Regge Finite Elements
with Applications in Solid Mechanics and Relativity

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Lizao Li

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advisor: Prof. Douglas N. Arnold

June, 2016
Acknowledgements

Acknowledge people.
Dedication

Dedicate to somebody.
Abstract

Abstract goes here.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>i</td>
</tr>
<tr>
<td>Dedication</td>
<td>ii</td>
</tr>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>1</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>2</td>
</tr>
<tr>
<td><strong>2 Generalized Regge finite element</strong></td>
<td>3</td>
</tr>
<tr>
<td>2.1 Interpretations of the classical Regge finite elements</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Definition of generalized Regge family</td>
<td>9</td>
</tr>
<tr>
<td>2.3 Basic properties</td>
<td>14</td>
</tr>
<tr>
<td>2.3.1 Bernstein decomposition for Lagrange elements</td>
<td>17</td>
</tr>
<tr>
<td>2.3.2 Geometric decomposition for Regge elements</td>
<td>20</td>
</tr>
<tr>
<td>2.4 Affine and approximation properties</td>
<td>27</td>
</tr>
<tr>
<td>2.4.1 Affine property</td>
<td>27</td>
</tr>
<tr>
<td>2.4.2 Approximation properties of the canonical interpolant</td>
<td>29</td>
</tr>
<tr>
<td>2.5 Coordinate representations and implementable degrees of freedom</td>
<td>36</td>
</tr>
<tr>
<td>2.5.1 Coordinate representations</td>
<td>36</td>
</tr>
<tr>
<td>2.5.2 Implementable degrees of freedom</td>
<td>38</td>
</tr>
<tr>
<td><strong>3 Geodesics on Generalized Regge metrics</strong></td>
<td>43</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>43</td>
</tr>
<tr>
<td>3.2 Review of the smooth geodesic theory</td>
<td>49</td>
</tr>
<tr>
<td>3.3 Global geodesics on Regge metrics</td>
<td>51</td>
</tr>
</tbody>
</table>
3.4 Local geodesics on Regge metrics: variational approach .......................... 53
3.5 Local geodesics on Regge metrics: geometric approach ......................... 58
3.6 Hamiltonian structures of local geodesics ............................................. 61
3.7 A robust algorithm for generalized local geodesics ............................... 64
3.8 Error analysis ................................................................. 68
3.9 Numerical examples: Kepler and Schwarzschild systems ..................... 77
  3.9.1 Kepler system ............................................................. 77
  3.9.2 Jacobi’s formulation ..................................................... 79
  3.9.3 Numerical examples for Keplerian orbits .................................... 80
  3.9.4 Schwarzschild system .................................................. 85

4 Rotated generalized Regge element and applications in solid mechanics 87
  4.1 Rotated generalized Regge element ............................................... 88
  4.2 Solving the biharmonic equation via the Hellan-Herrmann-Johnson mixed formulation ......................................................... 90
    4.2.1 Hellan-Herrmann-Johnson continuous mixed formulation ............ 91
    4.2.2 Hellan-Herrmann-Johnson discretization ................................ 92
    4.2.3 Discretization of biharmonic equation in higher dimensions using rotated Regge elements .................................................. 95
  4.3 Solving the elasticity equation via the Pechstein-Schöberl mixed formulation .......................................................... 101
    4.3.1 Continuous mixed formulation ....................................... 101
    4.3.2 Rotated Regge element discretization ................................ 105
    4.3.3 Numerical experiments .................................................. 107
  4.4 Connection with numerical relativity .............................................. 113
  4.5 Normal-normal continuous finite elements in all dimensions ............ 113

5 Model problems in relativity for studying discretization 114
  5.1 The fully nonlinear space-time Einstein equation ............................ 114
  5.2 Linearized space-time Einstein equation ...................................... 116
  5.3 Matrix calculus notation ...................................................... 118
  5.4 (1+3)-formulation of linearized initial-value problem ....................... 119
  5.5 Gauge freedom and simplified model problem ................................ 123
  5.6 Well-posedness on the full space and flat torus ................................ 128
  5.7 Elliptic steady state model problems and geometric theory of defects .... 129
  5.8 Fourier analysis on the flat torus: smooth case ............................. 132
6 Two failure modes of space-time Regge Calculus

6.1 Failure due to the infinite dimensional kernel
   6.1.1 Well-posedness at continuous level
   6.1.2 Discretization
   6.1.3 Numerical examples and discussion
   6.1.4 Implications for Regge Calculus

6.2 Failure due to the space-time scheme for the second-order time derivative
   6.2.1 Regge-calculus style derivation of the model problem
   6.2.2 Initial-value problem and well-posedness
   6.2.3 Regge calculus-like space-time discretization and finite element view

6.3 Numerical example
   6.3.1 von Neumann stability analysis
   6.3.2 Implication for Regge calculus

7 Conclusion

7.1 Summary and comments
7.2 Future works
# List of Tables

2.1 Bernstein-style Basis in 2D ......................................................... 17
2.2 Bernstein-style Basis in 3D ......................................................... 17

3.1 Convergence rates comparison for geodesic solvers. ......................... 49
3.2 Convergence rate for a fixed maximum time. $h_m$ is the mesh size. The rates in
the parenthesis are for the cases without turning $p$ at interior facets. .......... 81
3.3 The error growth rate in time $t$. The rates in the parenthesis are for the cases
without turning $p$ at interior facets. ................................................ 83

4.1 2D biharmonic degree 0 ............................................................. 98
4.2 2D biharmonic degree 1 ............................................................. 99
4.3 2D biharmonic degree 2 ............................................................. 99
4.4 3D biharmonic degree 0 ............................................................. 100
4.5 3D biharmonic degree 1 ............................................................. 101
4.6 3D biharmonic degree 2 ............................................................. 101
4.7 2D elasticity degree 1 .............................................................. 108
4.8 2D elasticity degree 2 .............................................................. 108
4.9 2D elasticity degree 3 .............................................................. 108
4.10 3D elasticity degree 1 ............................................................. 109
List of Figures

2.1 Degrees of freedom in 1D for \( r = 0, 1, 2, \ldots \) ................................................. 4
2.2 Degrees of freedom in 2D for \( r = 0, 1, 2, \ldots \) ................................................. 4
2.3 Degrees of freedom in 3D for \( r = 0, 1, 2, \ldots \) ................................................. 4
2.4 Interior degrees of freedom in 3D for \( r = 2, 3, 4, \ldots \) ........................................... 5
2.5 Subdivision-based degrees of freedom in 2D for \( r = 0, 1, 2, \ldots \) ......................... 5
2.6 A triangulated surface ........................................................................................................ 7
2.7 Mesh and non-meshes in 2D .............................................................................................. 9
2.8 REG\(^1\) assembly on a 3-triangle mesh ........................................................................... 12
2.9 Geometric decomposition of \( \mathcal{P}^3 \) on a triangle ............................................................ 19
2.10 Edge-based Bernstein decomposition for \( \mathcal{P}^3 \) on a triangle. The chosen edge is in red. Basis associated with edges are in black while those associated with cells are in blue. ................................................................. 24
2.11 Edge-based Bernstein decomposition for \( \mathcal{P}^3 \) on a tetrahedron. The chosen edge is thickened. Basis associated with edges are in red, those associated with triangles are in blue, and those associated with cells are in black. ................................. 24
2.12 Definition of \( h_c \) and \( \rho_c \) ...................................................................................... 31
2.13 A triangle with a small shape constant and one with a large shape constant .......... 32
2.14 Pictures for \( X^f_\hat{r} \) ....................................................................................................... 40
2.15 Subdivision-based degrees of freedom in 2D ................................................................. 40
2.16 Subdivision-based degrees of freedom in 3D ................................................................. 41
3.1 Illustration of the first geodesic idea ................................................................................. 45
3.2 Illustration of the second geodesic idea. In the middle two figures, the red line indicates the tangential direction while the blue indicates the normal direction. .......... 46
3.3 Keplerian orbits. Left to right, top to bottom \( r = 0, 1, 2, 3 \) ........................................... 47
3.4 Schwarzschild orbits. Left to right, top to bottom, \( r = 0, 1, 2, 3 \) ............................... 48
3.5 Failure of having a well-defined tangent space at a point ............................................... 54
6.6 Plot of the growth of the gradient and harmonic part of the numerical solution
to the nonlinear problem ................................................. 146
6.7 Plot of numerical and exact solution with projected initial data ................. 147
6.8 Plot of the growth of the gradient and harmonic part of the numerical solution
with projected initial data ................................................ 147
6.9 Illustration of the marching scheme ........................................ 152
6.10 Parallel marching scheme .................................................. 152
6.11 Uniform mesh ................................................................ 153
6.12 Randomly perturbed mesh ................................................... 154
6.13 Perturb every second spatial slice .......................................... 154
6.14 Perturb every third spatial slice ............................................. 155
6.15 Uniform mesh \( k = 0.5 \) and \( h = 1.0 \) ................................. 157
6.16 Uniform mesh \( k = 1.0 \) and \( h = 1.0 \) ................................. 157
6.17 Uniform mesh \( k = 1.0 \) and \( h = 0.5 \) ................................. 158
6.18 Perturb the middle node spatially ......................................... 158
6.19 Perturb the middle node temporally ........................................ 159
6.20 General perturbation of the middle node .................................. 159
6.21 Move all middle nodes temporally ........................................ 160
6.22 Move all middle nodes spatially .............................................. 160
6.23 Move all middle nodes spatially .............................................. 161
6.24 Uniform mesh \( k = 0.5 \) and \( h = 1.0 \) ................................. 162
Chapter 1

Introduction
Chapter 2

Generalized Regge finite element

The generalized Regge finite element family is the central object of this thesis. It is defined on simplicial meshes of dimension \( n \geq 1 \) for symmetric covariant 2-tensor fields. At each point, a covariant 2-tensor field takes two vectors at that point and returns a number. Hence, under the usual coordinate identifications, equivalently, it is a finite element family for symmetric \( n \times n \) matrix-valued functions. Its continuity properties make it very useful for solving differential equations in solid mechanics and geometry. In this chapter, its definition and various basic properties are studied in detail.

The notation \( \text{REG}^r \) will be used throughout this thesis to denote the generalized Regge finite elements of degree \( r \). To fix the ideas, a concrete and directly implementable description of \( \text{REG}^r \) is given here using coordinates. The reader should be reminded that the underlying object is coordinate-free. The more abstract definitions convenient for mathematical analysis will be the subject of Section 2.2. For a \( k \)-simplex \( f \), let \( \mathcal{P}^r(f) \) be the space of polynomials of degree \( r \) or less in \( k \) variables as functions on \( f \). For \( r < 0 \), it is understood that \( \mathcal{P}^r(f) = \{0\} \). For any line segment \( L \) in \( \mathbb{R}^n \) and any symmetric matrix \( u \in \mathbb{R}^{n \times n} \), define:

\[
\begin{align*}
u_L := t^T ut
\end{align*}
\]

where \( t \in \mathbb{R}^n \) is the coordinate difference between the end-points of \( L \). Clearly the sign of \( t \) does not affect the value and \( u_L \) is well-defined.

In 1D, a 1-by-1 matrix is a just scalar. On a line segment \( L \), the shape functions of \( \text{REG}^r(L) \) is \( \mathcal{P}^r(L) \). The degrees of freedom are integrals of \( u_L \) against \( \mathcal{P}^r(L) \). The degrees of freedom can be implemented by evaluating \( u_L \) for any function \( u \) on \( L \) at the points marked by the center of the green bars in Figure 2.1.
In this case, $u_L$ is just the value of $u$ times the squared Euclidean length of $L$. Note that all the points marked by the green bars are in the interior of $L$. Hence REG$^1$ is the same as the Discontinuous Lagrange element on $L$.

In 2D, let $T$ be a triangle in $\mathbb{R}^2$. The shape functions for REG$^r(T)$ consist of symmetric 2-by-2 matrix-valued functions whose components are in $P^r(T)$. Let $\{E_1, E_2, E_3\}$ be the three edges of $T$. The degrees of freedom are: for any symmetric 2-by-2 matrix-valued function $u$,

\[
\left\{
\begin{array}{c}
\text{on each } E_i, \text{ evaluate the 1D degrees of freedom,} \\
\text{on } T, \text{ integral of } u_{E_i} \text{ against } P^{r-1}(T) \text{ for } i = 1, 2, 3.
\end{array}
\right.
\]

The degrees of freedom associated with $T$ can be implemented by evaluating $u_{E_i}$ at the center of the blue triangles in Figure 2.2 for $i = 1, 2, 3$. Note that all the degrees of freedom associated with $T$ are interior to $T$ and that the first one of these showed up for degree $r = 1$.

In 3D, let $H$ be a tetrahedron in $\mathbb{R}^3$. The shape functions for REG$^r(H)$ consist of symmetric 3-by-3 matrix-valued functions whose components are in $P^r(H)$. Let $\{E_i\}_{i=1}^6$ be the six edges and $\{T_j\}_{j=1}^4$ be the four triangular faces of $H$. The degrees of freedom in 3D are: for any symmetric 3-by-3 matrix-valued function $u$,

\[
\left\{
\begin{array}{c}
\text{on each } E_i, \text{ evaluate the 1D degrees of freedom,} \\
\text{on each } T_j, \text{ evaluate the 2D degrees of freedom associated with the triangle,} \\
\text{on } H, \text{ integral of } u_{E_i} \text{ against } P^{r-2}(H) \text{ for } i = 1, \ldots, 6.
\end{array}
\right.
\]

The degrees of freedom can be implemented by evaluating $u_{E_i}$ at the center of the red tetrahedron in Figure 2.3 for $i = 1, \ldots, 6$. Again, all degrees of freedom associated with $H$ are interior to $H$. The first one shows up in degree $r = 2$. The pattern for further interior degrees of freedom are depicted in Figure 2.4.
The general pattern for $\text{REG}^r$ in dimension $n \geq 4$ is clear. A detailed description of this set of degrees of freedom can be found in Section 2.5.

The space $\text{REG}^r$ unifies and generalizes several discrete structures previously known in a wide variety of fields. The lowest degree element $\text{REG}^0$ in all dimensions $n \geq 1$, called \textit{Regge finite elements}, are well-known in the relativity, geometry, and finite element literature [23, 24, 27, 28, 95]. A review of its various interpretations is given in Section 2.1. In dimension $n = 2$, one can consistently rotate all the edge tangent vectors to normal vectors of the triangle. Under this, 2D $\text{REG}^r$ becomes the well-known Hellan-Herrmann-Johnson finite element [11, 18] for the biharmonic equation. This connection will be further studied and generalized in Chapter 2. In 2D, $\text{REG}^1$ is also equivalent to Pechstein-Schöberl’s lowest degree normal-normal stress finite element [87, 98, 99] for the linear elasticity equation. This connection will be further exploited to obtain stress elements for mixed elasticity in Chapter 3. Many further connections will be reviewed in the subsequent chapters.

There is another set of degrees of freedom for $\text{REG}^r$ which is equivalent to the one just described but is more geometrical and closer to the spirit of the original Regge finite element $\text{REG}^0$ used in Regge Calculus. This is most convenient to explain pictorially. In 2D, the original $\text{REG}^0$ assigns a number to each edge of the triangle. Each of these numbers has the meaning of the squared length of the corresponding edge measured by a constant metric. Here, $\text{REG}^r$ assigns a number to each of the small edges in the $r$-th subdivision of the triangle, depicted in Figure 2.5. Each of these numbers has the meaning of the squared lengths of the corresponding small edge measured using a polynomial degree $r$ metric. It
will be proven that this assignment of numbers determines a unique shape function of REG\(^r\).

Details of this set of degrees of freedom and its equivalence to the first one is derived in the second part of Section 2.5.

![Subdivision-based degrees of freedom in 2D for \(r = 0, 1, 2, \ldots\)](fig:subdivision_dof)

This chapter contains many fundamental results concerning properties of REG\(^r\). The first one is Theorem 2.2, which shows that the set of degrees of freedom for REG\(^r\) is unisolvent. Elements \(u\) of REG\(^r\) on a mesh have tangential-tangential continuity: for any simplex \(f\) of dimension \(\geq 1\) in the mesh and for any two vectors \(v\) and \(w\) parallel to \(f\), \(v^T u(x) w\) is single-valued at any \(x \in f\). This turns out to be the key property for its use in applications. The second main result of this chapter is Theorem 2.4, which states that this property in fact completely characterizes REG\(^r\), that is, any piecewise polynomial symmetric covariant 2-tensor fields of degree \(r\) or less has tangential-tangential continuity if and only if it belongs to REG\(^r\) on that mesh. The third main result is the locality property Theorem 2.3, which states that the tangential-tangential trace (the “value” which is guaranteed to be single-valued by the tangential-tangential continuity) on any face \(f\) in a mesh is completely determined by the degrees of freedom associated with that face \(f\). The fourth main result is Theorem 2.19 which shows that REG\(^r\) forms an affine family of finite elements for any fixed dimension \(\geq 1\). The locality and affine property together makes REG\(^r\) a canonical finite element with respect to tangential-tangential continuity in the language of [51], similar to many other well-known important finite elements. The fifth main result is the geometric decomposition Theorem 2.7. This seemingly technical result is of fundamental importance as it reveals the structure of REG\(^r\) and provides the tools for proving the unisolvency, characterization, and locality theorems just mentioned. The geometric decomposition further leads to the sixth main result: a Bernstein-style basis for REG\(^r\) described in Theorem 2.6. This is useful both theoretically and for software implementations. The seventh main result is Theorem 2.25, which proves the optimal approximation properties of the canonical interpolant of REG\(^r\). Finally, an elegant concrete implementation of REG\(^r\) is described in Section 2.5. This is the mathematical object underlies the software implementation of REG\(^r\) in the open source
software FEniCS by the author used in all the numerical experiments of this thesis.

The rest of this chapter is organized as follows. First, in Section 2.1, various historical interpretations of REG$^0$ are reviewed and compared. Then, in Section 2.2, REG$^r$ will be defined in a coordinate-free manner. Section 2.3 proves all the main theorems of REG$^r$ related to the geometric decomposition. Section 2.4 studies the affine properties of REG$^r$ and proves the optimal approximation theorems. This chapter ends with Section 2.5 describing two sets of useful coordinate choices for REG$^r$ with coordinate transformations and two sets of concrete degrees of freedom for REG$^r$. These are important for the software implementation of this finite element.

2.1 Interpretations of the classical Regge finite elements

The starting point is a triangulated surface like the one in Figure 2.6. Intuitively, geometric notions associated with smooth surfaces carry over. For example, this triangulated surface is apparently not flat, even though it is built from a finite number of flat triangles. This observation opens up the possibility of studying geometry on computers. One influential work in this direction was Regge Calculus [95] proposed by Tullio Regge in 1961, which is a discrete formulation of Einstein’s geometric theory for gravity. Regge elements are so named to acknowledge their roots in Regge Calculus.

Regge’s work exemplifies the geometrical view of Regge elements. The geometric object of interest is an $n$-dimensional polytope obtained by gluing together flat simplices along isometric boundary faces. For example, the surface in Figure 2.6 is a polygon obtained by gluing together flat triangles along edges of the same lengths (isometric edges). The study of triangulated polytopes in the Euclidean space is very old in mathematics as well. Since flat simplices are determined up to isometry by their edge lengths, they can be equivalently described as a polytope with a triangulation, that is, a mesh in numerical analysis, along with an assignment of lengths to the edges. This view is intuitive and transparent to implement on a computer. It remains the dominant view in the physics literature on numerical simulation.

Figure 2.6: A triangulated surface
using Regge Calculus (for a recent review, see [60]) and quantum gravity (for a recent review, see [114]). Such geometric objects are also called piecewise flat manifolds in differential topology [23].

The geometric view is perfectly fine as a discrete model of geometry on its own. However, it becomes inadequate when it is considered as an approximation to some smooth geometric object. The more advanced analytical view of Regge elements was first spelled out in detail in the work of Cheeger-Müller-Schrader [23, 24]. The main observation was that specifying the lengths of all edges is equivalent to prescribing on the polytope a piecewise constant metric such that shared faces are isometric. This “filling in”, interpolating numbers assigned to edges to a symmetric 2-tensor field on the entire polytope, leads to much more structure. For example, given a smooth surface and a sequence of triangulations, the quality of approximation by these piecewise constant metrics can be assessed by comparing them to the pullbacks of the smooth metric. Cheeger-Müller-Schrader further showed that notions of curvature carry over to these non-smooth metrics in a convergent manner, which will be examined in more detail in later chapters. The analytic view is frequently used in current mathematics literature on discrete geometry.

This idea of “filling in” interpolation to gain more structure is very powerful and has many parallels in the history of mathematics. A particularly relevant example is Whitney’s idea of interpolating simplicial cochains to piecewise linear differential forms [113]. This led to significant advances in differential topology and geometric measure theory. More important to applied mathematics, Whitney’s work led to the recent development of Finite Element Exterior Calculus (FEEC) [6, 8], which generalizes these piecewise linear forms to higher polynomial degrees and studies them in a Hilbert space framework. FEEC has been proven to play a central role in the numerical solution of differential equations in electromagnetism and solid mechanics. This thesis in a way tries to make the leap similar to the one from Whitney forms to FEEC for Regge calculus.

This leads to the finite element view of Regge elements pioneered by Christiansen [27, 28]. This view adds another layer of structure on top of the analytical one: Regge elements are not only just piecewise constant functions on a polytope, but also form a discrete Hilbert space which in a subtle sense discretizes a continuous Hilbert space, namely the function space of $L^2$-symmetric covariant 2-tensor fields with $H^{-1}$ distributional linearized Riemann curvature. This makes the rigorous numerical analysis possible. Indeed it is easier to study convergence in the Hilbert spaces context than to show an assignment of numbers to edges somehow converges to a smooth solution to a partial differential equation. This structure also makes it clear what properties need to be preserved when Regge's initial proposal is
generalized to polynomials of higher degrees. It is the author’s opinion that the finite element view of Regge elements is the most advantageous among the three views, at least for the purpose of numerical analysis.

### 2.2 Definition of generalized Regge family

In this section, the generalized Regge family is defined precisely. This is used as an opportunity to clarify the definitions of many basic concepts used in this thesis.

First, concepts related to the mesh are stated. In $\mathbb{R}^m$, the convex hull of $(n+1)$ points $\{v_0, \ldots, v_n\}$ of general position is called an $n$-simplex $c = [v_0, \ldots, v_n]$. Necessarily, $m \geq n$. This generalizes the notion of line segments (1-simplices), triangles (2-simplices), and tetrahedron (3-simplices) to all dimensions. Each $v_i$ is called a vertex of $c$. The convex hull of any $(k+1)$ vertices is a $k$-simplex by itself and is called a $k$-face of $c$. By convention, in an $n$-cell, 1-faces are also called edges, $(n-2)$-faces are also called bones, $(n-1)$-faces are also called facets, and $n$-faces are also called cells. A mesh $T$ is a finite collection of simplices in $\mathbb{R}^m$ satisfying:

- any face of a simplex in $T$ is a simplex in $T$,
- the intersection of any two simplices in $T$ is a face of both simplices,
- the union of all simplices in $T$ as a subset of $\mathbb{R}^m$ is a topological submanifold of dimension $n$.

The integer $m$ is called the geometric dimension of $T$ while $n$ is called the topological dimension of $T$. This nomenclature has its roots in the representation of a mesh on a computer as a list of coordinates for the vertices. In this thesis, mostly $m = n$. In this case $T$ is said to be a mesh of dimension $n$. Many alternative definitions of a mesh exist in the literature. This one is chosen for the ease and clarity of exposition and is developed from the definition of a geometric simplicial complex in [82, Section 7]. The manifold determined by $T$, called the carrier of $T$, is denoted by $|T|$. Let $M$ be any smooth manifold (possibly with boundary). If $T$ is a mesh with $|T|$ diffeomorphic to $M$, then $T$ is called a triangulation of $M$.

![Figure 2.7: Mesh and non-meshes in 2D](fig:mesh_non_mesh)

Second, some definitions from differential geometry are reviewed [111, Chapter 2]. Let $M$
be any smooth manifold (possibly with boundary). For any point $p \in M$, a covariant $k$-tensor at $p$ is a real $k$-linear form on the tangent space $T_pM$. A covariant $k$-tensor field on $M$ is a function on $M$ assigning to each point $p \in M$ a covariant $k$-tensor at $p$. A covariant 2-tensor field is called symmetric when its value at each point is a symmetric bilinear form. In this thesis, the space of all such symmetric covariant 2-tensor fields on $M$ is especially important and is denoted by $\mathcal{S}^m(M)$, where the superscript $m$ is used to emphasize the dimension $m$ of $M$. Let $N$ be another smooth manifold and $\phi : M \to N$ a smooth function. At every point $p \in M$, the differential $(d\phi)_p$ is a linear map from $T_pM$ to $T_{\phi(p)}N$ defined by the property that for any smooth $f : U \to \mathbb{R}$ on a neighborhood $U$ of $\phi(p)$ and any $v \in T_pM$:

$$(d\phi)_p(v)(f) := v(f \circ \phi).$$

This induces a map $(\phi^*)_p$ from covariant $k$-tensors at $\phi(p) \in N$ to covariant $k$-tensors at $p \in M$: for any covariant $k$-tensor $g$ at $\phi(p) \in N$ and any $k$ vectors $(u_1, \ldots, u_k)$ in $T_pM$,

$$(\phi^*)_p(g)(u_1, \ldots, u_k) := g(d\phi(u_1), \ldots, d\phi(u_k)).$$

Since $d\phi$ is well-defined over any point $p \in M$, any covariant $k$-tensor field $g$ on $N$ defines a covariant $k$-tensor field $\phi^*g$ on $M$ by applying $(\phi^*)_p$ in a pointwise fashion. This $\phi^*g$ is called the pullback of $g$ under $\phi$. In particular, $\phi^* : \mathcal{S}(N) \to \mathcal{S}^m(M)$ for any smooth $\phi$. Now let $c$ be an $n$-simplex and $f$ a $k$-face in $c$. Define $\iota_{f \to c}$ to be the inclusion of $f$ in $c$. In most situations, the cell $c$ is clear from the context and the notation is shortened to just $\iota_f$. By definition, for any $g \in \mathcal{S}(c)$, its pullback $\iota_f^*g \in \mathcal{S}^k(f)$ assigns to each point $p \in f$ a symmetric bilinear form on vectors tangent to $f$. Hence $\iota_f^*g$ is also called the tangential-tangential part of $g$ at face $f$ in the finite element literature [27, 28]. The term tangential-tangential part is preferred in this thesis to single out the pullback for covariant 2-tensors.

Third, some notations for polynomial spaces are introduced. For a $n$-simplex $c$, let $\mathcal{P}^r(c)$ be the space of polynomials of degree $r$ or less on $c$ as before. It is easy to show that [29, Equation (2.2.2)]:

$$\dim \mathcal{P}^r(c) = \binom{n + r}{n}.$$  

For the Euclidean space $\mathbb{R}^n$, the tangent space at different points are identified in a natural way and there is a canonical sense of constant vector fields (for the pedantic, take $\mathbb{R}^n$ with vector addition as a Lie group and then constant fields are left-invariant [111]). Any $n$-simplex $c$ is defined as a subset of the Euclidean space. So the notion of constant vector fields on $c$ is well-defined. Let $\mathbb{S}^n$ be the space of symmetric covariant 2-tensors at the origin in $\mathbb{R}^n$ and

$$\mathcal{P}^r \mathcal{S}(c) := \mathcal{P}^r(c) \otimes \mathbb{S}^n.$$  

10
Equivalently, $P^r S(c)$ can be characterized as the collection of all symmetric covariant 2-tensor fields on $c$ whose values on pairs of constant vector fields are polynomials of degree $r$ or less. The space $S^n$ is isomorphic to the space of symmetric $n$-by-$n$ matrices. Hence,

$$\dim S^n = \binom{n+1}{2}, \quad \dim P^r S(c) = \binom{n+r}{n} \binom{n+1}{2}. \quad \text{eq:dim_count}$$

Let $\mathcal{T}$ be any mesh of topological dimension $n$. Then the space of piecewise polynomial symmetric covariant 2-tensor fields of degree $r$ or less is defined by:

$$P^r \mathcal{I}(\mathcal{T}) := \{ u \in \mathcal{I}(\mathcal{T}) \mid \text{the restriction of } u \text{ to each cell } c \text{ of } \mathcal{T} \text{ is in } P^r S(c) \},$$

where $u$ might be multi-valued on cell boundaries. It should be noted that one cannot define $P^r \mathcal{I}(M)$ on a general smooth manifold $M$ because there is no canonical sense of constant vector fields on $M$. However, for any triangulation $\mathcal{T}$ of $M$, $P^r \mathcal{I}(\mathcal{T})$ is still well-defined.

Fourth, several notions related to finite elements used in this thesis are defined. A simplicial finite element is a triple $(c, V, \Sigma)$. The first component $c$ is a simplex. The second component $V$ is a finite-dimensional function space on $c$. The last component $\Sigma = \{(r_f, \Sigma_f)\}_{f \subset c}$ is a collection of ordered pairs indexed by faces $f$ of $c$, where for each face $f$, $r_f$ is a map from $V$ to some function space $V_f$ on $f$ and $\Sigma_f$ is a subspace of the dual space $V'_f$. This $\Sigma$ is further required to satisfy the unisolvency condition:

$$V' = \bigoplus_{f \subset c} \{(u \mapsto l(r_f(u))) \mid l \in \Sigma_f\}.$$

In a finite element $(c, V, \Sigma)$, the simplex $c$ is called the domain, elements of $V$ are called shape functions, and elements of $\Sigma$ are called degrees of freedom. When specifying a finite element, the unisolvency is usually the part which requires a non-trivial proof. This definition is based on the classical definition of Ciarlet [29, Section 2.3] with one difference. Traditionally, the set of degrees of freedom is simply given as a basis for the dual space. This, on one hand, specifies too much as their spans are enough to determine the crucial inter-element continuity properties of the finite element [6, Section 4]. For a particular software implementation, a basis $B_f$ can be fixed for each $\Sigma_f$. Then $\bigcup_f B_f$ leads to a dual basis which can be used to map an element of $V$ to a numeric array on a computer. This choice, however, does not affect its mathematical analysis. On the other hand, the classical definition does not make explicit the important aspect of finite elements that these basis are associated with faces of the simplex so they can be patched together on a mesh through the assembly process [8, Section 2.1].

A single finite element is rarely of any interest. A much more useful notion is a finite element family $F$, which is a function defined on a collection $D(F)$ of simplices and associates
to each simplex $c \in D(F)$ a finite element $F(c)$. Given a mesh $\mathcal{T}$, a finite element family $F$ is called *assemblable* on $\mathcal{T}$ if all the cells of $\mathcal{T}$ are in $D(F)$ and whenever two cells $c_1$ and $c_2$ intersect at a face $f$, both $F(c_1)$ and $F(c_2)$ give the same $r_f(V)$ and $\Sigma_f$ on $f$. In such a situation, a *finite element space* on $\mathcal{T}$, denoted by $F(\mathcal{T})$, can be obtained through the *finite element assembly process*: $F(\mathcal{T})$ is the collection of functions $u$ on $\mathcal{T}$ possibly multi-valued on cell boundaries such that:

- the restriction $u|_c$ to each cell $c$ is a shape function of $F(c)$,
- if any two cells $c_1$ and $c_2$ share a face $f$ then $l \circ r_f(u|_{c_1}) = l \circ r_f(u|_{c_2})$ for all $l \in \Sigma_f$.

![Figure 2.8: REG$^1$ assembly on a 3-triangle mesh.](image)

Finally, the *generalized Regge finite element* can be defined precisely. On an $n$-simplex $c$ in $\mathbb{R}^n$, the generalized Regge finite element of degree $r$, is given by the space of shape functions:

\begin{equation}
\mathcal{P}^r \mathcal{I}(c)
\end{equation}

and degrees of freedom assigned to each $k$-face $f$ of $c$ with $k \geq 1$:

\begin{equation}
r_f := t_f^r : \mathcal{P}^r \mathcal{I}(c) \to \mathcal{P}^r \mathcal{I}(f), \quad \Sigma_f := \{ u \to \int_f u : q \mid q \in \mathcal{P}^{r-k+1} \mathcal{I}(f) \},
\end{equation}

where the colon $:$ denotes the Frobenius inner product on $\mathbb{S}^k$. It will be proven in Theorem 2.2 that this set of degrees of freedom is unisolvent. In fact, any inner product on $\mathbb{S}^k$ can be used in place of the Frobenius one and the resulting finite element will be the same.

The dimensions of the space and the degrees of freedom are counted. The dimensions of $\mathcal{P}^r$ and $\mathcal{P}^r \mathcal{I}$ are already computed in (2.3). On one hand,

\[
\dim V = \dim \mathcal{P}^r \mathcal{I}(c) = \binom{n + r}{r} \binom{n + 1}{2}.
\]
On the other hand, it is an elementary count that the number of $k$-faces in an $n$-simplex is
\[
\#(k\text{-faces of } c) = \binom{n+1}{k+1}.
\]
Hence the total number of degrees of freedom is:
\[
\sum_{k=1}^{n} (\#(k\text{-faces of } c)) \left( \dim \mathcal{P}^{r-k+1} \mathcal{I}(f) \right) = \sum_{k=1}^{n} \binom{n+1}{k+1} \binom{r+1}{k} \binom{k+1}{2}.
\]
As a consequence of the unisolvency (Theorem 2.2), the following identity must hold:
\[
\binom{n+r}{r} \binom{n+1}{k+1} = \sum_{k=1}^{n} \binom{n+1}{k+1} \binom{r+1}{k} \binom{k+1}{2}.
\]  
**eq:unisolvency_count**

Identities of this form can be verified independently on a computer using Zeilberger’s Algorithm [90], which is implemented in many software packages. For example, in Wolfram Alpha (https://www.wolframalpha.com), typing the following line in the textbox verifies identity (2.5):
\[
\text{sum } \text{binom}(n+1,k+1) \times \text{binom}(k+1,2) \times \text{binom}(r+1,k) \text{ from } k=1 \text{ to } n
\]

To end the section, the generalized Regge finite element family of degree $r$, denoted by $\text{REG}_r$, is introduced. For any $r \geq 0$, it is a function defined on all simplices $c$ of dimension $n \geq 1$ which assigns the Regge finite element of degree $r$ on $c$ to each $c$.

**Theorem 2.1.** For any fixed $r \geq 0$, $\text{REG}_r$ is assemblable on any mesh $\mathcal{T}$ of topological dimension $n \geq 1$.

**Proof.** The conditions for assemblability are checked one by one. Each cell of $\mathcal{T}$ is of dimension $n$ so it is in the domain of $\text{REG}_r$. Suppose two cells $c_1$ and $c_2$ intersect at a $k$-face $f$. Then,
\[
\iota^*_f \circ c_1 \mathcal{P}^r \mathcal{I}(c_1) = \mathcal{P}^r \mathcal{I}(f) = \iota^*_f \circ c_2 \mathcal{P}^r \mathcal{I}(c_2).
\]
Finally, it is clear from the definition (2.4b) that $\Sigma_f$ only depends on $f$ and hence is the same in both $\text{REG}_r(c_1)$ and $\text{REG}_r(c_2)$. Hence $\text{REG}_r$ is assemblable on $\mathcal{T}$. \hfill $\square$

The resulting assembled space is denoted by $\text{REG}_r(\mathcal{T})$. This finite dimensional functional space plays a key role in the rest of this thesis.
2.3 Basic properties

The goal of this section is to establish several important results on generalized Regge elements. The first one is

**Theorem 2.2 (Unisolvency).** The set of degrees of freedom (2.4b) is unisolvent.

The second result shows that $\text{REG}^r$ satisfies the **locality property** defined in [51]. In the language of [51], this, along with the affine properties in the next section, implies that $\text{REG}^r$ is a canonical finite element for tangential-tangential continuous symmetric covariant 2-tensor fields.

**Theorem 2.3 (Locality).** Let $f$ be a $k$-face in an $n$-simplex $c$ with $k \geq 1$ and $u \in \mathcal{P}^r \mathcal{I}(c)$. Then $i_f^* u$ is completely determined by the subset of degrees of freedom associated with $f$ defined in (2.4b).

The third result characterizes $\text{REG}^r$ as a special subspace of piecewise polynomial symmetric covariant 2-tensor fields:

**Theorem 2.4 (Characterization).** Let $\mathcal{T}$ be a mesh of topological dimension $n \geq 1$. Suppose $u \in \mathcal{P}^r \mathcal{I}(\mathcal{T})$ is a piecewise polynomial covariant symmetric 2-tensor. Then $u \in \text{REG}^r(\mathcal{T})$ if and only if $i_f^* u$ is single-valued at each interior $k$-face $f$ of the mesh with $k \geq 1$.

On an simplex $c$, the vector space of tangential-tangential bubble functions $\mathcal{P}^r_\partial \mathcal{I}(c)$ is defined as

$$\mathcal{P}^r_\partial \mathcal{I}(c) := \{ u \in \mathcal{P}^r \mathcal{I}(c) | i_{\partial c}^* u = 0 \}.$$

The fourth result concerns the structure of these bubble functions.

**Theorem 2.5.** Let $c$ be an $n$-simplex. The dual space to the tangential-tangential bubble functions is:

$$[\mathcal{P}^r_\partial \mathcal{I}(c)]' = \left\{ u \mapsto \int_c u : q \in \mathcal{P}^{r-n+1} \mathcal{I}(c) \right\}.$$

In particular,

$$\dim \mathcal{P}^r_\partial \mathcal{I}(c) = \dim \mathcal{P}^{r-n+1} \mathcal{I}(c) = \binom{n}{2} \binom{r+1}{n}.$$

It turns out that the most fundamental theorem on $\text{REG}^r$ is a technical result called the **geometric decomposition.** It is most convenient to state and prove this via a **Bernstein-style basis**, that is, a basis for $\text{REG}^r$ indexed by faces of a simplex. Similar decompositions are useful both theoretically [7] and for software implementations [62]. To state these two, some preparations are needed.
Fix an arbitrary \( n \)-simplex \( c = [v_0 \ldots v_n] \). Let \((\lambda_j)_{j=0}^n\) be the barycentric coordinates on \( c \); these \( \lambda_i \) are linear functions determined by \( \lambda_i(v_j) = \delta_{ij} \). A **multi-index** \( \alpha \in \mathbb{N}^{0,n} \) is an array \( \alpha = (\alpha_0, \ldots, \alpha_n) \) of \( (n + 1) \) non-negative integers \( \alpha_i \geq 0 \). Set:

\[ \lambda^\alpha := \lambda_0^{\alpha_0} \cdots \lambda_n^{\alpha_n}, \quad |\alpha| := \sum_{i=0}^{n} \alpha_i. \]

The **support** \( [\alpha] \) of \( \alpha \) is defined to be:

\[ [\alpha] := \{ i \in \{0, \ldots, n\} | \alpha_i \geq 1 \} \tag{2.6} \]

For a face \( f = [v_{f_0} \ldots v_{f_k}] \), the **index set** \( I(f) \) contains the indices for the vertices of \( f \):

\[ I(f) := \{ f_0, \ldots, f_k \} \tag{2.7} \]

Under this notation, the (un-normalized) **Bernstein basis** \( B^r(c) \) for \( \mathcal{P}^r(c) \) is given by [73]:

\[ B^r(c) := \{ \lambda^\alpha | \alpha \in \mathbb{N}^{0,n}, |\alpha| = r \}. \]

Given two 1-forms \( l_1, l_2 \) on \( c = [v_0 \ldots v_n] \), their **symmetric tensor product** \( l_1 \otimes l_2 \) is a symmetric covariant 2-tensor given by

\[ (g_1 \otimes g_2)(u_1, u_2) := \frac{1}{2} [g_1(u_1)g_2(u_2) + g_1(u_2)g_2(u_1)] \]

for all pairs of vectors \( u_1 \) and \( u_2 \). On \( c \), each edge \( [v_i v_j] \) can be associated with a covariant 2-tensor:

\[ d\lambda_i \otimes d\lambda_j \in \mathbb{S}^n. \]

Note that because \( \{\lambda_i\} \) are linear functions, their differentials are constants. In the above, the usual identification of constants and constant functions are assumed. Due to the tensor product structure (2.2) of \( \mathcal{P}^r \mathcal{S}(c) \), for any \( p \in \mathcal{P}^r(c) \) and any edge \( [v_i v_j] \) of \( c \), \( p d\lambda_i \otimes d\lambda_j \in \mathcal{P}^r \mathcal{S}(c) \).

The following result gives bases for \( \mathcal{P}^r \mathcal{S}(c) \) and \( \mathcal{P}^r \mathcal{F}(c) \) in terms of barycentric coordinates.

**Theorem 2.6** (Basis). Let \( c = [v_0 \ldots v_n] \) be an \( n \)-simplex and \( \{\lambda_j\}_{j=0}^n \) be the barycentric coordinates. Define \( e(c) \) to be the collection of all the edges of \( c \). Then,

\[ B^r_{\mathcal{F}}(c) := \{ \lambda^\alpha d\lambda_i \otimes d\lambda_j \mid \alpha \in \mathbb{N}^{0,n}, |\alpha| = r, [v_i v_j] \in e(c) \} \]

forms a basis for \( \mathcal{P}^r \mathcal{F}(c) \) and

\[ B^r_{\mathcal{S}}(c) := \{ \lambda^\alpha d\lambda_i \otimes d\lambda_j \mid [v_i v_j] \in e(c), \alpha \in \mathbb{N}^{0,n}, |\alpha| = r, [\alpha] \cup \{i, j\} = I(c) \} \]

forms a basis for \( \mathcal{P}^r \mathcal{S}(c) \).
To state the geometric decomposition, an extension operator is needed to take polynomials on a face of a simplex to the simplex. Let \( f = [v_{f_0} \ldots v_{f_k}] \) be any \( k \)-face of \( c \). This \( f \) is a \( k \)-simplex on its own and has its own barycentric coordinates \( \{ \lambda^f_{i, f} \}_{i=0}^k \). Further, \( \mathcal{B}^r(f) \) defined using these \( \{ \lambda^f_{i, f} \} \) forms a basis for \( \mathcal{P}^r(f) \). There is a canonical map \( \mathcal{E}^r_{f\to c} : \mathcal{P}^r(f) \to \mathcal{P}^r(c) \), called the barycentric extension [7, Section 2.2], which simply replaces any appearance of \( \lambda^f_{i, f} \) with \( \lambda^c_{i, c} \) in the expansion of any \( \mathcal{P}^r(f) \) in basis \( \mathcal{B}^r(f) \). For example, if \( f = [v_0v_1v_2] \), then

\[
\mathcal{E}^2_{f\to c}(\lambda^f_{1, f} \lambda^f_{2, f}) = \lambda^c_{1, c} \lambda^c_{2, c}.
\]

The dependency of \( \mathcal{E}^r \) on \( r \) is significant. Recall that \( \sum \lambda_i = 1 \). So a polynomial of degree \( k \) has a different representation in \( \mathcal{B}^r \) for each \( r \geq k \). In the above, if \( \mathcal{E}^3_{f\to c} \) were applied, the result becomes

\[
\mathcal{E}^3_{f\to c}(\lambda^f_{1, f} \lambda^f_{2, f}) = \mathcal{E}^3_{f\to c}(\lambda^f_{1, f} \lambda^f_{2, f}(\lambda^f_0 + \lambda^f_1 + \lambda^f_2)) = \lambda^c_0 \lambda^c_1 \lambda^c_2 + (\lambda^c_1)^2 \lambda^c_2 + \lambda^c_1 (\lambda^c_2)^2,
\]

which is a cubic polynomial on \( c \).

For any face \( f = [v_{f_0} \ldots v_{f_k}] \) of an \( n \)-simplex \( c = [v_0 \ldots v_n] \), the barycentric extension \( \mathcal{E}^r_{f\to c} \) can be extended to a map from \( \mathcal{P}^r(f) \) to \( \mathcal{P}^r(c) \) naturally via the basis given in Theorem 2.6. For example, if \( f = [v_0v_1v_2] \) then for any edge \([v_1v_2]\) of \( f \),

\[
\mathcal{E}^r_{f\to c}[\lambda^f_0(\lambda^f_1)^2 \lambda^f_2 d \lambda^f_0 \circ d \lambda^f_1] := \lambda^c_0(\lambda^c_1)^2 \lambda^c_2 d \lambda^c_1 \circ d \lambda^c_1.
\]

The geometric decomposition theorem decomposes various spaces associated with a simplex according to faces of that simplex:

\[\text{thm:geometric_decomposition}\]

**Theorem 2.7** (Geometric decomposition). Let \( c \) be a simplex. Then,

\[
\mathcal{P}^r(f) = \bigoplus_{\dim f \geq 1} \mathcal{E}^r_{f\to c} \mathcal{P}^r(f).
\]

The basis decomposes accordingly:

\[
\mathcal{B}^r(f) = \bigcup_{\dim f \geq 1} \mathcal{E}^r_{f\to c} \mathcal{B}^r(f)
\]

where the union is disjoint. Moreover, the dual space also decomposes geometrically:

\[
[\mathcal{P}^r(f)]' = \bigoplus_{\dim f \geq 1} \left\{ u \to \int_f (v^*_f u) : q \in \mathcal{P}^{-\dim f + 1}(f) \right\}.
\]

For example, the unisolvency (Theorem 2.2) is a direct consequence of the dual space decomposition. Examples of this geometric decomposition of the basis in 2D and 3D are listed in Table 2.1 and Table 2.2.
TABLE 2.1: Bernstein-style Basis in 2D

<table>
<thead>
<tr>
<th>( r )</th>
<th>([v_i v_j])</th>
<th>([v_i v_j v_k])</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( d\lambda_j \odot d\lambda_j )</td>
<td>( d\lambda_j \odot d\lambda_k ), ( d\lambda_j \odot d\lambda_j )</td>
</tr>
<tr>
<td>1</td>
<td>( \lambda_i d\lambda_i \odot d\lambda_j ), ( \lambda_j d\lambda_i \odot d\lambda_j )</td>
<td>( \lambda_j d\lambda_i \odot d\lambda_k ), ( \lambda_i d\lambda_i \odot d\lambda_j )</td>
</tr>
<tr>
<td>2</td>
<td>( \lambda_i^2 d\lambda_i \odot d\lambda_j ), ( \lambda_j^2 d\lambda_j \odot d\lambda_k ), ( \lambda_k^2 d\lambda_i \odot d\lambda_j )</td>
<td>( \lambda_i \lambda_j d\lambda_i \odot d\lambda_j ), ( \lambda_j \lambda_k d\lambda_i \odot d\lambda_k ), ( \lambda_k \lambda_j d\lambda_i \odot d\lambda_k )</td>
</tr>
</tbody>
</table>

Table 2.1: Bernstein-style Basis in 2D

<table>
<thead>
<tr>
<th>( r )</th>
<th>([v_i v_j])</th>
<th>([v_i v_j v_k])</th>
<th>([v_i v_j v_k v_l])</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Same as 2D, 1 per edge</td>
<td>Same as 2D, 3 per triangle</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Same as 2D, 2 per edge</td>
<td>Same as 2D, 3 per triangle</td>
<td>( \lambda_i \lambda_j d\lambda_i \odot d\lambda_l ), ( \lambda_j \lambda_k d\lambda_i \odot d\lambda_i ), ( \lambda_k \lambda_l d\lambda_i \odot d\lambda_l ), ( \lambda_l \lambda_i d\lambda_i \odot d\lambda_k )</td>
</tr>
</tbody>
</table>

Table 2.2: Bernstein-style Basis in 3D

The proofs of these theorems are nontrivial and are divided into many steps. In the first subsection, relevant results on the Bernstein decomposition of the scalar \( \mathcal{P}^r \) is reviewed. In the second subsection, results on \( \text{REG}^r \) are proved one-by-one.

### 2.3.1 Bernstein decomposition for Lagrange elements

In order to derive the geometric decomposition of \( \mathcal{P}^r \mathcal{S}(c) \), related concepts and results for the scalar polynomial space \( \mathcal{P}^r(c) \) have to be reviewed first.

As mentioned in the introduction of this section, on an \( n \)-simplex \( c \), the un-normalized \textit{Bernstein basis} \( B^r(c) \) for \( \mathcal{P}^r(c) \) is given by [73]:

\[
B^r(c) := \{\lambda^\alpha | \alpha \in \mathbb{N}_0^n, |\alpha| = r\}.
\]

The normalization factor is dropped because it is not important for the discussion here.
It should be noted that the normalization, however, is important in software implementa-
tions [71, Chapter 4]. This basis has many advantages numerically, for example see [62].

The Bernstein basis has the following elementary property:

**Lemma 2.8.** Suppose \( p \in \mathcal{P}^r(c) \) is divisible by \( \lambda_k \) for some \( k \), then in the expansion of \( p \) in the Bernstein basis each summand contains a \( \lambda_k \) factor individually.

**Proof.** Suppose the claim is false. Then necessarily, a linear combination of
\[
\{ \lambda_0^a_0 \cdots \lambda_n^a_n \mid a \in \mathbb{N}_0^n, |a| = r, a_k = 0 \}
\]
equals \( \lambda_k q \) for some polynomial \( q \) of degree less than or equal to \((r-1)\). But this \( q \) can be represented in the Bernstein basis \( \mathcal{B}^{r-1}(c) \) again. In particular, each term of the thus expanded product in \( \lambda_k q \) is a basis element in \( \mathcal{B}'(c) \) and contains a \( \lambda_k \) factor. Thus a linear combination of elements in \( \mathcal{B}'(c) \) which do not contain the factor \( \lambda_k \) equals a linear combination of elements in \( \mathcal{B}'(c) \) which do contain the factor \( \lambda_k \). This contradicts the fact that elements of \( \mathcal{B}'(c) \) are linearly independent. \( \square \)

Let the support \( [\cdot] \) and index set \( \mathcal{I}(\cdot) \) be defined as in (2.6) and (2.7) respectively. For any face \( f \) of an \( n \)-simplex \( c \), set
\[
\mathcal{B}'_c(f) := \{ \lambda^a \mid a \in \mathbb{N}_0^n, |a| = r, [a] = \mathcal{I}(f) \}
\]
which is the subset of the Bernstein basis \( \mathcal{B}'(c) \) whose factors involve exactly \( \lambda_i \) associated with vertices of \( f \). It is clear that every element of \( \mathcal{B}'(c) \) is in a unique \( \mathcal{B}'_c(f) \) for some face \( f \) of \( c \). The trivial observation that the map \([v_{f_0} \ldots v_{f_k}] \mapsto \{v_{f_0} \ldots v_{f_k}\} \) is a bijection between faces of \( c \) and subsets of vertices of \( c \) implies:
\[
\mathcal{B}'(c) = \bigcup_{f \subset c} \mathcal{B}'_c(f),
\]
where the union is disjoint and is taken over all faces \( f \) of \( c \). Elements of \( \mathcal{B}'_c(f) \) vanishes on the boundary of \( f \). This can be made more local by introducing *bubble functions* on any simplex \( f \):
\[
\hat{\mathcal{P}}^r(f) := \{ p \in \mathcal{P}^r(f) \mid p|_{\partial f} = 0 \}.
\]
It turns out that following is a basis for \( \hat{\mathcal{P}}^r(f) \):
\[
\hat{\mathcal{B}}'(c) := \{ \lambda^a \mid a \in \mathbb{N}_0^n, |a| = r, [a] = \mathcal{I}(c) \},
\]
Indeed, each \( \lambda_i \) vanishes on the facet opposite to vertex \( v_i \) and every facet is opposite to a vertex, every element of \( \hat{\mathcal{B}}'(c) \) vanishes on the boundary of \( c \) and is thus in \( \hat{\mathcal{P}}^r(c) \). On the
other hand, if \( p \in \mathcal{P}(c) \), then \( p \) is divisible by \( \lambda_0 \cdots \lambda_n \). By Lemma 2.8, this implies that every term in the expansion of \( p \) in the Bernstein basis is in \( \mathcal{B}(c) \). Hence, \( \mathcal{B}(c) \) is a basis for \( \mathcal{P}(c) \) (also derived in [7, Equation (2.4)]).

Comparing formulae (2.9) and (2.10), it is clear that for any face \( f \) of a simplex \( c \):

\[
B_c^r(f) = E^r \mathcal{B}(f).
\]

Hence,

\[
\mathcal{B}(c) = \bigcup_{f \in c} E^r \mathcal{B}(f), \quad \mathcal{P}(c) = \bigoplus_{f \in c} E^r \mathcal{P}(f).
\]

This is called the Bernstein decomposition in [7]. An example of this for \( \mathcal{P}^3 \) on a triangle is shown in Figure 2.9.

This decomposition is useful for software implementations [43]. Moreover, this gives an elegant proof of unisolvency of degrees of freedoms for Lagrange elements. It is clear that the map from \( \mathcal{P}^{r-n-1}(c) \) to \( \mathcal{P}(c) \) given by

\[
p \mapsto p \lambda_0 \cdots \lambda_n
\]

is an isomorphism. Thus the space of functionals

\[
\left\{ p \mapsto \int_c p q \mid q \in \mathcal{P}^{r-n-1}(c) \right\}
\]

is isomorphic to the dual space \( [\mathcal{P}(c)]' \). The Bernstein decomposition then implies that

\[
[\mathcal{P}(c)]' = \bigoplus_{f \in c} \left\{ p \mapsto \int_f p q \mid q \in \mathcal{P}^{r-\dim f-1}(f) \right\}.
\]

This is known to be the degrees of freedom for Lagrange finite elements [7].
2.3.2 Geometric decomposition for Regge elements

This subsection is subtle. The main idea is to derive the Bernstein-style basis (Theorem 2.6) for \( \text{REG} \) first. This is used to give a constructive proof of the geometric decomposition (Theorem 2.7) and the bubble characterization (Theorem 2.5). Then all the other theorems in the introduction of this section follow. A road map is provided below for the reader:

1. Derive a basis for \( S^n \) in terms of barycentric coordinates on an \( n \)-simplex (Theorem 2.9).
2. Derive the Bernstein-style basis \( B^r_{ij}(c) \) for \( P^r \mathcal{I}(c) \).
3. Derive the action of pullback on basis elements (equation (2.16) and Theorem 2.11) and establish that \( B^r_{f \rightarrow c} \) is the right inverse of the pullback.
4. Derive the Bernstein-style basis \( \hat{B}^r_{ij}(c) \) for \( \hat{P}^r \mathcal{I}(c) \). The key was the basis \( B^r_{ij}(c) \subset B^r(c) \) associated to each edge \([v_i v_j] \) of \( c \) for polynomials vanishing on all facets containing that edge (Lemma 2.13 and Lemma 2.14). This combined with the second step proves Theorem 2.6.
5. Derive the geometric decomposition of \( P^r \mathcal{I}(c) \) (first part of Theorem 2.7). This is a constructive proof based on an edge-based Bernstein decomposition of \( P^r(c) \) (Lemma 2.16 and Lemma 2.17).
6. Prove the bubble characterization (Theorem 2.5) via an explicit bijection using the Bernstein-style basis. Then prove the dual geometric decomposition (Lemma 2.18). This along with the previous step proves Theorem 2.7.
7. Prove Theorem 2.2–2.4 as corollaries.

The first step is to recall the well-known connection [27, Proposition 3.2] between the barycentric coordinates and the space of piecewise constant symmetric covariant 2-tensors.

**Theorem 2.9.** Let \( c = [v_0 \ldots v_n] \) be an \( n \)-simplex, \( \{\lambda_i\}_{i=0}^n \) its barycentric coordinates, and \( e(c) \) the collection of all edges of \( c \). Then the set

\[
\{d\lambda_i \circ d\lambda_j | [v_i v_j] \in e(c)\}
\]

forms a basis for \( \mathbb{S}^n = \mathcal{P}^0 \mathcal{I}(c) \) under the identification of constants with constant functions.

**Proof.** First, because each \( \lambda_i \) is linear, \( d\lambda_i \) is constant. So is the symmetric tensor product \( d\lambda_i \circ d\lambda_j \). Hence \( d\lambda_i \circ d\lambda_j \in \mathcal{P}^0 \mathcal{I}(c) \). Second, elementary dimension counts show that

\[
\dim \mathcal{P}^0 \mathcal{I}(c) = \binom{n+1}{2} = \#(e(c)).
\]

Thus the only thing left to show is that these \( d\lambda_i \circ d\lambda_j \) are linearly independent. Note that the simplex \( c \) is in some Euclidean space and inherits its affine structure. For any two points \( p \) and \( q \) in \( c \), \( p - q \) can be identified as a constant vector field on \( c \). For a linear function \( f \),
the action of \((p - q)\) as a derivation is just:

\[(p - q)(f) = f(p) - f(q).\]

Hence, by definition of \(\lambda_i\):

\[d\lambda_i(v_j - v_k) = (v_j - v_k)(\lambda_i) = \lambda_i(v_j) - \lambda_i(v_k) = \delta_{ij} - \delta_{ik}.\]  

Then direct computation shows: for any \([v_k, v_l] \in e(c)\),

\[(d\lambda_i \circ d\lambda_j)(v_k - v_l, v_k - v_l) = (\delta_{lk} - \delta_{ij})(\delta_{jk} - \delta_{jl})
\[= \begin{cases} -1, & \text{if } (i = k \text{ and } j = l) \text{ or } (i = j \text{ and } k = l) \\ 0, & \text{otherwise.} \end{cases}\]

Thus the span of the given set has dimension at least \(\#(e(c))\). This proves the linear independence and hence the claim.

Let \(c\) be an \(n\)-simplex. By the tensor product structure (2.2), the previous theorem and the Bernstein basis (2.8) together imply that

\[B^r_{\mathcal{I}}(c) := \{ p d\lambda_i \circ d\lambda_j : p \in B^r(c), [v_i, v_j] \in e(c) \}
= \{ \lambda^\alpha d\lambda_i \circ d\lambda_j : \alpha \in \mathbb{N}_0^n, |\alpha| = r, [v_i, v_j] \in e(c) \} \]
forms a basis for \(\mathcal{P}^r\mathcal{I}(c)\). For any face \(f\) of \(c\), the pullback \(i_f^*\) is linear. Its action on equation basis element is just:

\[i_f^*(p d\lambda_i \circ d\lambda_j) = (p \circ i_f)(i_f^*(d\lambda_i \circ d\lambda_j)).\]  

Hence, the tensor part \(d\lambda_i \circ d\lambda_j\) and the polynomial part can be dealt with separately.

**Lemma 2.10.** Let \(c\) be any simplex and \(f\) any face of \(c\). For any \(u \in \mathcal{P}^r\mathcal{I}(f)\),

\[i_f^*B^r_{\mathcal{I}} \rightarrow u = u.\]

**Proof.** Suppose \(c = [v_0 \ldots v_n]\) and \(f = [v_{f_0} \ldots v_{f_k}]\). Let \(\{\lambda_i^c\}\) and \(\{\lambda_i^f\}\) be the barycentric coordinates on \(c\) and \(f\) respectively. Since \(f\) determines a unique affine subspace of the Euclidean space where \(c\) is in, there is a canonical identification the tangent space of \(f\) as a subspace of tangent space of \(c\). Under this, both the inclusion \(i_f\) and the differential \(d i_f\) are just identity maps. In particular,

\[\lambda_i^c \circ i_f = \lambda_i^f, \quad i_f^*(d\lambda_i^c) = d\lambda_i^f.\]

Hence \(i_f^*B^r_{\mathcal{I}} \rightarrow c\) is the identity on the basis elements in \(B^r_{\mathcal{I}}(f)\). Both maps are linear so this extends to \(\mathcal{P}^r\mathcal{I}(f)\). This proves the claim.

\[\square\]
The following theorem collects the key properties of the pullback of \(d\lambda_i \odot d\lambda_j\):

**Theorem 2.11.** Let \(c\) be a simplex, \(f\) any \(k\)-face of \(c\), and \([v_i v_j] \in e(c)\) an edge. Then \(\iota_f^* (d\lambda_i \odot d\lambda_j) \neq 0\) if and only if the edge \([v_i v_j]\) is part of \(f\). Further, the set

\[
\{ \iota_f^* (d\lambda_l \odot d\lambda_m) \mid [v_l v_m] \in e_c(f) \}
\]

forms a basis for \(\mathcal{P}^0 \mathcal{S}(f) = \mathbb{S}^k\), where \(\lambda_i\) are barycentric coordinates of \(c\) and \(e_c(f)\) is the collection of all edges of \(c\) contained in \(f\).

**Proof.** If the edge \([v_i v_j]\) is not part of \(f\), then either vertex must be outside of \(f\). Without loss of generality, say \(v_i\) is not in \(f\). Then, from the calculation (2.13), \(d\lambda_i\) vanishes on all tangent vectors of \(f\). Hence \(\iota_f^* (d\lambda_i \odot d\lambda_j)\) vanishes. On the other hand, if the edge \([v_i v_j]\) is part of \(f\), then equation (2.14) implies that \((d\lambda_i \odot d\lambda_j)(v_i - v_j, v_i - v_j) = -1\). So \(\iota_f^* (d\lambda_i \odot d\lambda_j)\) cannot vanish in this case. Moreover this shows that the elements of the set in the further part of the claim are linearly independent. Then the same dimension count (2.12) implies that that set forms a basis for \(\mathcal{P}^0 \mathcal{S}(f)\). \(\square\)

**Corollary 2.12.** Let \(c\) be an \(n\)-simplex and \(p \in \mathcal{S}(c)\) a function of the form

\[
p := \sum_{[v_i v_j] \in e(c)} p_{ij} d\lambda_i \odot d\lambda_j,
\]

where \(p_{ij} : c \to \mathbb{R}\) are arbitrary functions and the sum is over all edges of \(c\). Then the pullback to the boundary \(\iota_{\partial c}^* p\) vanishes if and only if the pullback to the boundary of every term in the sum vanishes individually.

**Proof.** Let \(f\) be any boundary facet of \(c\). Due to the tensor product structure,

\[
\iota_f^* p := \sum_{[v_i v_j] \in e(c)} p_{ij} \iota_f^* (d\lambda_i \odot d\lambda_j).
\]

For terms associated with edges \([v_i v_j]\) not contained in \(f\), the tensor part \(\iota_f^* (d\lambda_i \odot d\lambda_j)\) always vanishes. For terms associated with edges \([v_i v_j]\) contained in \(f\), by the second part of Theorem 2.11, these \(\iota_f^* (d\lambda_i \odot d\lambda_j)\) forms a basis for \(\mathcal{P}^0 \mathcal{S}(f)\). Hence each corresponding \(p_{ij}\) must vanish. Thus, for different boundary facets \(f\), the pullback of each summand vanishes individually for different reasons (either of the two mentioned here). Nevertheless, overall, the pullback of each summand to the boundary must vanish individually. \(\square\)

The next step is find a basis for the tangential-tangential bubble space \(\mathcal{P}^r \mathcal{S}(f)\). Let \(c = [v_0 \ldots v_n]\) be an \(n\)-simplex and \([v_i v_j] \in e(c)\) an edge. In light of Theorem 2.11 and the
pullback formula \((2.16)\), a basis element \(p d \lambda_i \otimes d \lambda_j \in \mathcal{D}^r \mathcal{I}(f)\) if \(p\) vanishes on the part of the boundary which does not contain the edge \([v_i v_j]\). More precisely, in an \(n\)-simplex \(c\), there are \((n + 1)\) facets (all facets are boundary facets in a simplex), of which exactly 2 facets, the one opposite to \(v_i\) and the one opposite to \(v_j\), do not contain the edge \([v_i v_j]\) and all the rest of the \((n − 1)\) facets contain that edge. Define \(B'_{ij}(c)\) to be the collection of elements of the Bernstein basis \(B'(c)\) which vanish on all the \((n − 1)\) facets which do contain \([v_i v_j]\).

**Lemma 2.14.** Let \(c\) be an \(n\)-simplex. Then

\[
\hat{B}'_r(c) := \{ p d \lambda_i \otimes d \lambda_j \mid p \in B'_{ij}(c), [v_i v_j] \in e(c) \}
\]

\[
= \{ \lambda^a d \lambda_i \otimes d \lambda_j \mid [v_i v_j] \in e(c), \alpha \in \mathbb{N}^{0,n}, |\alpha| = r, [\alpha] \cup \{i, j\} = I(c) \}
\]

forms a basis for bubbles \(\mathcal{D}^r \mathcal{I}(c)\).

**Proof.** First, the preceding discussion showed that every element of \(\hat{B}'_r(c)\) is in \(\mathcal{D}^r \mathcal{I}(c)\). Second, because both \(B'_{ij}(c)\) and \(\{d \lambda_i \otimes d \lambda_j \mid [v_i v_j] \in e(c)\}\) are linearly independent sets, elements in \(\hat{B}'_r(c)\) as their product are also linearly independent by the tensor product structure \((2.2)\). Finally, suppose \(q \in \mathcal{D}^r \mathcal{I}(c)\). Expand \(q\) in basis \(B'_{ij}(c)\) defined in \((2.15)\):

\[
q = \sum_{[v_i v_j] \in e(c)} \sum_{p \in B'_{ij}(c)} q_{p, i, j} p d \lambda_i \otimes d \lambda_j.
\]
By Corollary 2.12, for any edge \([v_i v_j] \in e(c)\), each
\[
\iota^*_c \left( \sum_{p \in B(c)} q_{p,i,j} p d\lambda_i \odot d\lambda_j \right) = \iota^*_c (d\lambda_i \odot d\lambda_j) \sum_{p \in B(c)} q_{p,i,j} (p \circ \iota_c) = 0.
\]
By Theorem 2.11, the polynomial
\[
\sum_{p \in B(c)} q_{p,i,j} (p \circ \iota_c)
\]
must vanish on the \((n-1)\) boundary facets containing edge \([v_i v_j]\). By Lemma 2.13, this is in the span of \(B'_{ij}(c)\). Thus the linear span of \(B'_{ij}(c)\) contains \(\mathcal{S}'(c)\). This proves the claim. □

The next step is to derive another geometric decomposition of the Bernstein basis \(B'(c)\) which are based on edges. Let \(c = [v_0 \ldots v_n]\) be an \(n\)-simplex and \([v_i v_j]\) be an edge. For any \(k\)-face \(f\) of \(c\) with \(k \geq 1\), let
\[
B_{c,i,j}^r(f) := \{ \lambda^\alpha d\lambda_i \odot d\lambda_j | \alpha \in \mathbb{N}^{0,n}, |\alpha| = r, [\alpha] \cap \{i,j\} = R(f) \},
\]
where the barycentric coordinates \(\{\lambda_i\}\) are for the cell \(c\).

Figure 2.10: Edge-based Bernstein decomposition for \(\mathcal{P}^3\) on a triangle. The chosen edge is in red. Basis associated with edges are in black while those associated with cells are in blue.

Figure 2.11: Edge-based Bernstein decomposition for \(\mathcal{P}^3\) on a tetrahedron. The chosen edge is thickened. Basis associated with edges are in red, those associated with triangles are in blue, and those associated with cells are in black.
Lemma 2.15. Let $c = [v_0 \ldots v_n]$ be an $n$-simplex. Then, for any fixed edge $[v_i v_j]$ of $c$,

$$B'(c) = \bigcup_{f \ni [v_i v_j]} B^r_{c,ij}(f),$$

where the union is disjoint and is taken over all faces $f$ of $c$ containing edge $[v_i v_j]$.

Proof. It is clear that each $B^r_{c,ij}(f)$ is a subset of the Bernstein basis $B'(c)$. Let the edge $[v_i v_j]$ be fixed. The condition $\alpha \cap \{i, j\} = I(f)$ implies that for different $f$, these $B^r_{c,ij}(f)$ are disjoint. On the other hand, suppose $p = \lambda^a$ is any element of $B'(c)$. Then, let $f$ be the face of $c$ determined by the vertices $\alpha \cup \{i, j\}$. It is clear that $p \in B^r_{c,ij}(f)$. Hence the union covers $B'(c)$. This proves the claim. \hfill \Box

Let $c$ be an $n$-simplex and $f$ any $k$-face of $c$ with $k \geq 1$. Further, let $E_{f \rightarrow c}^r$ be the barycentric extension defined before. Comparing the formula for $B^r_{c,ij}(f)$ in Lemma 2.13 with the definition of $B^r_{c,ij}(f)$ in equation (2.17), whenever $f$ contains the edge $[v_i v_j]$, clearly,

$$B^r_{c,ij}(f) = E_{f \rightarrow c}^r B^r_{i j}(f).$$

Therefore, there is an edge based geometric decomposition of the Bernstein basis:

Lemma 2.16. Let $c = [v_0 \ldots v_n]$ be an $n$-simplex. Then, for any edge $[v_i v_j]$ of $c$,

$$B'(c) = \bigcup_{f \ni [v_i v_j]} E_{f \rightarrow c}^r B^r_{c,ij}(f),$$

where the union is disjoint and is taken over all faces $f$ of $c$ containing edge $[v_i v_j]$.

This decomposition of the Bernstein basis leads to the desired geometric decomposition of the Regge finite element basis $B^r_{c}(c)$.

Lemma 2.17. Let $c$ be a simplex. Then,

$$B^r_{c}(c) = \bigcup_{\dim f \geq 1} E_{f \rightarrow c}^r \mathcal{B}^r_{c}(f) \quad \text{and} \quad \mathcal{P}^r \mathcal{I}(c) = \bigoplus_{\dim f \geq 1} E_{f \rightarrow c}^r \mathcal{P}^r \mathcal{I}(f),$$

where the first union is disjoint.

Proof. It is clear that the partition of the basis $B^r_{c}(c)$ implies the direct sum decomposition of $\mathcal{P}^r \mathcal{I}(c)$. So it is sufficient to prove the partition of the basis. By definition,

$$B^r_{c}(c) = \bigcup_{[v_i v_j] \in \mathcal{E}(c)} \{p d \lambda_i \otimes d \lambda_j \mid p \in B^r(c)\}. $$

25
Lemma 2.16 implies that
\[ B^r_{\mathcal{S}}(c) = \bigcup_{[v_i, v_j] \in \mathcal{E}(c) \cap \mathcal{F}(f)} \bigcup_{p \in B^r_{ij}(f)} \{ E^r_{f \leftarrow c} pd \lambda_i \otimes d \lambda_j \mid p \in B^r_{ij}(f) \}. \]

Exchange the order of the two unions:
\[ B^r_{\mathcal{S}}(c) = \bigcup_{\dim f \geq 1 [v_i, v_j] \in \mathcal{E}(f)} \bigcup_{p \in B^r_{ij}(f)} \{ E^r_{f \leftarrow c} pd \lambda_i \otimes d \lambda_j \mid p \in B^r_{ij}(f) \}. \]

Finally, Lemma 2.14 says that the inner union is exactly \( E^r_{f \leftarrow c} \hat{B}^r_{\mathcal{S}}(f) \). Hence,
\[ B^r_{\mathcal{S}}(c) = \bigcup_{\dim f \geq 1} E^r_{f \leftarrow c} \hat{B}^r_{\mathcal{S}}(f) \]
proves the claim.

The next step is to derive the geometric decomposition of the dual space.

**Lemma 2.18.** For any \( n \)-simplex \( c \),
\[ [\mathcal{D}^r \mathcal{I}(c)]' = \left\{ u \mapsto \int_c u : q \mid q \in \mathcal{P}^{r-n+1} \mathcal{I}(c) \right\}. \]
Further, the geometric decomposition of \([\mathcal{P} \mathcal{I}(c)]'\) in Theorem 2.7 holds.

**Proof.** Clearly, the map from Bernstein basis \( B^{r-n+1}(c) \) to the edge-associated Bernstein basis \( B^r_{ij}(c) \) given by
\[ p \mapsto (\lambda_0 \cdots \hat{\lambda}_i \cdots \hat{\lambda}_j \cdots \lambda_n) p \]
is a bijection. Hence, it induces a linear isomorphism between \( \mathcal{P}^{r-n+1}(c) \) and the span of \( B^r_{ij}(c) \). In particular, the dual relation holds:
\[ [\text{span} \hat{B}^r_{ij}(c)]' = \left\{ p \mapsto \int_c pq \mid q \in \mathcal{P}^{r-n+1}(c) \right\}. \]

Then the tensor product structure (2.2), the fact that the Frobenius inner product is an inner product on \( \mathbb{S}^n \), and the basis result Theorem 2.6 together implies the claim.

Finally, the geometric decomposition of \([\mathcal{P} \mathcal{I}(c)]'\) is derived. Dualize the geometric decomposition of \( \mathcal{P} \mathcal{I}(c)' \) in Lemma 2.17 gives:
\[ [\mathcal{P}^r \mathcal{I}(c)]' = \bigoplus_{\dim f \geq 1} [E^r_{f \leftarrow c} \mathcal{P}^r \mathcal{I}(f)]'. \]

By Lemma 2.10 \( i_f^* E^r_{f \leftarrow c} \) is identity on the bigger space \( \mathcal{P}^r \mathcal{I}(f) \). The first part of this lemma then implies for each \( k \)-face \( f \),
\[ [E^r_{f \leftarrow c} \mathcal{P}^r \mathcal{I}(f)]' = \left\{ u \mapsto \int_f (i_f^* u) : q \mid q \in \mathcal{P}^{r-k+1} \mathcal{I}(f) \right\}. \]

This proves the claim. □
Given all the previous results, the theorems at the beginning of this section follows easily. Indeed, the geometric decomposition of the dual space directly proves the unisolvency (Theorem 2.2).

Lemma 2.10, the geometric decomposition (Lemma 2.17), and the characterization of bubbles (Lemma 2.18) combined implies the locality result (Theorem 2.3).

Finally, the characterization result (Theorem 2.4) is proved. Suppose \( u \in \text{REG}^r(\mathcal{T}) \) for some mesh \( \mathcal{T} \). By locality (Theorem 2.3), on each cell \( c \) of the mesh, the degrees of freedom fixes \( \iota_f^*u \) for all \( k \)-faces \( f \) with \( k \geq 1 \). Then the finite element assembly process forces \( \iota_f^*u \) to be single-valued. On the other hand, suppose \( u \in \mathcal{P}^r_S(\mathcal{T}) \) with single-valued \( \iota_f^*u \) for all \( k \)-faces \( f \) with \( k \geq 1 \). Then the degrees of freedom can be evaluated on this \( u \) and obtained a \( u' \in \text{REG}^r(\mathcal{T}) \). By unisolvency, the restrictions \( u'|_c = u_c \) agree on each cell \( c \) of the mesh. Hence \( u = u' \in \text{REG}^r(\mathcal{T}) \).

### 2.4 Affine and approximation properties

There two main results of this section. The first one, Theorem 2.19, shows that \( \text{REG}^r \) forms an affine family of finite elements for any fixed dimension \( n \geq 1 \). Such affine families have many advantages [29, Section 2.3]. For example, for software implementations, this makes the assembly of bilinear forms involving such finite elements very efficient [72, Chapter 6]. Another consequence of the affine property is the second result of this section, stated in Theorem 2.24 and Theorem 2.25, which prove the optimal approximation properties of the canonical interpolant induced by the degrees of freedom (2.4b). Both results require some preparations to state precisely.

#### 2.4.1 Affine property

First affine properties of finite elements have to be defined. Two finite elements \((\mathring{c}, \mathring{V}, \mathring{\Sigma})\) and \((c, V, \Sigma)\) are called affine equivalent if there is an affine isomorphism \( \phi : \mathbb{R}^n \to \mathbb{R}^n \) such that \( c = \phi(\mathring{c}) \), \( V = \phi^*(V) \) under the appropriate pullback \( \phi^* \) for the function space, and for every face \( \mathring{f} \) of \( \mathring{c} \) and its corresponding face \( f := \phi(\mathring{f}) \), the associated degrees of freedom \((\mathring{r}_f, \mathring{\Sigma}_f) \in \mathring{\Sigma} \) and \((r_f, \Sigma_f) \in \Sigma \) satisfies

\[
\mathring{r}_f(V) = \phi^*(r_f(V)) \quad \text{and} \quad \phi_* (\mathring{\Sigma}_f) = \Sigma_f,
\]

where the \( \phi_* \) is defined naturally: for any \( \mathring{l} \in \mathring{\Sigma}_f \) and \( u \in r_f(V) \),

\[
(\phi_* \mathring{l})(u) = \mathring{l}(\phi^* u).
\]
This definition is adapted from the classical definition of equivalence of finite elements [15, Section 3.4]. A finite element family \( F \) is a called an affine family [29, Section 2.3], if all the finite elements in the image of \( F \) are affine equivalent to \( F(\hat{c}) \) for a fixed simplex \( \hat{c} \). This \( F(\hat{c}) \) is called the reference element of the affine family.

Note that all simplices of the same dimension can be mapped to each other via affine maps. More precisely, suppose \( c = [v_0, \ldots, v_n] \) and \( \hat{c} = [\hat{v}_0, \ldots, \hat{v}_n] \) are two \( n \)-simplices. By definition the vertices of both are of general position, therefore both \( \{v_1 - v_0, \ldots, v_n - v_0\} \) and \( \{\hat{v}_1 - \hat{v}_0, \ldots, \hat{v}_n - \hat{v}_0\} \) form basis for \( \mathbb{R}^n \). Let \( A \) be the invertible linear map which takes \( \{v_1 - v_0, \ldots, v_n - v_0\} \) to \( \{\hat{v}_1 - \hat{v}_0, \ldots, \hat{v}_n - \hat{v}_0\} \) and \( \phi : \mathbb{R}^n \to \mathbb{R}^n \) an affine map given by

\[
\phi(x) := A(x - v_0) + \hat{v}_0.
\]

It is clear that \( \phi \) maps \( c \) to \( \text{barc} \) bijectively. Further, its differential \( d\phi \) is just the constant matrix \( A \). Thus up to affine bijections, there is a unique \( n \)-simplex for each \( n \). For clarity, in this thesis, for each \( n \geq 0 \), the \( n \)-simplex \( \hat{c} = [\hat{v}_0, \ldots, \hat{v}_n] \) in \( \mathbb{R}^n \) with vertices

\[
\hat{v}_0 = [0, \ldots, 0], \quad \hat{v}_1 = [1, 0, \ldots, 0], \quad \hat{v}_2 = [0, 1, 0, \ldots, 0], \quad \ldots, \quad \hat{v}_n = [0, \ldots, 0, 1],
\]

(2.19)
is chosen to be the representative. This \( \hat{c} \) is referred to as the reference \( n \)-simplex.

Given these definitions, the affine property of REG\( r \) can be stated precisely:

**Theorem 2.19.** Fix any \( r \geq 0 \) and \( n \geq 1 \). Let \( \hat{c} \) be the reference \( n \)-simplex. For any \( n \)-simplex \( c \), REG\( r(c) \) is affine equivalent to REG\( r(\hat{c}) \). Thus the restriction of generalized Regge family to simplices of the same dimension forms an affine family.

To prove this theorem, the following lemma is needed:

**Lemma 2.20.** Let \( \hat{c} \) and \( c \) be two \( n \)-simplices and \( \phi \) the linear isomorphism mapping \( \hat{c} \) to \( c \) defined in equation (2.18). For any face \( \tilde{f} \) of \( \hat{c} \), \( \tilde{f} := \phi(\tilde{f}) \) is a face of \( c \). For any \( u \in \mathcal{F}(c) \),

\[
\phi^* i_{\tilde{f}}^* u = i_{\tilde{f}}^* \phi^* u.
\]

Moreover, for any \( u \in \mathcal{F}(c) \) and \( v \in \mathcal{F}(\hat{c}) \),

\[
\int_{\hat{c}} (\phi^* u) : v = \int_{\hat{c}} [u : ((\phi^*)^T v)](\det \phi)^{-1},
\]

where \( (\phi^*)^T \) is the transpose of \( \phi^* \) under the Frobenius inner product.

**Proof.** First since \( \phi \) maps vertices to vertices, the image \( f = \phi(\tilde{f}) \) is indeed a face of \( c \). Identify the tangent space to \( \tilde{f} \) (or \( f \)) as a subspace of that of \( \hat{c} \) (or \( c \)). Then both \( i_{\tilde{f}}^* \) and \( i_{f}^* \) are identity maps. This proves \( \phi^* i_{\tilde{f}}^* = i_{f}^* \phi^* \) on \( \mathcal{F}(c) \). For the last one, note that

\[
(\phi^* u) : v = (u : [(\phi^*)^T v]) \circ \phi
\]
where the term in the braces is a scalar function on $c$. The last claim then follows from the change of variable formula for integrals.

Proof of Theorem 2.19. Let $\phi$ be the affine map from the reference $n$-simplex $\hat{c}$ to $c$ given by equation (2.18). This fulfills $\phi(\hat{c}) = c$. Because the differential $d\phi$ is constant, elements of $\phi^*\mathcal{P}^r \mathcal{I}(c)$ are still polynomials of degree $r$. The invertibility of $d\phi$ then implies that $\phi^*\mathcal{P}^r \mathcal{I}(c) = \mathcal{P}^r \mathcal{I}(\hat{c})$. Finally, the conditions on degrees of freedom have to be checked. First, this $\phi$ acts as an affine isomorphism from all faces of $\hat{c}$ to $c$. Hence, for every face $\hat{f}$ of $\hat{c}$ and its corresponding face $f := \phi(\hat{f})$, the associated degrees of freedom $(\hat{r}_f, \hat{\Sigma}_f) \in \hat{\Sigma}$ and $(r_f, \Sigma_f) \in \Sigma$ satisfies

$$\hat{r}_f(\hat{V}) = \mathcal{P}^r \mathcal{I}(\hat{f}) = \phi^*[\mathcal{P}^r \mathcal{I}(f)] = \phi^*[r_f(V)].$$

Note that in Lemma 2.20, $(\phi^*)^T$ is an invertible constant matrix and $\det \phi$ is a nonzero constant. So the map

$$q \mapsto (\det \phi)^{-1}(\phi^*)^T q$$

is a bijection between $\mathcal{P}^s \mathcal{I}(\hat{f})$ and $\mathcal{P}^s \mathcal{I}(f)$ for any integer $s$. Thus by definition of the degrees of freedom (2.4b), $\phi_* \hat{\Sigma}_f = \Sigma_f$ as required. This proves the claim.

2.4.2 Approximation properties of the canonical interpolant

Here the optimal error rates for the canonical interpolant of the generalized Regge family are proved. In order to state the result, several definitions have to be clarified.

Let $\Omega$ be a Lipschitz polytope in $\mathbb{R}^n$ and $\mathcal{T}$ be a triangulation of $\Omega$. For any smooth $g \in \mathcal{H}(\Omega)$, the degrees of freedom for $\text{REG}^r(\mathcal{T})$ can be evaluated on $g$ to obtain an element $I^r g \in \text{REG}^r(\mathcal{T})$. This $I^r g$ is called the canonical interpolant of $g$ and the map $I^r$ is called the canonical interpolation operator. Let $I$ be the identity operator. The approximation property is a statement about $(I - I^r) g$ in some appropriate Sobolev norm.

In order to define the Sobolev spaces, a background Riemannian manifold is needed [10]. In numerical analysis, in the end, the mesh is always in some Euclidean space. Hence, the background Riemannian manifold is always assumed to be $\mathbb{R}^n$ with the Euclidean metric. This might cause some confusion. Symmetric covariant 2-tensor fields, like elements of $\text{REG}^r(\mathcal{T})$, can serve as Riemannian metrics if it is everywhere positive definite. In the geometry literature (for example [24]), Regge finite element is studied in the context of metric approximation where the difference between a smooth metric and its discrete approximation is measured under the smooth metric itself. In this thesis, however, the error is always measured in the background metric on the triangulation induced by its embedding in the Euclidean space. This, while being extrinsic, is a very convenient and meaningful choice for
numerical analysis. As will be shown in this thesis, REG(\mathcal{F}) have many applications where it is not used as a discrete metric. So the error measured in the Euclidean background is always available.

Let \( \Omega \) be a Lipschitz polytope in \( \mathbb{R}^n \). The standard Sobolev spaces [2] for scalar-valued functions on \( \Omega \) are denoted by \( W^{s,p}(\Omega) \). For a covariant 2-tensor field \( g \) on \( \Omega \subset \mathbb{R}^n \), its pointwise norm \( |g| \) is a scalar-valued function obtained from evaluating the Frobenius norm of \( g \) at each point of \( \Omega \). The \( W^{s,p}(\Omega) \)-seminorm and norm of \( g \) are defined as:

\[
|g|_{W^{s,p}(\Omega)} := \|g\|_{W^{s,p}(\Omega)}, \quad \|g\|_{W^{s,p}(\Omega)} := \|g\|_{W^{s,p}(\Omega)}.
\]

The Sobolev space of symmetric covariant 2-tensor fields on \( \Omega \), denoted by \( W^{s,p}(\Omega) \), is defined as the completion of the smooth \( \mathcal{F}(\Omega) \) in this norm [25,44,78].

The space \( W^{s,p}(\Omega) \) is defined in an invariant way. This is convenient in most applications. But for the scaling argument based proof of approximation properties, it is advantageous to use an equivalent but more concrete coordinate characterization. Since \( \Omega \subset \mathbb{R}^n \), elements in \( \mathcal{F}(\Omega) \) can be viewed as \( \mathbb{S}^n \)-valued functions. Take any orthonormal basis \( \{s_i\}_{i=1}^{n(n+1)/2} \) of \( \mathbb{S}^n \). Any function \( g \in \mathcal{F}(\Omega) \) has a unique representation in this basis

\[
g = \sum_{i=1}^{n(n+1)/2} g_is_i,
\]

where \( g_i \) are scalar functions. The pointwise norm and \( W^{s,p}(\Omega) \)-semi-norm of \( g \) are:

\[
|g| = \sqrt{\sum_{i=1}^{n(n+1)/2} g_i^2}, \quad |g|_{W^{s,p}(\Omega)} = \sqrt{\sum_{i=1}^{n(n+1)/2} g_i^2_{W^{s,p}(\Omega)}}.
\]

On the other hand, each \( g_i \) is just a scalar function. Let \( W^{s,p}(\Omega,\mathbb{S}^n) \) be the Bochner space of \( \mathbb{S}^n \)-valued \( W^{s,p} \)-functions. The \( W^{s,p}(\Omega,\mathbb{S}^n) \)-semi-norm of \( g \) is defined as:

\[
|g|_{W^{s,p}(\Omega,\mathbb{S}^n)} = \sqrt{\sum_{i=1}^{n(n+1)/2} |g_i|_{W^{s,p}(\Omega)}^2}.
\]

It is clear that this is independent of the orthonormal basis chosen. Basically the \( W^{s,p}(\mathcal{F}(\Omega))-norm \) is the \( W^{s,p} \)-norm of the pointwise Frobenius norm and the \( W^{s,p}(\Omega,\mathbb{S}^n)-norm \) is the Frobenius norm of the component-wise \( W^{s,p} \)-norm. The latter allows us to use all the standard results for scalar Sobolev spaces. It is good news that these two are equivalent:

\[
\text{Lemma 2.21. There exist constants } c \text{ and } C \text{ only depending on the dimension } n \text{ such that }
\]

\[
c|g|_{W^{s,p}(\Omega,\mathbb{S}^n)} \leq |g|_{W^{s,p}(\mathcal{F}(\Omega))} \leq C|g|_{W^{s,p}(\Omega,\mathbb{S}^n)}.
\]

In particular, the \( W^{s,p}(\mathcal{F}(\Omega))-norm \) and the \( W^{s,p}(\Omega,\mathbb{S}^n)-norm \) are equivalent.
Proof. Take any orthonormal basis for $\mathbb{S}^n$ and let $\{g_i\}$ be the components of $g$. The trick is to use the $l^p$-norm on components of $\mathbb{S}^n$ over each point of the domain. It turns out that the $l^p L^p$-norm and the $L^p l^p$-norm of a matrix-valued function are the same:

$$
\| \left( \sum_i |g_i|^p \right)^{1/p} \|_{L^p} = \left( \int_\Omega \sum_i |g_i|^p \right)^{1/p} = \left( \sum_i \int_\Omega |g_i|^p \right)^{1/p} = \left( \sum_i \|g_i\|_{L^p}^p \right)^{1/p},
$$

where the sum is over all the basis. This is a purely algebraic fact. So the same holds for the $l^p W^{s,p}$-semi-norm and the $W^{s,p} l^p$-semi-norm as well. Since $\mathbb{S}^n$ is a finite dimensional space, all norms on it are equivalent. In particular, over each point of $\Omega$, the $l^p$-norm on components of $\mathbb{S}^n$ is equivalent to the Euclidean Frobenius norm with constants only depending on $n$. This proves the claim.

In this view, $W^{s,p} \mathcal{F}(\Omega)$ is nothing more than $n(n+1)/2$ copies of $W^{s,p}(\Omega)$. Therefore all the standard Sobolev space theory carries over in a straightforward fashion. From the definition, $I_T^s u \in \mathcal{D}' \mathcal{F}(\mathcal{S})$ is a piecewise polynomial which is discontinuous across the interior facets. By the standard trace theorems [44, Theorem 1.5.1.2],

$$
I_T^s u \in W^{s,p} \mathcal{F}(\Omega), \quad \forall s \in \left[0, 1/p \right).
$$

On the other hand, from the definition of the degrees of freedom (2.4b), the canonical interpolant $I_T^s$ needs the integral of the function restricted to $k$-faces for $1 \leq k \leq n$. By the trace theorems again, $I_T^s$ can be extended from smooth $\mathbb{S}^n(\Omega)$ boundedly to

$$
I_T^s : W^{s,p} \mathcal{F}(\Omega) \rightarrow \text{REG}'(\mathcal{S}), \quad \forall s \in \left((n-1)/p, \infty \right].
$$

This establishes the space and norm where the error of the canonical interpolant is going to be assessed.

Moreover, some geometric quantities related to the mesh are needed to study approximations. For any simplex $c$, the size $h_c$ of $c$ is the Euclidean diameter of the circumscribing sphere of $c$, the inradius $\rho_c$ is the Euclidean diameter of the inscribing sphere of $c$, and the shape constant $\sigma_c$ is defined to be the ratio $h_c/\rho_c$. See Figure 2.12 for illustrations.

![Figure 2.12: Definition of $h_c$ and $\rho_c$](fig:geom_constants)
The shape constant quantifies how far away the simplex $c$ is from being degenerate (with vertices no longer of general position). See Figure 2.13 for visual examples.

These quantities are useful for estimating the norm of the differential.

**Lemma 2.22.** Let $c, c'$ be two $n$-simplices and $\phi : \mathbb{R}^n \to \mathbb{R}^n$ the affine bijective map from $c$ to $c'$ as defined in (2.18). Then, $d(\phi^{-1}) = (d\phi)^{-1}$ and

$$
\|d\phi\| \leq \frac{\sqrt{n}h_{c'}}{\rho_c}, \quad \|d\phi^{-1}\| \leq \frac{\sqrt{n}h_c}{\rho_c'},
$$

where the norm is the Euclidean Frobenius norm.

**Proof.** This result for the Euclidean operator 2-norm is well-known [29, Theorem 3.1.3]. A proof is reproduced below. First, both $d\phi = A$ and $d\phi^{-1} = A^{-1}$ are just constant linear maps. To prove the first inequality about $d\phi$, note that any vector of Euclidean length $\rho_c$ can be realized as the difference between two points in $c$. Their images are at most $h_{c'}$ apart under $\phi$, which proves the claim in the operator norm. All norms on a finite dimensional space are equivalent. For the invertible $d\phi$, let $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_n > 0$ be its singular values. It is well-known [42, Corollary 2.4.3] that the Euclidean operator norm of $d\phi$ is $\sigma_1$ while its Euclidean Frobenius norm is $\sqrt{\sigma_1^2 + \cdots + \sigma_n^2}$. This gives the $\sqrt{n}$ factor in the final result. For the inverse, the proof is similar.

The estimates on the differential can be used to estimate the pullback in the $W^{s,p}\mathcal{A}(\Omega)$-norm. This will be a key step in the proof of the approximation theorem.

**Theorem 2.23.** Let $\bar{c}, c$ be two $n$-simplices and $\phi : \mathbb{R}^n \to \mathbb{R}^n$ the affine isomorphism from $\bar{c}$ to $c$ as defined in equation (2.18). Let $g$ be any function in $W^{s,p}\mathcal{A}(c)$ and $\bar{g} := \phi^* g$. Then, there exists a constant $C = C(n,s)$ such that

$$
|\bar{g}|_{W^{s,p}\mathcal{A}(\bar{c})} \leq C\|d\phi\|^{s+2}\|\det(d\phi)|^{-1/p}|g|_{W^{s,p}\mathcal{A}(c)},
$$

$$
|g|_{W^{s,p}\mathcal{A}(c)} \leq C\|d\phi^{-1}\|^{s+2}\|\det(d\phi)|^{1/p}|\bar{g}|_{W^{s,p}\mathcal{A}(\bar{c})}.
$$
Proof. Fix the same arbitrary orthonormal basis \( \{e_i\}_{i=1}^n \) for \( \mathbb{R}^n \) for both \( \bar{c} \) and \( c \). This implicitly identifies the tangent space of \( \bar{c} \) and \( c \) at any point. Since the Euclidean Frobenius norm is invariant under orthogonal transformations, it does not matter which basis is chosen. It should be stressed that the norm on the pullback \( \phi^*g \) is not measured in the pullback metric but in the background metric on \( \bar{c} \). Otherwise it would be an isometry and there is no scaling at all. In this basis, \( \bar{g} \) and \( g \) are matrix-valued functions and \( d\phi \) is a constant matrix \( A \). By the definition of the pullback,

\[
\bar{g} = \phi^*g = A^T(g \circ \phi)A.
\]

Recall [42, Equation (2.3.3)] that the Frobenius norm is compatible with matrix product. So the point-wise norm satisfies:

\[
|\bar{g}| = |A^T(g \circ \phi)A| \leq \|d\phi\|^2|(g \circ \phi)|.
\]

Hence, the \( W^{s,p}(\mathcal{F}^\Omega) \)-semi-norm is estimated by:

\[
|\bar{g}|_{W^{s,p}(\mathcal{F}^\Omega)} \leq \|d\phi\|^2|(g \circ \phi)|_{W^{s,p}(\mathcal{F}^\Omega)}.
\]

Recall the classical scaling result [29, Theorem 3.1.2] for scalar-valued functions \( u \in W^{s,p}(c) \):

\[
|u \circ \phi|_{W^{s,p}(\mathcal{F}^\Omega)} \leq C\|d\phi\|^s|\det(d\phi)|^{-1/p}|u|_{W^{s,p}(c)},
\]

where \( C = C(s,n) \). Since each component of \( u \circ \phi \) is just a scalar function, using the homogeneity of the \( W^{s,p}(\cdot,\mathbb{S}^n) \)-norm and the norm equivalence Lemma 2.21,

\[
|g \circ \phi|_{W^{s,p}(\mathcal{F}^\Omega)} \leq C_1|g \circ \phi|_{W^{s,p}(\mathcal{F}^\Omega)} \leq C_2\|d\phi\|^s|\det(d\phi)|^{-1/p}|g|_{W^{s,p}(\mathcal{F}^\Omega)} \leq C_3\|d\phi\|^s|\det(d\phi)|^{-1/p}|g|_{W^{s,p}(\mathcal{F}^\Omega)},
\]

where the constant \( C_3 = C_3(s,n) \). This proves the claim for \( \phi^* \). The same result applied to the inverse gives the second estimate.

Compared with the classical scaling result [29, Theorem 3.1.2], this theorem contains an extra \( \|d\phi\|^2 \) due to the tensor pullback. This should also be compared to similar estimates for alternating multilinear form fields (differential forms) in [103, Theorem 5], which was used in [53] to derive estimates for Finite Element Exterior Calculus. It should be noted that in [53,103], the pullback estimates are proved in the Euclidean operator norm on differential forms but in the end the metric induced norms on differential forms are used in applications.

Given all these, an estimate for the canonical interpolant on any simplex is:
Because it is a finite dimensional space. Thus that,  

\[ \|u-c\|_{W^{r+1,p}(\mathcal{F}(c))} \leq C_{\sigma}^{r+\frac{3}{2}} h^{\frac{1}{2}} |g|_{W^{r+1,p}(\mathcal{F}(c))}, \quad \forall g \in W^{s,p}(\mathcal{F}). \]

Compared with classical results for scalar-valued functions [15, Theorem 4.4.4], the only difference is that for covariant 2-tensors the exponent for \( \sigma \) is \((t+2)\) while for scalar functions it is \( t \). This means that the approximation properties of \( \text{REG}^r(c) \) are more sensitive to the shape of \( c \). In particular, while for scalar functions, the \( L^p \)-estimates are independent of \( \sigma \), for \( \text{REG}^r(c) \) the \( L^p \)-estimates are still degraded if \( \sigma \) is large.

**Proof of Theorem 2.24.** The first step is to establish the claim on the reference \( n \)-simplex \( \hat{c} \). As before, without loss of generality, take any orthonormal basis \( \{e_i\}_{i=1}^n \) for \( \mathbb{R}^n \). The idea is again to estimate component by component. Clearly \( \hat{c} \) is a Lipschitz domain in \( \mathbb{R}^n \). The Bramble-Hilbert lemma (see [14] and [29, Theorem 3.1.1]) states that for all \( r \geq 0 \), there exists a constant \( C = C(r,n) \) (the dependency on \( n \) follows from its dependency on \( \hat{c} \)) such that for scalar-valued functions:

\[ \inf_{p \in \mathcal{S}(\hat{c})} \|u-p\|_{W^{r+1,p}(\hat{c})} \leq C |u|_{W^{r+1,p}(\hat{c})}, \quad \forall u \in W^{r+1,p}(\hat{c}). \]

This implies a similar result for the Bochner space \( W^{r+1,p}(\hat{c},\mathbb{S}^n) \):

\[ \inf_{p \in \mathcal{S}(\hat{c}) \otimes \mathbb{S}^n} \|u-p\|_{W^{r+1,p}(\hat{c},\mathbb{S}^n)} \leq C |u|_{W^{r+1,p}(\hat{c},\mathbb{S}^n)}, \quad \forall u \in W^{r+1,p}(\hat{c},\mathbb{S}^n). \]

Due to the norm equivalence (Lemma 2.21), this holds for \( W^{r+1,p}(\hat{c}) \) as well, with constants depending on \( r \) and \( n \). It was already shown before in (2.20) that for \( s > (n-1)/p \), the canonical interpolant \( I_\hat{c}^r : W^{s,p}(\hat{c}) \to \mathcal{P}(\mathcal{F}(\hat{c})) \) is bounded. The norm on \( \mathcal{P}(\mathcal{F}(\hat{c})) \) does not matter because it is a finite dimensional space. Thus \( \|I_\hat{c}^r\|_{W^{s,p}(\hat{c})} \to W^{r+1,p}(\hat{c}) \) is a constant depending only on \( r, n, s \). It is also clear that \( I_\hat{c}^r \) is a projection which preserves \( \mathcal{P}(\mathcal{F}(\hat{c})) \). Hence, whenever \( s \leq (n-1)/p, r+1 \), for any \( t \in [0,s] \), there exists a constant \( C = C(r,n,s,t) \) such that,

\[ \|u-I_\hat{c}^ru\|_{W^{r,p}(\mathcal{F}(\hat{c}))} = \inf_{p \in \mathcal{P}(\mathcal{F}(\hat{c}))} \|(I-I_\hat{c}^r)(u-p)\|_{W^{r,p}(\mathcal{F}(\hat{c}))} \]

\[ \leq \|(I-I_\hat{c}^r)(u-p)\|_{W^{s,p}(\mathcal{F}(\hat{c}))} \inf_{p \in \mathcal{P}(\mathcal{F}(\hat{c}))} \|(u-p)\|_{W^{r,p}(\mathcal{F}(\hat{c}))} \]

\[ \leq C |u|_{W^{s,p}(\mathcal{F}(\hat{c}))}, \quad \forall u \in W^{s,p}(\mathcal{F}(\hat{c})). \]
The next step is the scaling argument. Let $c$ be any $n$-simplex and $\phi : \mathbb{R}^n \to \mathbb{R}^n$ the affine bijection mapping $\hat{c}$ to $c$ defined in equation (2.18). For any $u \in W^{k,p}(\mathcal{F}(c))$ and $t \in [0,s]$, the second estimate in Theorem 2.23 implies
\[
\|u - I^t_c u\|_{W^{k,p}(\mathcal{F}(c))} \leq C_1 \|d\phi^{-1}\|^{t+2} |\det(d\phi)|^{1/p} |\phi^*(u - I^t_c u)|_{W^{k,p}(\mathcal{F}(\hat{c}))},
\]
where $C_1 = C_1(n,t)$. Crucially, the affine property (Theorem 2.19) implies that the canonical interpolation operator commutes with pullbacks $I^t_c \phi^* = \phi^* I^t_c$. Thus, using the estimate for $I^t_c$ in the previous step,
\[
\|u - I^t_c u\|_{W^{k,p}(\mathcal{F}(c))} \leq C_1 \|d\phi^{-1}\|^{t+2} |\det(d\phi)|^{1/p} |\hat{u} - I^t_c \hat{u}|_{W^{k,p}(\mathcal{F}(\hat{c}))} \leq C_1 C_2 \|d\phi^{-1}\|^{t+2} |\det(d\phi)|^{1/p} |\hat{u}|_{W^{k,p}(\mathcal{F}(\hat{c}))},
\]
where $C_2 = C_2(r,n,s,t)$. Applying the first estimate in Theorem 2.23,
\[
\|u - I^t_c u\|_{W^{k,p}(\mathcal{F}(c))} \leq C_1 C_2 C_3 \|d\phi^{-1}\|^{t+2} \|d\phi\|^s |u|_{W^{s,p}(\mathcal{F}(c))},
\]
where $C_3$ depends on $n$ and $s$. Finally, this, along with the estimates for $\|d\phi\|$ and $\|d\phi^{-1}\|$ in Lemma 2.22, leads to the estimate in the claim. \qed

Last, for a mesh $\mathcal{T}$, the mesh size $h(\mathcal{T})$ and the shape constant $\sigma(\mathcal{T})$ are defined as the maximum of the cell-wise $h_c$ and $\sigma_c$ over all cells $c$ in $\mathcal{T}$. The very useful approximation theorem for REG$'$ is:

**Theorem 2.25.** Let $\Omega$ be a bounded Lipschitz polytope in $\mathbb{R}^n$ and $\{\mathcal{T}_h\}$ be a sequence of triangulations of $\Omega$ indexed by mesh size $h$ with uniformly bounded shape constants $\sup_h \sigma(\mathcal{T}_h) =: \sigma < \infty$. The canonical interpolant $I^r_c$ for REG$'$$(\mathcal{T}_h)$ defined on smooth $\mathcal{F}(\Omega)$ can be extended boundedly to $W^{k,p}(\mathcal{F}(\Omega))$ for any $p \in [1,\infty]$ and $s \in ((n-1)/p,\infty]$. Further, for any $r \geq 0$, $t \in [0,1/p)$, and $s \in ((n-1)/p,r+1]$, there exists a constant $C = C(\sigma,n,r,t,s) > 0$ such that
\[
|g - I^r_c g|_{W^{s,p}(\mathcal{F}(\Omega))} \leq C h^{s-t} |g|_{W^{r,p}(\mathcal{F}(\Omega))}, \quad \forall g \in W^{s,p}(\mathcal{F}(\Omega)).
\]
This is optimal in the sense that it is as good as the best approximation in terms of order in $h$.

**Proof.** For a fixed $h$, apply Theorem 2.24 to each cell $c$ in $\mathcal{T}_h$ where $\sigma_c \leq \sigma$ is absorbed into the constant. Sum over all the cells in $\mathcal{T}_h$ leads to the estimate in the claim:
\[
|g - I^r_c g|_{W^{s,p}(\mathcal{F}(\Omega))}^p = \sum_{c \in \mathcal{T}_h} |g - I^r_c g|_{W^{r,p}(\mathcal{F}(c))}^p \leq C^p h^{p(s-t)} \sum_{c \in \mathcal{T}_h} |g|_{W^{r,p}(\mathcal{F}(c))}^p = C^p h^{p(s-t)} |g|_{W^{r,p}(\mathcal{F}(\Omega))}^p.
\]
\qed

35
2.5 Coordinate representations and implementable degrees of freedom

The first part of this section describes \( \text{REG}^r \) in coordinates. The most important results being equation (2.21) for the pullback in coordinates and Theorem 2.26 for a canonical coordinate basis. These are important for the software implementation of this finite element.

The second part describes the details of two sets of equivalent concrete degrees of freedom for \( \text{REG}^r \). The first set is the one actually used in the FEniCS implementation by the author. The second set is of geometric appeal and is closest to the original \( \text{REG}^0 \).

2.5.1 Coordinate representations

So far, \( \text{REG}^r \) is used in a coordinate-free fashion. For its concrete implementation on a computer, however, inevitably some coordinate basis has to be fixed. As will be shown in the rest of this thesis, there are also many applications where it is natural to use \( \text{REG}^r \) as concrete symmetric-matrix-valued functions. In this section, formulae for the coordinate representation are derived.

First, the standard coordinate identification (see for example [111]) is reviewed. In \( \mathbb{R}^n \), the canonical vector basis is the Euclidean basis \( \{ e_i \}_{i=1}^n \), where each \( e_i \) is the tangent vector to the coordinate function \( x_i \). Under this, vector fields on \( \mathbb{R}^n \) becomes \( \mathbb{R}^n \)-valued functions. The canonical basis for 1-forms consists of the differentials of the Euclidean coordinate functions \( \{ dx_i \}_{i=1}^n \). Under this, 1-forms are also identified with \( \mathbb{R}^n \)-valued functions. The evaluation of a 1-form \( l \) on a vector-field \( u \) is computed as

\[
l(u) = l^T u,
\]

where elements of \( \mathbb{R}^n \) are identified as column vectors. The basis choice for 1-forms induces a canonical choice of basis for covariant 2-tensors given by \( \{ dx_i \otimes dx_j \}_{1 \leq i, j \leq n} \). Under this, covariant 2-tensor fields are identified with \( n \)-by-\( n \) matrix-valued functions and elements of \( \mathcal{S}(\mathbb{R}^n) \) are identified with symmetric-matrix-valued functions. The evaluation of \( g \in \mathcal{S}(\mathbb{R}^n) \) on two vector fields \( u, v \) is then given by

\[
g(u, v) = u^T g v.
\]

For any subset of \( \mathbb{R}^n \), like a domain \( \Omega \) or a mesh \( \mathcal{T} \), all these identifications are inherited. For a mesh \( \mathcal{T} \) in \( \mathbb{R}^n \), this is global in the sense that the same basis are used for all the cells. Under this, \( \text{REG}^r(\mathcal{T}) \) becomes a space of symmetric-matrix-valued polynomials of degree \( r \) or less.
The next step is to derive the coordinate representation of the pullbacks. Suppose $U, U'$ are two open subsets in $\mathbb{R}^n$ and $\phi : U \to U'$ is a diffeomorphism. From the definition of the differential and chain rule, the coordinate representation of $d\phi$ is an $\mathbb{R}^{n \times n}$-valued function on $U$ with components:

$$[d\phi]_{i,j} = \partial_j \phi_i,$$

where following the usual notation the first index is the row index and the second index is the column index. Let $g \in \mathcal{S}(U')$. By definition, its pullback in this coordinates is a symmetric-matrix-valued function on $U$:

$$(\phi^* g)(x) = [d\phi(x)]^T[g \circ \phi(x)][d\phi(x)].$$

(2.21)

Note that these formulae are possible because both $U$ and $U'$ are open subsets of the same $\mathbb{R}^n$. In this case, there is a canonical way to identify the same Euclidean basis for both. This is, however, no longer true when, for example, $U'$ is a lower-dimensional subset of $U$.

Let $c = [v_0 \ldots v_n]$ be an $n$-simplex in $\mathbb{R}^n$ and $f$ be a $k$-face with $1 \leq k < n$. There is no natural Euclidean basis on $f$ which is compatible with the Euclidean basis on $c$. This is potentially problematic because then there are as many arbitrary choices as the number of faces of $c$ to be made. For 2-tensors and only for 2-tensors, however, there is a canonical barycentric system with appealing geometric associations. Let $\mathbb{V}_n$ be the space of symmetric 2-vectors in $\mathbb{R}^n$, that is, the span of $\{e_i \otimes e_j\}_{1 \leq i \leq j \leq n}$. This $\mathbb{V}_n$ is the dual to $\mathbb{S}_n$. The observation that the number of edges in an $n$-simplex $\binom{n+1}{2}$ equals the dimension of the space of symmetric 2-tensors $\mathbb{S}_n$ can be lifted into two related statements on vector spaces:

**Theorem 2.26.** Let $c = [v_0 \ldots v_n]$ be an $n$-simplex and $\{\lambda_i\}_{i=0}^n$ its barycentric coordinates. For any edge $e = [v_i v_j]$ of $c$, let

$$g_e := -dv_i \otimes dv_j, \quad v_e := v_j - v_i.$$

Then,

$$\{g_e | e \subset c\}$$

is a basis for $\mathbb{S}_n$, and

$$\{v_e \otimes v_e | e \subset c\}$$

is a basis for $\mathbb{V}_n$.

(2.22)

Further these two basis are dual to each other:

$$g_e(v_{e'}) = \begin{cases} 1, & \text{if } e = e', \\ 0, & \text{otherwise}. \end{cases}$$

(2.23)

**Proof.** The fact the $\{g_e\}$ forms a basis for $\mathbb{S}_n$ has already been proved in Theorem 2.9. For $\{v_e \otimes v_e\}$, it is enough to show that for any $g \in \mathbb{S}_n$, knowing the values

$$\{g(v_e, v_e) | e \subset c\}.$$
is enough to evaluate $g(v_i - v_0, v_j - v_0)$ for any pair $i$ and $j$, because $(v_i - v_0)_{i=1}^n$ forms a basis for $\mathbb{R}^n$. When $i = j$, this is already known. When $i \neq j$, by polarization identity for bilinear forms:

$$g(v_i - v_0, v_j - v_0) = \frac{1}{2}[g(v_i - v_j, v_i - v_j) - g(v_i - v_0, v_i - v_0) - g(v_j - v_0, v_j - v_0)].$$

All terms on the right-hand side are of the form $g(v_e, v_e)$ for some edges $e$. This proves that $\{v_e \odot v_e\}$ spans $\mathbb{V}^n$. The last claim follows immediately from the computations in the proof of Theorem 2.9.

Under the barycentric basis, the tangential-tangential pullback has a canonical representation. Indeed, each function $g \in \mathcal{S}(c)$ is an $\mathbb{R}^{n(n+1)/2}$-valued function where the components are indexed by edges of $c$. By the dual structure and Theorem 2.11, tangential-tangential pullback to a face $f$ of $c$ simply drops components indexed by edges which are not part of $f$, that is, a simple projection:

$$i_f^*\sum_{e \subset c} a_ee = \sum_{e \subset f} a_ee,$$

The classical REG$^0$ used in Regge Calculus is also given in this basis. The problem with the barycentric system is that it does not have a basis for anything else, for example, for vectors.

It is recommended that the barycentric basis system is used for mathematical analysis and software implementation internal to the generalized Regge finite element. The Euclidean basis system should be used for all other places.

### 2.5.2 Implementable degrees of freedom

In this subsection, two elegant and efficiently implementable degrees of freedom are derived for REG$^r$. The first one is the mathematical description of the REG$^r$ implemented in FEniCS [72] by the author as part of this thesis. The second one has a geometric interpretation that is closest to the original Regge finite element REG$^0$.

From the definition (2.4b) of the degrees of freedom, it comes down to choose a basis for

$$\left\{ u \mapsto \int_f (i_f^* u) : q | q \in \mathcal{P}^{r-k+1,s}(f) \right\}$$

for each $k$-face $f$ with $k \geq 1$ of an $n$-simplex $c$. The direct implementation of the above is not convenient because as mentioned in the previous subsection, when $u$ is identified as a symmetric matrix, the Euclidean basis is implicitly assumed but there is no good canonical representation of $i_f^* u$ in the Euclidean basis. Moreover, the numerical integrals are not efficient in implementations.
The two issues outlined above are dealt with separately. For the first one, note that for a $k$-face $f$, because $(S^k)' = V^k$,

$$\left\{ u \mapsto \int_f (s^*_f u) \cdot q \mid q \in P^{r-k+1}(f) \right\} = \left\{ u \mapsto \int_f p(s^*_f u) \cdot \phi \mid \phi \in V^k, p \in P^{r-k+1}(f) \right\},$$

where the dot denotes the duality pairing between $S^k$ and $V^k$. This is further simplified when the edge-based basis for $V^k$ in (2.22) is chosen because

$$(s^*_f u) \cdot (v_e \circ v_e) = u(v_e, v_e) = v_e^T u v_e,$$

and the pullback is obtained “for free”. Thus the following set can be used as a basis for the degrees of freedom associated with $f$:

$$\{ u \mapsto \int_f (v_e^T u v_e) p_i \mid \text{for every edge } e \text{ of } f \text{ and every element } p_i \text{ of a basis for } P^{r-k+1}(f) \}. $$

This is good for many purposes already. But for a concrete software implementation, the integral moments can be implemented more efficiently by pointwise evaluations at points which can fix an element of $P^{r-k+1}(f)$. Let $X^f_{r-k+1}$ be such a set of points in $f$. The final directly implementable degrees of freedom associated with $f$ are:

$$\{ u \mapsto (v_e^T u v_e)(x) \mid e \subset f, x \in X^f_{r-k+1} \}. $$

Examples of this are given in the introduction of this chapter (see Figure 2.2 and Figure 2.3).

There are many possible choices of $X^f_{r-k+1}$. The following is one particular choice which is used frequently in FEniCS. First, let $\hat{f}$ be the reference $k$-simplex defined in equation (2.19). The equi-distance $X^f_{r-k+1}$ is given by

$$X^f_{r-k+1} := \left\{ \left( \frac{m_1}{r+2}, \ldots, \frac{m_k}{r+2} \right) \mid m_j \in \mathbb{Z} \text{ and } m_j \geq 1 \text{ for } j = 1, \ldots, k. \sum_{j=1}^k m_j \leq r+1 \right\}. $$

Pictorially, this for various values of $k$ and $r$ are depicted in Figure 2.14.
On a general $k$-face $f$ of an $n$-simplex $c$, the equi-distance $X^f_r$ is defined as the image of $X^f_r$ under the affine isomorphism $\phi$ mapping $\hat{f}$ to $f$. It is well-known that pointwise evaluation of at points in $X^f_r$ are linearly independent and forms a dual basis to $\mathcal{P}^r(f)$ [29].

There is another choice of $X^f_r$ which is geometrically appealing. The idea is to take the mid-point of all the small edges in the subdivisions of the cell. This subdivision-based $X^f_r$ is best described with pictures. See examples for 2D in Figure 2.15 and for 3D in Figure 2.16.
The pattern for higher dimensions is clear.

The subdivision-based $X_f^r$ in fact leads to another set of degrees of freedom for REG'. Instead of pointwise evaluation, one can consider using the polynomial symmetric covariant 2-tensor as the metric to measure the squared lengths of these small edges in the subdivision as degrees of freedom. More precisely, for each small edge connecting point $p_1$ and $p_2$, one can associate a functional:

$$u \rightarrow \int_0^1 (p_2 - p_1)^T u(p_1 + t(p_2 - p_1))(p_2 - p_1) dt.$$

The union of such functionals over all the small edges in the $r$-th division of the cell $c$ forms another unisolvent degrees of freedom for REG'($c$). Up to scaling, this is in fact equivalent to the degrees of freedom defined by equation (2.23). Indeed, the edge whose mid-point is inside a face $f$ of $c$ must be parallel to one of the un-subdivided edges of $f$. In the interior of each $k$-face $f$, the integrals of a scalar function over all the small edges interior to $f$ forms a unisolvent set for $\mathcal{S}^{r-k+1}(f)$ as before. This proves the equivalence. Comparing pictures of the subdivision-based degrees of freedom in Figure 2.15 and Figure 2.16 and the previous degrees of freedom in Figure 2.2 and Figure 2.3, the equivalence is evident. This has a nice geometric interpretation: REG'($c$) assigns one number to each of the small edge in the $r$-th subdivision of $c$. These numbers have the meaning of the squared edge lengths. By the unisolvency, there are $\binom{n+1}{2}\binom{n+r}{r}$ small edges in the $r$-th subdivision and these numbers determines a unique element of REG'($c$). This is the most geometric interpretation that clearly shows that REG'($c$) generalizes REG$^0$($c$) used in Regge Calculus. Physicists studying quantum gravity have long searched for generalizations of Regge Calculus, with even ideas like area-based degrees of freedom \[108\]. This generalization is much more natural and elegant.

It should be noted that both choices of $X_f^r$ above are known to be not optimal \[110\]. The performance of choices of the degrees of freedom can be evaluated quantitatively by the Lebesgue constant. For a set of degrees of freedom $\Sigma$, let $I_\Sigma$ be the induced interpolant. The
Lebesgue constant $\Lambda_\Sigma$ is the smallest constant such that the following holds for all smooth $u$:

$$\|u - I_\Sigma u\| \leq (1 + \Lambda_\Sigma) \inf_p \|u - p\|,$$

where the infimum is taken over the shape functions and the norm should be appropriate for the function space being discretized. The optimal $X_r^f$ to control $\mathcal{P}^r(c)$ with the smallest possible Lebesgue constant is known [110]. Further, in the same paper, it was shown that the first choice of $X_r^f$ with the equi-distance lattice points (2.24) is not good for $r \geq 10$. In practice though, the optimal $X_r^f$ is messy to implement and for real problems, degree $r$ more than 3 is rarely used. So the easier equi-distance $X_r^f$ was chosen for the software implementation of REG' in FEniCS by the author.
Chapter 3

Geodesics on Generalized Regge metrics

3.1 Introduction

One of the main applications of symmetric covariant 2-tensor fields is in geometry, where they serve as Riemannian metrics. Similarly, on a mesh, everywhere positive definite functions in the generalized Regge space can serve as discrete Riemannian metrics on the mesh. In this sense REG\(r\) can be used to discretize Riemannian geometry.

The piecewise constant REG\(0\) has been studied extensively in the literature for this purpose. Historically, Riemannian metrics in REG\(0\) are important in the mathematical study of Euclidean polyhedrons and are referred to as polyhedral metrics [4]. In the theory of relativity in physics, Lorentzian metrics, which are symmetric covariant 2-tensor fields similar to Riemannian metrics, play a central role. Regge used Lorentzian metrics in REG\(0\) to derive a geometric discretization of the Einstein field equation in his influential work [95]. Subsequently, Lorentzian metrics in REG\(0\) are referred to as Regge metrics in the physics literature.

In this chapter, the focus is on the Riemannian case. The generalization of most results here to pseudo-Riemannian metrics, which contain both Riemannian metrics and Lorentzian metrics as special cases, is straightforward. In what follows, everywhere positive definite functions in REG\(r\) are called generalized Regge metrics, or simply REG\(r\) metrics, to acknowledge the inspiration from Regge's work and also to emphasize that these are piecewise polynomial Riemannian metrics different from the Regge metrics used in the physics literature.

In the context of Riemannian geometry, geodesics are basic objects for quantifying and characterizing geometry. This chapter examines various mathematical and computational
aspects of geodesics on generalized Regge metrics.

Geodesics on discrete geometries are of considerable practical interest. Geodesics on 2D triangulations embedded in 3D Euclidean space are important in computer graphics [50] and computer-aided design [70]. In relativity, geodesics model the trajectories of light rays and free-falling test particles [109]. After Regge’s initial work, physicists explored the interpretation and computation of geodesics on Regge metrics [16, 22, 116, 117]. Finally, in other parts of this thesis, REG will be used to solve PDEs in solid mechanics where the solutions either are or can be interpreted as Riemannian metrics. In such cases, geodesics can be used for visualizing these symmetric covariant 2-tensor fields [54].

The theory of smooth geodesics on smooth Riemannian manifolds is well-understood [32]. This will be reviewed in Section 3.2. Geodesics are essentially generalizations of straight lines in the Euclidean space to Riemannian manifolds. There are two aspects in the classical theory. One is of a global nature, where the “shortest” curve connecting two points is been sought after. This generalizes the notion of a line segment in the Euclidean space. The other one is of a local nature, given a position and a velocity, the “straightest” curve needs to be defined. This generalizes the notion of a ray in the Euclidean space. The study of the interplay between the two occupies substantial part of differential geometry. In this chapter, both will be studied for REG metrics.

It turned out that the key ingredient behind the global aspect is the distance structure, which can be quite non-smooth. This part is thus easy to generalize to REG metrics. In Section 3.3, this will be studied in detail. The main result is that REG metrics have a well-behaved length structure, in which geodesics are well-defined. Further, REG metrics are the least smooth (that is the most general) piecewise polynomial Riemannian metric for which the usual sense of curve length is preserved (Theorem 3.2).

The local theory of geodesics turned out be subtle for REG metrics. Indeed, the local theory uses “velocity”, which inevitably requires some differential structure. Intuitively, generalized Regge metrics are piecewise smooth, so problems can arise only when smooth geodesics in the interior of a simplex reach an interior facet in the mesh. In the case of Riemannian metrics in REG, two seemingly unrelated ideas for resolving this are popular in the literature. The first idea [4] is of a geometrical nature. Take a 2D REG metric for example. This can be identified as the metric for a triangulated surface like the one in Figure 3.1 embedded in some Euclidean space. Near an interior edge, the two triangles containing that edge can be cut off from the rest and then flattened in Euclidean $\mathbb{R}^2$. Then intuitively geodesics should connect any two points in different triangles by straight lines in $\mathbb{R}^2$. The geodesic on the REG metric can thus be obtained by pulling the straight lines back to the mesh via
the isometry between the two triangles and their flattened counterparts, giving a piecewise straight line. This idea readily generalizes to higher dimensions \[24\]. In Section 3.5, this idea will be generalized to \( \text{REG}^r \) for all \( r \geq 1 \). The general case is quite subtle. For \( \text{REG}^r \) metrics, the mesh can be given a metric-dependent piecewise smooth globally \( C^1 \) differential structure, under which it is a \( C^1 \)-manifold having singularities at low-dimensional faces with a \( C^0 \)-Riemannian metric on it. It will be shown that local geodesics can be defined as Carathéodory solutions to the geodesic equation away from the singularities.

The second idea \[116\] is motivated by physical interpretations. For \( \text{REG}^0 \), free falling test particles follow straight lines in the interior of each simplex in the mesh as usual because they are flat. When the trajectory crosses an interior facet, the part of the velocity tangential to the facet should not change due to the tangential-tangential continuity of the metric. It remains to define how the normal component should change. Physically, the energy (the squared length of the velocity vector measured in the metric) is conserved during a free fall. Hence it is reasonable to require that the squared length of the velocity measured in both sides of the facet to be equal. It will be shown later that this is equivalent to requiring the normal projection on both sides to have the same magnitude. Hence at the facet, the tangential part of the velocity remains the same while the normal part rotates to match the facet normal on the other side. This is illustrated in Figure 3.2. This variational approach readily generalizes to \( \text{REG}^r \) in all dimensions and for all \( r \geq 0 \). This is derived rigorously in Section 3.4. Under this the geodesics are usual smooth geodesics inside each cell and rotates in this way when they cross interior facets.

\[ \text{Figure 3.1: Illustration of the first geodesic idea.} \]
It will be proven that the geometric approach and the variational approach lead to the same definition for local geodesics on REG\(r\) metrics (Theorem 3.8). While the variation approach is easy to understand and use, the more abstract geometric approach reveals some subtle structure of the geodesics. In particular, it will be shown in Section 3.6 that geodesics on REG\(r\) metrics still have a symplectic structure. This, for example, suggests that symplectic discretizations should be used to compute geodesics in this case.

The situation becomes complicated when a local geodesic reaches a face of dimension \(\leq (n-2)\) in a mesh of dimension \(n\). It is a known pathology [4,92] that in general the geodesics becomes undefined in this case. In this thesis, the goal is to use REG\(r\) metrics as approximations to smooth Riemannian metrics. Therefore such pathologies are considered as artifacts and not features of the discrete geometry. The details will be discussed in Section 3.4.

In Section 3.7, a robust algorithm to compute local geodesics on REG\(r\) metrics is proposed. The basic idea is to use a symplectic collocation method to solve the Hamiltonian formulation of the geodesic equation inside each cell, then rotate the momentum crossing interior facets as prescribed in the variational picture. To make this useable, accurate, provably halt in finite time, and robust against various numerical issues is nontrivial. One feature of the algorithm is a robust way of dealing with the pathology above where the local geodesic comes close to a low-dimensional face, with which the numerical solution continues at the cost of error negligible error. Further, the algorithm is implemented in Python using FEniCS as a part of this thesis.

In Section 3.8, the error between the smooth geodesic on a smooth Riemannian metric and the geodesics on a sequence of REG\(r\) metrics approximating that metric is derived. Since
usually the metric approximation is a harder problem than solving ODEs, it is reasonable to assume that the ODEs solver error is small or of higher order compared to the error due to the metric approximation. Thus the results in that section essentially gives the practical a priori error estimates between the smooth geodesic on the smooth metric and the computed numerical geodesic on the approximating REG" metric.

Finally, in Section 3.9, computational examples using the geodesic code are given for Keplerian orbits and Schwarzschild orbits. Figure 3.3 shows the result of the computation of 50 periods of a Keplerian orbit on the same mesh with REG" for r = 0, 1, 2, 3. The exact orbit is periodic and follows an ellipse. The advantage of going to higher degrees is obvious.

Figure 3.3: Keplerian orbits. Left to right, top to bottom r = 0, 1, 2, 3.

Figure 3.4 shows the result of the computation of 50 periods of a Schwarzschild orbit on
the same mesh with $\text{REG}^r$ for $r = 0, 1, 2, 3$. The exact orbit is almost an ellipse with a slow precession (its major axis slowly rotates around the center). Because the discretization is almost symplectic, qualitatively, the numerical solution would precess even for the Keplerian case where there is no precession. Comparing to previous plots, it is clear that $\text{REG}^r$ with higher degrees performs much better than $\text{REG}^0$. For $\text{REG}^0$, it is even difficult to tell if the orbit is Keplerian or Schwarzschild. This showcases the need for higher degree $\text{REG}^r$ developed here for relativistic simulations.

![Schwarzschild orbits](fig:geodesic_schwarzschild_example)

**Figure 3.4:** Schwarzschild orbits. Left to right, top to bottom, $r = 0, 1, 2, 3$.

In the case of Keplerian orbits, the exact solution can be evaluated for arbitrary large time to arbitrary accuracy via analytical methods. This will be used to validate the error estimates proved in Section 3.8 and also to test the long-time behavior of the solver. To give
a sense of the result, the rates for the error in the position, energy, and momentum in terms of the fineness of the discretization and time $t$ are listed in Table 3.1. These are compared with the error estimates for standard ODE solvers from [47]. The first three are obtained by applying the standard ODE solvers using the smooth metric, where $h_s$ is the constant step size. For the last two, the Kepler problem is transformed into a geodesic problem and the relevant metric is interpolated into REG$^r$ on a unstructured mesh of size $h$ using the canonical interpolant. The geodesics were then computed using the proposed algorithm with a step size smaller but comparable to $h$. The naive one is obtained by applying the ODE solvers directly without taking the non-smoothness of REG$^r$ into consideration, that is, without the rotations at interior facets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Error in position</th>
<th>Error in energy</th>
<th>Error in momentum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit Euler</td>
<td>$t^2 h_s$</td>
<td>$t h_s$</td>
<td>$t h_s$</td>
</tr>
<tr>
<td>Implicit Euler</td>
<td>$t^2 h_s$</td>
<td>$t h_s$</td>
<td>$t h_s$</td>
</tr>
<tr>
<td>Collocation at Gauss(2r)</td>
<td>$t h_s^{2r}$</td>
<td>$h_s^{2r}$</td>
<td>0</td>
</tr>
<tr>
<td>Naive REG$^r$</td>
<td>$t^2 h^r$</td>
<td>$th^r$</td>
<td>$th^r$</td>
</tr>
<tr>
<td>REG$^r$</td>
<td>$(t + et^2)h^{r+1}$</td>
<td>$h^{r+1}$</td>
<td>$(1 + et)h^{r+1}$</td>
</tr>
</tbody>
</table>

Table 3.1: Convergence rates comparison for geodesic solvers.

Because the geodesics have to exit the cell at the cell boundary. Inevitably, the step size for the symplectic solver inside each cell cannot be completely uniform. This causes a slow loss of symplecticity of the geodesic solver, which shows up as the $\epsilon$-terms for REG$^r$ in the table. In practice, this effect is negligible, except for very very long-term computations. For example, for the Kepler problem, the quadratic term in the error in position is not observable even after 10000 orbits. The linear growth in the error of the momentum, however, is clearly observable for $r \geq 2$ but remains very small for a long time.

### 3.2 Review of the smooth geodesic theory

In this section, basic facts of geodesics on smooth Riemannian manifolds are reviewed.

Let $(M,g)$ be a smooth Riemannian manifold. A piecewise smooth curve is a continuous function $\gamma : [a,b] \rightarrow M$ with a finite partition $a = t_0 < \cdots < t_n = b$ such that each restriction $\gamma|_{(t_i, t_{i+1})}$ is smooth. The length of such $\gamma$ is defined as:

$$L(\gamma) := \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} \sqrt{g_{ij}(\gamma(t))\dot{\gamma}^i(t)\dot{\gamma}^j(t)} \, dt.$$  

(3.1)
The curve length is invariant under reparameterization. A natural choice is *reparametrization by arc length*, where the parameter value is required to equal its length along the curve:

\[ L(\gamma|_{[a,t]}) = t - a, \quad \forall t \in [a,b]. \]

It is convenient for later discussion to relax this a little bit. A curve is of *constant speed* if

\[ L(\gamma|_{[a,t]}) = c(t - a), \quad \forall t \in [a,b], \]

for some constant \( c > 0 \). Clearly, a curve is of constant speed if and only if the *kinetic energy* \( g_{ij} \dot{\gamma}^i \dot{\gamma}^j \) is constant along the curve. It is parameterized by arc length if \( c = 1 \).

The curve length induces a metric structure on the Riemannian manifold \( M \): for \( p, q \in M \), the *distance* between them is defined as:

\[ d(p, q) := \inf \{ \text{lengths of all piecewise smooth curves connecting } p \text{ and } q \}. \]

The minimizers \( \gamma \) which are of constant speed are called *global geodesics*. For such curves, by definition, for any \( t_1 \) and \( t_2 \) in its domain,

\[ d(\gamma(t_1), \gamma(t_2)) = L(\gamma|_{[t_1,t_2]}) = c|t_1 - t_2|. \]

If \( \gamma \) happens to be parameterized by arc length, that is \( c = 1 \), then it is called a *minimizing geodesic*.

Global geodesics are important in optimization and planning applications, for example see [80]. The global nature of these can be inappropriate for many other applications. This leads to another useful notion. A piecewise smooth curve is a *local geodesic* if every point on the curve has a neighborhood where equation (3.2) holds. In applications like mechanics and relativity, physical laws are generally assumed to be local. In this case, local geodesics are more meaningful.

To derive a usable local condition for local geodesics, it is convenient to introduce the *energy functional*:

\[ E(\gamma) := \frac{1}{2} \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} g_{ij}(\gamma(t)) \dot{\gamma}^i(t) \dot{\gamma}^j(t) dt. \]

It is similar to the length but without the square root in the integrand. The energy is not invariant under reparameterization of the curve, so its minimizers are more constrained. Further, Cauchy-Schwarz inequality implies that for piecewise smooth \( \gamma : [a,b] \to M \),

\[ L(\gamma)^2 \leq 2(b - a)E(\gamma). \]

The equal sign holds if and only if the kinetic energy is constant. The following theorem is well-known ( [32, Chapter 3 and 9]):
Theorem 3.1. Let \((M,g)\) be a smooth Riemannian manifold. A piecewise smooth curve \(\gamma: [a,b] \to M\) is a critical point of \(E\) if and only if \(\gamma\) is smooth and solves the geodesic equation:

\[
\ddot{\gamma}^i + \Gamma^i_{kl} \dot{\gamma}^k \dot{\gamma}^l = 0,
\]

where \(\Gamma^i_{jk}\) is the Christoffel symbol associated with \(g\) defined by:

\[
\Gamma^i_{jk} := \frac{1}{2} g^{il} (\partial_k g_{ij} - \partial_i g_{jk} + \partial_j g_{ki}).
\]

Moreover, such \(\gamma\) has constant kinetic energy. Thus, it is a critical point of the length \(L\) with constant speed, or equivalently, a local geodesic.

The geodesic equation is a second-order ordinary differential equation (ODE). This can be used to setup evolution problems given an initial position and velocity. Formally, given \(p \in M\) and \(v \in T_p M\), the local geodesic problem tries to find a smooth curve \(\gamma(t)\) satisfying

\[
\ddot{\gamma}^i + \Gamma^i_{kl} \dot{\gamma}^k \dot{\gamma}^l = 0, \quad \gamma(0) = p, \quad \dot{\gamma}(0) = v.
\]

In sum, the problem of finding a global geodesic given two points \(p,q\) is akin to a boundary value problem while the problem of finding a local geodesic given a point and a velocity vector is an initial-value problem. Both are interesting in applications.

3.3 Global geodesics on Regge metrics

In this section, global geodesics on generalized Regge metrics are defined rigorously. This is most natural under the framework of metric geometry \([19, 45]\). It is an elegant formulation of the relationship between distances and lengths of curves.

Let \((X,d)\) be a metric space. The length of a continuous curve \(\gamma: [a,b] \to X\) is defined in terms of the distance:

\[
L(\gamma) := \sup \sum_{[t_i]} d(\gamma(t_i), \gamma(t_{i+1})),
\]

where the supremum is taken over all finite partitions of \([a,b]\). When \(L(\gamma) < \infty\), \(\gamma\) is said to be rectifiable.

Global geodesics are defined as rectifiable curves \(\gamma(t)\) of constant speed, that is, for any \(t_1\) and \(t_2\) in its domain,

\[
d(\gamma(t_1), \gamma(t_2)) = c|t_1 - t_2|,
\]
for some constant \( c > 0 \). The curve is called \textit{minimizing} if \( c = 1 \). Clearly, by equation (3.2), smooth Riemannian global geodesics in the previous section are included as a special case of this.

The length functional induces another distance function on \( X \) called the \textit{intrinsic distance}:

\[
d_I(p,q) := \inf \{ \text{lengths of all rectifiable curves } \gamma \text{ connecting } p \text{ and } q \}. \tag{3.5} \]

Well-behaved metric spaces satisfies \( d_I = d \) and are called \textit{length spaces}.

One easy way to construct length spaces is to start with some well-behaved curve length functional and then define the distance as its associated intrinsic distance in equation (3.5). Indeed, a smooth Riemannian manifold described in the previous section leads to a length space this way.

Generalized Regge metrics is the natural analog of smooth Riemannian metrics among piecewise polynomial metrics in the context of length space:

\textbf{Theorem 3.2.} Let \( T \) be a mesh and \( g \) a piecewise polynomial Riemannian metric on \( T \). The length \( L \) defined in equation (3.1) is single-valued for piecewise smooth curves in \( T \) if and only if \( g \in \text{REG}^r(T) \). In this case, \( (T, d_g) \) is a length space, where \( d_g \) is the intrinsic distance induced by the curve length under \( g \).

\textit{Proof.} From definition, \( L(\gamma) \) is well-defined for a curve \( \gamma \) inside a \( k \)-face \( f \) of the mesh if and only if \( \iota^*_f g \) are single-valued on \( f \) pulling back from all cells containing \( f \). The characterization theorem of generalized Regge elements (cross-chapter ref) states that a piecewise polynomial covariant 2-tensor field on a mesh has single-valued pullbacks \( \iota^*_f g \) for all faces of dimension \( \geq 1 \) of the mesh if and only if \( g \in \text{REG}^r(T) \). This proves the first part. When \( g \in \text{REG}^r(T) \), the curve length functional \( L(\gamma) \) is the same as the Riemannian case. In particular, the induced distance \( d_g \) satisfies the requirements of being a distance function. Hence \( (T, d_g) \) is a length space by definition.

For the lowest degree case \( \text{REG}^0 \), each simplex is flat. Riemannian metrics in \( \text{REG}^0 \) can be realized geometrically as triangulated polytopes embedded in some Euclidean space [24]. The global geodesics in this case have been studied extensively in mathematics [4] and in computational geometry [50, 70]. From the discretization and approximation point of view, these works focus on the extrinsic polyhedral approximations of smooth embedded surfaces and the geodesics on the approximate surfaces. In this thesis, however, the intrinsic approximation of the metric will be the main focus instead. Convergence questions here can be reduced to approximation properties of the discrete metric, which has been addressed in Chapter (cross-chapter ref).
Another potentially interesting problem is the computation of the distance function on a generalized Regge metric. In the smooth case, this is equivalent to solving an Eikonal equation and can be discretized by the Fast Marching Method [61]. This is an extensively studied area in computational geometry. The generalized Regge case is not discussed here and will be future work.

### 3.4 Local geodesics on Regge metrics: variational approach

In this section, the definition and basic properties of local geodesics of a generalized Regge metric are studied. This is of particular interest in mathematical physics and numerical analysis.

Let $\mathcal{T}$ be a mesh and $g$ a generalized Regge metric on $\mathcal{T}$. As described in the previous section, $(\mathcal{T}, g)$ is also a length space. The local geodesics can therefore be defined again as curves which satisfies the geodesic condition locally. More precisely, a piecewise smooth curve $\gamma(t)$ in $\mathcal{T}$ is a local geodesic if and only if every point on it has a neighborhood where it is of constant speed:

$$d(\gamma(t_1), \gamma(t_2)) = c|t_1 - t_2|.$$ 

As discussed in the previous section, the length and energy of a piecewise smooth curve are well-defined on $(\mathcal{T}, g)$. By Cauchy-Schwarz inequality again, it is clear that the above definition of a local geodesic is equivalent to requiring $\gamma$ to be a critical point of the energy functional locally. This will be used to derive a local condition for local geodesics in Theorem 3.3. But before that, there is some subtlety which needs to be addressed.

While global geodesics of generalized Regge metrics are very similar to their smooth Riemannian counterparts, the local geodesics have some significant differences, due to the non-smooth nature of the metric. In particular, the crucial local geodesic initial-value problem does not carry over directly. These pathologies already show up for REG$^0$. Below examples of these are given. Then a more refined definition of a generalized local geodesic initial-value problem and its solution strategy will be described.

First, there is ambiguity about the tangent space when a point is at some interior faces. For example, consider the apex $p$ of the tetrahedron in Figure 3.5. It is clear that a meaningful initial velocity must belong to the tangent space of a particular triangle at $p$. Thus, unlike the smooth case, where the state of the system is specified by a point in the manifold and a velocity in the tangent space, on a mesh $\mathcal{T}$, the state of the system is specified by a cell $c$ of $\mathcal{T}$, a point $p \in c$, and a velocity vector $v \in T_p c$. 

53
Second, unlike a smooth local geodesic which can be extended indefinitely, a local geodesic on generalized Regge metrics in general cannot be extended further if the curve hits an interior face of dimension \( \leq (n - 2) \) in a mesh of dimension \( n \). For \( \text{REG}^0 \), this is known in the computer graphics literature [92,93], but does not seem to be known in the physics literature. An example of this is illustrated in Figure 3.6. On the left is a three-triangle tetrahedral tent embedded in \( \mathbb{R}^3 \). Suppose it is cut along the edge marked by \( \succ \) and then flatten in \( \mathbb{R}^2 \) to the shape on the right. This operation is an isometry and the geodesics are just straight line segments in the flattened figure. It is clear that no global geodesic passes through the apex (circled). Hence if a local geodesic hits the apex, it cannot be extended further. In general, for two-dimensional \( \text{REG}^0 \), when a local geodesic hits a vertex, either it cannot be extended further if the sum of angles around that vertex is smaller than \( 2\pi \) or it has an infinite family of extensions if the angle sum is greater than \( 2\pi \).

For the two-dimensional \( \text{REG}^0 \) case, various generalizations of the notion of local geodesics were proposed in the literature [4,92], where the curves are required to be “straight” in some other sense. These ideas do not generalize directly to higher dimensions or to higher degree \( \text{REG}^r \). In this thesis, the focus is on the case where the non-smooth metric is itself an approximation to some smooth metric. These pathologies are thus considered artifacts rather
than an interesting feature of the discrete geometry.

A generic curve (submanifold of dimension 1) cannot hit a face of dimension \( n - 2 \) almost surely. In particular, for numerical computations, one can always perturb the solution within the machine precision to get around a low dimensional face. For a generic generalized Regge metric, this is problematic because the geodesics are not stable near a face of low dimension, as shown in the tent example. However, when the generalized Regge metric is an approximation to some smooth metric, it will be shown later in this chapter that the error committed converges to zero as the mesh is refined. Hence for the purpose of this thesis, only local geodesics that do not intersect low dimensional faces need to be considered.

Other than the two pathologies just described, the local geodesics on \( \text{REG}^r \) are similar to their smooth counterparts. The next step is to prove an analog of Theorem 3.1 describing a local condition for local geodesics. In this case, there is nothing special about \( \text{REG}^r(\mathcal{T}) \), the theorem will be applicable to any piecewise smooth Riemannian metric \( g \) on \( \mathcal{T} \) with tangential-tangential continuity: for any interior facet \( f \) of \( \mathcal{T} \), \( \iota_f^* g \) is single-valued evaluated from any cell containing \( f \). In particular, this contains the space of smooth Riemannian metrics on \( \mathcal{T} \) as a special case. This slightly more general case will be considered because it makes the study of error analysis later easier.

Before stating the theorem, some convenient notations are introduced for a frequently arising situation depicted in Figure 3.7. Suppose \( g \) is a piecewise smooth Riemannian metric on some mesh. Let \( c^+ \) and \( c^- \) be two cells intersecting at a facet \( f \). Suppose a piecewise smooth curve \( \gamma \) crosses \( f \) at a point \( p \) in the interior of \( f \). Note that there is a natural identification of the subspace \( T_p f \subset T_p c^+ \) with the subspace \( T_p f \subset T_p c^- \) via the affine structure intrinsic to \( f \). This identification is assumed implicitly throughout this chapter. Other quantities are however discontinuous. In such a situation, \( g_{ij}^+ \) is defined to be the restriction of \( g \) in \( c^+ \), \( n^+_i \) the unit outward normal vector to the facet \( f \) at \( p \) under \( g_{ij}^+ \), and \( \dot{\gamma}^+_i \in T_p c^+ \) the velocity vector of \( \gamma \) at \( p \). Quantities like \( g_{ij}^-, n^- \), and \( \dot{\gamma}^-_i \) are similarly defined in \( c^- \).
Theorem 3.3. Let $\mathcal{T}$ be a mesh of dimension $n$ and $g$ a piecewise smooth Riemannian metric with tangential-tangential continuity. A piecewise smooth curve $\gamma : [a, b] \rightarrow \mathcal{T}$ which does not intersect any interior faces of dimension $\leq (n - 2)$ is a local geodesic if and only if it satisfies the geodesic equation (3.4) inside each cell and at each point $p$ where $\gamma$ intersects a facet $f$, the tangential projection of $\dot{\gamma}$ is the same on both sides: for all vectors $t^j \in T_p f$,

$$g_{ij}^+ t^i = g_{ij}^- t^i$$

and the normal projection has the same length on both sides:

$$g_{ij}^+ \dot{\gamma}^i n^j_+ + g_{ij}^- \dot{\gamma}^i n^j_- = 0.$$

In particular, the kinetic energy $g_{ij} \dot{\gamma}^i \dot{\gamma}^j$ is constant along any local geodesic (even when the curve crosses a facet). Moreover, $\gamma$ is $C^{0,1}$ globally. If $g$ happens to be in $C^k$ globally, $k \geq 0$, then $\gamma$ is in $C^{k+1,1}$ globally. If $g$ happens to be smooth, then $\gamma$ is smooth and solves the usual smooth geodesic equation everywhere.

Before proving this theorem, a corollary very useful for computations is given:

Corollary 3.4. Suppose $g$ is piecewise smooth with tangential-tangential continuity and $\gamma$ crosses an interior facet $f$ as depicted in Figure 3.7, then at point $p \in f$, the value of $\dot{\gamma}^i$ satisfies the following update formula:

$$\dot{\gamma}^- = \dot{\gamma}^+ - (g_{jk}^+ \dot{\gamma}^j n^k_+)(n^i_+ + n^i_-),$$

Proof. Set $a_\pm := g_{ij}^\pm \dot{\gamma}^i n^j_\pm$ and $t^i_\pm := \dot{\gamma}^i_\pm - a_\pm n^i_\pm$. The theorem implies that

$$t^i_+ = t^i_-, \quad a_+ + a_- = 0.$$
Thus, \[
\dot{\gamma}_+^i - \dot{\gamma}_-^i = (t_+^i + a_+ n_+^i) - (t_-^i + a_- n_-^i) = a_+ (n_+^i + n_-^i),
\]
which proves the identity in the claim.

Thus, analytically, the \textit{generalized initial-value problem} for local geodesics can be solved by alternating between solving the smooth geodesic equation inside each cell until the curve hits the cell boundary and applying equation (3.6) to move to the next cell. The procedure has to stop when the local geodesic hits a low-dimensional face.

In the literature, results similar to Theorem 3.3 for non-smooth metrics are derived through variational methods \cite{49,74}, or through Filippov’s theory for differential inclusions \cite{102}, or through the regularization of the metric \cite{75, 76}. In another direction, similar results for curved interface were derived in \cite{40}. The REG\textsuperscript{0} case was derived in \cite{116}. The case considered here has a simple proof and much stronger conclusions (namely uniqueness and regularity). To prove Theorem 3.3, the following lemma on the variation of the energy functional is needed:

\textbf{Lemma 3.5.} Let $\mathcal{T}$ be a mesh of dimension $n$ and $g$ a piecewise smooth Riemannian metric. Suppose $\gamma : [a, b] \to \mathcal{T}$ is a piecewise smooth curve which does not cross interior faces of dimension $\leq (n - 2)$ in $\mathcal{T}$. Let $\gamma_s(t) : ( - \epsilon, \epsilon) \times [a, b] \to \mathcal{T}$ be a smooth family of variations:

$\gamma_0(t) = \gamma(t)$ for all $t$, $\gamma_s(a) = \gamma(a)$, $\gamma_s(b) = \gamma(b)$ for all $s$, and $\gamma_s(t)$ is $C^\infty$ in $s$ for each $t$,

with $\epsilon > 0$ small enough that none of $\gamma_s(t)$ intersect any interior faces of dimension $\leq (n - 2)$. Let $v$ be the variational vector field of $\gamma_s$ relative to $\gamma$:

$v(t) := \frac{\partial}{\partial s} \gamma_s(t) \bigg|_{s=0}$.

Then the variation of the energy functional is:

\[
\frac{\partial}{\partial s} E(\gamma_s) \bigg|_{s=0} = \frac{1}{2} \sum_{i=1}^{n-1} \left( g_{ij}^+ \dot{\gamma}_+^i \dot{\gamma}_+^j - g_{ij}^- \dot{\gamma}_-^i \dot{\gamma}_-^j \right) \bigg|_{t=t_i} + \sum_{i=1}^{n-1} \left( g_{ij}^+ \gamma_s^i v_s^j - g_{ij}^- \gamma_s^i v_s^j \right) \bigg|_{t=t_i} - \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} (\dot{\gamma}^i + \Gamma^i_{kl} \dot{\gamma}^k) g_{ij} v^j dt,
\]

where $t_i$ are points in the domain of $\gamma$ where either $\dot{\gamma}$ is discontinuous or $\gamma$ crosses an interior facet.

\textbf{Proof.} This is just a direct computation from the definition:

\[
E(\gamma_s) = \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} g_{ij} \gamma_s^i \dot{\gamma}_s^j dt,
\]

and integration by parts.
Then the main theorem of this section follows:

**Proof of Theorem 3.3.** Lemma 3.5 gives the condition for $\gamma$ to be a critical point of the energy functional. At a point of discontinuity of $\dot{\gamma}$ in the interior of a cell, the situation is exactly the same as the smooth Riemannian case. These conditions forces $\gamma$ to be smooth and solves the geodesic equation in the interior of each cell. At a point $t_i$ where $\gamma$ crosses an interior facet $f$ at $p = \gamma(t_i)$, the conditions for critical points require:

$$g^{i}{}_{ij} \dot{\gamma}^i = g_{ij} \dot{\gamma}^j,$$

$$g^{i}{}_{ij} \dot{\gamma}^i w^j = g_{ij} \dot{\gamma}^i w^j, \text{ for all } w^j \in T_p f.$$

The first equation implies the first condition in Theorem 3.3 directly. The second condition follows from the fact that $s \mapsto \gamma_s(t_i(s))$ is by definition a curve in $f$ where $t_i(s)$ is the time $\gamma_s$ crosses that facet. So the corresponding variational vector field must be tangential to $f$.

Now, set $a_{\pm} := g_{ij} \dot{\gamma}_+^i n_\pm^j$ and $t_{i,\pm} := \dot{\gamma}_\pm^i - a_{\pm} n_\pm^i$. The second condition and tangential-tangential continuity together imply that $t_{i,\pm} = t_{i,\pm}$. Then the first condition implies that:

$$a_{+}^2 = a_{-}^2.$$

But $\gamma$ is leaving $c^+$ and entering $c^-$, which means $a_+$ and $a_-$ have opposite signs. Thus $a_+ + a_- = 0$. This proves the tangential projection and normal projection conditions in the theorem. This in particular shows that the critical points of the energy functional has constant kinetic energy. Thus they are critical points of the length with constant speed locally, or equivalently, local geodesics.

Finally, by standard ODE theory, $\gamma$ is smooth inside each cell. The two facet conditions and their derivatives imply that if $g$ is $C^k$ globally, then $\dot{\gamma}$ is $C^{k,1}$ globally. This proves the regularity claim.

**3.5 Local geodesics on Regge metrics: geometric approach**

In the introduction, two different intuitive approaches were given to compute the geodesics on $\text{Reg}^0$. In the previous section, the variational approach was generalized to handle higher degree $\text{Reg}^r$ cases. In this section, the cut-flatten-glue approach is generalized to piecewise smooth metrics with tangential-tangential continuity, which include $\text{Reg}^r$ as a special case. This is not as straightforward as the variational approach. Further, like the cut-flatten-glue approach before, this more abstract view is not useful directly for numerical computations.
Nevertheless, it offers crucial geometric insights into the structure of generalized Regge metrics. In particular, it is useful for understanding the symplectic structure in the next section.

Given a mesh $\mathcal{T}$ in $\mathbb{R}^n$ and a piecewise smooth Riemannian metric with tangential-tangential continuity $g$ on $\mathcal{T}$. As an embedded submanifold of $\mathbb{R}^n$, $\mathcal{T}$ is a smooth manifold with polygonal boundary. Under this, $(\mathcal{T}, g)$ can be viewed as a smooth manifold with a piecewise smooth Riemannian metric. But there is nothing special about the embedding of $\mathcal{T}$ in $\mathbb{R}^n$. The same information about the metric can be specified simplex by simplex independently.

A more intrinsic but subtle interpretations of $(\mathcal{T}, g)$ is known in the literature for REG$^0$ [24].

Take a 2D mesh for an example. Every triangle in the mesh can be isometrically embedded in Euclidean $\mathbb{R}^2$, with edge lengths given by $g$. Locally, the images of each pair of triangles sharing an edge in the mesh can be glued together to form a trapezoid in Euclidean $\mathbb{R}^2$ as a smooth Riemannian submanifold with polygonal boundary. This gluing operation can be done at all shared edges of the Euclidean triangles. A typical mental image of the result would be a 2D triangulated surface in 3D (in general a higher embedding dimension might be needed). A triangulated surface is no longer a smooth manifold. It has ridges and conic points. A more careful construction can get rid of the ridges by going to a higher dimension every time a new triangle is added (indeed, each pair of triangles can be glued together to a trapezoid without ridges). But the conic points will persist. Under this view, the 2D REG$^0(\mathcal{T})$ corresponds to an abstract Riemannian manifold, which is a smooth manifold with a constant Euclidean metric away from the vertices and is singular at the vertices. The proper framework for this is the theory of stratified manifolds. But, for this chapter, as discussed before, the vertices can be simply discarded. In general for REG$^0(\mathcal{T})$ of dimension $n$, let $\hat{\mathcal{T}}$ be the manifold obtained by removing faces of dimension $\leq (n-2)$ from the mesh $\mathcal{T}$. Then an abstract smooth manifold $\hat{\mathcal{T}}$ with the Euclidean metric can be obtained from REG$^0(\mathcal{T})$ using a similar construction [24].

For the general case, the following theorem will be proved in Chapter (cross-chapter ref) of this thesis:

**Theorem 3.6.** Let $\mathcal{T}$ be a mesh of dimension $n$ and $g$ a piecewise smooth Riemannian metric with tangential-tangential continuity on $\mathcal{T}$. There exists an atlas depending on $g$ for $\mathcal{T}$ which is piecewise smooth, globally $C^1$ on $\hat{\mathcal{T}}$, and singular at $\mathcal{T} - \hat{\mathcal{T}}$, under which the piecewise smooth metric $g$ can be extended to a globally $C^0$-Riemannian metric on $\hat{\mathcal{T}}$. Let $\hat{\mathcal{T}}^g$ denote the $C^1$-manifold obtained from the topological manifold $\hat{\mathcal{T}}$ with the aforementioned atlas. Then $g$ is a $C^0$-Riemannian metric on $\hat{\mathcal{T}}^g$ satisfying the condition that each cell in $\hat{\mathcal{T}}^g$ is isometric to its corresponding cell in $(\mathcal{T}, g)$ via a smooth map whose differential is identity on vectors tangential to the boundary facets of each interior cell. Further such $(\hat{\mathcal{T}}^g, g)$ is unique.
This has many geometric consequences which will be discussed in Chapter (cross chapter ref). Notice that for $\text{REG}^0$, as described before, $(\mathcal{F}, g)$ is a smooth manifold with a smooth (globally constant) Euclidean metric. For generalized Regge metrics $\text{REG}^r$ with $r > 0$, the abstract manifold is less smooth.

Nevertheless, this has enough regularity for geodesics. Indeed, a piecewise smooth and globally $C^0$ metric on a mesh is Lipschitz. Its Christoffel symbols, which depend on up to the first derivatives of the metric, are piecewise smooth but globally discontinuous functions. It turns out that the usual geodesic equation (3.4), though still does not make sense in the classical view, becomes well-posed in some general sense. Geodesics on Lipschitz metrics were studied in the physics literature with the application of geodesics in gravitational shock waves [68, 102]. In general, it was proved in [102] that the local geodesic problem has $C^1$-solutions on Lipschitz Riemannian metrics in the Filippov sense as a direct application of the theory of differential inclusions in [36]. The detailed discussion on this will not be pursued here.

Instead, an elementary treatment will be given for the special case here where the metric is further piecewise smooth and the local geodesics are required to be transverse to the interior facets.

**Theorem 3.7.** Let $M$ be a mesh of dimension $n$ with a piecewise smooth globally $C^1$ smooth structure which might be singular at faces of dimension $\leq (n - 2)$ and $g$ a piecewise smooth $C^0$-Riemannian metric on $M$. Suppose $q_0$ is a point in the interior of some cell $c$ in $M$ and $v_0 \in T_{q_0}c$. Starting with initial data $(q_0, v_0)$, construct a curve $\gamma : [0, T] \rightarrow M$ by alternating between solving the smooth geodesic equation inside a cell and move to the next cell by continuity of $\gamma$ and $\dot{\gamma}$. This process can go on as long as $\gamma$ exits cells transversely in the interior of a facet. Then $\gamma \in C^{1,1}$ and it solves the geodesic equation on $(M, g)$ almost everywhere (that is, a Carathéodory solution). In particular, it is the unique $C^{1,1}$ curve which satisfies the initial condition, crosses interior facets transversely, and solves the geodesic equation almost everywhere.

**Proof.** This is obvious. Inside each cell, the solution to the geodesic equation is smooth. Because $\dot{\gamma}$ is piecewise smooth and globally continuous, on a bounded interval, $\gamma \in C^{1,1}$. From the transverse condition, $\gamma$ can only intersect interior facets and fails to satisfy the geodesic equation at a null subset of $[0, T]$. The uniqueness follows from the uniqueness of the smooth geodesic in each cell and the continuity conditions. \qed
Local geodesics defined in this way agree with the local geodesics defined variationally in the previous section:

**Theorem 3.8.** Let $\mathcal{T}$ be a mesh of dimension $n$ and $g$ a piecewise smooth Riemannian metric with tangential-tangential continuity on $\mathcal{T}$. Let $(\mathcal{\hat{T}}^g, g)$ be the induced abstract Riemannian manifold and $\Phi : \mathcal{\hat{T}}^g \rightarrow \mathcal{T}$ be the piecewise smooth isometry in Theorem 3.6. Take any cell $c$ in $\mathcal{T}$, any point $p \in c \cap \mathcal{\hat{T}}^g$, and any vector $v \in T_p c$. Let $\gamma$ in $\mathcal{\hat{T}}^g$ be the curve defined in Theorem 3.7 for $(\mathcal{\hat{T}}^g, g)$ with initial data $(q, v)$ and $\gamma'$ in $\mathcal{T}$ be the local geodesic constructed using Theorem 3.3 with initial data $(c, q, v)$. Then $\Phi \circ \gamma = \gamma'$ as long as they are defined.

**Proof.** This can be proved cell by cell. In cell $c$, $\Phi$ is a smooth isometry. By definition, $\gamma$ and $\gamma'$ are solutions to the same smooth geodesic equation with the same initial data. By standard ODE theory, $\gamma$ and $\gamma'$ coincide. Both then exits $c$ at the same point in the interior of one of the boundary facets $f$ of $c$ with the same velocity. On $(\mathcal{\hat{T}}^g, g)$, because $g$ is $C^0$, the geodesic equation (3.4) implies that $\dot{\gamma}$ is at least $C^0$ and therefore the solution $\gamma$ is at least $C^1$. Hence, necessarily the kinetic energy and the facet tangential part of $\gamma$ are preserved crossing $f$. These two conditions determines the velocity $\dot{\gamma}$ on the other side of the facet uniquely in $(\mathcal{\hat{T}}^g, g)$. Both conditions are invariant under $\Phi$. The preservation of these two are exactly the conditions for local geodesics in $(\mathcal{T}, g)$ in Theorem 3.3. This proves the equivalence.  

### 3.6 Hamiltonian structures of local geodesics

Hamiltonian mechanics offers an elegant and efficient way to encapsulate many important properties of physical systems in a mathematical framework [9]. It is known that smooth local geodesics can also be formulated in the Hamiltonian framework [33, Section 28.3]. It has been recognized for a long time that the preservation of the Hamiltonian structure is of great importance for the discretization of such systems because this is crucial for retaining the correct qualitative behavior and leads to good long-time error properties [47]. In this section, it will be demonstrated that local geodesics of generalized Regge metrics or piecewise smooth Riemannian metrics with tangential-tangential continuity in general also have a Hamiltonian structure. This suggests that a symplectic discretization should be used for computing local geodesics in this case as well. The advantage of the use of a symplectic discretization scheme will also be demonstrated later through numerical examples.

First, the smooth case is reviewed. Let $g$ be a smooth Riemannian metric on a smooth manifold $M$. The *Hamiltonian for geodesics* is a functional on the cotangent bundle $H$:
$T^* M \to \mathbb{R}$ given by

$$H(p,q) := \frac{1}{2} g^{ij}(q)p_ip_j,$$

where $q \in M$ and $p \in T^*_q M$ so together $(p,q) \in T^* M$. The corresponding equation of motion is:

$$q^i = \frac{\partial H}{\partial p^i} = g^{ij} p_j,$$

$$\dot{p}_i = -\frac{\partial H}{\partial q^i} = -\frac{1}{2} p_j p_k \partial_i g^{jk}.$$

It is clear that under the substitution $\gamma(t) = q(t)$ and $\dot{\gamma}^i = g^{ij} p_j$, the Hamiltonian equation of motion (3.7) and the geodesic equation (3.4) are equivalent. This shows that a local geodesic on a smooth Riemannian manifold is equivalent to a Hamiltonian flow on the cotangent bundle. This makes the machinery from symplectic geometry available to the study of local geodesics.

There are several immediate geometric properties which are consequential for the discretization [47]. First is the conservation of the Hamiltonian, that is, $H$ is constant along any geodesics. This follows from the fact that geodesics have constant kinetic energy $g_{ij} \dot{\gamma}^i \dot{\gamma}^j$ (Theorem 3.1) and $\dot{\gamma}^i = g^{ij} p_j$. Second is reversibility: going forward in time with momentum $p$ is the same as going backward in time with momentum $-p$. Symbolically, let $\phi_t : T^* M \to T^* M$ be the solution map to equation (3.7) and $\rho : (q,p) \mapsto (q,-p)$, then,

$$\rho \circ \phi_t = \phi_{-t} \circ \rho.$$

This can be seen from equation (3.7): when the sign of $p$ is flipped, the right-hand side for $\dot{q}$ flips sign while the right-hand side for $\dot{p}$ is unchanged. This has important consequences for the dynamics of the system [47, Chapter V] (for example, the existence of period orbits). Last and most important is symplecticity, which is the fundamental property of a Hamiltonian system. To explain this, some symplectic geometry is needed. Using variables $p_i$ and $q_j$ for $T^* M$ as before, the symplectic form (on the cotangent bundle) $\omega$ is a 2-form on $T^* M$ given by:

$$\omega := \sum_{i=1}^n dq^i \wedge dp_i.$$

It is easy to verify that $\omega$ is closed $d\omega = 0$ and non-degenerate: $\omega(u,v) = 0$ for all $v$ if and only if $u = 0$ [9, Chapter 8]. Note that the cotangent bundle $T^* M$ is a manifold of dimension $2n$ on its own. Let $J_\omega : T(T^* M) \to T^*(T^* M)$ be a linear map induced by $\omega$: for $u \in T(T^* M)$,

$$[J_\omega(u))(v) := \omega(u,v), \quad \forall v \in T(T^* M).$$
Due to the non-degeneracy of $\omega$, $J_\omega$ is a linear isomorphism. A Hamiltonian $H$ is a real-valued smooth function on $T^*M$. The vector field $X_H := J_\omega^{-1}dH$ on $T^*M$ is called the Hamiltonian vector field. It has the nice property that $\omega$ is conserved along the flows of $X_H$:

$$L_{X_H}\omega = \iota_{X_H}d\omega + d\iota_{X_H}\omega = d[J_\omega(X_H)] = ddH = 0,$$

where $L_{X_H}$ is the Lie derivative, the first step uses Cartan’s magic formula, the second step uses the fact that $\omega$ is closed, the third and fourth step use the definition of contraction and $J_\omega$. The relevance of this to the current discussion is clear with a computation in coordinates. In the coordinates of $(p_i, q_j)$, $J_\omega$ becomes a $2n$-by-$2n$ block matrix [9, Chapter 8, 37C]:

$$J := \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix},$$

where $I$ is the $n$-by-$n$ identity matrix. Under this, the definition of a Hamiltonian vector field $X_H = J_\omega^{-1}dH$ reads:

$$\begin{bmatrix} \dot{p} \\ \dot{q} \end{bmatrix} = J^{-1}\nabla H(p, q) = \begin{bmatrix} -\partial_q H \\ \partial_p H \end{bmatrix},$$

which is exactly the equation of motion for Hamiltonian systems. In particular, under mild regularity conditions, it can be shown that a flow on $T^*M$ is the solution to the equation of motion for some Hamiltonian locally if and only if it preserves the symplectic form [47, Chapter VI Theorem 2.6]. Symplecticity is of great importance because of this. For example, suppose a discrete flow preserves the symplectic form as well. Then it is also the flow of some other Hamiltonian. If one can show this Hamiltonian is a perturbation of the Hamiltonian to be approximated, the whole machinery of Hamiltonian perturbation theory can be deployed to study the qualitative and long-term dynamics of the discrete flow with respect to the exact flow. Indeed, this is the key idea behind the explanation of the desirable properties of symplectic discretizations [47, Chapter X].

Let $\mathcal{T}$ be a mesh of dimension $n$ and $g$ be a piecewise smooth Riemannian metric with tangential-tangential continuity. It is clear that local geodesics on $(\mathcal{T}, g)$ still preserves the Hamiltonian and is reversible. The main result of this section is that local geodesics also have a symplectic structure. The general theory of non-smooth Hamiltonian systems was systematically studied by Marsden [75,76]. The case here fits in that framework. In fact, the situation here is sufficiently simple that an independent treatment with minimal modification to the smooth theory is needed and is given here.

Let $(\mathcal{\hat{T}}^g, g)$ be the abstract Riemannian manifold constructed in Theorem 3.6. Since $\mathcal{\hat{T}}^g$ is only $C^1$ globally, its cotangent bundle $T^*\mathcal{\hat{T}}^g$ is a $C^0$ manifold of dimension $2n$. In particular,
it does not make sense to talk about vector fields and differential forms on $T^*\mathcal{F}^g$ directly. Let $(q_i, p_j)$ be a local coordinate patch for $T^*\mathcal{F}^g$. The only problematic quantity is $dp_j$, which is a piecewise smooth and globally discontinuous function. This does not cause any problem. In the following, it is implicitly understood that the $p$-components of vector fields or differential forms on $T^*\mathcal{F}^g$ are only piecewise smooth. Because $g$ is piecewise smooth and globally $C^0$, using a similar argument to the one used in Theorem 3.7, it is clear that a unique Carathéodory solution $(q, p)$ to the Hamiltonian equation of motion (3.7) can be constructed. In particular, the equation is satisfied almost everywhere, $p$ is piecewise smooth and globally $C^0$ while $q$ is piecewise smooth and globally $C^1$. Let the symplectic form $\omega$ be defined on $\mathcal{F}^g$ using equation (3.8), which is now discontinuous in the $p$-components globally. Then it is still conserved along the flow because the Lie derivative identity (3.9) still holds in the distributional sense. It is in this sense that local geodesics on REG$^r(\mathcal{T})$ have a (metric dependent) symplectic structure.

Because the symplectic structure is defined with respect to the abstract metric dependent manifold $(\mathcal{F}^g, g)$, it is not immediately clear its explicit corresponding structure in the original computational coordinates $(\mathcal{T}, g)$. But this does show that on $(\mathcal{T}, g)$ it is possible to define a generalized Hamiltonian system using certain non-smooth theory for ODEs. This will not be pursued here. Given Theorem 3.8, for the purpose of computing local geodesics, because rotating the velocity as specified in equation (3.6) can be implemented exactly, a discretization is globally symplectic as long as it is symplectic in each cell.

### 3.7 A robust algorithm for generalized local geodesics

Given Theorem 3.3 and its corollary, the computation of generalized local geodesics is straightforward with an exact solver for the usual smooth geodesic equation. Indeed, one can repeat: solve the smooth geodesic equation (3.4) in a cell until the curve hits a facet and then move to the next cell and rotate the tangent vector according to the jump condition (3.6). The process would end when the curve hits a face of dimension $\leq (n - 2)$. In practice, however, there are many problems due to numerical issues and practical concerns. In what follows, a robust method is proposed and implemented for solving the geodesic initial-value problem on Riemannian REG$^r$. Given a mesh, a step size $h > 0$, and a position-momentum pair $(q_0, p_0)$, the algorithm repeats the following steps:

- Identify which cell the initial point is in.
- Solve the smooth Hamiltonian geodesic equation inside the cell using a symplectic collocation method with step size $h$. Step until the curve leaves the current cell.
• Solve for the intersection with the boundary of the cell. Truncate the last step at the boundary.
• Identify the cell for the next step.
• Rotate the momentum crossing to the next cell.

It stops either when the curve exits the computational domain or when a specified time $T > 0$ is reached. In particular, this algorithm does not stop the computation when the curve comes close to a face of dimension $\leq (n - 2)$.

For the first step, the algorithm finds all the cells which are numerically near the starting point $q$. If there is only one such cell, then it is chosen. If there are more than one cells, that is when $q$ is near a face of dimension $\leq (n - 1)$, the tie is broken in the following way. The momentum $p$ is flattened using the metric in each nearby cell $c$ to get an initial velocity $v \in T_p c$. Then for some fixed $\epsilon$ (for example $\epsilon = 0.01$), $q' := q + \epsilon hv$ is computed. The cell with the minimum distance to its $q'$ is chosen. If there are tied minimizers, a random choice is made.

![Figure 3.8: Possible bad initial conditions. The green cell is chosen.](image)

For the next step, the geodesic ODE needs to be solved in the interior of each cell. For $\text{REG}^0$ this is trivial since the geodesics are just straight lines. For higher degree elements, the geodesic equation is nonlinear which cannot be solved in closed form even for $\text{REG}^1$ and has to be solved numerically. As mentioned before, the Hamiltonian structure is frequently of physical importance in geodesic computations. Thus a symplectic discretization of the Hamiltonian geodesic equation is used. Overall, equation (3.7) is solved using Collocation method at Gauss points in the interior of cells. It is known that this implicit method is symmetric and symplectic [47]. There are several notable details. First, the metric $g$ is a piecewise polynomial. The inverse metric appearing in the Hamiltonian equation of motion (3.7) can-
not be represented accurately in a finite element space for symbolic derivative computation. Instead, the gradient of the inverse metric is evaluated exactly via:

\[ \partial_i g^{jk} = -g^{im}g^{kn}\partial_j g_{mn}. \]  \hspace{1cm} (3.10)

Second, due to performance concern, the collocation method is implemented via its equivalent Runge-Kutta method [47, Chapter II Theorem 1.4]. In practice, REG with \( r \geq 4 \) is rarely needed. So collection at 3 Gauss points was chosen as the default solver. This is an order 6 symplectic solver with many good properties. The nonlinear equation at each step is solved using a fixed point iteration with linear extrapolation from the previous step as the initial guess [47, Chapter VIII.6.1]. In practice, a step size \( h \) smaller than the radius of the inscribed sphere of a cell is accurate enough.

Once the discrete geodesic steps outside of its current cell, the facet intersection needs to be computed. To do this, the stage values of the Runge-Kutta method are used to construct the collocation interpolant. Since it is possible that the curve passes through the cell near a face of low-dimensions, special care is needed. In practice, the interval for the interpolant frequently varies from \( 10^{-1} \) to \( 10^{-10} \). A properly scaled robust Barycentric Lagrange interpolant [12] was implemented here for this purpose. Given the Euclidean coordinates of the vertices of a simplex \( c \), the distance from any point to \( c \) can be computed using standard robust routines [86]. Let \( \gamma(t) \) be the interpolant and \( d_c(p) \) be the distance function. A bisection method is implemented to find the first smallest \( t^* \) within some tolerance such that

\[ d_c(\gamma(t^*)) > 0. \]

A standard root finding routine for \( d_c(\gamma(t)) = 0 \) will fail here because it cannot guarantee the curve leaves the current cell beyond numerical tolerance, potentially leading to an infinite loop.

The next step is to identify the next cell to start the next round of the geodesic solver. First, the boundary facet \( f \) of \( c \) which is closest to \( \gamma(t^*) \) is chosen. The ties are broken by a random choice. If \( f \) lies on the domain boundary, the computation terminates. Otherwise, the next cell \( c' \) is the cell opposite to \( c \) at \( f \). Due to numerical issues for the rare situation where \( \gamma(t^*) \) is near a face of dimension \( \leq (n-2) \), a crucial check is needed. If the point \( \gamma(t^*) \) is outside of \( c' \), that is \( d_c(\gamma(t^*)) \) is greater than some small tolerance, then the solver is restarted using the first step to find a new starting cell. This is called a bad crossing. If the point \( \gamma(t^*) \) is inside of \( c' \), which is almost always the case, then the next cell is naturally \( c' \). This is called a good crossing.
Finally, the momentum is rotated using the following formula derived from equation (3.6):

$$p_i^{-} = g_{ik}^-(g_{+}^{kj}n_+^k n_+^j - n_+^k n_+^j)p_+^j.$$  \hfill eq:pgjump

It should be noted that for bad crossings, the momentum is rotated as if the curve is crossing from $c$ to $c'$ via $f$ and then a new cell instead of $c'$ is chosen. This commits an error which will be analyzed in the next section. Intuitively, when the discrete metric is a good approximation of some smooth metric, the error committed is proportional to the tolerance and is thus very small.

It should be noted that unfortunately this algorithm does not lead to a globally symplectic discretization. This is due to the known problem that nonuniform time-stepping degrades the performance of a symplectic integrator [47, Section VIII.3]. Because the curve has to hit the cell boundary, the last step in a cell cannot in general have the same step size as the previous steps. In particular, the correct step size for the last step is not known a priori. Thus current strategies for symplectic discretization with adaptive time stepping cannot be applied here. The problem of finding a fully symplectic implementation remains open. In practice, however, this is less of an issue. Because the metric approximation is the harder problem, the error due to the metric approximation is much larger than the error committed by the ODE solver. Thus, as will be demonstrated in the numerical example section, the errors associated with the violation of the symplectic structure will not dominate the total error except for extremely long-term simulations.

The robust algorithm outlined here is implemented in Python using FEniCS. The code repository can be found at (FIXME: code repository). All the numerical examples later in this chapter are computed using this code.
3.8 Error analysis

Let \((M, g)\) be a smooth Riemannian manifold and \(\gamma : [a, b] \to M\) a smooth geodesic. Suppose \(\{T_h\}\) is a sequence of triangulations of \(M\), on which \(g_h \in \text{REG}'(T_h)\) are Riemannian metrics and \(\gamma_h\) geodesics to \((T_h, g_h)\) with the same initial data as \(\gamma\). This section studies when \(\gamma_h\) is close to \(\gamma\) and how close the approximation is. In practice, it is reasonable to assume that the error in the ODE solver is comparable or of higher order compared to the error due to metric approximation (for example, through the use of time steps finer than the mesh size). Hence the results of this section gives the practical a priori error estimates for the errors between the true geodesic and the computed geodesics on the generalized Regge metrics.

First, the difference measure needs to be specified. This is completely arbitrary. When the mesh \(T\) is given as an embedded manifold in some \(\mathbb{R}^n\), the mesh size of \(T\) is measured in the Euclidean metric there, which is standard in the numerical analysis literature. Error related statements are usually made in terms of the mesh size. Hence it is natural to measure the difference in geodesics using the Euclidean distance between the coordinates of the curves. For the rest of this section, the single bar norm \(| \cdot |\) for tensor values denotes the norm under the Euclidean metric in the background \(\mathbb{R}^n\) coordinates. The Sobolev norms of tensor-valued functions are defined through the Sobolev norms on the point-wise \(| \cdot |\)-norm. For piecewise smooth tensor-valued functions \(u\) on \(T\), notations like \(\|u\|_{W^{s,p}(T)}\) mean the piecewise \(W^{s,p}\)-norm on each cells in \(T\) combined using the scaling of \(p\)-norms in the obvious way. For example, \(\|g\|_{W^{1,\infty}(T)}\) is the maximum over all cells of the \(W^{1,\infty}\)-norm of \(g\) restricted to these cells. When the norm is not taken piecewise, the domain \(T\) in the notation will be suppressed. For example, for a smooth metric \(g\) on \(T\), \(\|g\|_{W^{s,p}}\) is just the usual Sobolev norm. It should be noted that this differs from the convention in the geometry literature, where the differences are measured intrinsically in the smooth Riemannian metric being approximated. Here, this smooth Riemannian metric is usually the unknown in the metric approximation problem. In any case, for non-singular metrics on compact domains, the convergence rates remain the same for both the extrinsic and the intrinsic approach.

The main result of this section is the following theorem:

**Theorem 3.9.** Let \(M\) be a domain in \(\mathbb{R}^n\) and \(T_h\) a family of triangulations of \(M\) parameterized by the mesh size \(h\). Suppose \(g\) is a smooth Riemannian metric on \(M\) and \(g_h \in \text{REG}'(T_h)\) a family of Riemannian metrics satisfying \(\|g - g_h\|_{L^\infty} \leq \frac{\epsilon}{2} \|g^{-1}\|_{L^\infty}^{-1}\) uniformly in \(h\). Suppose \(\gamma : [0,T] \to M\) is a smooth geodesic under \(g\) and \(\gamma_h\) a family of geodesics under \(g_h\) with the same initial conditions as \(\gamma\). Moreover, assume the “no-stuck” condition: there exists a constant \(V > 0\) such that the time \(\gamma_h\) takes to traverse through a single mesh cell is bounded above by...
\[
\dot{\gamma}(t) - \dot{\gamma}_h(t) \leq C(g - g_h)_{W^{1,\infty}(\mathcal{F}_h)} + h^{-1} g - g_h_{L^\infty},
\]
\[
|\gamma(t) - \gamma_h(t)| \leq C(h g - g_h)_{W^{1,\infty}(\mathcal{F}_h)} + g - g_h_{L^\infty}.
\]

The “no-stuck” condition on the discrete metrics is quite intuitive. Basically, it excludes situations like the one depicted in Figure 3.10, where the geodesic is trapped in a single cell somehow. This is obviously necessary, because in this theorem, the only other assumption on the discrete metrics \(g_h\) is that \(g_h\) is close to \(g\) in \(L^\infty\)-norm with no control over the derivatives.

This following corollary shows the expected convergence rate in practice when the metric approximation is as good as the best approximation:

**Corollary 3.10.** Under the assumption of Theorem 3.9, suppose the shape constants of the meshes are bounded uniformly, \(g\) is known, and \(g_h\) are the Regge canonical interpolants. Then,
\[
|\dot{\gamma}(t) - \dot{\gamma}_h(t)| \leq C h^r, \quad |\gamma(t) - \gamma_h(t)| \leq C h^{r+1},
\]
where \(C\) depends on \(\|g\|_{W^{2,\infty}}, \|g^{-1}\|_{L^\infty}, V, T, |\dot{\gamma}(0)|, \) the degree \(r\), the dimension of the domain, and the shape constant bound of the meshes.

**Proof.** This follows from the previous theorem and the error estimates for the Regge canonical interpolant (cross-chapter reference).
The corollary below shows that the restarting strategy used in the robust algorithm in the previous section when the geodesics of the approximating generalized Regge metric goes near a face of low-dimension does not cause any problems:

**Corollary 3.11.** Under the assumption of Theorem 3.9, suppose $\gamma_h$ comes to a distance $\epsilon h$, $\epsilon < 1$, to a face of dimension $\leq (n-2)$ at time $t_\ast$. Restart the extension of $\gamma_h$ by keeping the Euclidean velocity vector $\dot{\gamma}_h(t_\ast)$ while moving its position to a point in another cell within the $\epsilon$-sphere. Still call this (discontinuous) curve $\gamma_h$ after $t_\ast$, and extend it as usual. Then, the error estimates still holds with an additional $\epsilon h$ error in both the position and velocity estimates.

**Proof.** It is clear that an extra error of $\epsilon h$ is incurred for the position at time $t = t_\ast$. For the velocity vector, the error is proportional to the difference between the values of $g_h$ at the two points. Using $g$, this difference is bounded by

$$2\|g - g_h\|_{L^\infty} + \epsilon h \|g\|_{W^{1,\infty}}.$$ 

After $t_\ast$, the original estimate applies to the restarted geodesic approximation problem to the smooth geodesic under $g$ with the same initial condition as $\gamma_h(t_\ast)$. The difference between this smooth geodesic and the original geodesic up to a fixed time $T$ can be bounded by $C\epsilon h$, using the standard ODE perturbation theorem (see Theorem 3.13 later). This proves the claim. 

In practice, $\epsilon h$ is close to machine precision. So this is negligible.
The proof of Theorem 3.9 is somewhat long. It is adapted from the standard technique for proving error estimates for ODE solvers. The main idea is captured in Figure 3.12.

![Figure 3.12: The black curve is the smooth geodesic. The blue curve is the geodesic on $g_h$.](fig:fan)

In each cell, consider auxiliary smooth geodesics under the smooth metric using the position and velocity of $\gamma_h$ when it enters that cell as the initial condition (the green curves in Figure 3.12). The final error is then bounded by the sum of the successive difference between all these green curves at the final time $T$. The difference between neighboring green curves comes from two sources. First in a cell, one curve is a geodesic under $g$ while the other is a geodesic under $g_h$ with the same initial data. When $g_h$ exists that cell, the velocity of $\gamma_h$ is further rotated. At this time, the difference between $\gamma_h$ and the auxiliary smooth geodesic is denoted by $e_i$ as in Figure 3.12. Then afterwards, the two green curves are both geodesics to $g$ but with difference $e_i$ in initial conditions. Note that in both cases, only geodesics to smooth metrics are considered and can be handled by standard theory. This is made more precise below. The proof uses several technical lemmas which are stated and proved immediately after this proof.

**Proof of Theorem 3.9.** Fix a particular $h$. Let $t_1, t_2, \ldots, t_n$ be the time $\gamma_h$ leaves the $n$-th cell it ever transverses such that at time $T$ it is still inside the $(n + 1)$-th cell. Set $t_0 = 0$ and $t_{n+1} = T$. Define a sequence of auxiliary curves $\lambda_k$ which morphs from $\gamma$ to $\gamma_h$ as depicted in Figure 3.12: for $k = 0, \ldots, (n + 1)$,

$$
\lambda_k(t) := \begin{cases} 
\gamma_h(t), & \text{for } t \in [0, t_k), \\
 f_k(t), & \text{for } t \in [t_k, T],
\end{cases}
$$

71
where \( f_k : [t_k, T] \rightarrow M \) is the geodesic under \( g \) with the initial condition

\[
f_k(t_k) = \gamma_h(t_k), \quad \dot{f}_k(t_k) = \dot{\gamma}_h(t_k+).
\]

For any \( t \in [0, T] \), let \( m = m(t) \) be the integer such that \( t_m \leq t \leq t_{m+1} \). It is clear from the definition that

\[
\lambda_0(t) = \gamma(t), \quad \lambda_{m+1}(t) = \gamma_h(t).
\]

Hence,

\[
|\gamma(t) - \gamma_h(t)| = |\lambda_0(t) - \lambda_{m+1}(t)| \leq \sum_{k=0}^{m} |\lambda_k(t) - \lambda_{k+1}(t)|.
\]

Each summand \( |\lambda_k(t) - \lambda_{k+1}(t)| \) goes through three phases. The first phase when \( t \in [0, t_k) \), it vanishes because \( \lambda_k(t) = \lambda_{k+1}(t) = \gamma_h(t) \). In the second phase when \( t \in [t_k, t_{k+1}) \), \( \lambda_k(t) = f_k(t) \) and \( \lambda_{k+1}(t) = \gamma_h(t) \) are two geodesics with the same initial data under the metric \( g \) and \( g_h \) respectively in the \((k+1)\)-th cell (\( g \) can go out, of course). At the end, define

\[
e_{k+1} := f_k(t_{k+1}) - g_h(t_{k+1}), \quad \dot{e}_{k+1} := \dot{f}_k(t_{k+1}) - \dot{g}_h(t_{k+1}).
\]

In the third phase, when \( t \in [t_{k+1}, T] \), \( \lambda_k(t) \) and \( \lambda_{k+1}(t) \) are geodesics of the same metric \( g \) with difference in initial data given by \( e_{k+1} \) and \( \dot{e}_{k+1} \) in position and velocity respectively. By standard ODE theory and Lemma 3.15, the difference at time \( t \) after \( t_{k+1} \) can be bounded: there exists a constant \( C_1 \) depending only on \( \|g\|_{W^{2,\infty}}, \|g^{-1}\|_{L^\infty}, |\dot{\gamma}(0)| \), and \( T \), such that

\[
|\lambda_k(t) - \lambda_{k+1}(t)| + |\dot{\lambda}_k(t) - \dot{\lambda}_{k+1}(t)| \leq C_1 (|e_{k+1}| + |\dot{e}_{k+1}|).
\]

Since the norms on \( g \) were taken over the maximum of the whole domain, globally,

\[
|\gamma(t) - \gamma_h(t)| + |\dot{\gamma}(t) - \dot{\gamma}_h(t+)| \leq C_1 \sum_{k=0}^{m} (|e_{k+1}| + |\dot{e}_{k+1}|)
\]

The right-hand side can be estimated by using Lemma 3.16 for the two geodesics with the same initial condition but different metric and then applying Lemma 3.17 for the rotation of the velocity at the interior facet. The result is:

\[
|\gamma(t) - \gamma_h(t)| + |\dot{\gamma}(t) - \dot{\gamma}_h(t+)| \leq C_2 \sum_{k=0}^{m} [e^{M(t_{k+1} - t_k)}h \|g - g_h\|_{W^{1,\infty}(c_{k+1})} + \|g - g_h\|_{L^\infty}],
\]

where \( c_{k+1} \) is the \((k+1)\)-th cell \( \gamma_h \) passes and \( C_2 \) and \( M \) has the same dependence as \( C_1 \). By the “no-stuck” assumption, \( t_{k+1} - t_k \leq h/V \). So the exponential term can be absorbed in a constant \( C_3 \) with the addition dependency on \( V \):

\[
|\gamma(t) - \gamma_h(t)| + |\dot{\gamma}(t) - \dot{\gamma}_h(t+)| \leq C_3 \sum_{k=0}^{m} [h \|g - g_h\|_{W^{1,\infty}(c_{k+1})} + \|g - g_h\|_{L^\infty}].
\]
On one hand, using the “no-stuck” assumption again, the number of summands is bounded by $TV/h$. Hence, there exists a constant $C$ depending on $\|g\|_{W^{2,\infty}}$, $\|g^{-1}\|_{L^\infty}$, $|\dot{\gamma}(0)|$, $V$, and $T$ such that

$$|\dot{\gamma}(t) - \dot{\gamma}(t^+)| \leq C(g - g_h \|g\|_{W^{2,\infty}} + h^{-1}g - g_h \|g\|_{L^\infty}).$$

On the other hand, integrate in time for each interval $[t_k, t_{k+1}]$,

$$|\gamma(t) - \gamma_h(t)| \leq C_3 \sum_{k=0}^{m} (t_{k+1} - t_k)(h\|g - g_h\|_{W^{1,\infty}(\mathcal{T})} + \|g - g_h\|_{L^\infty})$$

$$\leq C_3 T(h\|g - g_h\|_{W^{1,\infty}(\mathcal{T})} + \|g - g_h\|_{L^\infty}),$$

where in the last step the cell-wise norm is again bounded by the global maximum. This proves the theorem.

The rest of this section contains the proofs of all the lemmas used above. First, the following lemma bounds the Euclidean norm of the geodesics:

**Lemma 3.12.** Let $\mathcal{T}$ be a mesh in $\mathbb{R}^n$ and $g$ a piecewise smooth Riemannian metric with tangential-tangential continuity. Suppose $\gamma : [a, b] \to \mathcal{T}$ is a geodesic under $g$. Then

$$\|g^{-1}\|_{L^\infty}|\dot{\gamma}(0)| \leq |\gamma(t)| \leq \|g\|_{L^\infty}|\dot{\gamma}(0)|.$$ 

**Proof.** By Theorem (3.3), the speed of $\gamma$ measured in $g$ is constant along $\gamma$:

$$g_{ij}\dot{\gamma}^i(t)\dot{\gamma}^j(t) = g_{ij}\dot{\gamma}^i(0)\dot{\gamma}^j(0).$$

Then elementary linear algebra proves the claim.

A key result is the variation of constant theorem for ODEs which essentially is a stability estimate. This is known as the Alekseev-Gröbner Theorem [46, Corollary I.14.6]:

**Theorem 3.13.** Let $y(t, t_0, y_0)$ be the solution to

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0,$$

and $z(t)$ be the solution to a perturbed equation:

$$z'(t) = f(t, z(t)) + \delta(t, z(t)), \quad z(t_0) = z_0,$$

where $\partial_y f$ exists and is continuous. Then,

$$z(t) - y(t) = \int_0^t \frac{\partial y}{\partial y_0}(t, t_0, y_0 + s(z_0 - y_0))(z_0 - y_0)ds + \int_{t_0}^t \frac{\partial y}{\partial y_0}(t, s, z(s))\delta(s, z(s))ds.$$
In order to use this theorem, it is convenient to write the geodesic equation (3.4) in the following position-velocity form by defining \( q^i = \gamma^i \) and \( v^j = \dot{\gamma}^j \).

\[
\begin{align*}
q^i &= v^i, \\
\dot{v}^i &= -\Gamma^i_{kl}v^k v^l,
\end{align*}
\]

where the Christoffel symbol \( \Gamma^i_{kl} \) defined after equation (3.4) is a function of \( q^i \) and \( y := [q^i, v^j] \) is a curve in the tangent bundle.

This lemma bounds the error estimate in the Christoffel symbol:

**Lemma 3.14.** Let \( M \) be a smooth manifold. Suppose \( g, g_1, g_2 \) are three smooth Riemannian metrics on \( M \) and \( \Gamma, \Gamma_1, \Gamma_2 \) are their corresponding Christoffel symbols (with the indices suppressed). Then, for integer \( s \geq 0 \), there exists a constant \( C \) depending only on \( \|g^{-1}\|_{L^\infty} \) and \( \|g\|_{W^{s+1,\infty}} \) such that

\[
\|\Gamma\|_{W^{s,\infty}} \leq C.
\]

Suppose \( g_1 \) and \( g_2 \) sufficient close satisfying \( \|g_1 - g_2\|_{L^\infty} < \frac{1}{2}\|g_2^{-1}\|_{L^\infty}^{-1} \), then,

\[
\|\Gamma_1 - \Gamma_2\|_{L^\infty} \leq C\|g_1 - g_2\|_{W^{1,\infty}},
\]

where the constant \( C' \) depends only on \( \|g_2\|_{W^{1,\infty}} \) and \( \|g_2^{-1}\|_{L^\infty} \).

**Proof.** From the definition of the Christoffel symbol, the derivative of inverse metric formula (3.10), and chain rule, clearly,

\[
\begin{align*}
\|\Gamma\|_{L^\infty} &\leq \|g^{-1}\|_{L^\infty} \|g\|_{W^{1,\infty}}, \\
\|\Gamma\|_{W^{1,\infty}} &\leq \|g^{-1}\|_{L^\infty}^2 \|g\|_{W^{1,\infty}}^2 + \|g^{-1}\|_{L^\infty} \|g\|_{W^{2,\infty}}, \\
\end{align*}
\]

This proves the first claim. For the second one,

\[
\Gamma_1 - \Gamma_2 = g_1^{-1}(\partial g_1) - g_2^{-1}(\partial g_2) = (g_1^{-1} - g_2^{-1})(\partial g_2) + g_1^{-1}(\partial g_1 - \partial g_2),
\]

where \( (\partial g_i) \) is the lazy notation for the first-derivative terms in the definition of the Christoffel symbol. By assumption, \( \|g_2^{-1}\|_{L^\infty} \|g_1 - g_2\|_{L^\infty} < \frac{1}{2} \). Standard linear perturbation theorem [59, I.4.24] implies that

\[
\|g_1^{-1} - g_2^{-1}\|_{L^\infty} \leq \frac{\|g_1 - g_2\|_{L^\infty} \|g_2^{-1}\|_{L^\infty}^2}{1 - \|g_2^{-1}\|_{L^\infty} \|g_1 - g_2\|_{L^\infty}} \leq 2\|g_2^{-1}\|_{L^\infty} \|g_1 - g_2\|_{L^\infty} \leq \|g_2^{-1}\|_{L^\infty}.
\]

This also shows that \( \|g_1^{-1}\|_{L^\infty} \leq 2\|g_2^{-1}\|_{L^\infty} \). This proves the second estimate. \( \square \)
This lemma gives a crude stability bound for the smooth geodesic equation:

**Lemma 3.15.** Let \( \gamma(t,t_0,y_0) \) be the solution to equation (3.12) with initial data \( \gamma(t_0) = y_0 \). Then, there exists a constant \( C \) depending on \( \|g\|_{W^{2,\infty}}, \|g^{-1}\|_{L^\infty} \), and \( |y_0| \) such that

\[
\left\| \frac{\partial y}{\partial y_0}(t,t_0,y_0) \right\| \leq e^{C(t-t_0)}.
\]

**Proof.** Let \( \Phi(t) := \frac{\partial y}{\partial y_0}(t,t_0,y_0) \). By the standard ODE theory [46, Theorem I.14.3], \( \Phi \) solves the linear ODE:

\[
\Phi(t) = \frac{\partial F}{\partial y}(t,y(t,t_0,y_0))\Phi(t), \quad \Phi(t_0) = I,
\]

where \( I \) is the identity matrix of the correct size. Using the definition of \( F \),

\[
\frac{\partial F}{\partial y} = \begin{bmatrix}
0 & I \\
-\partial_i \Gamma_{kl} v^k v^l & -2\Gamma^i_{ii} v^l
\end{bmatrix}.
\]

Then Lemma 3.14 applies to the \( \Gamma \)-terms and Lemma 3.12 applies to the \( v \)-terms. Hence, there exists a constant \( C \) with the dependency as stated in the claim of this lemma such that:

\[
\left\| \frac{\partial F}{\partial y} \right\| \leq C,
\]

for all \( t \geq t_0 \). Then standard ODE comparison theorem proves the claim. \( \square \)

Given the previous lemmas, the differences between geodesics to different metrics with the same initial condition can be bounded. A form useful to the case here is stated below:

**Lemma 3.16.** Let \( c \) be an \( n \)-simplex in \( \mathbb{R}^n \) of Euclidean diameter \( h \). Suppose \( \tilde{g} \) is a Riemannian metric on \( c \) and \( \tilde{\gamma} \) a geodesic with initial condition \( \gamma(t_0) = q_0 \in c \) and \( \tilde{\gamma}(t_0) = \gamma_0 \). Suppose \( g \) is any Riemannian metric on \( c \) with \( \|g - \tilde{g}\|_{L^\infty} \leq \frac{1}{2} \|\tilde{g}^{-1}\|_{L^\infty} \). Let \( \gamma \) be the geodesic under \( g \) with the same initial data \( (q_0, \gamma_0) \). Set \( y := [q^i, v^j] \) for \( \gamma \) as before and define \( \tilde{y} \) similarly. Then, before \( \gamma \) exits \( c \), there exist constants \( C \) and \( M \) depending only on \( \|\tilde{g}\|_{W^{2,\infty}}, \|\tilde{g}^{-1}\|_{L^\infty} \), and \( |\gamma_0| \) such that

\[
|\gamma(t) - \tilde{\gamma}(t)| \leq C e^{M(t-t_0)} h \|g - \tilde{g}\|_{W^{1,\infty}}.
\]

**Proof.** By Theorem 3.13 and the definition of \( F(y) \) in equation (3.12),

\[
|\gamma(t) - \tilde{\gamma}(t)| \leq \int_{t_0}^t \frac{\partial \tilde{y}}{\partial y_0}(\Gamma^i_{jk} - \tilde{\Gamma}^i_{jk}) v^j v^k ds \leq (\sup_t |v(t)|_{L^\infty}) \left\| \frac{\partial \tilde{y}}{\partial y_0} \right\| \|\Gamma - \tilde{\Gamma}\|_{L^\infty} \int_{t_0}^t v ds.
\]

Because \( \gamma \) cannot exit \( c \),

\[
\int_{t_0}^t v ds = |q(t) - q(t_0)| \leq h.
\]

75
By Lemma 3.14, there is a constant $C_1$ depending only on $\|\bar{g}\|_{W^{1,\infty}}$ and $\|\bar{g}^{-1}\|_{L^{\infty}}$ such that

$$\|\Gamma - \bar{\Gamma}\|_{L^{\infty}} \leq C_1 \|g - \bar{g}\|_{W^{1,\infty}}.$$ 

By Lemma 3.15, there is a constant $C_2$ depending on $\|\bar{g}\|_{W^{2,\infty}}$, $\|\bar{g}^{-1}\|_{L^{\infty}}$, and $|v_0|$, such that

$$\left\| \frac{\partial \bar{y}}{\partial y} \right\| \leq e^{C_2(t-t_0)}.$$ 

Moreover, on finite dimensional spaces, the $| \cdot |$-norm controls the $l^{\infty}$-norm by a constant,

$$\sup |v(t)|_{l^{\infty}} \leq C_3 \sup |v(t)| \leq C_3 C_4 |v_0|,$$

where $C_3$ only depends on the dimension of $v$ and $C_4$ is from Lemma 3.12. Combining all these estimates, one obtains

$$|y(t) - \bar{y}(t)| \leq C_1 C_3 C_4 e^{C_2(t-t_0)} |v_0| h \|g - \bar{g}\|_{W^{1,\infty}}.$$ 

\[\square\]

The rotation of the velocity vector at the interior facets can be bounded by the jump in the unit normal vector across the facet. This jump is estimated by the following lemma:

**Lemma 3.17.** Fix an $(k-1)$-dimensional hyperplane $H$ in $\mathbb{R}^k$ and any basis $\{t_1, \ldots, t_{k-1}\}$ for vectors parallel to $H$. Let $\bar{g}$ be a $k$-by-$k$ symmetric positive definite matrix. Suppose $g$ is any $k$-by-$k$ symmetric positive definite matrix satisfying $|g - \bar{g}| \leq \frac{1}{2} |\bar{g}^{-1}|^{-1}$. Let $\bar{n}$ and $n$ be the outward (with respect to the origin) unit vectors normal to $H$ under $\bar{g}$ and $g$ respectively. Then, there exists a constant $C$ depending only on $H$ and $|\bar{g}^{-1}|$ such that

$$|n - \bar{n}| \leq C |g - \bar{g}|.$$

**Proof.** Let $u(s) = (1-s)\bar{g} + sg$ for $s \in [0,1]$. Because the space of positive definite matrices is convex, $u(s)$ is positive definite for all $s$. With a computation of the Neumann series similar to that at the end of the proof of Lemma 3.14, it can be shown that

$$|u^{-1}| \leq |\bar{g}^{-1}|,$$

uniformly in $s$. Let $T$ be the constant $n \times (n-1)$ matrix $[t_1, \ldots, t_{n-1}]$. Then $n(s)$ solves:

$$T^T un = 0, \quad n^T un = 1.$$ 

The Euclidean norm of $n$ is therefore bounded by a constant depending on $\{t_i\}$ and the norm of $u^{-1}$ and in turn $\bar{g}^{-1}$, uniformly in $s$. Differentiate the equations with respect to $t$,

$$T^T u' + T^T un' = 0, \quad 2n^T un' + n^T u'n = 0.$$ 

76
This is a linear system. Solve for $n'$,

$$n' = - T^{-T} u^{-1} \begin{bmatrix} T & T \end{bmatrix}^T u' n.$$

Because columns of $T$ and $n$ are $u$-orthogonal, the first term is bounded uniformly in $s$. The rest of the terms in the above other than $u'$ are bounded uniformly in $s$ as well. Then,

$$|n - \bar{n}| = |n(1) - n(0)| = \left| \int_0^1 n'(t) dt \right| \leq C \int_0^1 |u'| dt = C|g - \bar{g}|,$$

where $C$ depends only on $|\bar{g}^{-1}|$ and the tangent vectors.

3.9 Numerical examples: Kepler and Schwarzchild systems

3.9.1 Kepler system

Kepler system is the classical Newtonian description of planetary motion under the gravity of a central star. In natural units, the problem is, find $q : [0, T] \to \mathbb{R}^2$ such that

$$\ddot{q} = - \frac{q}{|q|^3}, \quad q(0) = q_0, \quad \dot{q}(0) = v_0.$$

This has a known exact solution, which is derived below. First it is easy to check that the energy $H$ and the angular momentum $L$ defined below are conserved quantities [47, Equation (2.5)]:

$$H := |\dot{q}|^2/2 - 1/|q|, \quad L := q \times \dot{q} = q_1 \dot{q}_2 - q_2 \dot{q}_1.$$

Switch to polar coordinates $q = (r \cos \theta, r \sin \theta)$. The above becomes:

$$H = (r^2 + r^2 \dot{\theta}^2)/2 - 1/r, \quad L = r^2 \dot{\theta}.$$  \hspace{1cm} (3.13)

After a tedious elementary computation, it can be shown that the trajectories are ellipses:

**Lemma 3.18** (Equation (2.10) of [47]). Let $e = \sqrt{1 + 2HL^2}$. Then, $r$ and $\theta$ satisfies:

$$r = \frac{L^2}{1 + e \cos(\theta - \theta_0)}.$$

That is, the trajectories are ellipses with eccentricity $e$.

**Proof.** This is a well-known result. A direct proof is outlined here. Take $r$ as a function of $\theta$. Then $\dot{r}(t) = r'(\theta) \dot{\theta}(t)$. Substituting the second part of equation (3.13) $\dot{\theta} = Lr^{-2}$ into the first equation for $H$, after some algebra, one gets

$$\frac{\sqrt{1 - e^2} dr}{e \sqrt{e^2 - (2Hr + 1)^2}} = d\theta,$$

77
The substitution \( u := (L^2/r - 1)/e \) leads to:

\[
- \frac{du}{\sqrt{1-u^2}} = d\theta \implies u = \cos(\theta - \theta_0),
\]

which proves the claim.

The time dependency still has to be solved. Use \( L = r^2 \dot{\theta} \) in equation (3.13) to eliminate \( r \):

\[
\frac{L^3}{(1 + e \cos(\theta - \theta_0))^2} d\theta = dt. \tag{3.14}
\]

A nontrivial change of coordinates has to be used to integrate this. Without loss of generality, set \( \theta_0 = 0 \). The function

\[
r = \frac{L^2}{1 + e \cos(\theta)} \tag{3.15}
\]

describes an ellipse with semi-major axis \( a := L^2/(1 - e^2) \) in polar coordinates where the origin is at the right focus. The new coordinate system in Figure 3.13 has the center of the ellipse as the origin. For a point \( P \) on the ellipse, let \( R \) be its projection down to the \( x \)-axis, and \( Q \) be the intersection of the ray \( RQ \) with the circle of radius \( a \) centered at the origin. The new angle variable \( E := \angle QOR \) is called the eccentric anomaly.

![Figure 3.13: Definition of the eccentric anomaly](fig:ecc_ano)

In Figure 3.13, the length of the segment \( PC \) is \( r \), \( \angle PCR = \theta \), and the length of \( OC \) is \( ea \) is the focal length. The fact that \( OR = OC + CR \) then implies that

\[
a \cos E = ea + r \cos \theta \implies \cos \theta = \frac{a}{r}(\cos E - e).
\]

78
By equation (3.15) and the definition \( a = L^2/(1 - e^2) \), the above becomes
\[
\cos \theta = \frac{\cos E - e}{1 - e \cos E}.
\]
Substituting this back into equation (3.14), one gets the Kepler's equation [37, Equation (4.59)]:
\[
E + e \sin E = \frac{(1 - e)^{3/2}}{L^3}(t - t_0).
\]
To evaluate the exact solution, for each \( t \), the above equation is solved using Newton's method to obtain \( E \), which can in turn be used to evaluate \( \cos \theta \) and \( r \) and then \( q \) in the original equation. This can be done to any precision for arbitrarily large \( t \) and will be used to evaluate the long-time properties of the geodesic solver.

### 3.9.2 Jacobi's formulation

The Kelper's system can be formulated as a geodesic problem using the Jacobi's formulation. Recall the following classical theorem [1, Theorem 3.7.7]:

**Theorem 3.19.** Let \((M, g)\) be a Riemannian manifold and \( V : M \to \mathbb{R} \). A stationary point \( \gamma : [a, b] \to M \) to the Lagrangian
\[
\int_a^b \frac{1}{2} g_{ij} \dot{\gamma}^i \dot{\gamma}^j - V \, dt,
\]
with total energy \( \mathcal{E} \) is a geodesic \( \gamma(s) \) of the Riemannian manifold \((M, \tilde{g})\) with the Jacobi metric \( \tilde{g} := 2(\mathcal{E} - V)g \) under the reparameterization
\[
s(\tau) = 2 \int_0^\tau \mathcal{E} - V(\gamma(t)) \, dt.
\]

The Kelper's system corresponds to a Lagrangian on the Euclidean space \((\mathbb{R}^2, \delta_{ij})\) with \( V(q) = -|q|^{-1} \). Its Jacobi metric is thus
\[
g_{ij} = 2(\mathcal{E} + |q|^{-1}) \delta_{ij}.
\]
Here the potential \( V \) is always negative and is normalized so that \( V \to 0 \) at infinity. Therefore, when \( \mathcal{E} \geq 0 \), the trajectories are unbounded. When \( \mathcal{E} < 0 \), the trajectories are trapped inside the region where \( \mathcal{E} - V \) remains positive. Within this region, the Jacobi metric \( g_{ij} \) is Riemannian. From the discussion in the previous subsection, the trajectories are in fact ellipses. The corresponding geodesic equation in the symplectic formulation is:
\[
\dot{q}^i = \frac{p^i}{2(\mathcal{E} + |q|^{-1})}, \quad \dot{p}_i = -\frac{|p|^2 q_i}{4(\mathcal{E} + |q|^{-1})^2 |q|^3}.
\]
The solution \( q(s) \) to this system is related to the exact solution \( q(t) \) before via the reparameterization:

\[
s(\tau) = \int_0^\tau \left( 2(\delta^2 + |q(t)|^{-1}) \right) dt.
\]

In the numerical experiments, the Jacobi metric (3.16) is used to find Kelperian orbits.

### 3.9.3 Numerical examples for Kelperian orbits

For all the numerical experiments, parameters \( H = -1.5 \) and \( L = 0.5 \) were chosen for the Kepler’s system. An elliptic annulus domain slightly bigger than the exact orbit is triangulated using the FEniCS package \texttt{mshr}. A visualization of the discrete Kepler metric is given in Figure 3.14.

![Figure 3.14: Plot of a discrete Kepler metric. The color indicates the pointwise Euclidean norm of the metric.](fig:kepler_metric)

Examples of plots of the numerical solution can be found in the introduction (see Figure 3.3).

First, the convergence rates for a fixed maximum time are tested. For this set of numerical experiments, the generalized geodesic equation is solved on a sequence of refiner and refiner meshes for 1.65 period with the canonical Regge interpolant of the Jacobi metric as the metric. For all the mesh sizes, the solver step size \( h_s \) is chosen to be \( 2 \times 10^{-5} \), which is smaller than the smallest mesh size \( h_m \approx 7 \times 10^{-5} \). This ensures the error convergence rate is due to the better approximation of the metric. After each computation, the \( L^\infty \)-errors in the position \( q \), the energy \( H \), and the momentum \( L \) are estimated from the maximum error of the computed solution at points uniformly sampled at a density of 200 points per period.
The results are summarized in Table 3.2. There the error rates without turning of $p$ at the cell boundaries are included as well, which corresponds to using the existing ODE geodesic solver directly on the Regge metric pretending it is continuous. Detailed plots of the errors are found in Figures 3.15, 3.16, and 3.17.

<table>
<thead>
<tr>
<th>Metric</th>
<th>mesh sizes</th>
<th>max error in position</th>
<th>max error in $H$</th>
<th>max error in $L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>REG$^0$</td>
<td>[64,128,256,512,1024]</td>
<td>$h_m^1$ (1)</td>
<td>$h_m^1$ (1)</td>
<td>$h_m^1$ (1)</td>
</tr>
<tr>
<td>REG$^1$</td>
<td>[64,128,256,512,1024]</td>
<td>$h_m^2$ ($h_m^2$)</td>
<td>$h_m^2$ ($h_m^2$)</td>
<td>$h_m^2$ ($h_m^2$)</td>
</tr>
<tr>
<td>REG$^2$</td>
<td>[32,64,128,256,512]</td>
<td>$h_m^3$ ($h_m^2$)</td>
<td>$h_m^3$ ($h_m^2$)</td>
<td>$h_m^3$ ($h_m^2$)</td>
</tr>
<tr>
<td>REG$^3$</td>
<td>[16,32,64,128,256]</td>
<td>$h_m^4$ ($h_m^{3.5}$)</td>
<td>$h_m^4$ ($h_m^{3.5}$)</td>
<td>$h_m^4$ ($h_m^{3.5}$)</td>
</tr>
</tbody>
</table>

Table 3.2: Convergence rate for a fixed maximum time. $h_m$ is the mesh size. The rates in the parenthesis are for the cases without turning $p$ at interior facets.

For the lowest degree, the turning $p$ step is obviously important as the derivative of the metric vanishes in the interior of all cells. From the above, this step is important even for higher degree Regge elements in order to get clean optimal convergence rates.

![Figure 3.15](image_url)

Figure 3.15: Blue: log-log plot of mesh size against position error for degree 0,1,2,3. Red: reference slope for convergence of order 1,2,3,4.
Figure 3.16: Blue: log-log plot of mesh size against error in the energy for degree 0, 1, 2, 3.
Red: reference slope for convergence of order 1, 2, 3, 4.

Figure 3.17: Blue: log-log plot of mesh size against momentum error for degree 0, 1, 2, 3.
Red: reference slope for convergence of order 1, 2, 3, 4.
In the second sets of numerical experiments, the long time behavior of the error is assessed. The Kepler Jacobi metric is interpolated into \( \text{REG}^r \) and the generalized geodesic equation is solved for 100 orbits for \( r = 0 \) and 300 orbits for \( r = 1, 2, 3 \). Then the computed solutions are sampled uniformly at a density of 200 points per period and compared with the exact solution. The growth of the error in the position, energy, and momentum are recorded. The results are summarized in Table 3.3. There the error growth rates without turning of \( p \) at interior facets are included as in the previous numerical experiment. Detailed plots of the errors are found in Figures 3.18, 3.19, and 3.20.

<table>
<thead>
<tr>
<th>Metric</th>
<th>mesh size</th>
<th>error in position</th>
<th>error in ( H )</th>
<th>error in ( L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{REG}^0 )</td>
<td>160</td>
<td>( t^1 ) ((t^2))</td>
<td>( 1 ) ((t^1))</td>
<td>( 1 ) ((t^1))</td>
</tr>
<tr>
<td>( \text{REG}^1 )</td>
<td>96</td>
<td>( t^1 ) ((t^2))</td>
<td>( 1 ) ((t^1))</td>
<td>( 1 ) ((t^1))</td>
</tr>
<tr>
<td>( \text{REG}^2 )</td>
<td>96</td>
<td>( t^1 ) ((t^2))</td>
<td>( 1 ) ((t^1))</td>
<td>( 1 ) ((t^1))</td>
</tr>
<tr>
<td>( \text{REG}^3 )</td>
<td>48</td>
<td>( t^1 ) ((t^2))</td>
<td>( 1 ) ((t^1))</td>
<td>( \epsilon t^1 ) ((t^1))</td>
</tr>
</tbody>
</table>

Table 3.3: The error growth rate in time \( t \). The rates in the parenthesis are for the cases without turning \( p \) at interior facets.

The observed rates agree with the expectation. The energy \( H \) is conserved for all time. There should in fact be a small constant times \( t^1 \) in the error in \( L \) for degree \( r \geq 1 \). This is due to the occasional variable step size. This becomes obvious only for \( r \geq 2 \). For physical problems, \( r \geq 2 \) would be rather rare for 3D problems due to memory constraints. So this should not be an issue for most applications. It is also interesting to note that without the turning \( p \) step, the error grows one order faster in \( t \). Thus even for medium length simulations, the turning \( p \) step is crucial. It should also be noted that the long time error behavior for \( \text{REG}^0 \) is somewhat sporadic.
Figure 3.18: Plot of time against the error in position for degree $0, 1, 2, 3$.

Figure 3.19: Plot of time against the relative error in energy for degree $0, 1, 2, 3$. 
Figure 3.20: Plot of time against the relative error in momentum for degree 0,1,2,3.

3.9.4 Schwarzschild system

The Schwarzschild metric is the most general static spherically symmetric solution to the Einstein field equation in general relativity [13]. It can be used as a model for the gravitational field around a star, to which the Newtonian mechanics used in the Kepler system is a classical approximation [109, Chapter 6]. In spherical coordinates, the metric for a star of mass $M$ in natural units has the form [109, Equation 6.1.43]:

$$ds^2 = -\left(1 - \frac{2M}{r}\right)dt^2 + \left(1 - \frac{2M}{r}\right)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2).$$

It can be shown that the Jacobi metric for a particle of mass $m$ and total energy $E$ in this system is given by [41, Equation 3.1]:

$$ds^2 = \left(E^2 - m^2 + \frac{2Mm^2}{r}\right)\left[\frac{dr^2}{(1 - \frac{2M}{r})^2} + \frac{r^2}{1 - \frac{2M}{r}}(d\theta^2 + \sin^2\theta d\phi^2)\right],$$

where $E \leq m$. In this numerical example, planar orbits are computed. By spherical symmetry, without loss of generality, set $\theta = 2\pi$. Then the Jacobi metric becomes:

$$ds^2 = \left(E^2 - m^2 + \frac{2Mm^2}{r}\right)\left[\frac{dr^2}{(1 - \frac{2M}{r})^2} + \frac{r^2d\phi^2}{1 - \frac{2M}{r}}\right].$$
It is known that the orbits are almost ellipses with precession (that is, the major axis of the ellipses rotates). A plot of the metric is shown in Figure 3.21. The mesh is obtained from \texttt{mshr}. The red part corresponds to the singularity of the metric at the star while the deep blue circle is where the Jacobi metric vanishes. Technically, the Jacobi metric is defined only inside this circle. The computed curves always stays inside the circle so this does not cause any problems.

![Figure 3.21: Plot of the discrete Schwarzschild metric. The color indicates the pointwise Euclidean norm of the metric.](fig:schwarzschild_metric)

Example orbit plots can be found in the introduction (see Figure 3.4).
Chapter 4

Rotated generalized Regge element and applications in solid mechanics

In 2D, it is clear that a finite element with tangential continuity, like the Nédéléc edge element of the first kind [83], can be transformed into a finite element with normal continuity, like the Raviart-Thomas element [94], via a simple rotation of 90° (see Figure 4.1). Using this rotation, for example, rotated Nédéléc edge elements can be used to discretize $H(\text{div})$ and rotated Raviart-Thomas elements can be used to discretize $H(\text{curl})$ in 2D. However, this, as will be shown in the first section, is an entirely 2D phenomenon.

![Figure 4.1: Rotation of 2D Nédéléc edge elements to Raviart-Thomas elements](fig:2d_rotate)

Generalized Regge elements, having tangential-tangential continuity, can also be “rotated” to form finite elements with normal-normal continuity. In 2D, this transforms the generalized Regge elements to the well-known Hellan-Herrmann-Johnson (HHJ) elements [11, 18], where normal-normal continuous finite elements were used for the bending moment tensor in a mixed formulation for the biharmonic equation. To the best of the author’s knowledge,
generalizations of the HHJ elements to higher dimensions are not known in the literature. Remarkably, this transformation works in $n$-dimension for all $n \geq 2$. In particular, the rotated generalized Regge elements can be used to directly in the HHJ mixed formulation of the biharmonic equation in all dimensions. In another direction, Pechstein-Schöberl [87–89] proposed a mixed formulation of the elasticity equation in 2D and 3D using normal-normal continuous finite elements. In 2D, their stress elements were equivalent to the HHJ elements and thus equivalent to rotated Regge as well. In 3D, the rotated Regge elements are different from and form a strict subspaces of the Pechstein-Schöberl elements. The rotated Regge elements can also be used to discretize Pechstein-Schöberl mixed formulation of the elasticity equation. In this chapter, we explore these two applications.

The rest of the chapter is organized as follows. First we introduce the rotated generalized Regge elements and prove their key properties: in 2D, they form the space of normal-normal continuous finite elements while in $n$D, $n \geq 3$, they form a strict subspace of normal-normal continuous finite elements. Second, we establish the connection between rotated Regge element and HHJ in 2D and use it to discretize the biharmonic equation in $n$D. Third, we review the Pechstein-Schöberl mixed formulation of linear elasticity and propose a method using generalized Regge element for the stress in $n$D. In both cases, we will state the formulation and study the stability and convergence properties numerically. Further analysis of these methods will be explored in future works. Finally we briefly explain how the discretizations of the biharmonic equation and elasticity equation are related to numerical relativity.

### 4.1 Rotated generalized Regge element

Let $R$ be the clockwise $90^\circ$-rotation matrix in 2D:

$$ R := \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. $$

It is clear that this matrix rotates tangential vectors to every 1-plane in $\mathbb{R}^2$ to normal vectors. This, however, does not work in higher dimensions.

**Theorem 4.1.** For $m \geq 3$, there does not exist a nonzero linear map which maps tangential vectors to every $(m-1)$-plane in $\mathbb{R}^m$ to normal vectors. In particular, there is no nonzero linear map mapping piecewise smooth tangential continuous vector fields on a mesh to piecewise smooth normal continuous vector fields.

**Proof.** Suppose $A$ is such a map. Let $\{e_1, \ldots, e_m\}$ be the Euclidean basis for $\mathbb{R}^m$. First, it is clear that $e_i^TAe_i = 0$ for all $i$ so $A$ has zeros on the diagonals. Take any $1 \leq i < j \leq m$. For
\( m \geq 3 \), we can find an \((m - 1)\)-plane containing both \( e_i \) and \( e_j \). Thus \( e_i^T A e_j = e_j^T A e_i = 0 \) by assumption. This implies that all off diagonal entries of \( A \) are zero too. Hence \( A = 0 \).

The case for symmetric matrix fields is quite different. Let \( P \) be an \((n - 1)\)-plane in \( \mathbb{R}^n \) and \( \{t_1, \ldots, t_{n-1}, n\} \) be an orthonormal basis adapted to \( P \) such that \( t_i \) are tangent to \( P \) while \( n \) is normal to \( P \). For a symmetric matrix field \( u \) in \( \mathbb{R}^m \), its tangent-tangent part is the \((m - 1) \times (m - 1)\) symmetric matrix field

\[
    u_P := [t_1 \cdots t_{m-1}]^T u [t_1 \cdots t_{m-1}],
\]

and its normal-normal part is the scalar field \( n^T u n \) on \( P \). Similar to the vector case, with respect to a mesh, a piecewise smooth symmetric matrix field is tangent-tangent continuous or normal-normal continuous if the tangent-tangent parts or normal-normal parts are single-valued at all interior facets.

**Theorem 4.2.** Suppose \( m \geq 2 \). Let \( S \) be the linear map from symmetric matrix fields to symmetric matrix fields in \( \mathbb{R}^m \) given by:

\[
    Su := u - I n^T u.
\]

The map \( S \) maps tangent-tangent continuous symmetric matrix fields to normal-normal continuous symmetric matrix fields.

**Proof.** Let \( f \) be any interior facet of the mesh and \( \{t_1, \ldots, t_{n-1}, n\} \) an orthonormal basis for \( \mathbb{R}^n \) adapted to \( f \). By definition,

\[
    n^T (Su)n = n^T u n - n^T u n.
\]

The trace can be computed via:

\[
    n^T u n + \sum_{i=1}^{m-1} t_i^T u t_i.
\]

Hence,

\[
    n^T (Su)n = - \sum_{i=1}^{m-1} t_i^T u t_i = -tr u_f,
\]

is just the trace of the tangent-tangent part of \( u \) to \( f \), which is continuous across interior facets by assumption.

The definition of \( S \) is most intuitive in 2D. Suppose \( x, y \) are normal vectors. Then \( Rx \) and \( Ry \) are tangential vectors. Hence if \( u \) is tangent-tangent continuous, \( R^T u R \) must be normal-normal continuous. A direct computation shows that

\[
    R^T u R = -(Su)^T. \tag{4.1}
\]

89
Let $\mathcal{T}$ be a mesh in $\mathbb{R}^m$. We use $\text{REG}^r(\mathcal{T})$ to denote the space of generalized Regge elements of degree $r$ on $\mathcal{T}$. Let $\text{NN}^r(\mathcal{T})$ be the space of normal-normal continuous piecewise polynomial symmetric matrix fields of degree $\leq r$ on $\mathcal{T}$.

**Theorem 4.3.** In dimension $m$, $m \geq 2$,

$$S(\text{REG}^r(\mathcal{T})) \subset \text{NN}^r(\mathcal{T}).$$

In particular, the two spaces are equal in 2D. For $m \geq 3$, $S(\text{REG}^r(\mathcal{T}))$ is a strict subspace.

**Proof.** By Theorem 4.2, the first claim holds. Due to equation (4.1), in 2D, $S$ maps tangential-tangential continuous symmetric matrix fields to normal-normal continuous symmetric matrix fields and vice versa. Hence the two spaces must be equal. For $m \geq 1$, it is clear that $S$ is invertible and

$$S^{-1}u = u - \frac{1}{m-1}\text{tr}u.$$

Suppose $v \in \text{NN}^r(\mathcal{T})$ and $m \geq 3$. $v$ satisfies only 1 condition at every interior facet. However, in order to be tangential-tangential continuous, a symmetric matrix field needs to satisfy $(n-1)$ conditions at every interior facet. Because of this, $S^{-1}v$ is not in $\text{REG}^r(\mathcal{T})$ in general. Hence $S(\text{REG}^r(\mathcal{T}))$ is only a strict subspace. $\square$

More properties of $\text{NN}^r(\mathcal{T})$ will be studied at the end of this chapter.

### 4.2 Solving the biharmonic equation via the Hellan-Herrmann-Johnson mixed formulation

Let $\Omega$ be a bounded Lipschitz domain in $\mathbb{R}^n$. The biharmonic problem is, given $f : \Omega \to \mathbb{R}$, find $u : \Omega \to \mathbb{R}$ such that

$$\Delta \Delta u = f, \quad \text{in } \Omega, \quad \text{eq:biharmonic}$$

$$u = \partial_n u = 0, \quad \text{on } \partial \Omega.$$

This is a classic problem with many applications including 2D Kirchhoff-Love plate models [106], potential formulations in 3D elasticity [97], and stationary Cahn-Hilliard phase separation models [112].

The continuous theory for the 2D biharmonic equation is well-established. In particular, it can be shown that given $f \in H^{-1}$ there is a unique solution $u \in H^2$. An exposition of the existence and regularity theory on Lipschitz domains can be found in [44, Chapter 7]. Its finite element discretization is also very mature. A survey can be found in [29, Chapter 6]. The theory for the 3D case, however, is less developed.
In this section, we first review the Hellan-Herrmann-Johnson (HHJ) mixed discretization of equation (4.2). We then show that \( S(\text{REG}_r) \) is equivalent to the HHJ element for the symmetric matrix field variable. After that we show how to use \( S(\text{REG}_r) \) to solve the biharmonic equation in nD and study the convergence properties numerically in 3D [5,11,18].

4.2.1 Hellan-Herrmann-Johnson continuous mixed formulation

First, the biharmonic problem is put into a Hilbert space context via a mixed formulation. Let \( \mathbb{S}^n \) be the space of symmetric \( n \times n \) matrices and

\[
H(\text{div div}) := \{ u \in L^2 \otimes \mathbb{S}^n \mid \text{div div} u \in H^{-1} \},
\]

where the first div is taken by rows and the second div is the usual vector divergence. This space is given the graph norm.

The mixed formulation of equation (4.2) is, given \( f \in H^{-1} \), find \((\sigma,u) \in H(\text{div div}) \times \dot{H}^1\) such that

\[
(\sigma, \tau) - \langle u, \text{div div} \tau \rangle = 0, \quad \forall \tau \in H(\text{div div}),
\]

\[
\langle \text{div div} \sigma, v \rangle = \langle f, v \rangle, \quad \forall v \in \dot{H}^1,
\]

where \( \langle \cdot, \cdot \rangle \) is the duality pairing between \( H^{-1} \) and \( \dot{H}^1 \).

The following two theorems are well-known in the literature [66] (see also [11,18]). Although they were proven only in 2D, the same proofs work in any dimension. The proofs are reproduced here for the convenience of the reader.

First, the mixed system (4.3) itself is well-posed:

**Theorem 4.4** (Theorem 2.2 of [66]). Given \( f \in H^{-1} \), there exists a unique pair \((\sigma,u)\) solving system (4.3). Further there exists a constant \( C > 0 \) depending only on \( \Omega \) such that

\[
\|\sigma\|_{H(\text{div div})} + \|u\|_{\dot{H}^1} \leq C\|f\|_{H^{-1}}.
\]

**Proof.** This follows from Brezzi’s theorem [17]. It is clear that \((\sigma,\tau)\) is coercive over the kernel of \( \text{div div} \). It remains to show the inf-sup condition for the bilinear form \( \langle v, \text{div div} \tau \rangle \). Note

\[
\text{div div} I v = \text{div} \nabla v = \Delta v.
\]

For any \( v \in \dot{H}^1 \), let \( \tau = -I v \in H^1 \otimes \mathbb{S}^n \subset H(\text{div div}) \). Then,

\[
\langle v, \text{div div} \tau \rangle = \langle v, -\Delta v \rangle = (\nabla v, \nabla v) \geq c\|v\|_{\dot{H}^1}^2,
\]

where \( c \) depends on the Poincaré constant for \( \Omega \) and

\[
\|\tau\|_{H(\text{div div})}^2 = \|\tau\|_{L^2}^2 + \|\text{div div} \tau\|_{H^{-1}}^2 = \|I v\|_{L^2}^2 + \|\Delta v\|_{H^{-1}}^2 \leq (n^2 + 1)\|v\|_{H^1}^2,
\]

91
where \( n \) is the dimension of \( \Omega \). Thus the inf-sup constant is bounded below by a constant depending only on the domain.

Second, the mixed system (4.3) can be used to solve the biharmonic equation (4.2).

**Theorem 4.5** (Corollary 2.3 of [66]). Given \( f \in H^{-1} \), suppose \((\sigma, u)\) is a solution to system (4.3), then \( \sigma = \nabla \nabla u \) and \( u \in \dot{H}^2 \) solves the biharmonic equation (4.2) as a distribution.

**Proof.** The biharmonic equation (4.2) has a unique solution, say \( w \in \dot{H}^2 \) with \( \Delta \Delta w = f \). It is clear that \( \nabla \nabla w \in H(\text{divdiv}) \). Once we show that \((\sigma, u) := (\nabla \nabla w, w)\) solves system (4.3), the theorem is then proved by the well-posedness of system (4.3). First, for test functions \( y \),

\[
\langle y, \text{divdiv} \tau \rangle = \langle \nabla \nabla y, \tau \rangle.
\]

Since the set of test functions is dense in \( \dot{H}^2 \), the same holds for \( w \). Hence,

\[
\langle w, \text{divdiv} \tau \rangle = \langle \nabla \nabla w, \tau \rangle = (\sigma, \tau),
\]

which shows that the first equation of system (4.3) holds. Similarly, by definition, \( \text{divdiv} \sigma = \text{divdiv} \nabla \nabla w = \Delta \Delta w = f \) as a distribution. Since the set of test functions is also dense in \( \dot{H}^1 \) the second equation of system (4.3) also holds.

### 4.2.2 Hellan-Herrmann-Johnson discretization

Let \( \Omega \) be a Lipschitz polyhedral domain in \( \mathbb{R}^n \) as before and \( \mathcal{T}_h \) a triangulation of \( \Omega \) with mesh size \( h \). Set

\[
V := \{ \sigma \in L^2 \otimes \mathbb{S}^n \text{ is piecewise } H^1 \text{ with normal-normal continuity} \},
\]

\[
W := \{ u \in \dot{H}^1 \text{ is piecewise } H^2 \}.
\]

We also need some additional convenient notations. For \( \sigma \in V \), on a facet with unit normal \( n \), define

\[
\sigma_{nn} := n^T \sigma n, \quad \sigma_{nt} := \sigma n - n \sigma_{nn}.
\]

Similarly, for \( u \in W \), let

\[
\partial_n u := n \cdot \nabla u, \quad \partial_t u := \nabla u - n \partial_n u.
\]

We give \( V \) and \( W \) the following mesh dependent norms:

\[
\| \sigma \|_V^2 := \sum_c \| \sigma \|_{L^2(\mathcal{c})}^2 + h \| \sigma \|_{L^2(\partial \mathcal{c})}^2,
\]

\[
\| u \|_W^2 := \sum_c \| u \|_{H^2(\mathcal{c})}^2 + h^{-1} \| \partial_n u \|_{L^2(\partial \mathcal{c})}^2.
\]
where both sum over all the cells \( c \) in the mesh \( T_h \). Using the same notation, we define a mesh-dependent \( \text{div}\div h \) by:

\[
\langle \text{div}\div h \sigma, v \rangle := \sum_c \left( \int_c \sigma : \nabla \nabla v - \int_{\partial c} \sigma_{nn} \partial_n v \right) = \sum_c \int_c \sigma : \nabla \nabla v - \sum_f \int_f \sigma_{nn} [\partial_n v],
\]

where \( \sum_f \) means sum over all facets \( f \) of \( T_h \) and the \textit{jump} \( [\partial_n v]_f \) is defined as the difference of \( \partial_n v \) on both sides of \( f \) if \( f \) is an interior facet and just \( \partial_n v \) if \( f \) is a boundary facet. Clearly from the definition, there exists a constant independent of \( h \) such that

\[
|\langle \text{div}\div h \sigma, v \rangle| \leq C \| \sigma \|_V \| u \|_W.
\]

The HHJ discretization chooses the following discrete subspaces of \( V \) and \( W \):

\[
V_h := \text{NN}^r(T_h), \quad W_h := \text{CG}^{r+1}(T_h) \cap \dot{H}^1.
\]

The \( \text{NN}^r \) finite element space in 2D is referred to as the Hellan-Herrmann-Johnson element in this thesis. Let \( T \) be a triangle. Then \( \text{NN}^r(T) \) is defined by the shape functions

\[
\mathcal{P}^r(T) \otimes S^2
\]

and the degrees of freedom

\[
\sigma \rightarrow \int_e (n^T \sigma n) q, \quad \forall q \in \mathcal{P}_r(e) \text{ and all edges } e \text{ of } T,
\]

\[
\sigma \rightarrow \int_T \sigma : \tau \quad \forall \tau \in \mathcal{P}_{r-1}(T) \otimes S^2,
\]

where \( n \) is the outward unit normal vector to \( T \).

Let \( R \) be the \( 90^\circ \)-rotation matrix defined before and \( t := Rn \) the unit tangent vector to the edges of \( T \). The generalized Regge element \( \text{REG}^r(T) \) in 2D is given by the same shape functions

\[
\mathcal{P}^r(T) \otimes S^2
\]

but tangential degrees of freedom

\[
\sigma \rightarrow \int_e (t^T \sigma t) q, \quad \forall q \in \mathcal{P}_r(e) \text{ and all edges } e \text{ of } T,
\]

\[
\sigma \rightarrow \int_T \sigma : \tau \quad \forall \tau \in \mathcal{P}_{r-1}(T) \otimes S^2.
\]

Note that in 2D, we have

\[
S^{-1} = S.
\]

Because \( S \) maps \( \mathcal{P}^r(T) \times S^2 \) bijectively into itself and

\[
\int_e [n^T (S \sigma) n] q = \int_e (n^T R^T \sigma R n) q = \int_e (t^T \sigma t) q,
\]

93
we conclude that $S : \text{REG}'(T) \rightarrow \text{NN}'(T)$ is an isomorphism of finite elements. Since we proved in early chapters of this thesis that $\text{REG}'$ is unisolvent, $\text{NN}'(T)$ defined here is unisolvent too. All the other properties carry over as well.

Given the discrete spaces, the discrete mixed problem is thus: given $f \in H^{-1}$, find $(\sigma, u) \in V_h \times W_h$ satisfying:

$$
\begin{align*}
(\sigma, \tau) - \langle u, \text{div} \text{div}_h \tau \rangle &= 0, \quad \forall \tau \in V_h, \quad \text{eq: discrete biharmonic} \\
\langle \text{div} \text{div}_h \sigma, v \rangle &= \langle f, v \rangle, \quad \forall v \in W_h,
\end{align*}
$$

First, we have consistency.

**Theorem 4.6.** Suppose $u \in H^3 \cap H^2$ solves the biharmonic equation (4.2). Let $\sigma := \nabla \nabla u$. Then $(\sigma, u)$ satisfies the discrete system (4.5).

**Proof.** First, $u \in W$ so $\langle u, \text{div} \text{div}_h \tau \rangle$ makes sense. Because $u \in H^2$, $[\partial_n u] = 0$ at all facets of the mesh. Hence, the first equation of (4.2) reads:

$$
\sum_c \left( \int_c \sigma : \tau - \int_c \tau \cdot \nabla \nabla u \right) = 0, \quad \forall \tau \in V_h.
$$

This certainly holds because $\sigma = \nabla \nabla u$ by definition. Second, $u \in H^3$ implies that $\sigma \in H^1 \otimes \mathbb{S}^n$. Hence $\sigma \in V$ and $\langle \text{div} \text{div}_h \sigma, v \rangle$ still makes sense. Then, for an interior facet $f$, $[\partial_n \sigma] = 0$ at $f$ because $\sigma$ is continuous across facets. On the other hand, for boundary facets $f$, $\partial_n v = 0$ because $\tau \in H^1$. Hence,

$$
\sum_c \int_{\partial c} \sigma_{nt} \partial_t v = \sum_f \int_f [\sigma_{nt}] \partial_t v = 0.
$$

Thus, by the identity $\sigma_n \cdot \nabla v = \sigma_{nn} \partial_n v + \sigma_{nt} \partial_t v$ and integration by parts:

$$
\langle \text{div} \text{div}_h \sigma, v \rangle = \sum_c \left( \int_c \sigma \cdot \nabla v - \int_{\partial c} \sigma_{nn} \partial_n v \right) = \sum_c \left( \int_c \sigma \cdot \nabla v - \int_{\partial c} \sigma_n \cdot \nabla v \right) = \sum_c \int_c -\text{div} \sigma \cdot \nabla v.
$$

Sum over the cells and integrate by parts again:

$$
\langle \text{div} \text{div}_h \sigma, v \rangle = \int_\Omega -\text{div} \sigma \cdot \nabla v = \int_\Omega v \text{div} \sigma - \int_\Omega n v \cdot \text{div} \sigma = \int_\Omega v \text{div} \sigma,
$$

where the last equality follows from the fact that $v = 0$ on the boundary. By definition, $\text{div} \sigma = \Delta \Delta u = f$. So the second equation of (4.5) is also satisfied. This proves the claim. 

In [11], the following stability and convergence theorem was proved:

**Theorem 4.7.** Suppose the domain is a convex polygon. Let $u \in H^3$ be a solution to the biharmonic equation (4.2) and $\sigma = \nabla \nabla u$. The discrete system (4.5) has a unique solution $(\sigma_h, u_h) \in \text{NN}^r \times \text{CG}^{r+1} \cap H^1$. This pair satisfies:

$$
\|\sigma - \sigma_h\|_{L^2} + \|u - u_h\|_{H^1} \leq C h \|u\|_{H^3}.
$$

94
Moreover, if \( u \) is smooth, then
\[
\|\sigma - \sigma_h\|_{L^2} \leq Ch^{r+1}\|u\|_{H^{r+3}},
\]
and for \( r = 0 \),
\[
\|u - u_h\|_{H^1} \leq Ch\|u\|_{H^3}, \quad \|u - u_h\|_{L^2} \leq Ch^2\|u\|_{H^4},
\]
while for \( r \geq 1 \),
\[
\|u - u_h\|_{H^1} \leq Ch^{r+1}\|u\|_{H^{r+2}}, \quad \|u - u_h\|_{L^2} \leq Ch^{r+2}\|u\|_{H^{r+3}}.
\]

### 4.2.3 Discretization of biharmonic equation in higher dimensions using rotated Regge elements

In dimension \( n, n \geq 3 \), the form of the continuous biharmonic equation (4.2), the continuous mixed formulation (4.3), and the mesh-dependent \( \text{div div} \) (4.4) remain the same as those in 2D. We noted that \( S(\text{REG}'r) \), which is defined in dimension \( n \) for all \( n \geq 2 \), is a discrete subspace of the infinite-dimensional mesh-dependent space \( V \). This opens up the possibility of using the pair
\[
V_h = S(\text{REG}'r), \quad W_h = \text{CG}^{r+1} \cap \hat{H}^1,
\]
eq:sreg_biharmonic
in higher dimensions for the discretization (4.5) to solve the biharmonic equation. In this subsection, we first validate that in 2D \( S(\text{REG}'r) \) can be used to solve the biharmonic equation in place of \( \text{HHJ}'r \) in practical implementations. Then we study the convergence of the discrete space choice (4.6) in 3D for solving the 3D biharmonic equation.

The finite element choice (4.6) can be implemented practically by a simple substitution in (4.5): given \( f \), find \( (\mu, u) \in \text{REG}'r \times W_h \) satisfying:
\[
(S\mu, S\rho) - \langle u, \text{div div}_h S\rho \rangle = 0, \quad \forall \rho \in \text{REG}'r,
\]
\[
\langle \text{div div}_h S\mu, v \rangle = \langle f, v \rangle, \quad \forall v \in W_h.
\]
Direct computation shows that:
\[
(S\mu, S\rho) = (\mu, \rho) + (n - 2)(\text{tr} \mu, \text{tr} \rho),
\]
is still coercive. The operator \( \text{div div} S \) also arises in numerical relativity. This connection will be explained in the later part of this chapter.

First for the 2D test, the author implemented \( \text{HHJ}'r+1 \) as part of this thesis in FEniCS.
To make this numerical test more realistic and interesting, the biharmonic equation:
\[
\Delta\Delta u = f,
\]
95
is solved on the non-convex cracked domain formed by deleting the triangle determined by 
\{(2,0.8),(2,0),(2.5,0)\} from the rectangle \([0,3] \times [0,2]\). The mesh is shown in Figure 4.2.

![Figure 4.2: Domain and mesh of the comparison test](fig:mesh_comparison)

The boundary conditions are as follows: \( u \) is clamped \( u = 0 \) and \( \partial_n u = 0 \) at all of the 
boundary except at the right edge where it is simply supported \( u = 0 \) and \( \nabla \nabla u = 0 \). The load 
is given by

\[
f(x,y) = \begin{cases} 
1, & \text{if } (x - 1.5)^2 + (y - 1)^2 < 0.2, \\
0, & \text{otherwise}.
\end{cases}
\]

The script `rotated_regge/demo_biharmonic_2d.py` in the companion repository of this thesis implemented the HHJ mixed formulation using quadratic \( S(\text{REG}^2) \) to solve this problem. A plot of the solution is given in Figure 4.3.
Then in script `rotated_regge/sreg_vs_hhj_2d.py`, the same problem is solved with HHJ$^2$ and the discrete displacement variable $u_h$ computed using $S(\text{REG}^2)$ and HHJ$^2$ are compared. The difference in $L^2$-norm is $1.080041764330992 \times 10^{-13}$, which shows that there is practically no difference.

Finally we test the convergence rates of the HHJ mixed formulation with $S(\text{REG}^r)$ numerically. This is implemented by the script `rotated_regge/biharmonic_conv.py` in the companion repository. The 2D case is done first to verify empirically the optimal convergence rates stated in the previous subsection. For this, the biharmonic equation is solved on the unit square with the following sinusoidal exact solution

$$u = \sin^2(\pi x) \sin^2(\pi y) \in C^\infty \cap H^2.$$

A sequence of unstructured meshes are generated using the FEniCS package `mshr`, which in turn internally uses CGAL [104] to generate the mesh. `mshr` takes a parameter “mesh size” which scales inversely with the diameter of the mesh, that is, doubling the mesh size is very close to half the diameter of the mesh. An example of the output of `mshr` for the unit square with mesh size 20 is shown in Figure 4.4.

Figure 4.3: 2D biharmonic equation demo

![Figure 4.3: 2D biharmonic equation demo](fig:demo_2d_biharmonic)
Table 4.1, Table 4.2, and Table 4.3 show the convergence test results for 2D with $r = 0, 1, 2$, where $\| \cdot \|$ means the $L^2$-norm. It is clear that the optimal convergence rates for both the $\sigma$ and $u$ are observed.

<table>
<thead>
<tr>
<th>Mesh size</th>
<th>$|u - u_h|$</th>
<th>Rate</th>
<th>$|\nabla (u - u_h)|$</th>
<th>Rate</th>
<th>$|\sigma - \sigma_h|$</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1.996271e-02</td>
<td>4.736657e-01</td>
<td>4.670851e+00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>5.287603e-03</td>
<td>1.97</td>
<td>2.171961e-01</td>
<td>1.15</td>
<td>2.370708e+00</td>
<td>1.00</td>
</tr>
<tr>
<td>32</td>
<td>1.291838e-03</td>
<td>1.98</td>
<td>1.074595e-01</td>
<td>0.99</td>
<td>1.210166e+00</td>
<td>0.94</td>
</tr>
<tr>
<td>64</td>
<td>3.269980e-04</td>
<td>1.97</td>
<td>5.399691e-02</td>
<td>0.99</td>
<td>6.086878e-01</td>
<td>0.99</td>
</tr>
<tr>
<td>128</td>
<td>8.137206e-05</td>
<td>1.99</td>
<td>2.694561e-02</td>
<td>1.00</td>
<td>3.037696e-01</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 4.1: 2D biharmonic degree 0
We then carry out the similar study for the convergence rates of the 3D biharmonic equation. The biharmonic equation is solved on the unit cube with the following sinusoidal exact solution

\[ u = \sin^2(\pi x)\sin^2(\pi y)\sin^2(\pi z) \in C^\infty \cap \tilde{H}^2. \]

A sequence of randomly perturbed meshes are generated in the following way. Given a mesh size, \( m \), we first create a uniform triangulation with \( m \) nodes per edge. Then, we perturbed the position of each internal mesh vertex by a 3D gaussian with zero mean and 10% of the diameter of the uniform mesh as standard deviation. An example of the perturbed mesh is given in Figure 4.5.
Table 4.4, Table 4.5, and Table 4.6 shows the convergence test results for 3D with $r = 0, 1, 2$. Due to the scale of the 3D problems and the memory limitation of the LU solver, especially for higher degrees, only relatively small meshes were tested. For $r = 0$, it seems that the method leads to a convergent approximation. The convergence rate, however, might be sublinear. For $r \geq 1$, it seems that the discrete solution converges to the true solution but the rates are suboptimal. It seems $\|u - u_h\| \sim h^r$ is one order suboptimal, $\|\nabla(u - u_h)\| \sim h^r$ is optimal, $\|\sigma - \sigma_h\| \sim h^{1+r/2}$ is also suboptimal.

<table>
<thead>
<tr>
<th>Mesh size</th>
<th>$|u - u_h|$</th>
<th>Rate</th>
<th>$|\nabla(u - u_h)|$</th>
<th>Rate</th>
<th>$|\sigma - \sigma_h|$</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6.563758e-02</td>
<td></td>
<td>9.228645e-01</td>
<td></td>
<td>8.053509e+00</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3.423797e-02</td>
<td>0.98</td>
<td>5.265519e-01</td>
<td>0.84</td>
<td>5.906622e+00</td>
<td>0.47</td>
</tr>
<tr>
<td>12</td>
<td>2.830453e-02</td>
<td>0.49</td>
<td>3.745720e-01</td>
<td>0.88</td>
<td>5.054238e+00</td>
<td>0.40</td>
</tr>
<tr>
<td>16</td>
<td>2.703180e-02</td>
<td>0.16</td>
<td>3.054689e-01</td>
<td>0.69</td>
<td>4.628145e+00</td>
<td>0.30</td>
</tr>
<tr>
<td>20</td>
<td>2.784529e-02</td>
<td>-0.14</td>
<td>2.733648e-01</td>
<td>0.51</td>
<td>4.491225e+00</td>
<td>0.14</td>
</tr>
<tr>
<td>24</td>
<td>2.640807e-02</td>
<td>0.45</td>
<td>2.401061e-01</td>
<td>1.11</td>
<td>4.290418e+00</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Table 4.4: 3D biharmonic degree 0 announcetafter
4.3 Solving the elasticity equation via the Pechstein-Schöberl mixed formulation

The linear elasticity equation is: on a bounded Lipschitz domain $\Omega$ in $\mathbb{R}^n$, given a vector field $f$, the body force on $\Omega$, find another vector field $u$, the displacement such that

$$\begin{align*}
\text{div} C \epsilon u &= -f, & \text{in } \Omega, \\
u &= 0, & \text{on } \partial \Omega,
\end{align*}$$

where the compliance tensor $C$ is given such that $(C \cdot \cdot, \cdot \cdot)$ is an inner product for symmetric 2-tensor fields. This equation is of great importance in solid mechanics. Many textbooks on this equation and applications exist, for example [77].

The well-posedness and regularity theory for this equation is well-understood. In particular, for smooth $C$, given $f \in H^{-1} \otimes \mathbb{R}^n$, there exists a unique $u \in \tilde{H}^1 \otimes \mathbb{R}^n$ solving the problem. An exposition of the regularity theory on various types of domains can be found in [44].

4.3.1 Continuous mixed formulation

The linear elasticity equation is put into a Hilbert space context via a mixed formulation suitable for discretization. Here we use the TDNNS formulation first proposed in [87, 88, 98, 101].
We review their continuous results and make the necessary changes to generalize them to all dimensions.

In the TDNNS mixed formulation, we will use $H(\text{div}\text{div})$ for the stress variable $\sigma = C\epsilon u$. This is where rotated Regge elements would fit. We still need another space to pair with $\text{div}\sigma$. For this, we need a Hilbert space for vector fields which is between $L^2 \otimes \mathbb{R}^n$ and $H^1 \otimes \mathbb{R}^n$ derived from the de Rham complex:

$$H^1 = \{ u \in L^2 \otimes \mathbb{R}^n | \partial_i u_j - \partial_j u_i \in L^2 \text{ for all } i,j \}.$$  

This is a Hilbert space under the graph inner product

$$(u,v)_{H^1} = \sum_{1 \leq k \leq n} (u_k,v_k) + \frac{1}{4} \sum_{1 \leq i < j \leq n} (\partial_i u_j - \partial_j u_i, \partial_i v_j - \partial_j v_i).$$

In 2D, $H^1$ is the space $H(\text{rot})$. In 3D, $H^1$ is the space $H(\text{curl})$. In general, this is the space of $L^2$-differential 1-forms \cite{6, 8}. It can be shown using integration by parts and a density argument \cite[page 19]{6} that elements of $H^1$ has a well-defined tangential trace to the boundary. More precisely, there is a bounded linear map $H^1 \rightarrow H^{-1/2}(\partial\Omega) \otimes \mathbb{R}^n$. Let $\hat{H}^1$ be the subspace of $H^1$ with vanishing tangential trace. We will show that there is a duality pairing $\langle \text{div}\tau, v \rangle$ for $\tau \in H(\text{div}\text{div})$ and $v \in \hat{H}^1$. This requires several steps.

First, we recall the following the regular decomposition result (Lemma 5 in \cite{31} with $k = 1$):

**Theorem 4.8.** On a bounded Lipschitz domain $\Omega$, for all $u \in \hat{H}^1$, there exists $\phi \in \hat{H}^1, z \in \hat{H}^1 \otimes \mathbb{R}^n$ such that $u = \nabla \phi + z$ with $\|\phi\|_{H^1} + \|z\|_{H^1} \leq M\|u\|_{H^1}$ for some constant $M$ depending only on $\Omega$.

Second, using a similar argument to Lemma 2.1 of \cite{87}, we show the following duality result:

**Theorem 4.9.** On a bounded Lipschitz domain, let

$$H^{-1}(\text{div}) := \{ u \in H^{-1} \otimes \mathbb{R}^n | \text{div} u \in H^{-1} \}. $$

Then, the dual space of $H^{-1}(\text{div})$ is:

$$(H^{-1}(\text{div}))' = \hat{H}^1.$$

In particular, $\text{div} H(\text{div}\text{div}) \subset H^{-1}(\text{div})$, therefore the pairing $\langle \text{div}\tau, v \rangle$ makes sense for $\tau \in H(\text{div}\text{div})$ and $v \in \hat{H}^1$ and leads to a bounded bilinear form on this pair of spaces.
Proof. By Theorem 4.8, we have the following equivalence of norms on distributions:
\[
\|f\|_{(\dot{H}^1)^\prime} = \sup_{u \in \dot{H}^1} \frac{\langle f, u \rangle}{\|u\|_{\dot{H}^1}} = \sup_{\phi, z} \frac{\langle f, \nabla \phi + z \rangle}{\|\phi\|_1 + \|z\|_1} \sim \sup_{\phi, z} \frac{\langle f, \nabla \phi \rangle}{\|\phi\|_1} + \sup_z \langle f, z \rangle = \|\nabla f\|_1 + \|f\|_1.
\]

By definition, \((\dot{H}^1)^\prime\) is the space of distributions with bounded dual norm. This implies the first claim. Finally, it is clear that for \(\sigma \in H(\text{div})\), \(\text{div} \sigma \in H^{-1} \otimes \mathbb{R}^n\) and \(\text{div} \nabla \sigma \in H^{-1}\). This proves the last claim.

Let \(A := C^{-1}\) be the *stiffness tensor*. The TDNNS continuous formulation is: given \(f \in H^{-1}(\text{div})\), find \(\sigma \in H(\text{div})\), \(u \in \dot{H}^1\) such that
\[
(A \sigma, \tau) + \langle u, \text{div} \tau \rangle = 0, \quad \forall \tau \in H(\text{div}),
\]
\[
\langle \text{div} \sigma, v \rangle = -\langle f, v \rangle, \quad \forall v \in \dot{H}^1.
\]

First we show that this system is well-posed. The theorem below largely follows the arguments in Theorem 2.3 of [87]. The proof there has a gap where they only proved \(\langle \text{div} \sigma, v \rangle = (w, v)_{H(\text{curl})}\) for \(v \in \dot{H}^1 \otimes \mathbb{R}^n\) but in the end took \(v = w\) where \(w \in H(\text{curl})\) is from a bigger space. Here we give the correct proof with more details and greater generality.

**Theorem 4.10.** On a bounded Lipschitz domain \(\Omega\) in \(\mathbb{R}^n\), there exists a unique solution \((\sigma, u)\) to system (4.8). Further there exists a constant \(M\) depending only on \(\Omega\) and the coefficient \(C\) such that
\[
\|\sigma\|_{H(\text{div})} + \|u\|_{\dot{H}^1} \leq M\|f\|_{(\dot{H}^1)^\prime}.
\]

Proof. This follows from Brezzi’s theorem [17]. We only need to show the inf-sup condition for \(\langle \text{div} \tau, v \rangle\). Fix any \(v \in \dot{H}^1\). Let \(w \in \dot{H}^1 \otimes \mathbb{R}^n\) be the solution to
\[
(C \epsilon w, \epsilon y) = (v, y)_{H^1}, \quad \forall y \in \dot{H}^1 \otimes \mathbb{R}^n.
\]
The left-hand side is a bounded coercive bilinear form. Hence such unique solution \(w\) exists. Define \(\tau := C \epsilon w\). Then \(\tau \in L^2 \otimes \mathbb{R}^n\) with
\[
\|\tau\|_{L^2} = \|C \epsilon w\|_{L^2} \leq M_1\|w\|_{H^1},
\]
where \(M_1\) is a constant depending on \(\Omega\) and \(C\). By Korn’s inequality and the equation defining \(w\),
\[
\|w\|_{H^1}^2 \leq M_2(C \epsilon w, \epsilon w) = M_2(v, w)_{H^1} \leq M_2\|v\|_{H^1} \|w\|_{H^1} \leq M_2\|v\|_{H^1} \|w\|_{H^1},
\]
for some constant \(M_2\) depending on \(\Omega\) and \(C\). Hence,
\[
\|w\|_{H^1} \leq M_2\|v\|_{H^1}, \quad \|\tau\|_{L^2} \leq M_1 M_2\|v\|_{H^1}.
\]

103
Moreover, by definition
\[ (\tau, \epsilon y) = (v, y)_{H^1}, \quad \forall y \in \dot{H}^1 \otimes \mathbb{R}^n. \]
\[ \text{eq:divtau} \]

Take any test function \( \rho \). We note that \( \nabla \rho \) has the property:
\[ \partial_i \partial_j \rho - \partial_j \partial_i \rho = 0. \]
Thus if we choose \( y = \nabla \rho \), we get:
\[ (\tau, \epsilon \nabla \rho) = (v, \nabla \rho)_{H^1} = (v, \nabla \rho) \leq \|v\|_{L^2} \|\rho\|_{H^1}. \]
\[ \text{eq:divtau} \]

Notice \( \epsilon \nabla \) is just the Hessian. Thus by definition, \( \text{div} \text{div} \tau \) is a distribution in \( H^{-1} \) with:
\[ \|\text{div} \text{div} \tau\|_{H^{-1}} \leq \|v\|_{L^2}. \]

Hence \( \sigma \in H(\text{div} \text{div}) \) with
\[ \|\tau\|_{H(\text{div} \text{div})} \leq M_3 \|v\|_{H^1}. \]

Now take the regular decomposition of \( v = \nabla \phi + z \) for some \( \phi \in \dot{H}^1 \) and \( z \in \dot{H}^1 \otimes \mathbb{R}^3 \) with
\[ \|\phi\|_{H^1} + \|z\|_{H^1} \leq M_4 \|v\|_{H^1}. \]
The \( z \)-part can be plug into equation (4.9) by choosing \( y = z \),
\[ \langle -\text{div} \tau, z \rangle := (\tau, \epsilon z) = (v, z)_{H^1}. \]

Note that equation (4.10) can be extended to \( \rho \in \dot{H}^1 \) by density. Hence, we can choose \( \rho = \phi \),
\[ \langle -\text{div} \tau, \nabla \phi \rangle := (\tau, \epsilon \nabla \phi) = (v, \nabla \phi)_{H^1}. \]

Adding these two up, we get
\[ \langle -\text{div} \tau, v \rangle = \|v\|_{H^1}^2. \]

The \( H(\text{div} \text{div}) \)-norm bound derived before the implies the inf-sup for the bilinear form \( \langle \text{div} \tau, v \rangle \). This proves the theorem.

The mixed system (4.8) solves the linear elasticity equation (4.7) when the body force \( f \) is in \( H^{-1}(\text{div}) \).

**Theorem 4.11.** On a bounded Lipschitz domain, suppose \( f \in H^{-1}(\text{div}) \). Let \( (\sigma, u) \) be the unique solution to the mixed system (4.8). Then \( u \in \dot{H}^1 \otimes \mathbb{R}^n \) and its solves the elasticity equation (4.7).

**Proof.** The elasticity equation has a unique solution in \( \dot{H}^1 \otimes \mathbb{R}^n \) when \( f \) is from a bigger space \( H^{-1} \otimes \mathbb{R}^n \). Given this special \( f \), let \( w \) be that unique solution in \( \dot{H}^1 \otimes \mathbb{R}^n \). Since the mixed system has a unique solution, we have proven the theorem if we can show that \( (Cw, w) \)
solves the mixed system. Let $\sigma := C\epsilon w$. It is clear that $\sigma \in L^2 \otimes \mathbb{S}^n$. The fact that $w$ solves the elasticity equation implies that $\text{div} \sigma = -f \in H^{-1}(\text{div})$, that is, $\text{div} \text{div} \sigma \in H^{-1}$. Hence $\sigma \in H(\text{div} \text{div})$ and satisfies the second equation of system (4.8). For vector-valued test functions $y$ we have,

$$\langle y, \text{div} \tau \rangle = -\langle \tau, \epsilon y \rangle, \quad \forall \tau \in H(\text{div} \text{div}).$$

By density, the above holds for $y = w \in \dot{H}^1 \otimes \mathbb{R}^n$ as well. Hence,

$$\langle w, \text{div} \tau \rangle = \langle \tau, -\epsilon w \rangle = -\langle \tau, A(C\epsilon w) \rangle = -\langle \tau, A\sigma \rangle.$$

This shows that the first equation of system (4.8) is satisfied as well.

### 4.3.2 Rotated Regge element discretization

In this subsection, we show how to discretize the mixed formulation (4.8) using generalized Regge elements. We will state an implementable method, prove its consistency, and test it numerically in the next subsection. The proof for stability and error estimates will be future work. The approach here follows closer to the mesh-dependent norm analysis framework of [11]. A different analysis approach for a different finite element discretization of the same mixed formulation (4.8) is given in [88].

The relationship between $H(\text{div} \text{div})$ and piecewise normal-normal continuous finite elements were already studied in the biharmonic section of this chapter. We still use $S(\text{REG'})$ to discretize $H(\text{div} \text{div})$. The finite element theory for the space $H\Lambda^1$ is well-understood [6, 8]. We use the FEEC element $\mathcal{P}'\Lambda^1$ to discretize $H\Lambda^1$. In dimension 2 and 3, $\mathcal{P}'\Lambda^1$ is Nédéléc edge elements of the second kind, which is widely used. The only thing remains here is to derive the formula for the pairing $\langle \text{div} \tau, v \rangle$. It is more natural to define this in mesh dependent spaces, in a fashion very similar to the that of the biharmonic case.

Let $\Omega$ be a Lipschitz polyhedral domain in $\mathbb{R}^n$ and $\mathcal{T}_h$ a mesh of size $h$. Define

$$V := \{\text{piecewise } H^1 \text{ symmetric matrix fields with normal-normal continuity} \},$$

$$W := \{\text{piecewise } H^1 \text{ vectors fields in } \dot{H}\Lambda^1 \text{ with tangential continuity} \}.$$

Note that piecewise $H^1$ vector fields with tangential continuity is already a subspace of $H\Lambda^1$. Hence the condition $\dot{H}\Lambda^1$ in the definition of $W$ simply means that elements of $W$ have vanishing tangential trace on the boundary. We make $V$ and $W$ Hilbert spaces by giving them mesh-dependent norms:

$$\|\sigma\|_V^2 = \sum_c \|\sigma\|_{L^2(c)}^2 + h \|\sigma\|_{L^2(\partial c)}^2.$$
\[ \|u\|_W^2 = \sum_c \|u\|_{H^1(c)}^2 + h^{-1}\|u\|_{L^2(\partial)}^2, \]

where both sums are over all the cells \(c\) in mesh \(\mathcal{T}_h\). We define a mesh-dependent \(\text{div}_h\) operator: for any \((\tau, v) \in V \times W\),

\[
\langle \text{div}_h \tau, v \rangle := \sum_c \int_c -\tau : \epsilon v + \int_{\partial c} \tau_{nn} v_n = \sum_c \int_c -\tau : \epsilon v + \sum_f \int_f \tau_{nn} [v_n], \tag{4.11} \]

where \(n\) is the unit outward normal to a cell \(c\), the second sum is over all facets \(f\) of the mesh, and as before \(\tau_{nn} := n^T \tau n\) and \(v_n := v \cdot n\). It is clear that this is well-defined. Further, it is a bounded bilinear form: there is a constant \(M\) independent of \(h\) such that

\[ |\langle \text{div}_h \tau, v \rangle| \leq M \|\tau\|_V \|v\|_W. \]

We now introduce our finite element choices as subspaces of \(V\) and \(W\). For \(r \geq 1\), let

\[ V_h := S(\text{REG}^r), \quad W_h := \mathcal{P}^r \Lambda^1 \cap \hat{H} \Lambda^1. \]

The discrete problem corresponding to mixed system (4.8) is: given \(f \in H^{-1}(\text{div})\), find \((\sigma, u) \in V_h \times W_h\), such that

\[
\langle A\sigma, \tau \rangle - \langle u, \text{div}_h \tau \rangle = 0, \quad \forall \tau \in V_h, \tag{eq:discrete_tdnns} \]

\[
\langle \text{div}_h \sigma, v \rangle = -\langle f, v \rangle, \quad \forall v \in W_h. \tag{4.12} \]

An obvious question is, given that \(\text{div}_h\) is not really \(\text{div}\) and \(V_h \times W_h\) does not have an apparent relationship to \(H(\text{div} \cdot \text{div}) \times \hat{H} \Lambda^1\), how is system (4.12) a discretization of the mixed system (4.8) at all. This situation is the same as the HHJ discretization of the biharmonic equation. This all makes sense if we have consistency, which means that solutions of the linear elasticity equation satisfies (4.12) in some sense, and discrete stability, which means that (4.12) itself is well-posed uniformly in \(h\). We prove the consistency here and leave the stability as future work.

We need more regularity than the minimal for this consistency theorem to hold. It is known that the \(H^2\) regularity for the elasticity equation (4.7) holds for smooth \(C\) on convex polyhedral or \(C^2\) domains [44] for \(f \in L^2 \otimes \mathbb{R}^n\).

**Theorem 4.12.** Suppose \(u \in (\mathcal{H}^1 \cap H^2) \otimes \mathbb{R}^n\) solves the elasticity equation (4.7). Let \(\sigma := C\epsilon u\). Then \((\sigma, u)\) satisfies system (4.12).

**Proof.** First, it is clear that \(u \in W\) and \(\sigma \in H^1 \otimes \mathbb{S}^n \in V\). Hence the equations in system (4.12) still make sense for this continuous \((\sigma, u)\). Since \(u \in H^1\) globally, \([u_n] = 0\). Thus,

\[
\langle u, \text{div}_h \tau \rangle = \int_{\Omega} -\tau : \epsilon u = \int_{\Omega} -A\tau : C\epsilon u = \int_{\Omega} -A\tau : \sigma = -\langle A\sigma, \tau \rangle. \]
This proves that the first equation is satisfied. Second, because \( \sigma \in H^1 \otimes \mathbb{S}^n \), we have \( [\sigma_{nt}] = 0 \) at all interior facets. On the other hand, \( \nu_T = 0 \) at all boundary facets. So overall,

\[
\sum_c \int_{\partial_c} \sigma_{nt} \nu_T = \sum_f \int_f [\sigma_{nt}] \nu_T = 0.
\]

Because \( u \in H^2 \otimes \mathbb{R}^n \), we have \( f \in L^2 \otimes \mathbb{R}^n \). Thus,

\[
(-f, v) = (\text{div} \sigma, v) = \sum_c \int_c (\text{div} \sigma) v = \sum_c -\int_c \sigma : \varepsilon v + \int_{\partial_c} \sigma_n \cdot v
\]

\[
= \sum_c -\int_c \sigma_n v + \int_{\partial_c} \sigma_{nn} v_n = (\text{div}_h \sigma, v),
\]

where the second to last equation used the decomposition \( \sigma_n \cdot v = \sigma_{nn} v_n + \sigma_{nt} v_T \) and the previous identity. This proves that the second equation is satisfied as well.

For software implementation of system (4.12) in an environment where \( \text{REG}^r \) is already implemented, as is the case for FEniCS, the following equivalent formulation should be used:

\[
\text{find } (\rho, u) \in \text{REG}^r \times \mathbb{W}_h \text{ such that }
\]

\[
(AS\rho, S\tau) - \langle u, \text{div}_h S\tau \rangle = 0, \quad \forall \tau \in \text{REG}^r, \quad \text{eq:regge_elasticity}
\]

\[
\langle \text{div}_h S\rho, v \rangle = -(f, v), \quad \forall v \in \mathbb{W}_h.
\]  

\[\text{(4.13)}\]

### 4.3.3 Numerical experiments

In this subsection, we show through numerical experiments that discretization (4.13) is stable but converges with suboptimal \( L^2 \) error rates in the \( \sigma \) variable. We further show that this method can be implemented to handle different boundary conditions in both 2D and 3D and do not suffer from locking phenomenon.

First, we look at the convergence test for 2D. The elasticity equation (4.7) is solved on the unit square with the following exact solution:

\[
u = \begin{bmatrix} \sin(\pi x) \sin(\pi y) \\ 15x(1-x)y(1-y) \end{bmatrix} \in C^\infty \cap H^1.
\]

A sequence of unstructured meshes are again generated using the FEniCS package mshr. The procedure is essentially the same as the one used for the biharmonic case. The implementation can be found in the script `rotated_regge/tnns_conv.py` in the companion repository.

Table 4.7, Table 4.8, and Table 4.9 show the convergence test results for 2D with \( r = 1, 2, 3 \), where \( \| \cdot \| \) means the \( L^2 \)-norm. It is clear that the \( L^2 \) convergence rate is optimal for \( u \) but 1 order suboptimal for \( \sigma \).
Table 4.7: 2D elasticity degree 1

<table>
<thead>
<tr>
<th>Mesh size</th>
<th>$|u - u_h|$</th>
<th>Rate</th>
<th>$|\sigma - \sigma_h|$</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>7.094244e-03</td>
<td></td>
<td>1.687453e-01</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1.742923e-03</td>
<td>2.08</td>
<td>8.163971e-02</td>
<td>1.08</td>
</tr>
<tr>
<td>32</td>
<td>4.548902e-04</td>
<td>1.88</td>
<td>4.239306e-02</td>
<td>0.92</td>
</tr>
<tr>
<td>64</td>
<td>1.130114e-04</td>
<td>2.00</td>
<td>2.114090e-02</td>
<td>1.00</td>
</tr>
<tr>
<td>128</td>
<td>2.834921e-05</td>
<td>1.98</td>
<td>1.062154e-02</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Table 4.8: 2D elasticity degree 2

<table>
<thead>
<tr>
<th>Mesh size</th>
<th>$|u - u_h|$</th>
<th>Rate</th>
<th>$|\sigma - \sigma_h|$</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1.945019e-03</td>
<td></td>
<td>3.559110e-02</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2.633687e-04</td>
<td>2.87</td>
<td>8.278951e-03</td>
<td>2.10</td>
</tr>
<tr>
<td>16</td>
<td>3.096338e-05</td>
<td>3.17</td>
<td>1.910408e-03</td>
<td>2.17</td>
</tr>
<tr>
<td>32</td>
<td>3.931342e-06</td>
<td>2.89</td>
<td>4.794675e-04</td>
<td>1.94</td>
</tr>
<tr>
<td>64</td>
<td>5.015333e-07</td>
<td>2.95</td>
<td>1.224961e-04</td>
<td>1.96</td>
</tr>
</tbody>
</table>

Table 4.9: 2D elasticity degree 3

<table>
<thead>
<tr>
<th>Mesh size</th>
<th>$|u - u_h|$</th>
<th>Rate</th>
<th>$|\sigma - \sigma_h|$</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.346217e-03</td>
<td></td>
<td>1.646650e-02</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.031960e-04</td>
<td>4.06</td>
<td>2.660197e-03</td>
<td>2.88</td>
</tr>
<tr>
<td>8</td>
<td>7.203260e-06</td>
<td>3.83</td>
<td>3.054329e-04</td>
<td>3.11</td>
</tr>
<tr>
<td>16</td>
<td>4.397091e-07</td>
<td>4.14</td>
<td>3.678944e-05</td>
<td>3.13</td>
</tr>
<tr>
<td>32</td>
<td>2.857008e-08</td>
<td>3.83</td>
<td>4.766975e-06</td>
<td>2.87</td>
</tr>
<tr>
<td>64</td>
<td>1.779997e-09</td>
<td>3.98</td>
<td>5.977013e-07</td>
<td>2.98</td>
</tr>
</tbody>
</table>

We then study the convergence rates in 3D. The linear elasticity equation is solved on the unit cube with the following exact solution:

$$u = \begin{bmatrix} \sin(\pi x) \sin(\pi y) \sin(\pi z) \\ 15x(1-x)y(1-y)z(1-z) \\ 7x(1-x)\sin(\pi y)\sin(\pi z) \end{bmatrix} \in C^\infty \cap \dot{H}^1.$$
A sequence of randomly perturbed meshes, like the one in Figure 4.5, are generated in the same way as in the 3D biharmonic case.

Table 4.10 shows the convergence test results for 3D with $r = 1$. Due to a regression bug in FEniCS, the bilinear form fails to assemble for $r \geq 2$. It seems that what was observed in 2D still holds, that the $L^2$ convergence rate is optimal for $u$ but 1 order suboptimal for $\sigma$.

<table>
<thead>
<tr>
<th>Mesh size</th>
<th>$|u - u_h|$</th>
<th>Rate</th>
<th>$|\sigma - \sigma_h|$</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.391169e-01</td>
<td></td>
<td>1.832970e+00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.049142e-01</td>
<td>1.86</td>
<td>1.048096e+00</td>
<td>0.89</td>
</tr>
<tr>
<td>6</td>
<td>4.901894e-02</td>
<td>1.54</td>
<td>7.042752e-01</td>
<td>0.81</td>
</tr>
<tr>
<td>8</td>
<td>2.892313e-02</td>
<td>2.01</td>
<td>5.382773e-01</td>
<td>1.02</td>
</tr>
<tr>
<td>10</td>
<td>1.908383e-02</td>
<td>2.07</td>
<td>4.314971e-01</td>
<td>1.10</td>
</tr>
</tbody>
</table>

Table 4.10: 3D elasticity degree 1

We then look at a more interesting 2D example. The domain and its unstructured mesh is shown in Figure 4.6. It is given by the rectangle $[0, 3] \times [0, 1]$ with three disks removed, one of radius 0.2 centered at $(0.4, 0.3)$, one of radius 0.375 centered at $(1.5, 0.5)$, and one of radius 0.3 centered at $(2.4, 0.6)$. The material is isotropic and homogeneous, that is, the stress and the strain are related by

$$ \epsilon u = \frac{(1 + \nu)\sigma - \nu I \text{tr} \sigma}{E}, $$

where the Young’s modulus $E = 10$ and the Poisson’s ratio $\nu = 0.2$. The boundary condition is given as follows. It is clamped on the left-side $u = [0, 0]$ and compressed on the right-side $u = [-1, 0]$. The top-side, bottom-side, along with the holes are traction-free $\sigma n = 0$. No external force is applied to this body.
We note that in the TDNNS formulation, the tangential part of the displacement \( u_\tau \) is an essential boundary condition, while the normal part is a natural boundary condition. Suppose \( u_n = g_n \) on the part of the boundary \( \Gamma_N \). Then we get

\[
\int_{\Gamma_N} \tau_{nn} g_n
\]

in the right-hand side of the first equation of (4.12). Similarly, the normal-normal traction is an essential boundary condition, while the normal-tangential traction is a natural boundary condition. The normal-tangential traction leads to an analogous additional boundary integral term in the right-hand side of the second equation of (4.12).

This problem is solved with degree \( r = 1 \). A plot of the solution is shown in Figure 4.6. Here the domain is deformed using the displacement vector field and colored by the von Mises stress, which is proportion to \( \sigma_d : \sigma_d \) where the deviatoric stress \( \sigma_d := \sigma - \frac{1}{2} I \text{tr} \sigma \).
In linear elasticity, it is well-known that the primal method (using the displacement alone as the main variable) suffers from the locking phenomenon: when the Poisson's ratio is close to 0.5, the quality of the numerical solution degrades substantially. Mixed methods should not suffer from this. In Figure 4.7, we show the solution of the same problem when the Poisson’s ratio is $\nu = 0.499999$. It is clear that the numerical solution is free of artifacts and is only slightly different from the previous case as expected. This confirms that this method does not suffer from locking.
We then look at a more interesting problem in 3D. The domain is the box $[0, 4] \times [0, 2] \times [0, 1]$ with two cylindrical holes, one along the $y$-axis centered at $(4/3, 0, 1/2)$ with radius $0.3$ and another one along the $z$-axis centered at $(8/3, 1, 0)$ with radius $0.7$. A mesh is created from this domain using mshr. The domain and the mesh are shown in Figure 4.8 and Figure ??.
The material is again isotropic and homogeneous with Young's modulus $E = 1.0$ and Pois-
son’s ratio $\nu = 0.2$. The boundary condition is as follows. The left-end is clamped $u = [0, 0, 0]$, while the right-end is been rotated by $\pi/6$. There is no external force. Figure ?? shows a visualization of the numerical solution. Again the domain is deformed using the displacement vector field and colored by the von Mises stress which is proportion to $\sigma_d : \sigma_d$ where the 3D deviatoric stress $\sigma_d := \sigma - \frac{1}{3} \text{tr} \sigma$.

Figure 4.11: Visualization of the 3D solution with Poisson ratio $\nu = 0.2$.

Figure ?? shows a visualization of the numerical solution when the Poisson’s ratio is $\nu = 0.499999$ instead. Again we observe that the solution is free of artifacts and fairly similar to the previous solution as expected. This confirms that this method does not suffer from locking in 3D as well.
4.4 Connection with numerical relativity

We end this chapter by describing the connection of the two problems studied in this chapter to numerical relativity. It will be shown in the model problems chapter of this thesis that the linearized Einstein equation (around the Minkowski metric) reads:

\[
\begin{align*}
\text{div} \text{div} S &\gamma = 0, \\
\text{div} S\gamma' + \text{curl} \text{curl} \beta &\equiv 0, \\
S\gamma'' + 2\text{ein} \gamma + S\nabla\nabla \alpha - 2S\varepsilon \beta' &\equiv 0.
\end{align*}
\]

where $\alpha$ is a scalar field, $\beta$ is a vector field, $\gamma$ is a symmetric matrix field, primes indicate the time derivatives, and ein is the linearized (Euclidean) Einstein tensor. This system has the structure of a constrained evolution equation, where the first two equations are constraints and the last equation is the evolution equation.

We would eventually want to solve this equation using the generalized Regge elements for the variable $\gamma$. For example, we would at least want to know to what extent the constraints are satisfied. We then need to look at $\text{div} \text{div} S$ and $\text{div} S$ of functions in REG$^\ast$. Notice that these two are exactly the main operators in the HHJ mixed formulation of the biharmonic equation and the TDNNS formulation of the linear elasticity equation we studied in this chapter.
chapter.

In some sense, these two are the most natural equations derived from the operators \( \text{div} \text{div} S \) and \( \text{div} S \). If an operator \( L \) is invertible, then we naturally would study the solution of \( L\sigma = f \). When the operator \( L \) has a nontrivial kernel, like the two operators here, it is natural to study the regularized problem:

\[
\min \|\sigma\|, \quad \text{subject to } L\sigma = f.
\]

Indeed the mixed formulations are just Lagrange multiplier versions of these regularized problems. The studies of these problems reveal a lot of useful information on the discretization these operators.

As will be shown later in this thesis, the Einstein equation as given has robustness issues and its discretization requires regularization. One of the most promising approaches is to add functions of the constraints into the evolution equation to regularize it. Hence the understanding of discretization \( \text{div} \text{div} S \) and \( \text{div} S \) provides useful information for the discretization of the Einstein equation as well. For example, what we learned in this chapter would suggest that the constraint equation involving \( \text{div} \text{div} S\gamma \) is likely to hold in the discrete sense, when tested against \( CG^{r+1} \), while the constraint involving \( \text{div} S\gamma' \) is likely to hold when tested against \( P^r \Lambda^1 \). These ideas will be explored in future work.
Chapter 5

Model problems in relativity for studying discretization

In this chapter, the main goal is to motivate, introduce, and analyze two continuous linear model problems, namely Problem 5.2 and Problem 5.4, from general relativity. The finite element discretization and analysis of these two problems will be the first step towards a rigorous theory of using Regge Calculus for numerical relativity.

The chapter is organized as follows. First, we review some relevant material on the Einstein equation. Second, we linearize the Einstein equation at the Minkowski metric and argue why this is a good model problem. Then we decompose the space-time linear problem into the spatial and temporal part and formulate an initial-value problem. After that, we give a complete treatment of the gauge freedom in the linear case and derive the hyperbolic model problem (Problem 5.2). We also prove the well-posedness of this model problem. Then we look at the corresponding steady state problem and formulate the elliptic model problem (Problem 5.4). We also look at the application of this elliptic model problem in the geometric theory of defects in solid mechanics. Finally, we end the chapter with a Fourier analysis of the two model problems on the flat torus. This reveals much of the structure of these two problems in a clear way and provides us with the expectation of the results in the general theory in later chapters.

5.1 The fully nonlinear space-time Einstein equation

The Einstein field equation [34] is a well-established model for large-scale structures in the universe. Many excellent textbooks on this topic exist, for example [48, 79, 109]. For the purpose of this thesis, we only need to know that it is a nonlinear second-order partial dif-
ferential equation on symmetric 2-tensor fields with many applications in physics. In this section, this aspect of the Einstein equation is recalled for the convenience of the reader.

We will write down the Einstein equation in coordinates following the notation in [3]. The 4 dimensions of the spacetime are labeled by integers 0, 1, 2, 3, where 0 is for the time. When used as indices, lower case Greek letters $\alpha, \beta, \ldots$ are for the spacetime and can take the values $0, 1, 2, 3$, while lower case Latin letters $i, j, \ldots$ are for the spatial part only and can take the values 1, 2, 3. Einstein's summation convention (repeated indices are always summed over) is assumed.

The unknown of the Einstein equation is the spacetime metric $g_{\alpha\beta}$, which is a pseudo-Riemannian metric of the signature $(-, +, +, +)$. Its associated Christoffel symbol is a nonlinear function containing its first-order derivatives [3, Equation (1.8.12)]:

$$\Gamma^\gamma_{\beta\gamma} := \frac{1}{2} g^{\alpha\mu}(\partial_\gamma g_{\beta\mu} + \partial_\beta g_{\gamma\mu} - \partial_\mu g_{\beta\gamma}),$$

where $g^{\alpha\beta}$ is the inverse of $g_{\mu\nu}$, that is, $g^{\alpha\beta}g_{\beta\mu} = \delta^\alpha_\mu$. With that, the Riemann curvature tensor is defined as a nonlinear function containing the first-order derivatives of the Christoffel symbol [3, Equation (1.9.2)]:

$$R^a_{\beta\mu\nu} := \partial_\mu \Gamma^a_{\beta\nu} - \partial_\nu \Gamma^a_{\beta\mu} + \Gamma^a_{\rho\mu} \Gamma^\rho_{\beta\nu} - \Gamma^a_{\rho\nu} \Gamma^\rho_{\beta\mu}.$$

Finally, the Einstein tensor is defined in terms of the Riemann tensor [3, Equation (1.10.4)]:

$$G_{\mu\nu} := R^a_{\mu\alpha\nu} - \frac{1}{2} g_{\mu\nu} g^{\alpha\beta} R^a_{\beta\alpha\gamma} + Q_{\mu\nu},$$

where $Q_{\mu\nu}$ is a polynomial in $g_{\alpha\beta}, g^{\alpha\beta}$, and $\partial_\lambda g_{\alpha\beta}$ which is still exactly quadratic in $\partial_\lambda g_{\alpha\beta}$.

\begin{equation}
G_{\mu\nu} = 0. \tag{5.1}
\end{equation}

Our main equation, the vacuum Einstein field equation is:

The above definition is not very transparent. The following theorem clears it up a bit. Following the convention, we simplify the notation by using the metric $g_{\alpha\beta}$ to raise and lower the indices implicitly: for example, given $u_{\alpha\beta}, u^\mu_\beta$ is defined as $u_{\alpha\beta} g^{\mu\alpha}$.

**Theorem 5.1.** The Riemann tensor, in terms of the metric, splits into a second-order principal part and lower-order terms as:

$$R^a_{\beta\mu\nu} = \frac{1}{2} g^{\alpha\lambda}[\partial_\mu(\partial_\beta g_{\lambda\nu} - \partial_\lambda g_{\beta\nu}) - \partial_\nu(\partial_\beta g_{\lambda\mu} - \partial_\lambda g_{\beta\mu})] + P^a_{\beta\mu\nu}, \tag{5.2}
$$

where $P^a_{\beta\mu\nu}$ is a polynomial in $g^{\alpha\beta}$ and $\partial_\lambda g_{\alpha\beta}$ which is exactly quadratic in the latter. The Einstein tensor also splits as a second-order principal term and lower-order terms:

$$G_{\mu\nu} = \frac{1}{2} \left( -\partial^\lambda \partial_\lambda g_{\mu\nu} + \partial_\mu \partial^\lambda g_{\lambda\nu} + \partial_\nu \partial^\lambda g_{\lambda\mu} - g^{\alpha\beta} \partial_\mu \partial_\nu g_{\alpha\beta} - g_{\mu\nu} g^{\alpha\beta} \partial_\lambda g_{\alpha\beta} + g_{\mu\nu} g^{\alpha\beta} \partial_\lambda g_{\alpha\beta} \right) + Q_{\mu\nu}, \tag{5.3}
$$

where $Q_{\mu\nu}$ is a polynomial in $g_{\alpha\beta}, g^{\alpha\beta}$, and $\partial_\lambda g_{\alpha\beta}$ which is still exactly quadratic in $\partial_\lambda g_{\alpha\beta}$.
Proof. Just plugging in the definition of Christoffel symbol and get

\[ R^\alpha_{\beta\mu\nu} = \frac{1}{2} g^{\alpha\lambda} \{ \partial_\mu (\partial_\beta g_{\lambda\nu} - \partial_\lambda g_{\beta\nu}) - \partial_\nu (\partial_\beta g_{\lambda\mu} - \partial_\lambda g_{\beta\mu}) \} + P^\alpha_{\beta\mu\nu}, \]

where the following was used to move derivatives from the inverse metric to the metric:

\[ 0 = \partial_\nu (\delta^\alpha_{\mu}) = \partial_\nu (g^{\alpha\beta} g_{\beta\mu}) = g_{\beta\mu} \partial_\nu g^{\alpha\beta} + g^{\alpha\beta} \partial_\nu g_{\beta\mu}, \]

and the resulting \( P^\alpha_{\beta\mu\nu} \) is the lower-order term as claimed. Taking the \( \alpha\mu \) trace, we get,

\[ R^\alpha_{\beta\alpha\nu} = -\frac{1}{2} \partial_\lambda \partial_\nu g_{\lambda\nu} + \frac{1}{2} (\partial_\beta \partial_\lambda g^{\alpha\lambda} + \partial_\nu \partial_\lambda g_{\alpha\beta}) - \frac{1}{2} g^{\alpha\lambda} \partial_\lambda \partial_\nu g + S_{\beta\nu}, \]

\[ g^{\beta\nu} R^\alpha_{\beta\alpha\nu} = \partial_\nu \partial_\lambda g_{\lambda\nu} - g^{\beta\nu} \partial_\nu \partial_\lambda g_{\alpha\beta} + g^{\beta\nu} S_{\beta\nu}, \]

where \( S_{\beta\nu} \) is a quadratic polynomial in \( g^{\alpha\beta} \) and \( \partial_\lambda g_{\alpha\beta} \). Plugging these into the definition of the Einstein tensor proves the claim.

This in particular shows that the map from the metric \( g_{\mu\nu} \) to its Einstein tensor \( G_{\mu\nu} \) is a nonlinear second-order differential operator.

5.2 Linearized space-time Einstein equation

The full nonlinear Einstein equation has a fairly complicated form as shown in the previous section. To understand this better, we look at some meaningful simplified model problems. From the perspective of numerical analysis of Galerkin methods, broadly speaking, the following diagram commutes:

Continuous nonlinear problem \( \xrightarrow{\text{Galerkin discretization}} \) Discrete nonlinear problem

\[ \downarrow \text{Iterative linearization} \quad \downarrow \text{Iterative linearization} \]

Continuous linear problem \( \xrightarrow{\text{Galerkin discretization}} \) Discrete linear problem

where iterative linearization means solving nonlinear problems via solving a sequence of linearized problems (for example, Newton's method). Hence the ability to solve the linearized problem is crucial. This leads us to study the linearized Einstein equation.

It is known that the Minkowski metric \( \eta_{\mu\nu} = \text{diag}(-1, 1, 1, 1) \) is a solution the Einstein equation. As the simplest possible model problem, we study the Einstein equation linearized around the Minkowski metric. Physically, this models the propagation of the gravitational waves through the empty space. The background metric \( \eta_{\mu\nu} \) is used to raise and lower indices.
**Theorem 5.2.** Let $s$ be a scalar parameter and $g_{\mu\nu} = \eta_{\mu\nu} + sh_{\mu\nu}$ for some symmetric 2-tensor field $h_{\mu\nu}$. Let $G_{\mu\nu}$ be the Einstein tensor of $g_{\mu\nu}$. Define

$$(\text{ein} h)_{\mu\nu} := \frac{d}{ds} G_{\mu\nu} \bigg|_{s=0}$$

to be the linearized Einstein operator at $\eta_{\mu\nu}$. Then

$$(\text{ein} h)_{\mu\nu} = \frac{1}{2} \left( -\partial^3 \partial_\lambda h_{\mu\nu} + \partial_\mu \partial^3 h_{\lambda\nu} + \partial_\nu \partial^3 h_{\lambda\mu} - \partial_\mu \partial_\nu h^{\alpha}_\alpha - \eta_{\mu\nu} \partial^3 \partial_\lambda h^\alpha_\alpha + \eta_{\mu\nu} \partial^3 \partial_\lambda h^a_a \right).$$  

(5.4)

In particular, the linearized Einstein equation reads:

$$(\text{ein} h)_{\mu\nu} = 0.$$  

(5.5)

**Proof.** Plug $g_{\mu\nu} := \eta_{\mu\nu} + sh_{\mu\nu}$ into equation (5.3) of Theorem 5.1 as the metric and compute to the first-order in $s$. The $Q_{\mu\nu}$ part, being exactly quadratic in the first derivative of the metric, vanishes.

It is straightforward to check that to the first-order in $s$, the inverse metric of $g_{\mu\nu}$ is:

$$g^{\mu\nu} = \eta^{\mu\nu} - sh^{\mu\nu} + o(s^2).$$

The first three principal terms of the Einstein tensor gives:

$$\frac{s}{2} \left( -\partial^3 \partial_\lambda h_{\mu\nu} + \partial_\mu \partial^3 h_{\lambda\nu} + \partial_\nu \partial^3 h_{\lambda\mu} \right) + o(s^2),$$

where the $s^2$ term comes from the fact that in the nonlinear formula $g^{\mu\nu}$ is used to raise the index in $\partial^3$. The computation for the rest three principal terms are also elementary. For example: the $g_{\mu\nu}g^{\alpha\beta}\partial^3 \partial_\lambda g_{a\beta}$ term becomes:

$$(\eta_{\mu\nu} + sh_{\mu\nu})(\eta^{\alpha\beta} - sh_{a\beta})(\eta^{\lambda\tau} - sh_{\lambda\tau}) \partial_\tau \partial_\lambda (\eta_{a\beta} + sh_{a\beta})$$

$$= s\eta_{\mu\nu} \eta^{\alpha\beta} \eta^{\lambda\tau} \partial_\tau \partial_\lambda h_{a\alpha} + o(s^2) = s\eta_{\mu\nu} \partial^3 \partial_\lambda h_{a\alpha} + o(s^2).$$

The rest three principal terms contributes:

$$\frac{s}{2} \left( -\partial_\mu \partial_\nu h^a_a - \eta_{\mu\nu} \partial^a \partial_\beta h_{a\beta} + \eta_{\mu\nu} \partial^3 \partial_\lambda h^a_a \right) + o(s^2).$$

Combining these, we get the claim.

Comparing the form of the linearized Einstein tensor (5.4) and the form of the full nonlinear Einstein tensor (5.3), we see that the linearized Einstein retains all the principal terms. In fact, if we were to linearized at some other solutions to the nonlinear Einstein equation, the form of the principal part of the linearized equation would be the exactly same, except that the indices are raised by a different background metric. Hence, up to lower-order terms and variable coefficients in the principal part, the linearized Einstein tensor at the Minkowski metric is the most general form of linearized Einstein tensor. This justifies the use of equation (5.5) as a model problem for relativity.
5.3 Matrix calculus notation

Although the index notation is very flexible and convenient for tensor computations, it does not convey meaning (in terms of familiar operators in calculus) efficiently. Both the domain and the image of the linearized Einstein operator are symmetric 2-tensor fields. The background Minkowski metric establishes a canonical Euclidean coordinate system, under which symmetric 2-tensor fields are just symmetric matrix fields. In this section, we review some notations from solid mechanics for matrix calculus and rewrite the linearized Einstein equation in terms of familiar differential operators in calculus.

First, we define some operators which are independent of the metric and the dimension. For any scalar field \( u \) and any vector field \( v \), we have the gradient, hessian, and the symmetric gradient:

\[
(\nabla u)_a := \partial_a u, \quad (\nabla\nabla u)_{a\beta} := \partial_a \partial_\beta u, \quad (ev)_{a\beta} := \frac{1}{2}(\partial_a v_\beta + \partial_\beta v_a).
\]

Given a metric \( g \) as a symmetric matrix field, we define the divergence and the Laplacian:

\[
(div g u) := g_{a\beta} \partial_a u^\beta, \quad (div g v)_\beta := g^{a\lambda} \partial_\lambda v_{a\beta}, \quad \Delta_g := g^{a\beta} \partial_a \partial_\beta.
\]

The Laplacian can act on tensor fields of any shape component by component. We further define some algebraic operators: the trace, the two operators frequently used in relativity \( J \) and \( S \): for a matrix field \( u \),

\[
(tr g u) := g^{a\beta} u_{a\beta}, \quad J_g u := u - \frac{1}{2} g(tr g u), \quad S_g u := u - g(tr g u).
\]

To further simply the notation, the subscript for the metric is usually omitted when the metric is obvious from the context.

Under Minkowski metric \( \eta \), the linearized Einstein operator (5.4) in this notation is thus:

\[
ein u = \frac{1}{2}(-\Delta u + 2\varepsilon \div u - \nabla \nabla tr u - \eta \div \div u + \eta \Delta tr u).
\]

We immediately see that if \( u \) happens to be divergence-free and trace-free, then \( \ein u \) is the same as half of the d’Lambertian of \( u \). In that case, the linearized Einstein equation is just a component-wise wave equation.

We get more with the matrix notation. The following theorem can be proved by a direct computation:

\section*{thm:identities}
Theorem 5.3. In dimension $m$, under any constant pseudo-Riemannian metric $g$ (for example, Minkowski metric and Euclidean metric), the following identities hold:

$$J^{-1}u = u - \frac{1}{m-2}g(\text{tr}u), \quad S^{-1}u = u - \frac{1}{m-1}g(\text{tr}u),$$

$$\text{div} g = \nabla, \quad \text{tr} \nabla = \text{div},$$

$$\text{div} \epsilon = \frac{1}{2} \Delta + \frac{1}{2} \nabla \text{div}, \quad \text{div} J \epsilon = \frac{1}{2} \Delta,$$

$$\text{div} S \nabla \nabla = 0, \quad \text{div} \text{div} S \epsilon = 0.$$

Corollary 5.4. We have a more compact formula for the linearized Einstein:

$$\text{ein} = - \frac{1}{2} J \Delta + J \epsilon \text{div} J.$$  \hspace{1cm} \text{(5.6)}

This is valid in any dimension under any constant pseudo-Riemannian metric.

Corollary 5.5.

$$\text{div ein} = 0, \quad \text{ein} \epsilon = 0.$$

Proof. This is a direct consequence of $\text{div} J \epsilon = \frac{1}{2} \Delta$ in Theorem 5.3:

$$\text{div ein} = - \frac{1}{2} \text{div} J \Delta + (\text{div} J \epsilon) \text{div} J = - \frac{1}{2} \text{div} J \Delta + \frac{1}{2} \Delta \text{div} J = 0.$$

$$\text{ein} \epsilon = - \frac{1}{2} \Delta J \epsilon + J \epsilon (\text{div} J \epsilon) = - \frac{1}{2} \Delta J \epsilon + \frac{1}{2} J \epsilon \Delta = 0.$$

Note that the Laplacian commutes with any linear operators.

The second identity in the above reveals the gauge freedom of the linearized Einstein equation. If a symmetric matrix field $u$ is a solution to $\text{ein} u = 0$, then $u + \epsilon \phi$ is also a solution for any vector field $\phi$. Mathematically, this means that $\text{ein}$ has an infinite dimensional kernel, just like curl in 3D calculus. This will have a major impact in the well-posedness of the initial-value problem at the continuous level and the choice of elements at the discrete level. These topic will be examined later in this thesis.

5.4 (1+3)-formulation of linearized initial-value problem

The Einstein field equation is a statement about a law of nature. In order to use it for anything, we need to set up a closed system of equations based on it. As discussed before, the principal part of the Einstein equation can be understood as a d’Alambertian plus some unhelpful terms. It is thus a hyperbolic equation and can be used to setup initial-value
problems or Cauchy problems. How to setup a well-posed hyperbolic problem in the fully nonlinear case is well-understood. For example [25,96] are very good references on this topic. The general case is complicated by the lack of a background manifold, which necessitates building the manifold and the metric with the correct causal structure. These are out of the scope of this thesis.

Here we are only interested in the linearized problem. In the presence of a background manifold, the four-dimensional Euclidean space, it is much easier to setup an initial-value problem. This will be described here.

The first step is to describe the background manifold. This is always a subset of $\mathbb{R}^4$ which can be decomposes into a direct product of the temporal part and the spatial part. We have many choices. The temporal part can be either $[0,T]$ for some terminal time $T > 0$ or simply $[0,\infty)$. Since the equation is linear, it is easy to find global solutions. So this choice does not matter. We will look at three cases of spatial domains $\Omega$. The first one is the full space $\mathbb{R}^3$. The second one is the flat torus $\mathbb{T}^3$, that is, a cube with side length $2\pi$ and periodic boundary condition. The third case is a bounded smooth domain or Lipschitz polyhedron in $\mathbb{R}^3$. In this case, the Dirichlet boundary condition is assumed: the trace of the metric perturbation to the spatial boundary (the tangential-tangential component of the matrix field) vanishes.

The second step is to decompose the temporal and spatial part of the matrix fields involved according to the decomposition of the background manifold $[0,T] \times \Omega$. This will make the hyperbolic nature of the linearized Einstein equation more transparent. In the literature, this is called (1+3)-decomposition. To make the notation less confusing, we will use the subscript (4) to remind us that the operator or variable is in 4D and no special notation for 3D objects. When in 4D, the background metric is always the Minkowski metric while in 3D the background metric is always the Euclidean metric $I := \text{diag}(1,1,1)$. We use the matrix notation for the decomposition:

$$h_{(4)} = \begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix},$$

where $\alpha$ is a scalar field for $g_{00}$, $\beta$ is a 3D vector field for $g_{0i}$, and $\gamma$ is a 3D symmetric matrix field for $g_{ij}$. All these fields are still defined on $[0,T] \times \Omega$ and are now interpreted as time-dependent functions. Similarly, 4D vector fields are also decomposed in the matrix notation:

$$u_{(4)} = \begin{bmatrix} \phi \\ w \end{bmatrix},$$

where $\phi$ is a scalar field for $u_0$ and $w$ is a 3D vector for $u_i$. The time derivative $\partial_0$ will be denoted by prime $'$ while spatial differential operators will continue be denoted using the
matrix calculus notation. Straightforward computations give the following theorem:

**Theorem 5.6.** In the (1 + 3) matrix notation, we have

\[
\nabla(4)u = \begin{bmatrix} u' \\ \nabla u \end{bmatrix}, \quad \nabla(4)u = \begin{bmatrix} u'' & \ldots \\ \nabla u' & \nabla u \end{bmatrix}, \quad \varepsilon(4) = \begin{bmatrix} \phi' & \ldots \\ \frac{1}{2}(\nabla \phi + w') & \epsilon w \end{bmatrix}
\]

\[
\text{tr}(4) \begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix} = -\alpha + \text{tr} \gamma, \quad \text{div}(4) \begin{bmatrix} \phi' \\ w \end{bmatrix} = -\phi + \text{div} w, \quad \text{div}(4) \begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix} = \begin{bmatrix} -\alpha' + \text{div} \beta \\ -\beta' + \text{div} \gamma \end{bmatrix},
\]

\[
J(4) \begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix} = \begin{bmatrix} \frac{1}{4}(a'' - \Delta a + \text{tr} \gamma'' - \Delta \text{tr} \gamma) & \ldots \\ \frac{1}{2}(\beta'' - \Delta \beta) & \frac{1}{2}(J\gamma'' - \Delta J\gamma) + \frac{1}{4}I(a'' - \Delta a) \end{bmatrix},
\]

where “…” parts are omitted by symmetry.

Given this, we can compute the (1 + 3) decomposition of ein:

**Theorem 5.7.**

\[
ein(4) = \frac{1}{2} \begin{bmatrix} \text{div div} \nabla S \gamma & \ldots \\ \text{div} \nabla S \gamma' + (\nabla \text{div} - \Delta) \beta & S\gamma'' + 2\text{ein} \gamma + S\nabla \alpha - 2\epsilon \beta \end{bmatrix}.
\]

**Proof.** This is a long computation using the previous theorem and formula (5.6). First,

\[
-\frac{1}{2}J\Delta(4) \begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix} = \begin{bmatrix} \frac{1}{4}(a'' - \Delta a + \text{tr} \gamma'' - \Delta \text{tr} \gamma) & \ldots \\ \frac{1}{2}(\beta'' - \Delta \beta) & \frac{1}{2}(J\gamma'' - \Delta J\gamma) + \frac{1}{4}I(a'' - \Delta a) \end{bmatrix}.
\]

Second,

\[
J\epsilon \text{div} J(4) \begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix} = J\epsilon \text{div}(4) \begin{bmatrix} \frac{1}{2}(a + \text{tr} \gamma) & \ldots \\ \beta & J\gamma + \frac{1}{2}Ia \end{bmatrix} = J\epsilon(4) \begin{bmatrix} -\frac{1}{2}(\alpha' + \text{tr} \gamma') + \text{div} \beta \\ -\beta' + \text{div} J\gamma + \frac{1}{2}\nabla \alpha \end{bmatrix}
\]

\[
= \begin{bmatrix} \frac{1}{4}(-\alpha'' - \text{tr} \gamma'' + \Delta a) + \frac{1}{2} \text{div div} J\gamma & \ldots \\ \frac{1}{2}(\beta'' + \text{div} \gamma' - \nabla \text{tr} \gamma' + \nabla \text{div} \beta - \epsilon \beta' + \epsilon \text{div} J\gamma + \frac{1}{2}\nabla \alpha) & J\epsilon \text{div} J\gamma + \frac{1}{2}J\nabla \alpha - \frac{1}{4}I(a'' + \text{tr} \gamma'') - S\epsilon \beta' \end{bmatrix}
\]

Combining these two parts, we get the claim. \(\square\)

Identity (5.7) is valid in (1 + m)-dimension for all \(m \geq 1\). Here the spatial part has dimension 3. We have some extra identities which are quite convenient. The following can be verified by a direct computation:

124
Theorem 5.8. In dimension 3 under the Euclidean metric, for a vector field $u$ and a symmetric matrix field $w$,

\[ J^{-1} = S, \quad \nabla \text{div} u - \Delta u = \text{curl} \text{curl} u, \quad \text{div} S \epsilon u = \frac{1}{2} \text{curl} \text{curl} u, \quad 2\epsilon \text{ein} w = \text{curl} (\text{curl} w)^T, \]

where curl of a matrix is defined row by row.

In 3D, $2\epsilon \text{ein}$ is also known as the Saint-Venant's operator or the incompatibility operator in the solid mechanics literature. Its properties will be explored further in the rest of this thesis.

We thus arrived at the $(1 + 3)$ linearized Einstein equation. A triple $(\alpha, \beta, \gamma)$ of time-dependent scalar, vector, symmetric matrix fields on $\Omega$ are components of a solution to the linearized Einstein equation if and only if:

\[
\begin{align*}
\text{div} \text{div} S \gamma &= 0, \\
\text{div} S \gamma' + \text{curl} \text{curl} \beta &= 0, \\
S \gamma'' + 2\epsilon \text{ein} \gamma + S \nabla \nabla \alpha - 2S \epsilon \beta' &= 0. 
\end{align*}
\]

This should be interpreted as a constrained evolution system, where the first two equations are constraints while the last one is the evolution equation. This is justified by the following theorem:

Theorem 5.9. The evolution equation propagates the constraints: suppose $(\alpha(t), \beta(t), \gamma(t))$ solves the evolution equation and satisfies the two constraint equations at $t = 0$, then it satisfies the two constraints for all $t$.

Proof. Recall that by Theorem 5.3 and its corollaries $\text{div} S \nabla \nabla = 0$ and $\text{div} \epsilon \text{ein} = 0$. Take the divergence of the evolution equation:

\[ \text{div} S \gamma'' - 2\text{div} S \epsilon \beta' = 0. \]

By Theorem 5.8, this is equivalent to the time derivative of the second constraint equation:

\[ (\text{div} S \gamma' + \text{curl} \text{curl} \beta)' = 0. \]

Hence if the second constraint is satisfied at a time, it is satisfied at all times. Now take the divergence of the second constraint equation:

\[ \text{div} \text{div} S \gamma' = 0. \]

This is the time derivative of the first constraint equation. Hence if the first constraint is satisfied at a time, it is also satisfied at all times. \qed
Finally we are ready to state the full \((1+3)\) initial-value problem for linearized relativity.

**Problem 5.1.** Let \(\Omega\) be one of \(\mathbb{R}^3\), \(T^3\), or a bounded smooth domain or Lipschitz polyhedron in \(\mathbb{R}^3\). Given a scalar field \(a\), a vector field \(b\), and two symmetric matrix fields \(g, h\) satisfying the compatibility conditions:

\[
\text{div div } S g = 0, \quad \text{div } S h + \text{curl curl } b = 0,
\]

find a triple \((\alpha(t), \beta(t), \gamma(t))\) of scalar, vector, symmetric matrix fields on \(\Omega\), such that \(\alpha(0) = a\), \(\beta(0) = b\), \(\gamma(0) = g\), \(\gamma'(0) = h\), and for all \(t > 0\):

\[
S \gamma'' + 2 \epsilon \gamma + S \nabla \alpha - 2 S \epsilon \beta' = 0.
\]

### 5.5 Gauge freedom and simplified model problem

At the end of Section 5.2, we noted the existence of gauge freedom \(\epsilon = 0\). Due to this, Problem 5.1 cannot be well-posed because of a lack of uniqueness of solution. In this section, we first study the structure of this gauge freedom and then derive a better model problem.

Two symmetric matrix fields \(g_{(4)}\) and \(h_{(4)}\) are called equivalent under the gauge condition, denoted by \(g_{(4)} \sim h_{(4)}\), if there exists a vector field \(u_{(4)}\) such that \(g_{(4)} = h_{(4)} + \epsilon u_{(4)}\). The structure of this is most clear in \((1+3)\) formulation. It turned out that to study this properly, we need some well-posedness results on elliptic equations. These will be recalled here.

Given a domain \(\Omega\), we will be dealing with the scalar and vector Poisson problem: given a smooth scalar field \(f\) on \(\Omega\), find a scalar field \(u\) on \(\Omega\) such that

\[
- \Delta u = f,
\]

and given a smooth vector field \(g\) on \(\Omega\), find a vector field \(w\) on \(\Omega\) such that

\[
- \Delta w = g.
\]

The results are recalled for different domains. First \(\Omega = \mathbb{R}^3\). This is the simplest case and can be understood with the classical potential theory. Basically, as long as we restrict to the space of decaying function \(u \to 0\) as \(|x| \to \infty\), both equations have a unique solution. Details of these elementary results can be found in, for example [35, Section 2.2]. Second \(\Omega = T^3\). This can be analyzed easily using Fourier series, which will be done in Section 5.8.

For now it is sufficient to know that as long as we restrict ourselves to functions of zero averages, both equations have a unique solution. Third \(\Omega\) is a bounded smooth domain or Lipschitz polyhedron in \(\mathbb{R}^3\). This is also a well-understood case. For the scalar Laplacian,
the homogeneous Dirichlet boundary condition guarantees the existence and uniqueness of a solution. For the vector Laplacian, we need an electric-field-type boundary condition:

\[ u \times n = 0, \quad \text{div} \, u = 0, \]

where \( n \) is the normal vector to \( \Omega \). The well-posedness of the vector Laplacian equation can be established using the elliptic theory of 2-form Hodge Laplacian in 3D [8, Section 4.2.2]. This is important because of its compatibility with the boundary condition on the metric:

**Lemma 5.10.** Let \( \Omega \) be a domain in \( \mathbb{R}^3 \) and \( n \) its unit normal vector on the boundary. If a vector field \( u \) satisfies \( u \times n = 0 \) on the boundary, then the tangential-tangential component of the symmetric matrix field \( \epsilon u \) also vanishes on the boundary.

**Proof.** The proof is most transparent in the index notation. A symmetric matrix field \( w \) has vanishing tangential-tangential part if and only if:

\[ \epsilon^{ilp} \epsilon^{jkq} n_l n_k w_{pq} = 0, \]

where \( \epsilon^{ilp} \) is the Levi-Civita symbol (this \( \epsilon \) is different from the symmetric gradient \( \epsilon \)). Plug in the definition of the symmetric gradient: \( \epsilon u \) has vanishing tangential-tangential component if and only if

\[ 0 = \epsilon^{ilp} \epsilon^{jkq} n_l n_k \partial_p u_q = \epsilon^{ilp} n_l \partial_p (\epsilon^{jkq} n_k u_q) \]

where in the first step the symmetrization is not necessary because \( \epsilon^{ilp} \epsilon^{jkq} n_l n_k \) already symmetrizes in the \( p, q \) indices. Convert back to vector notation:

\[ \epsilon^{jkq} n_k u_q = (n \times u)^j. \]

Hence indeed \( u \times n = 0 \) ensures that \( \epsilon u \) has vanishing tangential-tangential trace. □

It turned out that we can change \( \alpha \) and \( \beta \) arbitrarily along with certain components of \( \gamma \) at one time slice while stay in the same equivalence class:

**Theorem 5.11.** Let \( \Omega \) be one of \( \mathbb{R}^3 \), \( \mathbb{T}^3 \), or a bounded smooth domain or Lipschitz polyhedron in \( \mathbb{R}^3 \). Given any symmetric matrix field \( g_{(4)} \), scalar field \( \alpha(t) \), vector field \( \beta(t) \) on \([0, T] \times \Omega \) and any scalar field \( k \), vector field \( h \) on \( \Omega \), there exists a symmetric matrix field \( \gamma(t) \) on \([0, T] \times \Omega \) such that

\[ g_{(4)} \sim \begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix}, \quad \text{div} \, J \gamma(0) = h, \quad \text{tr} \, \gamma'(0) = k. \]
In particular, every equivalence class \([g(4)]\) under the gauge freedom has a unique representative of the form:

\[
g(4) \sim \begin{bmatrix} 0 & 0 \\ 0 & \gamma \end{bmatrix},
\]

where \(\gamma(t)\) satisfies:

\[
\text{div} J \gamma(0) = 0, \quad \text{tr} \gamma'(0) = 0.
\]

**Proof.** Suppose

\[
g(4) = \begin{bmatrix} \alpha_g & \beta_g^T \\ \beta_g & \gamma_g \end{bmatrix}.
\]

For any scalar field \(\xi_0\) and vector field \(\phi_0\) on \(\Omega\), we can define a scalar field \(\xi(t)\) and a vector field \(\phi(t)\) on \([0, T] \times \Omega\) by:

\[
\xi(t) := \xi_0 + \int_0^t (\alpha_g(s) - \alpha(s)) ds,
\]

\[
\phi(t) := \phi_0 + \int_0^t (2\beta_g(s) - 2\beta(s) - \nabla \xi(s)) ds.
\]

Direct computation shows that

\[
\begin{bmatrix} \alpha_g & \beta_g^T \\ \beta_g & \gamma_g \end{bmatrix} - \epsilon(4) \begin{bmatrix} \xi \\ \phi \end{bmatrix} = \begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma_g - \epsilon \phi \end{bmatrix}.
\]

Let \(\gamma := \gamma_g - \epsilon \phi\). Once we find \(\xi_0, \phi_0\) such that \(\text{div} J \gamma(0) = h\) and \(\text{tr} \gamma'(0) = k\), we are done proving this theorem.

Recall from Theorem 5.3 that \(\text{div} J c = \frac{1}{2} \Delta\). The condition \(\text{div} J \gamma(0) = 0\) implies a vector Laplace equation in \(\phi_0\):

\[-\Delta \phi_0 = -2 \text{div} J c \phi = 2h - 2 \text{div} J \gamma_g(0).
\]

The condition \(\text{tr} \gamma'(0) = k\) implies a scalar Laplace equation in \(\xi_0\):

\[-\Delta \xi_0 = \text{tr} \gamma'_g - k - 2 \text{div}(\beta_g(0) - \beta(0)).
\]

By assumption, \(\Omega\) is a domain such that both elliptic problems are well-posed. Define \(\phi_0\) and \(\xi_0\) to be the solutions to their corresponding equations. This proves the theorem.

The theorem above gives a complete characterization of the gauge freedom for linearized Einstein equation. In particular, it suggests that we only need to be able to solve an even more restricted problem where only some representative of the equivalence class is sought after. This is called *gauge fixing* in the physics literature.
The representatives specified in Theorem 5.11 require

\[ \alpha \equiv 0, \quad \beta \equiv 0, \quad \text{div} J \gamma(0) = 0, \quad \text{tr} \gamma'(0) = 0. \]

At time slice \( t = 0 \), the compatibility condition further requires that

\[ \text{div div} S \gamma(0) = 0, \quad \text{div} S \gamma'(0) + \text{curl curl} \beta(0) = 0. \]

Combining this two, at \( t = 0 \), we have

\[ \Delta \text{tr} \gamma(0) = 0, \quad \text{div} J \gamma(0) = 0, \]

\[ \text{tr} \gamma'(0) = 0, \quad \text{div} \gamma'(0) = 0, \]

where the following lemma is used:

**Lemma 5.12.** If \( \text{div div} Su = 0 \) and \( \text{div} Ju = 0 \) then \( \Delta \text{tr} u = 0 \).

**Proof.** By definition of \( J \) and \( S \):

\[ 0 = \text{div} Ju = \text{div} u - \frac{1}{2} \nabla \text{tr} u, \quad 0 = \text{div div} Su = \text{div div} u - I \Delta \text{tr} u. \]

Subtracting the divergence of the first from the second proves the claim.

This leads to the following linearized Einstein initial-value problem with fixed gauge:

**Problem 5.2.** Let \( \Omega \) be one of \( \mathbb{R}^3 \), \( \mathbb{T}^3 \), or a bounded smooth domain or Lipschitz polyhedron in \( \mathbb{R}^3 \). Given two symmetric matrix fields \( g \) and \( h \) satisfying:

\[ \Delta \text{tr} g = 0, \quad \text{div} J g = 0, \quad \text{tr} h = 0, \quad \text{div} h = 0, \]

find a symmetric matrix field \( \gamma(t) \) such that \( \gamma(0) = g \), \( \gamma'(0) = h \), and for all \( t > 0 \):

\[ S \gamma'' + 2 \text{ein} \gamma = 0. \tag{5.9} \]

This evolution equation also propagates the constraints on the initial data:

**Theorem 5.13.** The following identities hold:

\[ \text{trein} = -\frac{1}{2} \text{div div} S \tag{5.10} \]

Moreover, the solution \( \gamma(t) \) to Problem 5.2 satisfies for all \( t \geq 0 \):

\[ \Delta \text{tr} \gamma = 0, \quad \text{div} J \gamma = 0, \quad \text{tr} \gamma' = 0, \quad \text{div} \gamma' = 0. \]
Proof. The first identity is just a direct computation using the definition of ein. The trace of the evolution equation becomes:

\[ 0 = \text{tr} S \gamma'' + 2 \text{trein} \gamma = -2 \text{tr} \gamma' - \text{div} \text{div} S \gamma. \]

The initial data here is compatible for Problem 5.1 with \( \alpha = 0 \) and \( \beta = 0 \). Hence Theorem 5.9 applies:

\[ \text{div} \text{div} S \gamma = 0, \quad \text{div} S \gamma' = 0, \]

for all time. Thus the trace of the evolution equation implies that \( \text{tr} \gamma'' = 0 \). Since \( \text{tr} \gamma'(0) = 0 \), we have \( \text{tr} \gamma' = 0 \) for all time. Given this, \( \text{div} S \gamma' = 0 \) implies \( \text{div} \gamma' = 0 \). Since the trace and the divergence of \( \gamma \) are constant in time, the first two identities in the claim hold at \( t = 0 \) implies that they hold for all time.

There is no loss of generality to go from Problem 5.1 to Problem 5.2.

**Theorem 5.14.** Suppose \((a,b,g,h)\) is a compatible initial data for Problem 5.1. Define a scalar field \( \xi \) and a vector field \( \phi \) as the solutions to the Poisson problems:

\[ -\Delta \xi = \text{tr} h - 2 \text{div} b, \quad -\Delta \phi = -2 \text{div} Jg. \]

Set \( \bar{g} := g - c\phi \) and \( \bar{h} := h - 2cb + \nabla \nabla \xi \). Let \( \bar{\gamma}(t) \) be the solution to Problem 5.2 with initial data \((\bar{g}, \bar{h})\). Then

\[ a(t) := a, \quad \beta(t) := b, \quad \gamma(t) := \bar{\gamma}(t) + c\phi + t(2cb - \nabla \nabla \xi) - \frac{1}{2} t^2 \nabla \nabla a, \]

is one out of an infinite number of solutions to Problem 5.1.

**Proof.** First we check that \( \bar{g} \) and \( \bar{h} \) can be used as initial data for Problem 5.2. Indeed, by Theorem 5.3,

\[ \text{div} J \bar{g} = \text{div} Jg - \text{div} Jc\phi = \text{div} Jg - \frac{1}{2} \Delta \phi = 0. \]

Due to equation (5.10) and \( \text{ein} \epsilon = 0 \), we have

\[ \text{div} \text{div} S \epsilon = 0. \]

Hence the assumption that \( \text{div} \text{div} Sg = 0 \) implies that \( \text{div} \text{div} S \bar{g} = 0 \). By Lemma 5.12, we have \( \Delta \text{tr} \bar{g} = 0 \). This shows that \( \bar{g} \) can be used as initial data. We then check \( \bar{h} \):

\[ \text{tr} \bar{h} = \text{tr} h - 2 \text{div} b + \Delta \xi = 0. \]

Further, using the fact that \( h \) is compatible and \( c\nabla = \nabla \nabla \):

\[ \text{div} S \bar{h} = \text{div} Sh - 2 \text{div} Scb + \text{div} S \nabla \nabla \xi = \text{div} Sc \nabla \nabla \xi = 0. \]
Now $\text{tr} \tilde{h}$ and $\text{div} S \tilde{h}$ together implies that $\text{div} \tilde{h} = 0$. Hence $\tilde{h}$ can be used as the initial data as well.

Finally we check the defined $(a(t), \beta(t), \gamma(t))$ solves Problem 5.1. From the definition, it is clear that

$$a(0) = a, \quad \beta(0) = b, \quad \gamma(0) = g, \quad \gamma'(0) = h.$$ 

We note that $\gamma$ and $\tilde{\gamma}$ only differ by a symmetric gradient:

$$\gamma = \tilde{\gamma} + \epsilon(\phi + 2tb - t\nabla \xi - \frac{1}{2} t^2 \nabla a).$$

Hence $\text{ein} \gamma = \text{ein} \tilde{\gamma}$ and $\text{div div} S \gamma = \text{div div} S \tilde{\gamma}$. It is then straightforward to check that both of the constraint equations and the evolution equation are satisfied.

### 5.6 Well-posedness on the full space and flat torus

In this section we study the well-posedness of Problem 5.2.

We note that if $\Omega = \mathbb{R}^3$ or $\Omega = \mathbb{T}^3$, using the uniqueness of the Poisson problem, the condition $\Delta \text{tr} g = 0$ on the initial data is equivalent to $\text{tr} g = 0$. In these two cases, the conditions on $g$ become:

$$\text{tr} \gamma(0) = 0, \quad \text{div} \gamma(0) = 0.$$ 

In this case, the well-posedness is easy to establish and will be proven in this section. On a bounded domain $\Omega$, however, $\text{tr} \gamma(0)$ does not necessarily vanish. Indeed, only the tangential-tangential part of $\gamma$ vanishes on the boundary. Hence $\text{tr} \gamma = n^T \gamma n$ on the boundary, which can be arbitrary. The well-posedness in this case is more difficult and need the proper Hilbert space framework which will be discussed only in the later part of this thesis.

For now, we only consider the case $\Omega = \mathbb{R}^3$ or $\Omega = \mathbb{T}^3$. A symmetric matrix field $u$ is called divergence-free trace-free (DT) if and only if:

$$\text{tr} u = 0, \quad \text{div} u = 0.$$ 

On these two domains, Problem 5.2 is equivalent to the following more concise problem:

**Problem 5.3.** Let $\Omega$ be either $\mathbb{R}^3$ or $\mathbb{T}^3$. Given two DT symmetric matrix fields $g$ and $h$, find a symmetric matrix field $\gamma(t)$ such that $\gamma(0) = g$, $\gamma'(0) = h$, and for all $t > 0$:

$$S \gamma'' + 2\text{ein} \gamma = 0.$$ 

It is well-known that the component-wise wave equation is well-posed on $\Omega$. The above problem is in fact equivalent to the wave equation.
**Theorem 5.15.** Suppose $\gamma(t)$ is a solution to Problem 5.3 with DT initial data $(g, h)$. Then $\gamma$ also satisfies the wave equation:

$$\gamma'' - \Delta \gamma = 0.$$ 

Let $\tilde{\gamma}(t)$ be the solution to the above wave equation with initial data $\tilde{\gamma}(0) = g$ and $\tilde{\gamma}'(0) = h$. Then, $\tilde{\gamma}$ also solves Problem 5.3. Due to these, Problem 5.3 is well-posed.

**Proof.** By Theorem 5.13, the solution $\gamma$ is DT for all time. For DT $\gamma$, $S\gamma = J\gamma = \gamma$ and

$$\text{ein} \gamma = -\frac{1}{2} \Delta J\gamma + J\epsilon \div J\gamma = -\frac{1}{2} \Delta \gamma.$$ 

Hence the evolution equation $S\gamma'' + 2\text{ein} \gamma = 0$ is equivalent to the wave equation for DT $\gamma$. For $\tilde{\gamma}$, it is clear that the wave equation propagates the DT condition. Hence $\tilde{\gamma}$ is DT and therefore also satisfies the evolution equation. \qed

If $\Omega$ is a bounded domain in $\mathbb{R}^3$ and we only put the tangential-tangential boundary condition on $\gamma$, then the model problem 5.2 is fundamentally different from a component-wise wave equation.

### 5.7 Elliptic steady state model problems and geometric theory of defects

**sec:elliptc_prob**

A further simplification can be made by removing the time-dependence altogether. This leads us to study the steady state problem for the linearized Einstein equation. It is well-known in the finite element literature that the understanding of the corresponding elliptic part is the first step in analyzing the discretization of time-dependent problems (for example [105]). Here it turns out that the steady state problem has applications in solid mechanics and is of independent interest as well.

Setting the time derivative to zero, the steady state equation corresponding to the linearized evolution equation (5.9) is:

$$2\text{ein} \gamma = 0.$$ 

It is easy to check that if a symmetric matrix field $\gamma$ satisfies the above, then the constant metric

$$\begin{bmatrix} 0 & 0 \\ 0 & \gamma \end{bmatrix}$$ 

solves the linearized Einstein equation. The source problem is thus, given a symmetric matrix field $f$, find a symmetric matrix field $u$ such that

$$2\text{ein} u = f.$$ 

132
It is clear that this problem cannot be well-posed because \( \text{div} \epsilon = 0 \) is an obstruction to existence while \( \epsilon \epsilon = 0 \) is an obstruction to uniqueness. From these considerations, we look at this problem:

\[
\text{Problem 5.4. Let } \Omega \text{ be one of } \mathbb{R}^3, T^3, \text{ or a bounded smooth domain or Lipschitz polyhedron in } \mathbb{R}^3. \text{ Given a divergence-free symmetric matrix field } f, \text{ find a divergence-free symmetric matrix field } u \text{ such that }
\]
\[
2 \epsilon \epsilon u = f.
\]

On \( \mathbb{R}^3 \), \( u \) is further required to satisfy the decaying condition. On \( T^3 \), both \( u \) and \( f \) are required to have zero averages. When \( \Omega \) is a bounded domain in \( \mathbb{R}^3 \), \( u \) is required to satisfy the Dirichlet boundary condition: the tangential-tangential part of \( u \) vanishes on \( \partial \Omega \).

This is a highly nontrivial elliptic problem. Here we will focus on its interpretation and application in solid mechanics. In the next section, we will study its properties on \( T^3 \) using Fourier analysis. Similar analysis can be done on the full space \( \mathbb{R}^3 \) as well. The general bounded domain case requires sophisticated machinery from geometry and functional analysis and will be the subject of a full chapter in the later part of this thesis.

The geometric theory for defects and plasticity is an elegant theory with a long history in mathematics, solid mechanics, and physics. It started with Volterra's paper [107] on plasticity. Cartan [20] used dislocations in materials to motivate the notion of torsion in geometry. These works were subsequently picked up and developed further by engineers and physicists with a substantial literature. For good surveys, see [63, 65, 67, 85]. Recently it received more attention in the mathematics and physics literature, for some reviews [58, 64, 118]. The author is not aware of any treatment of this problem in numerical analysis yet. It is the author's opinion that its wider use in engineering was hampered largely by the current inability to efficiently solve the associated equations numerically. This thesis hopes to contribute to this from the numerical analysis perspective. The solution of Problem 5.4 is a first step in this direction. A high-level overview of the relevant ideas is reviewed here.

In solid mechanics, the deformation of a solid body is modeled by a symmetric matrix field, \( e \), called the strain tensor. It is essentially the difference of two Riemannian metrics in some ambient Euclidean space, one on the reference shape and another on the deformed shape, measuring the change between relative distances between points in the body. We have two types of deformations. The first one is elasticity deformation, which is a temporary self-reversing shape change. This can be modeled by a smooth displacement vector field \( u \), which measures to which direction and how much each point in the body moves. When the
deformation is small, we have \( \epsilon = \epsilon u \). This is usual elasticity model. Further discussions can be found in many good textbooks, for example [77].

The other kind of deformation is plastic deformation. This can be thought of as doing some cut-and-paste on the body itself. For example, we can deform a disk by cutting a small wedge out from it and soldering the two cut surfaces together. The resulting disk would have different mechanical properties from the original disk. There is no smooth displacement field which can model this process because some of the points are just gone. In this case, in the small deformation regime, we have a strain \( \epsilon \) which is not the symmetric gradient of any smooth vector field. Recall that \( \epsilon \epsilon = 0 \). Thus \( \epsilon \epsilon \) can be used to measure how far the strain \( \epsilon \) is from being elastic and \( \epsilon \epsilon \) has the interpretation of the effective strength of the plastic deformation.

A strain, elastic or plastic, creates a stress in the body. This is modeled by another a symmetric matrix field \( \sigma \), called the stress tensor, measuring various types forces at each point in the body. In the linear regime, the material itself is characterized through a coefficient \( C \), called the stiffness tensor, mapping symmetric matrix fields to symmetric matrix fields, \( \sigma = Ce \). Finally, the balance law closes the system \( \text{div} \sigma = -f \) where \( f \) is the external force on the body.

The following table taken from [67] nicely summarizes the models problems in solid mechanics and compares them with those in electromagnetism models:

<table>
<thead>
<tr>
<th>Electrostatics</th>
<th>Elastostatics</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{div} D = \rho )</td>
<td>( \text{div} \sigma = -f )</td>
</tr>
<tr>
<td>( \text{curl} E = 0 )</td>
<td>( \epsilon \epsilon = 0 )</td>
</tr>
<tr>
<td>( D = \epsilon E )</td>
<td>( \sigma = Ce )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Magnetostatics</th>
<th>Plastostatics</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{div} B = 0 )</td>
<td>( \text{div} \sigma = 0 )</td>
</tr>
<tr>
<td>( \text{curl} H = j )</td>
<td>( \epsilon \epsilon = \eta )</td>
</tr>
<tr>
<td>( B = \mu H )</td>
<td>( \sigma = Ce )</td>
</tr>
</tbody>
</table>

Problem 5.4 is the plastostatics problem with trivial coefficients. The interpretation of the plastostatics problem is as follows: \( \eta \) measures the effective strength of defects in the material generated by the plastic deformation. \( \text{div} \sigma = 0 \) means that the material is relaxed and static after the plastic deformation. This stress \( \sigma \) is called the residual stress. This problem arises, for example, in the modeling of grow of blood vessel walls [57]. Here \( \eta \) is related to the
grow rate of the blood vessel (cells grow more and bigger creating more material, opposite to the cutting a wedge out of a disk example) and $\sigma$ measures the residual stress caused by this growth.

5.8 Fourier analysis on the flat torus: smooth case

In this final section, we analyze the hyperbolic Problem 5.2 and elliptic Problem 5.4 on $T^3$ via Fourier analysis. This establishes some expectations for what to come in the later part of this thesis.

On $T^3$, a smooth scalar field $u$ can be represented as an infinite sum:

$$u(x) = \sum_{k \in \mathbb{Z}^3} u_k e^{ik \cdot x},$$

where $u_k \in \mathbb{C}$ are constants. For linear differential operators, it is sufficient to study their behavior for each $k$ individually.

Let $k \in \mathbb{Z}^3$, $k \neq 0$ be fixed. Let $m := k/\|k\|$, $n$ any unit vector orthogonal to $k$, and $l := m \times n$. The triple $(m, n, l)$ establishes an orthonormal basis for $\mathbb{R}^3$ adapted to $k$. We use the following matrix notation in this coordinate:

$$[a] := ae^{ik \cdot x}, \begin{bmatrix} a \\ b \\ c \end{bmatrix} := (am + bn + cl)e^{ik \cdot x},$$

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & j \end{bmatrix} := (amm^T + bmn^T + cml^T + dnn^T + enn^T + fnl^T + glm^T + hln^T + jll^T)e^{ik \cdot x}.$$  

Because the basis is orthonormal, algebraic operations like trace and transpose work directly in the matrix notation:

$$\text{tr} \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & j \end{bmatrix} = [a + e + j], \quad \begin{bmatrix} a & b & c \end{bmatrix}^T = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & j \end{bmatrix}.$$  

Matrix calculus is reduced to matrix algebra in this notation. This following theorem is a direct computation:
Theorem 5.16. The following holds:

\[
\begin{align*}
\nabla \left[ \begin{array}{c} a \\
0 \\
0 
\end{array} \right] &= i|k| \left[ \begin{array}{c} a \\
0 \\
0 
\end{array} \right], \\
\text{curl} \left[ \begin{array}{c} b \\
c \\
0 
\end{array} \right] &= i|k| \left[ \begin{array}{c} b \\
c \\
0 
\end{array} \right], \\
\text{div} \left[ \begin{array}{c} b \\
c \\
0 
\end{array} \right] &= i|k| \left[ \begin{array}{c} a \\
0 \\
0 
\end{array} \right],
\end{align*}
\]

For example, it is obvious that:

\[
\text{curl} \nabla = 0, \quad \text{div} \text{curl} = 0, \quad \epsilon \text{ein} = 0, \quad \text{div} \epsilon \text{ein} = 0.
\]

We have the Hodge decompositions for vector fields and symmetric matrix fields:

Theorem 5.17. On \( \mathbb{T}^3 \), for every smooth vector field \( u \), there exists a scalar field \( \phi \), another vector field \( v \), and a constant vector field \( p \) such that

\[
u = \nabla \phi + \text{curl} \, v + p.
\]

For every smooth matrix field \( \gamma \), there exists a vector field \( v \), another smooth matrix field \( \tau \), and a constant matrix field \( q \) such that

\[
\gamma = \epsilon \, v + \text{ein} \, \tau + q.
\]

Proof. \( k = 0 \) corresponds to the constant part. When \( k \neq 0 \), it is obvious in the matrix notation that the image of \( \nabla, \text{curl} \) in the vector case and the image of \( \epsilon, \text{ein} \) in the symmetric matrix case span the full space.

In the matrix notation, the Laplacian becomes multiplication by \(-|k|^2\). Hence both the scalar and vector Poisson problems are well-posed. The wave equation is a second-order ODE for each \( k \), which is also well-posed.

Theorem 5.18. On \( \mathbb{T}^3 \), the elliptic Problem 5.4 is well-posed.

Proof. By assumption, \( f \) is a divergence-free symmetric matrix field with zero average. For a fixed \( k \neq 0 \), suppose in the matrix notation,

\[
f_k = \left[ \begin{array}{ccc} 
0 & 0 & 0 \\
0 & a & b \\
0 & b & c 
\end{array} \right].
\]
Then it is clear that
\[
 u_k = \frac{1}{|k|^2} \begin{bmatrix}
 0 & 0 & 0 \\
 0 & -c & b \\
 0 & b & -a
\end{bmatrix}
\]
is the unique divergence-free symmetric matrix field with zero average such that ein \( u_k = f_k \).
By linearity, the claim holds.

The linearized Einstein equation is more interesting. We have already proven that on \( \mathbb{T}^3 \), Problem 5.2, or equivalently Problem 5.3 is well-posed. We look at the structure of the equation in more detail here.

Let \( \gamma \) be a symmetric matrix field for a fixed nonzero \( k \in \mathbb{Z}^3 \) of the form:
\[
 \gamma = \begin{bmatrix}
 a & b & c \\
 b & d & e \\
 c & e & f
\end{bmatrix}.
\]
The linearized Einstein evolution equation (5.9), \( S\gamma'' + 2\text{ein } \gamma = 0 \), reads:
\[
\begin{bmatrix}
 -d-f & b & c \\
 b & -a-f & e \\
 c & e & -a-d
\end{bmatrix}'' + |k|^2 \begin{bmatrix}
 0 & 0 & 0 \\
 0 & -f & e \\
 0 & e & -d
\end{bmatrix} = 0.
\]
Equivalently, it is an ODE system:
\[
 b'' = 0, \quad c'' = 0, \quad (d + f)'' = 0,
\]
\[
 e'' + |k|^2 e = 0,
\]
\[
 (a + f)'' + |k|^2 f = 0,
\]
\[
 (a + d)'' + |k|^2 d = 0.
\]
Define \( g = (d - f)/2 \) and \( h = (d + f)/2 \). Eliminate \( d \) and \( f \) from the above using \( g \) and \( h \), we get:
\[
 b'' = 0, \quad c'' = 0, \quad h'' = 0,
\]
\[
 e'' + |k|^2 e = 0,
\]
\[
 g'' + |k|^2 g = 0,
\]
\[
 a'' + |k|^2 h = 0.
\]
Equation (5.11)

It is clear that only the two oscillatory components \( e \) and \( g \) are bounded in time.
The constraint \( \text{div} \text{div } S\gamma = 0 \) reads:
\[
 h = 0,
\]
137
and $\text{div} S \gamma' = 0$ reads:

$$h' = 0, \quad b' = 0, \quad c' = 0.$$ 

With the constraints, the ODE system (5.11) becomes:

$$h = 0, \quad b' = 0, \quad c' = 0, \quad a'' = 0,$$
$$e'' + |k|^2 e = 0,$$
$$g'' + |k|^2 g = 0.$$ 

In this case, all but the $a$ components are bounded in time. Finally the DT gauge implies that:

$$a = 0, \quad b = 0, \quad c = 0, \quad a + 2h = 0.$$ 

Thus in the DT gauge, only the two good components are nontrivial. The solution $\gamma$ in the matrix notation is of the form:

$$\begin{bmatrix}
0 & 0 & 0 \\
0 & g & e \\
0 & e & -g
\end{bmatrix} = 0,$$

where both $g$ and $e$ oscillates at angular frequency $|k|$. In the physical literature, $g$ is called the $+$ polarization while $e$ is called $\times$ polarization of the gravitational wave.
Chapter 6

Two failure modes of space-time Regge Calculus

In 1961, Regge [95] proposed Regge Calculus as a space-time geometric discretization of the Einstein field equation. Only a decade later, Sorkin [101], proposed the first Regge Calculus-based scheme to solve the initial-value problem for the Einstein equation. Sorkin’s scheme was further developed and modified by physicists ever since. Two comprehensive review papers are [39, 115]. As of today, all these methods bear a very similar structure: the 4D discrete Regge-Einstein equation is used to form a marching scheme on a space-time mesh. We will refer to these scheme as Regge-Sorkin schemes. However, it is known that Regge-Sorkin schemes are unstable, see for example [91, Section 3.4].

In this chapter, we illustrate some essential features of Regge-Sorkin scheme using simpler model problems and replicate the observed known failure modes. The goals are two. First, we want to explain why this method fails. Second, given the known positive results in the mathematical literature regarding the Riemannian Regge Calculus [24, 28], that is Regge Calculus for the spatial part only, we propose that a (1+3) finite element approach has a high chance of success. Indeed, solutions to both failure modes mentioned here are well-understood for similar problems in the numerical relativity and finite element literature.

This chapter has two sections, one on failure due to the infinite dimension kernel and the other on failure due to the space-time scheme for the second-order time derivative. Both sections are organized in the same way. First, we introduce the continuous model problem and show how it is related to the Einstein equation. Second, we prove the continuous well-posedness of the model problem. We then introduce a seemingly reasonable discretization in some aspect resembling the Regge-Sorkin scheme. After that, we show through numerical
examples that these schemes fail in a way similar to how Regge-Sorkin scheme fails. Finally, we analyze how the failures happen mathematically, argue why Regge-Sorkin scheme has the same problem, and list well-known ways to fix these problems in the literature.

The conclusion of this chapter is, that the good way to solve the initial-value problem for the Einstein field equation should:

1. use a (1 + 3) approach to separate space and time,
2. regularize the evolution equation so that even constraint violating solutions are guaranteed to be bounded in time,
3. use a method of lines approach to discretize the regularized evolution equation,
4. use generalized Regge finite elements to discretize the spatial part of the metric where ein is the main operator.

To achieve these goals is beyond the scope of this thesis. However, this serves as the motivation for the next chapter, where we at some first steps towards this direction.

6.1 Failure due to the infinite dimensional kernel

The model problem of this section is the Maxwell wave equation. To start, recall the Maxwell equations in natural units:

\[
\begin{align*}
\text{div} E &= \rho, \\
\text{div} B &= 0, \\
\text{curl} E + B' &= 0, \\
\text{curl} B - E' &= j,
\end{align*}
\]

where \( E \) is the electric field, \( B \) is the magnetic field, \( \rho \) is the charge density, and \( j \) is the current. The right-hand side is required to satisfy the conservation law:

\[ \rho' + \text{div} j = 0. \]

We look at a simple case where \( \rho = 0 \). This implies \( \text{div} j = 0 \). Taking the time derivative of the fourth equation and eliminating \( B \) using the third equation, we get a constrained evolution system:

\[
\begin{align*}
\text{div} E &= 0, \\
E'' + \text{curl} \text{curl} E &= -j'.
\end{align*}
\]
This is the vector analog of the linearized Einstein equation from the previous chapter:

\[
\begin{align*}
\text{div} \gamma &= \text{div} \gamma' = 0, & \text{tr} \gamma &= \text{tr} \gamma' = 0, \\
S \gamma'' + 2 \text{ein} \gamma &= 0.
\end{align*}
\]

The key point of this analogy is that both curl curl and ein has an infinite dimensional kernel. So both evolution equations cannot ensure all components of the solutions to be bounded in time. The divergence-free or divergence-free trace-free constraints get rid of the part which grows in time. This way, constrained evolution systems exhibit the correct oscillatory behavior.

### 6.1.1 Well-posedness at continuous level

The Maxwell wave equation leads to our model problem. For simplicity, let the domain be the flat 3-torus \( \mathbb{T}^3 \), that is, a cube of side length \( 2\pi \) with periodic boundary condition. Given smooth vector fields \( a, b : \mathbb{T}^3 \to \mathbb{R}^3 \) and \( f : [0, T] \times \mathbb{T}^3 \to \mathbb{R}^3 \) satisfying the compatibility conditions:

\[
\begin{align*}
\int_{\mathbb{T}^3} a &= \int_{\mathbb{T}^3} b = 0, & \int_{\mathbb{T}^3} f(t) &= 0, \forall t \in [0, T], \\
\text{div} a &= \text{div} b = 0, & \text{div} f &= 0, \forall t \in [0, T],
\end{align*}
\]

find \( u : [0, T] \times \mathbb{T}^3 \to \mathbb{R}^3 \) such that \( u(0) = a, u'(0) = b, \) and

\[ u'' + \text{curl curl} u = f. \tag{6.1} \]

The zero spatial average conditions get rid of global constant functions on the flat torus, which ensures the well-posedness of the problem. This will be explained further in a later part of this section.

This evolution problem solves the Maxwell problem because equation (6.1) propagates the divergence-free constraint:

**Theorem 6.1.** Let \( u \) be any smooth solutions to (6.1) with compatible data. Then \( u \) is divergence-free and has zero average for all time.

**Proof.** Taking the divergence of the evolution equation, we get

\[ \text{div} u'' = 0. \]

By assumption \( \text{div} u(0) = \text{div} u'(0) = 0 \). Hence any solution \( u \) also satisfies \( \text{div} u \equiv 0 \). Integrating the evolution equation on \( \mathbb{T}^3 \), by Stokes’ theorem, we get

\[ \frac{d^2}{dt^2} \int_{\mathbb{T}^3} u = 0. \]
Again by assumption $u(0)$ and $u'(0)$ have zero averages, hence any solution $u$ also has zero average for all $t > 0$.

The evolution problem (6.1) with compatible data is related to the component-wise wave equation. Recall the vector identity:

$$\text{curl}\,\text{curl} - \nabla\text{div} = -\Delta,$$

where $\Delta$ is the component-wise Laplace operator. Thus the solution $u$ in the above theorem will also satisfy the component-wise wave equation:

$$u'' - \Delta u = f.$$  

This works in the reverse direction as well. Suppose $v$ is any smooth solution to the wave equation (6.3) with $v(0) = a$, $v'(0) = b$, and right-hand side $f$, where $a, b, f$ are compatible. Then it is clear that the divergence and the average of $v$ satisfy the homogeneous scalar wave equation with zero initial data. Hence $v$ is divergence-free with zero average for all time. By the same vector identity (6.2), this $v$ also solves the curl-curl wave equation (6.1). It is well-known that the wave equation is well-posed on $\mathbb{T}^3$ and has a unique smooth solution given smooth data (we proved this in the last chapter via Fourier analysis). We therefore proved the following theorem:

**Theorem 6.2.** The curl-curl wave equation (6.1) with compatible data is well-posed.

To better understand our model problem, we need the Hodge decomposition proved in the previous chapter: a smooth vector field on $\mathbb{T}^3$ can be decomposed into a sum of a gradient of a scalar field, the curl of another vector field, and a harmonic vector field:

$$C^\infty \otimes \mathbb{R}^3 = \text{curl}(C^\infty \otimes \mathbb{R}^3) \oplus \nabla(C^\infty) \oplus \mathbb{R}^3,$$

where the three components are orthogonal under the Euclidean inner product and the harmonic form part $\mathbb{R}^3$ consists of vector-valued global constant functions on $\mathbb{T}^3$. The first component is divergence-free. The last two components form the kernel of curl. The structure of equation (6.1) with compatible data becomes clear: the evolution equation is linear and operates on each component independently. On the curl part, the evolution equation is equivalent to the wave equation, due to vector identity (6.2). On the gradient part, the equation is equivalent to the ordinary differential equation (ODE):

$$w'' = 0, \quad w(0) = w'(0) = 0,$$

142
which is trivially solvable by $w = 0$. On the harmonic form part, the equation is trivial $0 = 0$. This also explains the zero average conditions we required: on the initial data, the zero average condition ensures the uniqueness of the solution, while on the right-hand side, it ensures the existence of a solution. The analogy between the structure of this problem and the linearized Einstein equation at the end of last chapter is very clear.

### 6.1.2 Discretization

In this section, we directly discretize the curl-curl evolution equation (6.1) using the standard method of lines.

The spatial part is discretized using the Nédélec edge elements [84], which is known to be a good spatial discretization for problems involving the curl operator [8, 52, 81]. Let $\text{NED}^1$ be the finite element space of Nédélec edge elements of degree 1 on a uniform mesh of $\mathbb{T}^3$. We solve the problem: given $a_h, b_h \in \text{NED}^1$ and $f : [0, T] \times \mathbb{T}^3 \to \mathbb{R}^3$, find $u_h : [0, T] \to \text{NED}^1$ satisfying $u_h(0) = a_h, u'_h(0) = b_h$, and

$$ (u''_h, w) + (\text{curl} u_h, \text{curl} w) = (f, w), \quad \forall w \in \text{NED}^1, \quad \text{eq: curl-curl_semidiscrete}$$

where $(\cdot, \cdot)$ denotes the $L^2$-inner product.

Then the temporal part is discretized using the Crank-Nicolson scheme [30], which is an implicit time stepping scheme known to be unconditionally stable and second-order accurate. For this problem, we introduce an auxiliary variable $v = u'$ and rewrite the semi-discrete equation as: $u_h(0) = a_h, v_h(0) = b_h$,

$$ (u'_h, y) - (v_h, y) = 0, \quad \forall y \in \text{NED}^1, $$
$$ (v'_h, w) + (\text{curl} u_h, \text{curl} w) = (f, w), \quad \forall w \in \text{NED}^1. $$

Finally, the Crank-Nicolson scheme is applied to this system: for time step size $k$,

$$ \begin{align*}
\left( u'^{n+1}_h - u'_h \right)_k &= (u_{h, k}^{n+1}, y) - (u_{h, k}^{n}, y) = 0, \quad \forall y \in \text{NED}^1, \\
\left( v'^{n+1}_h - v'_h \right)_k &= (\text{curl} u_{h, k}^{n+1}, w) + (\text{curl} u_{h, k}^{n}, w) = (f(n+1)k, w) + (f(nk), w), \quad \forall w \in \text{NED}^1,
\end{align*} $$

with initial data $u_0^h = a_h$ and $v_0^h = b_h$.

This fully discretized system can be solved using a Schur complement approach. First, we rewrite it in the matrix notation. We use the capital letters like $U$ to denote the coefficient vector corresponding to the finite element function $u$ in the basis representation. Let $M$ be the mass matrix and $A$ the stiffness matrix, that is,

$$ (u, v) = V^T M U, \quad (\text{curl} u, \text{curl} v) = V^T A U. $$

143
Define vectors $F^n$ via the identity:

$$U^T F^n = (f(nk), u).$$

The fully discrete system becomes:

$$\frac{U^{n+1} - U^n}{k} - \frac{1}{2}(V^{n+1} + V^n) = 0,$$

$$\frac{V^{n+1} - V^n}{k} + \frac{1}{2}A(U^{n+1} + U^n) = \frac{1}{2}(F^{n+1} + F^n).$$

Solve for $V^{n+1}$ in the first equation gives:

$$V^{n+1} = \frac{2}{k}(U^{n+1} - U^n) - V^n.$$

Substitute this into the second equation to eliminate $V^{n+1}$:

$$\frac{2}{k}(U^{n+1} - U^n) - V^n = \frac{1}{2}(F^{n+1} + F^n).$$

Notice that the above equation has only values known at step $n$ on the right-hand side. So at each time step, we solve the above equation for $U^{n+1}$. Then we use it to evaluate $V^{n+1}$ directly using the previous equation. When $k$ is substantially smaller than the mesh size, the matrix $(4 + k^2 A)$ is a small perturbation of four times the identity matrix. In this case, the $U^{n+1}$ equation can be solved efficiently using the algebraic multigrid method.

### 6.1.3 Numerical examples and discussion

For all numerical examples of this section, we use a travelling wave on $\mathbb{T}$ as the exact solution:

$$u = \begin{bmatrix} \sin(z-t) \\ \sin(x-t) \\ \sin(y-t) \end{bmatrix}.$$  

It is clear that $u$ is divergence-free and has zero average. Further, it satisfies the curl-curl wave equation (6.1) with right-hand side $f = 0$.

The fully discretized solver described in the previous section was implemented in FEniCS. All the source code can be found under the `curlcurl_wave_equation` directory in the companion repository of the thesis.

The numerical examples in this section are all carried out on an $8 \times 8 \times 8$ uniform mesh of $\mathbb{T}^3$. The solution is computed up $T = 200$ with a time step size $k = 0.1$.

For the first numerical experiment, we interpolate the initial data into the finite element space and run the solver. Figure 6.1 shows a plot of the numerical solution and the exact solution at the point $(0,0,0)$. 

144
It is clear that this method does not work. Though the oscillatory behavior was correctly captured, there is a linear growing trend in the center of the oscillation. The exactly same behavior was observed in the numerical experiments using Sorkin-style Regge Calculus for a linear wave solution (see Figure 3.33 and Figure 3.34 of [91]). The reason for this is easy to explain. It is known [6, 8] that our discrete space Nédélec also has a discrete Hodge decomposition:

\[ \text{NED}^1 = \nabla \text{CG}^1 \oplus \mathbb{R}^3 \oplus V_h, \]

where CG$^1$ is the space of Lagrange elements of degree 1 on the same mesh and $V_h$ is a subspace of NED$^1$ which is orthogonal to the first two components under the Euclidean inner product. The structure of the semi-discrete equation (6.4) is the same as the continuous equation (6.1). Again, the three components above evolve separately in time. In particular, on the $\nabla \text{CG}^1$ part, the equation is just an ODE: $w'' = 0$. On the harmonic form part $\mathbb{R}^3$ the equation trivially holds. On the $V_h$ part, the equation is a well-posed semi-discrete hyperbolic equation, which is second-order in both space and time. The problem here is that when we interpolate a divergence-free function with zero average into NED$^1$, its interpolant is not entirely in $V_h$. Figure 6.2 shows growth the norm of the $\nabla \text{CG}^1$ and $\mathbb{R}^3$ component of the numerical solution in time. In particular, the $\nabla \text{CG}^1$ part grows linearly as it should for the ODE $w'' = 0$ with non-zero initial data. This leads to the linear trend we observe and causes
a fast loss of accuracy.

![Figure 6.2: Plot of the growth of the gradient and harmonic part of the numerical solution](fig:curl_linear_org_vio)

One way to deal with this problem is to project the initial data into the $V_h$ space. This can be done in the following way. First, solve an auxiliary problem: given $u \in \text{NED}^1$, find $\phi \in \text{CG}^1$ satisfying:

$$(\nabla \phi, \nabla \psi) = (u, \nabla \psi), \quad \forall \psi \in \text{CG}^1.$$  

Then set:

$$Q u := u - \nabla \phi - \int_{T^3} u.$$  

It is clear that $Q : \text{NED}^1 \to V_h$. After we interpolate the initial data into $\text{NED}^1$, we further use $Q$ to project the discrete initial data into the $V_h$ space. If this projected initial data is used, the method has much better accuracy. Figure 6.3 again shows the plot of the value of the exact and numerical solution at $(0,0,0)$ and Figure 6.4 shows the plot of the growth of the norms of the $\nabla(\text{CG}^1)$ and $\mathbb{R}^3$ component of the solution. Two observations are made. First, the amplitude of the numerical solution is smaller. This is because the projection removed part of the energy in the oscillation. Second, the non-$V_h$ part of the solution still grows linearly but just with a much smaller initial data. The correct solution, the $V_h$ part, however has constant amplitude. So eventually the solution will still be dominated by the bad part.
This problem becomes much more severe for nonlinear problems. Next, we look at a
nonlinear perturbed problem:

\[ u'' + \text{curl}(1 + \epsilon u \times \text{curl}) = f, \]

where \( \epsilon \) is a small positive number. The form of the perturbation here is very similar to the nonlinear Einstein equation. With minimal modifications to the discretization and numerical scheme explained before, we can solve this perturbed problem with the same exact solution and mesh. First we use the interpolant of the initial data directly and solve the nonlinear perturbed problem with \( \epsilon = 0.1 \). Figure 6.5 shows the plot of the value of the exact and numerical solution at \((0,0,0)\) and Figure 6.6 shows the plot of the growth of the norms of the \( \nabla (CG^1) \) and \( \mathbb{R}^3 \) component of the solution. The numerical solution blows up at time \( t = 71.3 \). This is because the linear drift term moved the solution off the stability regime of this nonlinear equation. This should be compared to Figure 3.38 of [91] showing the blow-up of the Sorkin-style space-time Regge calculus.

![Figure 6.5: Plot of numerical and exact solution for the nonlinear problem](fig:curl_nonlinear_orig_sol)
For the nonlinear problem, the projection of the initial data is no longer sufficient. With the projected initial data, Figure 6.7 shows the plot of the value of the exact and numerical solution at (0,0,0) and Figure 6.8 shows the plot of the growth of the norms of the $\nabla(\text{CG}^1)$ and $\mathbb{R}^3$ component of the solution. The numerical solution blows up at a slightly later time $t = 100.7$. This time it is the growth of the harmonic part that drives the solution off its stability regime.
Figure 6.7: Plot of numerical and exact solution with projected initial data

Figure 6.8: Plot of the growth of the gradient and harmonic part of the numerical solution with projected initial data
6.1.4 Implications for Regge Calculus

Regge Calculus, directly applied to the Einstein field equation where the metric is a small perturbation of the Minkowski metric, is very similar to the situation here. As already shown by numerical experiments in the literature [91], the behavior is indeed very similar. This shows that due to the infinite dimensional kernel of the Einstein tensor, Regge Calculus is not a viable numerical method.

If, however, a method-of-lines approach is used, there is a chance that Regge calculus can be salvaged. Indeed, there are two well-established ways to deal with this problem. Both involve regularizing the evolution itself. One approach is Chorin’s projection method [26]. This basically means that we apply the projection operator $Q$ at each time step of the evolution. This is an expensive method because an elliptic equation has to be solved at each time step. Another approach is to regularize the evolution equation. For example, the curl-curl wave equation can be regularized as:

\[
\begin{align*}
\sigma' &= -\text{div} u, \\
u' &= v, \\
v' &= -\text{curl}\text{curl} u + \nabla \sigma + f.
\end{align*}
\]

Intuitively, by taking the time derivative of the last equation and substituting in the previous two equations, we get a full component-wise wave equation for $v$. This way the evolution equation itself has control over all components of the solution and exhibit the correct oscillatory behavior, even without any constraints. At the discrete level, constraint violating components oscillates at a small amplitude and does not significantly pollute the solution even for large time. Both methods are well-known and used in the numerical relativity literature using the (1 + 3) decomposition approach. The same should happen for Regge calculus as well.

It should be stressed here that in this example it is the continuous curl-curl wave equation itself that is bad and not suitable for direct discretization, no matter which discretization method is used. It is the evolution equation itself that needs regularization.

6.2 Failure due to the space-time scheme for the second-order time derivative

The model problem of this section is the scalar wave equation:

\[
u'' - \Delta u = 0.
\]
This can be written equivalently in the space-time form:

\[ \Box u = \text{div}(\eta \nabla u) = 0, \]

where \( \Box \) is the d'Lambertian, \( \eta \) is the 4D Minkowski metric, \( u \) is interpreted as scalar fields on the space-time. This is an simpler space-time model problem for the 4D space-time linearized Einstein equation:

\[ \text{ein} g = 0. \]

As a companion to his space-time Regge Calculus paper [101], Sorkin [100] also proposed methods to discretize matter fields in a way that is compatible with the Regge Calculus discretization of the Einstein equation. In particular, a space-time discretization of the scalar wave equation using essentially the Lagrange finite elements was proposed as an analog of the space-time Regge Calculus scheme. In this section, we show that this method also fails, albeit in a subtle way.

The space-time scalar wave equation is a simpler problem than the Einstein equation because there is no infinite dimensional kernel in the previous section in this case. It will be argued that the Sorkin-style space-time still fails due to the way the second-order time derivative is discretized. Hence even with arbitrarily high precision floating point arithmetics along with discrete initial data somehow perfectly lie in the correct space for the linearized Einstein equation, Regge calculus will still fail. In that case, the discrete equations are equivalent to the component-wise wave equation with a discretization of the time derivative similar to that in the discrete space-time scalar wave equation here.

Since the point here is that the discretization of the time derivative is bad, for the ease of discussion and visualization, the scalar wave equation in \((1+1)\) dimensions will be used. The Sorkin-style space-time discretization of the scalar wave equation fails in \((1+n)\) dimensions for all \(n \geq 1\). It works only for the uninteresting case \(n = 0\), where the equation is an ordinary differential equation (ODE) and the corresponding method is the well-known finite element in time method for second-order ODEs.

### 6.2.1 Regge-calculus style derivation of the model problem

We derive the space-time discretization using a Regge calculus-style approach. Let \( \Omega = [0,1] \) be our spatial domain and \( T > 0 \) some positive real number. The space-time is \([0,T] \times \Omega\).

For scalar fields \( u \) and \( v \) on the space-time, we use single parenthesis for the spatial \(L^2\)-inner product and double parenthesis for the space-time \(L^2\)-inner product:

\[
(u,v) := \int_{\Omega} uv, \quad ((u,v)) := \int_{[0,T] \times \Omega} uv.
\]
The action $S$ for a scalar field $u$ on the space-time is given by:

$$S(u) := \int_{[0,T] \times \Omega} \eta_{\alpha \mu} (\partial^\mu u)(\partial^\alpha u) = \frac{1}{2} [-(u', u') + ((\nabla u, \nabla u))].$$  

We derive its equation of motion by taking the first variation:

$$-(u', v') + ((\nabla u, \nabla v)) = 0,$$

for any scalar field $v$ on the space-time. For test functions $v$, integrate by parts, we indeed get the scalar wave equation:

$$u'' - \Delta u = 0.$$  

Again, to formulate an initial-value problem from this, we need to specify more information.

### 6.2.2 Initial-value problem and well-posedness

Our model problem is the initial-boundary value problem for the scalar wave equation: given a scalar field $\alpha$ on $\Omega$, find $u$ on $[0,T] \times \Omega$ satisfying $u(0) = 0$, $u'(0) = \alpha$, and

$$u'' - \Delta u = 0, \quad u = 0,$$

in $[0,T] \times \Omega$, on $[0,T] \times \partial \Omega$.

The initial data $u(0) = 0$ is set merely for simplicity. The homogeneous spatial boundary conditions are used here to further exclude the possibility that the harmonic forms pollute the solution as in the $T^3$ case studied previously.

To properly formulate a continuous problem, we will need the Hilbert space framework. Space-time function spaces like $L^2([0,T], H^1(\Omega))$ are abbreviated as $L^2H^1$. Let

$$X := H^1L^2 \cap L^2\hat{H}^1, \quad X^0 = \{ u \in X | u(0) = 0 \}, \quad X^T = \{ u \in X | u(T) = 0 \},$$

where $\hat{H}^1$ is the space of functions which are in spatial $H^1$ and vanish on the spatial boundary.

From the form of the variation of the action (6.6), we define a bilinear $B : X^0 \times X^T \to 0$ by

$$B(u, v) := -((u', v')) + ((\nabla u, \nabla v)).$$

The continuous weak form of the scalar wave equation is: given $a \in H^1$, find $u \in X^0$ such that

$$B(u, v) = (b, v(0)), \quad \forall v \in X^T.$$  

**Theorem 6.3.** Problem (6.7) is well-posed. Further the solution $u$ satisfies the scalar wave equation as a distribution.
Proof. Restrict \( v \) to test functions. The equation implies that \( u \) satisfies the scalar wave equation as a distribution. Testing against \( v \in X^T \) further shows that \( u'' \in L^2H^{-1} \). The well-posedness of the wave equation in this case can be established using standard semi-group approaches, see for example [21, Section 2.6.4].

6.2.3 Regge calculus-like space-time discretization and finite element view

In this section, we first derive Regge calculus-like space-time discretization of our model problem. Then we interpret this discretization as a finite element method.

Let \( \mathcal{T} \) be a uniform mesh of \([0,T] \times \Omega \) with some temporal mesh size \( k \) and spatial mesh size \( h \). A discrete scalar field is a continuous piecewise linear function on \( \mathcal{T} \), parameterized by its values at the nodal points. We write down the discrete action, which is the same as the continuous one (6.6): for a discrete scalar field \( u \),

\[
S_h(u) := \int_{[0,T] \times \Omega} \eta_{\alpha\mu}(\partial^{\alpha}u)(\partial_{\mu}^0 u) = \frac{1}{2}[-(u',u') + ((\nabla u, \nabla u))].
\]

We derive its equation of motion by taking the first variation, which again leads to the same equations

\[-((u',v')) + ((\nabla u, \nabla v)) = 0\]

for any discrete scalar field \( v \) on \( \mathcal{T} \). This is the Regge calculus-style discrete wave equation.

This is very natural from the finite element point of view. Let \( \text{CG}^1 \) be the space of Lagrange elements of degree 1 on \( \mathcal{T} \). It is well-known that \( \text{CG}^1 \subset H^1H^1 \). Define discrete subspaces with temporal and spatial boundary conditions:

\[
X_h := \text{CG}^1 \cap L^2H^1, \quad X_h^0 = \{u \in X_h | u(0) = 0\}, \quad X_h^T = \{u \in X_h | u(T) = 0\},
\]

Clearly,

\[
X_h \subset X, \quad X_h^0 \subset X^0, \quad X_h^T \subset X^T.
\]

We thus get a conforming discretization of equation (6.7) via the Galerkin projection: find \( u \in X_h^0 \) such that

\[
B(u,v) = (b,v(0)), \quad \forall v \in X_h^T.
\]

This is straightforward to implement. The resulting linear system can actually be solved locally. This has a marching structure very similar to Sorkin-style space-time Regge calculus. It can be best described through Figure 6.9. There the values at the purple nodes are already known from the spatial boundary condition. The first two layers of solid green nodes are known from the initial data. First, we take the test function centered at the blue node as the test function. Its support is marked by the thickened lines. Out of the 7 nodes in the support
of this test function, the value of the solution at only one node, at the dashed circle, is not yet known. The discrete equation (6.8) can be applied here to solve for the value there. Once this is done, we can choose the tent function centered at the node immediately to the left of the blue node. The situation is the same and we can use equation (6.8) again to fill in one more value in the third temporal slice. Repeating this, we fill the whole third temporal slice. The situation will look exactly the same as we started but with three layers of green nodes. We can thus repeat this and fill in the entire mesh to obtain the solution.

![Illustration of the marching scheme](fig:marching)

This scheme can be highly parallelized. Figure 6.10 shows that after the computing the value at the second node in the third layer, we can already start to fill the fourth layer, at the same time the third layer is filled.

![Parallel marching scheme](fig:parallel)

### 6.3 Numerical example

The behavior of the discrete problem (6.8) is subtle. In this section, we show some intriguing numerical examples. In the next section, we give a full explanation and analysis.
The discrete problem (6.8) is implemented in FEniCS. All the source code can be found under the `spacetime_wave_equation` directory in the companion repository of the thesis.

We choose the following standing wave as the exact solution:

\[ u(t,x) = \sin(3\pi x)\sin(3\pi t). \]

It is clear that \( u \in X^0 \). The right-hand side is just 0. We always compute to \( T = 15.0 \).

First, we take a uniform mesh with spatial size \( h = 0.02 \) and temporal size \( k = 0.01 \). The gives a Courant-Friedrichs-Lewy (CFL) number of 0.5, far from the critical value 1. Figure 6.11 shows the result. The upper panel shows a zoomed in view of the mesh. The middle panel plots the numerical solution as a heat map. The bottom panel plots the spatial \( L^2 \)-norm in time. In this case, this method works and the numerical solution is very close to the exact solution.

![Uniform mesh](fig:cg_uniform)

Now, we take the same uniform mesh as before, then randomly perturbed each internal mesh points by at most 14% of the radius of inscribed circle of a triangle in the original mesh. Figure 6.12 shows the result in the same format as before. This is a fairly small perturbation. Yet, it is clear that the method is unstable.
We then take the good uniform mesh in the first time and move all the internal nodes in every second spatial slices by $0.2h$ together (deterministically). Figure 6.13 shows the result in the same format as before. The method seems stable.
Finally, we do the same as the previous experiment, but move all the internal nodes in every third spatial slices by $0.2h$ together (deterministically). Figure 6.13 shows the result in the same format as before. The method is extremely unstable and blows up quickly (notice the scale of the $y$-axis on the bottom panel).

![Figure 6.13: Perturb every third spatial slice](fig:cg_three)

### 6.3.1 von Neumann stability analysis

In this section, we explain the reason behind the previous numerical experiments.

As mentioned previously, the discretization (6.8) can be understood as a marching scheme. Set $U_n^m$ to be the value of the numerical solution at the $n$-th node in space and $m$-th node in time. We thus only need to understand the situation for the one patch:

![Diagram of one patch](image)
where all but $U_{n}^{m+1}$ are known. We take the tent function centered at $U_{n}^{m}$ as the test function.

On a uniform mesh with temporal size $k$ and spatial size $h$, the equation (6.8) can be assembled by hand. This is a tedious computation. The result is:

$$\frac{U_{n}^{m+1} - 2U_{n}^{m} + U_{n}^{m-1}}{k^2} - \frac{U_{n-1}^{m} - 2U_{n}^{m} + U_{n+1}^{m}}{h^2} = 0.$$  \tag{6.9}$$

Thus on a uniform grid, this method is exactly the central finite difference method for the scalar wave equation. It is stable as long as the CFL condition is satisfied:

$$k/h \leq 1.$$  \tag{6.10}$$

This explains why the method works nicely on a uniform mesh. Notice that on the uniform mesh, all though the two diagonal nodes $U_{n-1}^{m-1}$ and $U_{n+1}^{m+1}$ are in the support of the test function, they do not enter the final equation due to cancellations by symmetry. This leaves us with the classical 5-point stencil.

When the mesh is any perturbation of the uniform mesh, this is no longer the case. In particular, both $U_{n-1}^{m-1}$ and $U_{n+1}^{m+1}$ enters the equation. Since the mesh is a small perturbation from the uniform mesh, we can still use the numbering as before. We can analyze this method using standard von Neumann stability analysis [69, Section 9.6]. The idea is simple: we check the behavior of the method by marching a discrete function of the form

$$U_{n}^{m} = r^{m} e^{i\theta}.$$  

For plug the above formula into the marching equation similar to equation (6.9) and then solve for the magnification factor $r$ as a function of $\theta$. Our method is stable if $|r| \leq 1$ for all $\theta$.

Now the hand assembly approach becomes quite unwieldy. The marching equation similar to equation (6.9) can be evaluated numerically. The relevant code can be found in the python notebook `spacetime_wave_equation/analysis.ipynb`. The results are summarized here.

Because the patch involves three time levels, we have $r^{-1}, 1, r$ in the equation for the magnification factor. This means that $r$ is a quadratic polynomial in $e^{i\theta}$. In Figure 6.15, we show the situation on the uniform mesh when $k = 0.5$ and $h = 1.0$. The left panel shows the weight for each node in the marching equation. The right panel shows the image of $g(\theta)$ for $\theta \in [0, 2\pi]$. The orange and blue colors stand for the two different roots $g_1, g_2$ of the quadratic equation for $g$. In this case, it is already at the boundary of stability because the right-most part of the curve touches 1.
Figure 6.15: Uniform mesh \( k = 0.5 \) and \( h = 1.0 \)

Figure 6.16 shows the same information on the uniform mesh when \( k = 1.0 \) and \( h = 1.0 \). It is clear that it is still stable.

Figure 6.16: Uniform mesh \( k = 1.0 \) and \( h = 1.0 \)

Figure 6.16 shows the situation on the uniform mesh when \( k = 1.0 \) and \( h = 0.5 \). In this case the CFL condition fails. It is clear from the magnification factor plot that it is unstable.
In the next few figures, we will look at different types of perturbations of the good uniform mesh $k = 0.5$ and $h = 1.0$. First, Figure 6.18 shows the situation where the very middle node is perturbed by a small amount in the spatial direction. This is not stable.

Figure 6.18 shows the situation where the very middle node is perturbed by a large amount in the spatial direction.

Figure 6.18 shows the situation where the very middle node is perturbed by a large amount in the spatial direction.
amount in the temporal direction. This, however, is still stable.

Figure 6.19: Perturb the middle node temporally

Figure 6.20 shows the situation where the very middle node is perturbed a small amount in both the spatial and temporal direction. It turned out that it is stable when $\Delta x \leq \Delta t$ for the perturbation. This includes the previous two situations as special cases.

Figure 6.20: General perturbation of the middle node

Now we look at a different type of perturbation. Figure 6.21 shows the situation where all
the middle nodes for the same spatial position are moved by a small amount in the temporal
direction. This is unstable.

Figure 6.21: Move all middle nodes temporally

Figure 6.22 shows the situation where all the middle nodes for the same spatial position
are moved by a small amount in the spatial direction. This is also unstable.

Figure 6.22: Move all middle nodes spatially

Figure 6.23 shows the situation where all the nodes on the middle spatial slice are moved
by a small amount in the spatial direction. This is again unstable.

Figure 6.23: Move all middle nodes spatially

163
In sum, the stability of the discretization (6.8) depends on how the mesh is perturbed but not how big the perturbation is. In particular, there are many ways we can perturb a uniform mesh very far away from the CFL limit such that the method is unstable for arbitrarily small amount of perturbation.

Finally, we explain why perturbing every other spatial slice does not blow up but perturbing every third spatial slice does. The reason becomes clear in Figure 6.24. The patch involves three spatial slices. If we perturb every other slice, the perturbation at one spatial slice is exactly the opposite of that on the next slice. Hence the effect cancels each other. However, there is no such cancellation when we perturb every third spatial slice. In this case, the perturbation is of the type depicted in Figure 6.23 every three spatial slice. The method blows up exponentially because the magnification factor $|g| > 1$. 

Figure 6.23: Move all middle nodes spatially!
6.3.2 Implication for Regge calculus

We have shown that Sorkin-style space-time method for the scalar wave equation is unstable on general meshes. However, stable space-time finite element methods for hyperbolic equations abound, for example [38, 55, 56]. The main difference between these ones and the one here is how the second-order time-derivative is handled. The stable schemes discretize it as:

\[ ((u', v)) - ((p, v)) = 0, \quad \forall v \]
\[ ((p', q)) + A(u, q) = 0, \quad \forall q, \]

where \( A(u, q) \) is some bilinear form for the spatial part. The discrete spaces are constructed so that we can choose \( v = p' \) and \( q = u' \) as test functions. Adding the two equations together, we get

\[ ((p, p')) + A(u, u') = \frac{1}{2} [((p, p)) + A(u, u')]' = 0, \]

which is the natural energy estimate for this equation.

In Sorkin-style space-time methods, this is however formulated as:

\[ -((u', v')) + A(u, v) = 0, \quad \forall v \]

It is not clear if it is even possible to get an energy estimates by a choice of test function.

It is clear that in space-time Regge calculus, the situation is very similar to the unstable discretization (6.8). Because of this, the space-time aspect in the Sorkin's Regge Calculus scheme is very unlikely to work. We note that here again it is the form of the discrete equation that is bad. It matters little which finite element was used. In order to get out of this problem, we need to abandon the direct space-time approach.
Chapter 7

Conclusion

7.1 Summary and comments

7.2 Future works
Bibliography


168


173


174


