Springer Science & Business Media, 2010.
- [235] L. Younes, F. Arrate, and M. I. Miller. “Evolutions equations in computational anatomy”. In: *NeuroImage* 45.1 (2009), S40–S50.
- [236] A. Ženíšek. “Interpolation polynomials on the triangle”. In: *Numerische Mathematik* 15.4 (1970), pp. 283–296.
- [237] *Software used in 'On inverse problems and finite elements in shape analysis'*. May 2020. DOI: 10.5281/zenodo.3832141. URL: <https://doi.org/10.5281/zenodo.3832141>.
- [238] M. Zhang and P. T. Fletcher. “Finite-dimensional Lie algebras for fast diffeomorphic image registration”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2015, pp. 249–260.
- [239] W. P. Ziemer. *Weakly differentiable functions: Sobolev spaces and functions of bounded variation*. Vol. 120. Springer Science & Business Media, 2012.
- [240] J. Zitelli et al. “A class of discontinuous Petrov–Galerkin methods. Part IV: The optimal test norm and time-harmonic wave propagation in 1D”. In: *Journal of Computational Physics* 230.7 (2011), pp. 2406–2432.

# APPENDIX A:

## INNER PROBLEM CONVERGENCE RESULTS FOR $d = 2$

We also construct a test suite for problem 4 where  $d = 2$  similar to what was done in section 4.1.2.1 for  $d = 1$ . A globally continuous piecewise linear finite element space is used. Note that this finite element space is defined in space-time.

Table A.1 shows three manufactured solutions with convergence rates shown in table A.2, A.3 and A.4. Here, we use the same setup as in chapter 4 and these simulations take less than 8 minutes on a 2014 MacBook Pro with a 2.5GHz Intel Core i7 processor and 16GB of 1600MHz DDR3 RAM. These rates are also depicted in figure A.1. Similar to the  $d = 1$  case we have added forcing terms to the bilinear form so the computed solutions match the analytical ones in table A.1. Again, the solutions in table A.1 are defined up to a linear function catering for periodicity of the spatial domain.

TABLE A.1: Manufactured solutions to problem 4 where  $d = 2$ .

| No. | $I(x, t)$                                          | $u(x, t)$    |
|-----|----------------------------------------------------|--------------|
| 0   | $e^{-25((x-0.3(1-t)-0.6t)^2+(y-0.3(1-t)-0.6t)^2)}$ | $(0.3, 0.3)$ |
| 1   | $e^{-25((x-0.3(1-t)-0.7t)^2+(y-0.5)^2)}$           | $(0.4, 0)$   |
| 2   | $e^{-25((y-0.3(1-t)-0.7t)^2+(x-0.5)^2)}$           | $(0, 0.4)$   |

TABLE A.2: Convergence rates for the examples in table A.1.

| $h^{-1}$ | $L^2$ error | Order | Energy error | Order | CPU time (s) |
|----------|-------------|-------|--------------|-------|--------------|
| 5        | 4.0895e-02  | -     | 2.3317e-01   | -     | 0.0677       |
| 10       | 2.2204e-02  | 0.88  | 9.7651e-02   | 1.26  | 0.091        |
| 15       | 1.3367e-02  | 1.25  | 5.1744e-02   | 1.57  | 0.184        |
| 20       | 8.7581e-03  | 1.47  | 3.1688e-02   | 1.70  | 0.645        |
| 25       | 6.1231e-03  | 1.60  | 2.1277e-02   | 1.78  | 1.9852       |
| 30       | 4.4968e-03  | 1.69  | 1.5222e-02   | 1.84  | 5.3470       |
| 35       | 3.4307e-03  | 1.76  | 1.1406e-02   | 1.87  | 1.2873       |
| 40       | 2.6977e-03  | 1.80  | 8.8536e-03   | 1.90  | 2.8791       |
| 45       | 2.1737e-03  | 1.83  | 7.0653e-03   | 1.92  | 6.0720       |

TABLE A.3: Convergence rates for example 1 from table A.1.

| $h^{-1}$ | $L^2$ error | Order | Energy error | Order | CPU time (s) |
|----------|-------------|-------|--------------|-------|--------------|
| 5        | 3.6966e-02  | -     | 2.1786e-01   | -     | 0.048        |
| 10       | 2.0623e-02  | 0.84  | 8.9662e-02   | 1.28  | 0.092        |
| 15       | 1.2475e-02  | 1.24  | 4.7457e-02   | 1.57  | 0.190        |
| 20       | 8.1835e-03  | 1.47  | 2.9032e-02   | 1.71  | 0.622        |
| 25       | 5.7241e-03  | 1.60  | 1.9480e-02   | 1.79  | 1.976        |
| 30       | 4.2056e-03  | 1.69  | 1.3930e-02   | 1.84  | 5.508        |
| 35       | 3.2102e-03  | 1.75  | 1.0436e-02   | 1.87  | 13.053       |
| 40       | 2.5259e-03  | 1.80  | 8.0996e-03   | 1.90  | 28.917       |
| 45       | 2.0367e-03  | 1.83  | 6.4637e-03   | 1.92  | 66.384       |

TABLE A.4: Convergence rates for example 2 from table A.1.

| $h^{-1}$ | $L^2$ error | Order | Energy error | Order | CPU time (s) |
|----------|-------------|-------|--------------|-------|--------------|
| 5        | 3.6966e-02  | -     | 2.1786e-01   | -     | 0.062        |
| 10       | 2.0623e-02  | 0.84  | 8.9662e-02   | 1.28  | 0.118        |
| 15       | 1.2475e-02  | 1.24  | 4.7457e-02   | 1.57  | 0.250        |
| 20       | 8.1835e-03  | 1.47  | 2.9032e-02   | 1.71  | 0.822        |
| 25       | 5.7241e-03  | 1.60  | 1.9480e-02   | 1.79  | 2.563        |
| 30       | 4.2056e-03  | 1.69  | 1.3930e-02   | 1.84  | 6.584        |
| 35       | 3.2102e-03  | 1.75  | 1.0436e-02   | 1.87  | 1.680        |
| 40       | 2.5259e-03  | 1.80  | 8.0996e-03   | 1.90  | 4.252        |
| 45       | 2.0367e-03  | 1.83  | 6.4637e-03   | 1.92  | 8.924        |

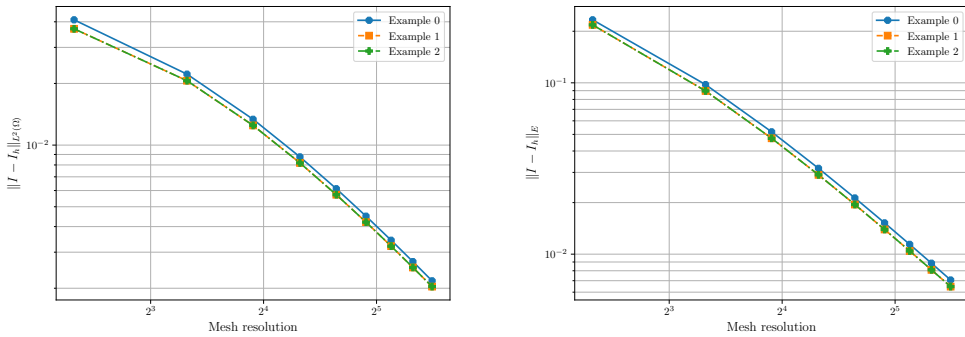


FIGURE A.1: Convergence as a function of mesh resolution  $h^{-1}$  for the examples in table A.1. Left:  $L^2$  errors. Right:  $E(\mathbf{b}; \Omega)$  errors.

## APPENDIX B:

# CONVERGENCE RESULTS FOR THE ROBUST XU-WU ELEMENT

[149] describes a general approach to implementation of FEM for advanced elements such as  $\tilde{V}_{th}^d$ . The author of [149], Robert Kirby, carried out the implementation of  $\tilde{V}_{th}^d$  in Firedrake [191, 19, 18, 70] that we use in our simulations. The simulations below and in chapter 3 have, however, been implemented solely by the author of this thesis. In addition to the Zenodo entry [237], table B.1 below provides branch and revision information for the packages that comprise the Firedrake framework. The reader should refer to the previous references for details of each of these components. The simulations below were run on a 2014 MacBook Pro with a 2.5GHz Intel Core i7 processor and 16GB of 1600MHz DDR3 RAM.

TABLE B.1: Branch and revision information for the Firedrake components.

| Package         | Branch    | Revision |
|-----------------|-----------|----------|
| COFFEE          | master    | 72a4464  |
| FInAT           | wuxu      | 0a68934  |
| PyOP2           | master    | 99d10ca  |
| fiat            | toriesz   | e1e3c21  |
| firedrake       | master    | b4e9c63  |
| h5py            | firedrake | 6768237  |
| libspatialindex | master    | 4768bf3  |
| libsupermesh    | master    | 832d3fb  |
| loopy           | firedrake | b9c84fe  |
| petsc           | firedrake | 4232e4a  |
| petsc4py        | firedrake | b735071  |
| pyadjoint       | master    | aa8c92b  |
| tsfc            | wuxu      | 862e867  |
| ufl             | wuxu      | 56cc4ac  |

We verify the Firedrake implementation of  $\tilde{V}_{th}^d$  via manufactured solutions to three tri-Laplacian equations. Similar to Examples 1 and 2 from [231] we compute discrete solutions to the following problems:

$$-\Delta^3 u = 0 \quad \text{in } [0, 1]^2, \quad (\text{B.1})$$

with boundary conditions given by the exact solution  $u = \exp(\pi y) \sin(\pi x)$ , and:

$$-\Delta^3 u = 0 \quad \text{in } D := [-1, 1]^2 \setminus [0, 1] \times [0, -1], \quad (\text{B.2})$$

with exact solution  $u = r^{2.5} \sin(2.5\theta)$ . Owing to the singularity,  $u \in H^{3+1/2}(D)$ . The results for different mesh resolutions can be seen in tables B.2 and B.3. The convergence rates are aligned with [231]; the solution converges linearly in the  $H^3$  norm while the singular solution converges with order  $1/2$ .

TABLE B.2: Convergence results for problem (B.1)

| $h^{-1}$ | $L^2$ error | Order | $H^1$ error | Order | $H^2$ error | Order | $H^3$ error | Order | CPU time (s) |
|----------|-------------|-------|-------------|-------|-------------|-------|-------------|-------|--------------|
| 4        | 4.7406e-02  | -     | 3.3149e-01  | -     | 1.1401e+01  | -     | 1.2895e+03  | -     | 0.025        |
| 8        | 1.6625e-02  | 1.51  | 1.0572e-01  | 1.65  | 4.6464e+00  | 1.29  | 1.0851e+03  | 0.25  | 0.047        |
| 16       | 3.8320e-03  | 2.12  | 2.3924e-02  | 2.14  | 1.4645e+00  | 1.67  | 6.9968e+02  | 0.63  | 0.050        |
| 32       | 9.3654e-04  | 2.03  | 5.7993e-03  | 2.04  | 4.1280e-01  | 1.83  | 3.9762e+02  | 0.82  | 0.163        |
| 64       | 2.3363e-04  | 2.00  | 1.4435e-03  | 2.01  | 1.0968e-01  | 1.91  | 2.1195e+02  | 0.91  | 0.652        |
| 128      | 5.8605e-05  | 2.00  | 3.6229e-04  | 1.99  | 2.8272e-02  | 1.96  | 1.0941e+02  | 0.95  | 2.890        |

TABLE B.3: Convergence results for problem (B.2)

| $h^{-1}$ | $L^2$ error | Order | $H^1$ error | Order | $H^2$ error | Order | $H^3$ error | Order | CPU time (s) |
|----------|-------------|-------|-------------|-------|-------------|-------|-------------|-------|--------------|
| 4        | 5.0748e-02  | -     | 3.3780e-01  | -     | 3.6276e+00  | -     | 5.4517e+01  | -     | 0.026        |
| 8        | 2.7191e-02  | 0.90  | 1.4243e-01  | 1.25  | 1.3545e+00  | 1.42  | 4.1543e+01  | 0.39  | 0.025        |
| 16       | 1.3709e-02  | 0.99  | 6.9938e-02  | 1.03  | 6.0983e-01  | 1.15  | 3.3153e+01  | 0.33  | 0.029        |
| 32       | 7.4075e-03  | 0.89  | 3.3062e-02  | 1.08  | 2.9160e-01  | 1.06  | 2.5779e+01  | 0.36  | 0.050        |
| 64       | 3.9324e-03  | 0.91  | 1.6324e-02  | 1.02  | 1.3826e-01  | 1.08  | 1.8807e+01  | 0.45  | 0.162        |
| 128      | 1.8512e-03  | 1.09  | 7.5369e-03  | 1.11  | 6.1964e-02  | 1.16  | 1.3277e+01  | 0.50  | 0.652        |

Table B.4 shows the convergence for the problem:

$$-\Delta^3 u = f \quad \text{in } [0, 1]^2, \quad (\text{B.3})$$

with exact solution  $u = 100x^4(1-x)^4y^4(1-y)^4$  where  $f$  is computed from  $u$ .

TABLE B.4: Convergence results for problem (B.3)

| $h^{-1}$ | $L^2$ error | Order | $H^1$ error | Order | $H^2$ error | Order | $H^3$ error | Order | CPU time (s) |
|----------|-------------|-------|-------------|-------|-------------|-------|-------------|-------|--------------|
| 4        | 9.6591e-05  | -     | 6.5273e-04  | -     | 6.9816e-03  | -     | 1.4920e-01  | -     | 0.033        |
| 8        | 1.6192e-05  | 2.58  | 1.3086e-04  | 2.32  | 2.0107e-03  | 1.80  | 8.8160e-02  | 0.76  | 0.042        |
| 16       | 3.1888e-06  | 2.34  | 3.5849e-05  | 1.87  | 5.8457e-04  | 1.78  | 4.7917e-02  | 0.88  | 0.072        |
| 32       | 8.6818e-07  | 1.88  | 9.8161e-06  | 1.87  | 1.6294e-04  | 1.84  | 2.4451e-02  | 0.97  | 0.263        |
| 64       | 2.2668e-07  | 1.94  | 2.5321e-06  | 1.95  | 4.2461e-05  | 1.94  | 1.2286e-02  | 0.99  | 1.053        |
| 128      | 5.7406e-08  | 1.98  | 6.3850e-07  | 1.99  | 1.0749e-05  | 1.98  | 6.1502e-03  | 1.00  | 4.531        |

Finally, table B.5 shows the same for the following biharmonic problem from [150] reproduced in (B.4):

$$-\Delta^2 u = f \quad \text{in } [0, 1]^2, \quad (\text{B.4})$$

with homogeneous Dirichlet and Neumann conditions, where  $f$  is computed from the exact solution  $u = 100(x(1-x)y(1-y))^2$ . The finite element space used here is simply  $\tilde{V}_h$  defined in chapter 3, but without forcing the degrees of freedom on the boundary to be zero. This space is robust for fourth-order equations as shown by the convergence rates. Improved  $L^2$  convergence rates for this element are, to the best of the author's knowledge, an open problem.

TABLE B.5: Convergence results for problem (B.4)

| $h^{-1}$ | $L^2$ error | Order | $H^1$ error | Order | $H^2$ error | Order | CPU time (s) |
|----------|-------------|-------|-------------|-------|-------------|-------|--------------|
| 8        | 5.4956e-04  | -     | 8.6657e-03  | -     | 4.3903e-01  | -     | 0.072        |
| 16       | 1.7601e-04  | 1.64  | 1.5847e-03  | 2.45  | 1.3690e-01  | 1.68  | 0.094        |
| 32       | 4.8253e-05  | 1.87  | 3.6481e-04  | 2.12  | 5.6496e-02  | 1.28  | 0.229        |
| 64       | 1.2317e-05  | 1.97  | 8.9131e-05  | 2.03  | 2.6179e-02  | 1.11  | 0.842        |
| 128      | 3.0887e-06  | 2.00  | 2.2129e-05  | 2.01  | 1.2728e-02  | 1.04  | 4.440        |
| 256      | 7.7188e-07  | 2.00  | 5.5200e-06  | 2.00  | 6.2952e-03  | 1.02  | 18.400       |

## APPENDIX C:

# TOOLS AND TECHNIQUES FOR ANALYSIS OF NONCONFORMING ELEMENTS

### C.1 PROOF SKETCHES OF (3.22a) AND (3.22b)

We reproduce the inequalities for convenience:

$$\begin{aligned} |e|^{-1} \|f\|_{0,e}^2 &\lesssim h_K^{-2} \|f\|_{0,K}^2 + |f|_{1,K}^2, \quad f \in H^1(K) \\ |e| \|\llbracket w \rrbracket\|_{0,e}^2 &\lesssim h_K^2 \sum_{K \in \mathcal{K}_e} \|w\|_{1,K}^2, \quad w|_K \in H^1(K). \end{aligned}$$

These rely mainly on trace-type results and scaling arguments cf. appendix D.

For the first we have:

$$\|\hat{f}\|_{0,\hat{e}}^2 \lesssim \|\hat{f}\|_{1,\hat{K}}^2 = \frac{|\hat{K}|}{|K|} (\|f\|_{0,K}^2 + h_K^2 |f|_{1,K}^2),$$

by the trace theorem and a scaling argument. Rearranging:

$$\frac{|K|}{|\hat{K}| h_K^2} \|\hat{f}\|_{0,\hat{e}}^2 \lesssim h_K^{-2} \|f\|_{0,K}^2 + |f|_{1,K}^2.$$

Now since  $h_K^2$  is proportional to the area of  $K$  then:

$$\|\hat{f}\|_{0,\hat{e}}^2 \lesssim h_K^{-2} \|f\|_{0,K}^2 + |f|_{1,K}^2.$$

Noticing that  $\|\hat{f}\|_{0,\hat{e}}^2 = |e|^{-1} \|f\|_{0,e}^2$  completes the result.

For the second:

$$\begin{aligned} \|\llbracket f \rrbracket\|_{0,e}^2 &= |\det B_e| \|\llbracket \hat{f} \rrbracket\|_{0,e}^2 \\ &\leq |\det B_e| \sum_{K \in \mathcal{K}_e} |\hat{f}|_{1,\hat{K}}^2 \quad [42, \text{Lemma 7.5.26}] \\ &\leq |\det B_e| \sum_{K \in \mathcal{K}_e} |\det B_k|^{-1} h_K^2 |f|_{1,K}^2 \\ &= \sum_{K \in \mathcal{K}_e} \frac{|e|}{|\hat{e}|} \frac{|\hat{K}|}{|K|} h_K^2 |f|_{1,K}^2. \end{aligned}$$

Now since  $|e|^2 \simeq |K|$  by the formula for the area of a triangle, multiplying by  $|e|$  on both sides yields:

$$|e| \|\llbracket f \rrbracket\|_{0,e}^2 \lesssim \sum_{K \in \mathcal{K}_e} h_K^2 |f|_{1,K}^2.$$

## C.2 ERROR ANALYSIS FOR NONCONFORMING $\mathcal{P}^1$ FINITE ELEMENTS

We demonstrate how to obtain a linear convergence rate for nonconforming  $\mathcal{P}^1$  finite elements (see [42, Equation 10.3.2]) for a simple Poisson problem in 2 dimensions where the source term is given by a mesh-aligned function. Letting  $\Omega$  denote the unit square, and  $\Omega_h$  a conforming shape-regular triangulation thereof and  $\mathcal{E}_h$  its mesh skeleton, and  $Q$  denote the subset of  $\mathcal{E}_h$  highlighted in figure 3.2. Then, for simplicity, we define the functional  $f$  by

$$v \mapsto f(v) = \int_Q pv \, dx,$$

where  $p$  is a smooth polynomial defined on  $Q$ . In this way  $f$  is a similar object to the sources terms observed in (3.6b) and (3.11b). The variational system we will study here is:

$$\langle \nabla u, \nabla v \rangle_{0,\Omega} = f(v), \quad \forall v \in H_0^1(\Omega). \quad (\text{C.1})$$

which corresponds to the strong form equation:

$$-\Delta u = 0 \quad \text{in } \Omega \setminus Q, \quad (\text{C.2a})$$

$$u = 0 \quad \text{on } \partial\Omega, \quad (\text{C.2b})$$

$$[\![\nabla u]\!] = p \quad \text{on } Q. \quad (\text{C.2c})$$

where the last equation is understood in an integral sense, and jump across  $Q$  is defined as  $[\![v]\!]_Q = \eta_Q^+ v^+ + \eta_Q^- v^-$  and  $\eta|_{\partial\Omega}$  is the outward normal to  $\Omega$ . Now we define  $\Omega^-$  and  $\Omega^+$  by the domains inside and outside  $Q$ . The connection between these two is seen by a simple integration by parts:

$$-\int_{\Omega} \Delta u v \, dx = \int_{\Omega^-} \nabla u \cdot \nabla v \, dx + \int_{\Omega^+} \nabla u \cdot \nabla v \, dx - \int_Q [\![\nabla u]\!] v \, ds \quad (\text{C.3})$$

$$= \int_{\Omega^-} \nabla u \cdot \nabla v \, dx + \int_{\Omega^+} \nabla u \cdot \nabla v \, dx - \int_Q p v \, ds \quad (\text{C.4})$$

$$= 0. \quad (\text{C.5})$$

Letting  $V_h$  denote the nonconforming  $\mathcal{P}^1$  finite element space defined over the triangulation  $\Omega_h$ . The discrete problem we are looking at is:

$$\sum_{K \in \Omega_{th}} \langle \nabla u_h, \nabla v_h \rangle_{0,K} = f(\{v_h\}), \quad \forall v_h \in V_h. \quad (\text{C.6})$$

Since functions in  $V_h$  are not globally  $H^1$  the trace is not unique on  $Q$  hence the average operator present in the right-hand side. Skipping all well-posedness theory we jump straight to the error analysis. Letting  $a, a_h$  denote the usual bilinear forms we have for  $w \in V_h$ :

$$\begin{aligned}
a_h(u - u_h, w_h) &= \sum_{K \in \Omega_{th}} \langle \nabla u_h, \nabla w_h \rangle_{0,K} - \int_Q p\{w_h\} ds \\
&= \sum_{K \in \Omega_{th}} [\langle -\Delta u_h, w_h \rangle_{0,K} + \int_{\partial K} \eta \cdot \nabla u w_h ds] - \int_Q p\{w_h\} ds \\
&= \sum_{e \in \mathcal{E}_h \setminus Q} \left[ \int_e \nabla u \cdot \llbracket w_h \rrbracket ds \right] + \int_Q \llbracket \nabla u w \rrbracket ds - \int_Q p\{w_h\} ds.
\end{aligned}$$

Recall the following identity:

$$\llbracket ab \rrbracket = \{a\} \llbracket b \rrbracket + \llbracket a \rrbracket \{b\}, \quad (\text{C.7})$$

so:

$$\int_Q \llbracket \nabla u w_h \rrbracket ds = \int_Q \{ \nabla u \} \llbracket w_h \rrbracket + \llbracket \nabla u \rrbracket \{ w_h \} ds,$$

so

$$a_h(u - u_h, w_h) = \sum_{e \in \mathcal{E}_h \setminus Q} \left[ \int_e \nabla u \cdot \llbracket w_h \rrbracket ds \right] + \int_Q \{ \eta \cdot \nabla u \} \llbracket w_h \rrbracket ds, \quad (\text{C.8})$$

since  $\llbracket \nabla u \rrbracket \{ w_h \} = p\{ w_h \}$  on  $Q$ . We now apply the same methods from [42, Chapter 10]. Let  $e$  be an edge of some  $K \in \Omega_h$ . By [42, Equation 10.3.8, 10.3.9] there exists some  $C > 0$  such that:

$$|e|^{-1} \|\xi\|_{0,e}^2 \leq C(h_K^{-2} \|\xi\|_{0,K}^2 + |\xi|_{H^1(K)}^2), \quad \forall \xi \in H^1(K).$$

and we also use the result:

$$|e| \|\llbracket w \rrbracket\|_{0,e}^2 \leq C \sum_{K \in \Omega_{th}} h_K^2 |w|_{H^1(K)}^2.$$

The first term on the right-hand side of (C.8) now becomes:

$$\begin{aligned}
\sum_{e \in \mathcal{E}_h \setminus Q} \int_e \nabla u \cdot \llbracket w_h \rrbracket ds &= \sum_{e \in \mathcal{E}_h \setminus Q} \int_e (\eta \cdot \nabla u - c_e) \nabla \cdot \llbracket w_h \rrbracket ds \\
&\leq \sum_{e \in \mathcal{E}_h \setminus Q} |e|^{-1/2} \|\eta \cdot \nabla u - c_e\|_{0,e} |e|^{1/2} \|\llbracket w \rrbracket\|_{0,e} \\
&\leq \left( \sum_{e \in \mathcal{E}_h \setminus Q} |e|^{-1} \|\eta \cdot \nabla u - c_e\|_{0,e}^2 \right)^{1/2} \left( \sum_{e \in \mathcal{E}_h \setminus Q} |e| \|\llbracket w \rrbracket\|_{0,e}^2 \right)^{1/2}.
\end{aligned}$$

Now using the previous results and letting  $\mathcal{K}_e$  define the elements adjacent to  $e$  the notation :

$$\begin{aligned}
\sum_{e \in \mathcal{E}_h \setminus Q} \int_e \nabla u \cdot \llbracket w_h \rrbracket ds &\leq C \left( \sum_{e \in \mathcal{E}_h \setminus Q} \min_{K \in \mathcal{K}_e} (h_K^{-2} \|\eta \cdot \nabla u - c_e\|_{0,e}^2 + |u|_{H^2(K)}^2) \right)^{1/2} \\
&\quad \times \left( \sum_{e \in \mathcal{E}_h \setminus Q} \sum_{K \in \mathcal{K}_e} h_K^2 |w_h|_{H^1(K)}^2 \right)^{1/2}.
\end{aligned}$$

Selecting  $c_e$  as the average of  $\nabla u \cdot \eta$  on  $K$  and letting  $\|w_h\|_h^2 = a_h(w_h, w_h)$ :

$$\sum_{e \in \mathcal{E}_h \setminus Q} \int_e \nabla u \cdot \llbracket w_h \rrbracket ds \leq Ch \left[ \sum_{K \in \Omega_{th}} \|\Delta u\|_{0,K}^2 \right]^{1/2} \|w_h\|_h.$$

Here we have used [89, Proposition 1.135] to say

$$\|\eta \cdot \nabla u - c_e\|_{0,e} \lesssim h_K |u|_{2,K}.$$

Now we treat the second term on the right-hand side of (C.8) which we write as:

$$\int_Q \{\nabla u\} \cdot \llbracket w_h \rrbracket ds = \sum_{e \in Q} \int_e \frac{\nabla u^+ + \nabla u^-}{2} \cdot \llbracket w_h \rrbracket ds.$$

Now we apply the same technique from before to each term involving  $u^+$  and  $u^-$ , respectively. We denote by  $\Omega_h^+$  the set of triangles  $K \in \Omega_h$  located outside the curve  $Q$  ( $\Omega_h^-$  denoting those on the inside). Also let  $\mathcal{K}_e^+$  be the set of triangles in  $\Omega_h^+$  having  $e$  as an edge and such that  $K \in \mathcal{K}_e^+$  lies on the outside of  $Q$  with a similar definition for  $\mathcal{K}_e^-$ . For  $u^+$ :

$$\begin{aligned} \sum_{e \in Q} \int_e \nabla u^+ \cdot \llbracket w_h \rrbracket ds &\leq C \left( \sum_{e \in Q} \min_{K \in \mathcal{K}_e^+} (h_K^{-2} \|\eta \cdot \nabla u^+ - c_e\|_{0,e} + |u^+|_{H^2(K)}^2) \right)^{1/2} \\ &\quad \times \left( \sum_{e \in \mathcal{E}_h \setminus Q} \sum_{K \in \mathcal{N}(e)} h_K^2 |w_h|_{H^1(K)}^2 \right)^{1/2} \\ &\leq Ch \left[ \sum_{K \in \Omega_h^+} \|\Delta u^+\|_{0,K}^2 \right]^{1/2} \|w_h\|_h. \end{aligned}$$

Similarly:

$$\sum_{e \in Q} \int_e \nabla u^- \cdot \llbracket w_h \rrbracket ds \leq Ch \left[ \sum_{K \in \Omega_h^-} \|\Delta u^-\|_{0,K}^2 \right]^{1/2} \|w_h\|_h.$$

Combining these results we can write:

$$\int_Q \{\nabla u\} \cdot \llbracket w_h \rrbracket ds \leq \frac{C}{2} h \left[ \sum_{K \in \Omega_{th}} \|\Delta u\|_{0,K}^2 \right]^{1/2} \|w_h\|_h.$$

Now returning to the error estimate we get:

$$a_h(u - u_h, w_h) \lesssim h \left[ \sum_{K \in \Omega_{th}} \|\Delta u\|_{0,K}^2 \right]^{1/2} \|w_h\|_h.$$

By a Strang type argument we know that, for integers  $j \geq 1$ :

$$\begin{aligned} \|u - u_h\|_h &\lesssim h^j \left[ \sum_{K \in \Omega_{th}} \|u\|_{2,K} \right]^{1/2} + h \left[ \sum_{K \in \Omega_{th}} \|\Delta u\|_{0,K}^2 \right]^{1/2} \\ &\lesssim h \left[ \sum_{K \in \Omega_{th}} \|u\|_{2,K} \right]^{1/2}, \end{aligned}$$

because  $u$  is smooth locally on  $K \in \Omega_h$  (approximation) and convergence is linear in  $h$  (consistency).

# APPENDIX D:

## SCALING ARGUMENTS

Let  $\hat{K}$ ,  $K$  be convex simplices in  $\mathbb{R}^d$  such that  $\text{diam}(\hat{K}) = 1$  and  $\psi_K : \hat{K} \rightarrow K$  be an affine diffeomorphism with Jacobian  $B_K$  and that:

$$|\det B_K| = \frac{|K|}{|\hat{K}|}, \quad \|B_K\| \leq \frac{h_K}{\rho_{\hat{K}}}, \quad \|B_K^{-1}\| \leq \frac{h_{\hat{K}}}{\rho_K},$$

where  $h_K = \text{diam}(K)$  and  $\rho_K$  is the diameter of the largest ball inscribed in  $K$ , see [89, Lemma 1.100] and where the norms are sup-norms over the components of the matrices. For shape-regular meshes this means that there exists a constant  $c > 0$  such that  $\|B_K\| \|B_K^{-1}\| \leq c$ .

For  $v \in H^k(K)$  and  $\hat{v} = v \circ \psi_K$  we have:

$$|\hat{v}|_{m,\hat{K}} \lesssim \|B_K\|^m |\det B_K|^{-1/2} |v|_{m,K}, \quad (\text{D.1a})$$

$$|v|_{m,K} \lesssim \|B_K^{-1}\|^m |\det B_K|^{1/2} |\hat{v}|_{m,\hat{K}}. \quad (\text{D.1b})$$

We sketch the proof of (D.1b) for  $m = 0, 1$ , see the previous reference for more details.

$m = 0$ :

$$\begin{aligned} \|v\|_{0,K}^2 &= \int_{\hat{K}} |\hat{v}|^2 |\det B_k| \, d\hat{x} \\ &= |\det B_k| \|\hat{v}\|_{0,\hat{K}}^2. \end{aligned}$$

$m = 1$ :

$$\begin{aligned} |v|_{1,K}^2 &= \int_{\hat{K}} [(\nabla_x v) \circ \psi_K]^2 |\det B_k| \, d\hat{x} \\ &\lesssim |\det B_k| \|B_K^\top\| \int_{\hat{K}} [\nabla_{\hat{x}} \hat{v}]^2 \, d\hat{x} \\ &= |\det B_k| \|B_K^\top\| \|\hat{v}\|_{1,\hat{K}}^2. \end{aligned}$$

These estimates are typically used in estimating interpolation error. Indeed, let  $\pi_K v$  be the standard local interpolant of polynomial degree  $m - 1$ , for instance,

for  $v \in H^m(K)$ :

$$\begin{aligned}
\|v - \pi v\|_{m-1,K} &\lesssim |\det B_k|^{1/2} \|B_K^{-1}\| \|\hat{v} - \hat{\pi} \hat{v}\|_{m-1,\hat{K}} \\
&\lesssim |\det B_k|^{1/2} \|B_K^{-1}\|^{m-1} \|\hat{v}\|_{2,\hat{K}} && \text{(Bramble-Hilbert)} \\
&\lesssim |\det B_k|^{1/2} \|B_K^{-1}\|^{m-1} |\det B_k|^{-1/2} \|B_K^{-1}\|^m \|\hat{v}\|_{m,\hat{K}} \\
&\simeq \|B_K^{-1}\| \|v\|_{m,K} \\
&\lesssim h_K \|v\|_{m,K}.
\end{aligned}$$

# APPENDIX E:

## LIMITING PROCEDURE

For completeness we show a standard procedure in analysis namely the limiting procedure also sometimes called a density argument.

For  $O \in \mathbb{R}^d$  and let  $\phi \in C^\infty(\bar{O}) = X$  be dense in a Banach space  $G$ . The following inequality holds for some  $\phi \mapsto T\phi$ ,  $T : X \rightarrow Y$ , where  $Y$  is Banach:

$$\|T\phi\|_Y \lesssim \|\phi\|_X, \quad (\text{E.1})$$

and our goal is to define  $T$  on  $G$ . By density, there exists a sequence  $\Phi = \{\phi_n\}_{n \geq 0}$  such that  $\phi_n \rightarrow \psi$ , for all  $\psi \in G$ . Formally we define the limit:

$$T\psi = \lim_{n \rightarrow \infty} T\phi_n, \quad (\text{E.2})$$

although we have yet to determine if this limit is well-defined. From (E.1) we get:

$$\|T\phi_n - T\phi_{n+j}\|_Y \lesssim \|\phi_n - \phi_{n+j}\|_X, \quad (\text{E.3})$$

for integers  $n, j \geq 0$ . Since  $\Phi$  is convergent, the right-hand side of (E.3) converges since  $X$  is Banach. This implies that  $\{T\phi_n\}_{n \geq 0}$  is Cauchy in  $Y$  so the left-hand side of (E.3) converges and (E.2) is well-defined. Using continuity of the norm we verify (E.1) for this limit:

$$\begin{aligned} \|T\psi\|_Y &= \left\| \lim_{n \rightarrow \infty} T\phi_n \right\|_Y \\ &= \lim_{n \rightarrow \infty} \|T\phi_n\|_Y \\ &\lesssim \lim_{n \rightarrow \infty} \|\phi_n\|_Y \\ &= \|\psi\|_Y. \end{aligned}$$