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

May 2018

Acknowledgements

I would like to express my sincere gratitude to my advisor Prof. Douglas Arnold, who taught me how to be a mathematician: diligence in thought and clarity in communication (I am still struggling with both). I am also grateful for his continuous guidance, help, support, and encouragement throughout my graduate study and the writing of this thesis.

Abstract

This thesis proposes a new family of finite elements, called generalized Regge finite elements, for discretizing symmetric matrix-valued functions and symmetric 2-tensor fields. We demonstrate its effectiveness for applications in computational geometry, mathematical physics, and solid mechanics. Generalized Regge finite elements are inspired by Tullio Regge's pioneering work on discretizing Einstein's theory of general relativity. We analyze why current discretization schemes based on Regge's original ideas fail and point out new directions which combine Regge's geometric insight with the successful framework of finite element analysis. In particular, we derive well-posed linear model problems from general relativity and propose discretizations based on generalized Regge finite elements. While the first part of the thesis generalizes Regge's initial proposal and enlarges its scope to many other applications outside relativity, the second part of this thesis represents the initial steps towards a stable structure-preserving discretization of the Einstein's field equation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

We describe a new family of finite elements called *generalized Regge finite elements*, for discretizing symmetric matrix-valued functions (coordinate formulation) and symmetric 2-tensor fields (coordinate-free formulation). We demonstrate its effectiveness for applications in computational geometry and solid mechanics. As the name suggests, this new finite element family is a generalization of the classical Regge finite element, which has its roots in a discretization of Einstein's theory of General Relativity called Regge Calculus. We claim that Regge Calculus fails as a numerical method and study why such failure occur. Further, we derive well-posed linear models problems from General Relativity and propose their discretizations based on generalized Regge finite elements.

1.1 From Regge Calculus to Regge finite elements: a brief introduction

The starting point is a mesh like the one in Figure 1.1.

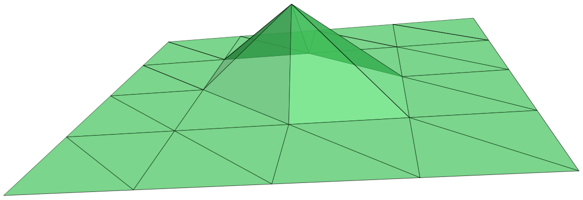


Figure 1.1: A triangulated surface

Apparently, part of this mesh is not flat, even though it is built from a finite number of flat triangles. This "non-flatness" can be quantified by looking at the sum of angles around each vertex. Around an apparently flat vertex, the sum of angles around it is exactly 2π, as shown in Figure 1.2.

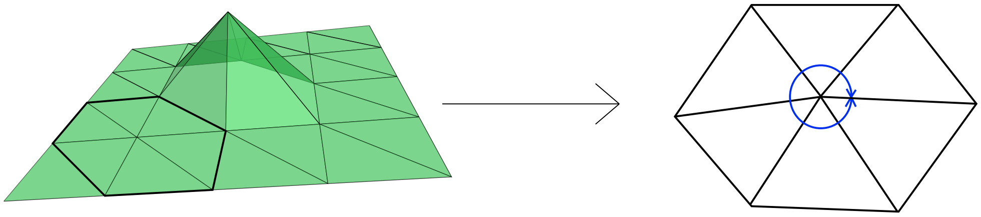


Figure 1.2: Flat: sum of angles equals 2π.

However, around an apparently nonflat vertex, the sum of angles around it is either greater or smaller than 2π, as shown in Figure 1.3.

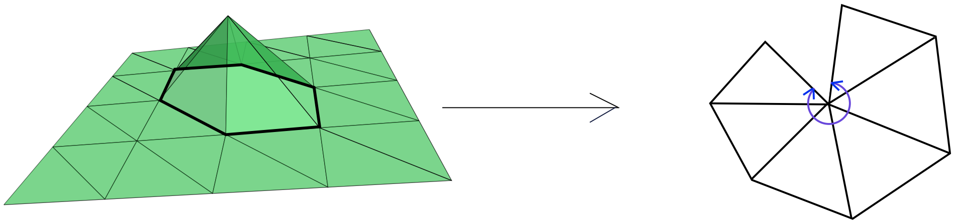


Figure 1.3: Nonflat: sum of angles does not equal 2π.

At each vertex, the difference between the sum of angles around it and 2π is called the *angle deficit* at that vertex. In 1961, Tullio Regge in his influential paper [96] generalized the notion of angle deficit to higher dimensions and linked it to the smooth notion of scalar curvature in differential geometry. He further derived a discrete formulation of Einstein's geometric theory for gravity based on simplicial meshes and angle deficits and speculated that such discrete models can be used on computers to approximate the notoriously difficult Einstein's field equation. Later literature referred to this discretization as *Regge Calculus*. In many ways, this started the field of studying geometry on the computer. We will see that Regge's discrete geometric model is a special case of the generalized Regge elements proposed in this thesis. In fact, generalized 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 1.1 is a polygon obtained by gluing together flat triangles along edges of the same lengths. Since flat simplices are determined up to isometry by their edge lengths, they can be equivalently described as a mesh 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 using Regge Calculus (for a recent review, see [61]) and quantum gravity (for a recent review, see [116]). 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 [24, 25]. The main observation was that specifying the lengths of all edges is equivalent to prescribing on the mesh 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 mesh, 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. With this view, Cheeger-Müller-Schrader further proved that certain curvatures, including Regge's discrete scalar curvature, on these non-smooth metrics converge to the their smooth counterparts in a subtle way, when a suitable sequence of Regge finite elements are used to approximate a smooth Riemannian metric. The analytic view still dominates in the current mathematics literature on discrete geometry.

This idea of "filling in" 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 [115]. 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) [8, 10], which generalizes these piecewise linear forms to higher polynomial degrees and studies their approximation properties and use in finite element methods in a Hilbert space framework. FEEC has been proven to be an effective framework for 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 [28,29]. It adds another layer of structure on top of the analytical one: Regge elements are not only just piecewise constant functions on a mesh, 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 a Hilbert spaces context than to show an assignment of numbers to edges somehow converges to a smooth solution to a partial differential equation.

The finite element view is the starting point of this thesis. First, we generalize Regge's initial proposal to piecewise polynomials of all degrees. Second, we apply the resulting finite elements to various applications other than numerical relativity. Third, we derive linear model problems from general relativity and propose discretizations based on generalized Regge elements. Fourth, we study the failure of the space-time Regge Calculus as a way to discretize Einstein's field equation and show that the methods proposed in this thesis does not suffer from the same problems.

1.2 Outline

The rest of the thesis can be roughly divided into 3 parts.

Part 1 consists of Chapter 2 alone. We give a precise definition of generalized Regge finite element family and prove many of its basic properties including unisolvency of the degrees of freedom and optimal approximation rates. We also give two implementable bases, one of which is used in the author's implementation of generalized Regge elements for 2D and 3D in the open-source Python finite element software FEniCS.

Part 2 consists of Chapter 3 and Chapter 4. Here we study applications of generalized Regge elements outside of numerical relativity, in computational geometry and solid mechanics.

In Chapter 3, we demonstrate that elements of generalized Regge finite elements, when interpreted as discrete non-smooth Riemannian metrics, retain many geometric properties of smooth Riemannian metrics. Further they can serve as effective discrete approximations to smooth Riemannian metrics. In particular, we propose and implement a robust algorithm for computing geodesics on these discrete metrics and analyze the error when the discrete metric is an approximation to some smooth metric.

In Chapter 4, we look at applications of generalized Regge finite elements in solid mechanics. In particular, we propose discretizations of the biharmonic equation and linear elasticity equation in all dimensions using generalized Regge elements and demonstrate their effectiveness via numerical examples. We will also note how these two applications are related to numerical relativity.

Part 3 consists of Chapter 5 and 6. We shift our focus to numerical relativity. We have two goals in mind. First, we give three strong reasons why Regge Calculus fails as a numerical method for solving Einstein's field equation in General Relativity. Second, we show that all these three issues have parallels in other contexts of finite element analysis. Further, these similar problems have been solved in the finite element literature in their corresponding domains. So one reasonable way moving forward is to adapt these known effective solutions to general relativity. We claim that this is possible through the use of generalized Regge finite elements for certain regularized system of differential equations derived from Einstein's field equation which does not treat space and time on the same footing. The full program is a huge undertaking and this thesis only serves as a starting point in that direction.

In Chapter 5, we look at basic problems in numerical relativity. For simplicity, we derive linearized problems which still capture features of the equation essential for its discretization. In particular, we see that the Einstein field equation, which Regge Calculus directly discretizes, needs regularization due to weak hyperbolicity. This weak hyperbolicity is well-known in numerical relativity literature, but it is not so far recognized as a fatal problem for Regge Calculus. We then propose linear models problems regularizing the linearized Einstein equation. Moreover, we prove that these problems are well-posed and propose discretizations based on generalized Regge elements.

In Chapter 6, we study the failure of Regge Calculus as a numerical method. In particular, we study two modes of failure in detail under the finite element framework, namely the infinite dimensional kernel of the curvature operator and the space-time scheme. We also discuss their parallels in the finite element literature and the corresponding solutions. This shows why the space-time unregularized approach should be abandoned in favor of regularized $(1+3)$ approach in Chapter 5.

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-by-n matrix-valued functions. In this chapter, its definition and various basic properties are studied in detail.

We use REG^r to denote the generalized Regge finite elements of degree r, which we will define for integer $r \geq 0$. To fix the ideas, we start with a directly implementable description of REG^r in 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 2.1. For a k-simplex f, let $\mathscr{P}^r(f)$ be the space of polynomials of degree r or less in k variables as functions on f. For $r < 0$, it is understood that $\mathscr{P}^r(f) = \{0\}$. For any line segment L in $\mathbb{R}^n$ and any symmetric matrix $u \in \mathbb{R}^{n \times n}$, define:

$$u_L := t^T u t$$

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 $\mathscr{P}^r(L)$. The degrees of freedom are integrals of u_L against $\mathscr{P}^r(L)$. 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.

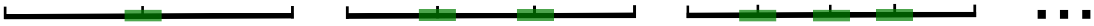


Figure 2.1: Degrees of freedom in 1D for $r = 0, 1, 2, \ldots$.

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^r is the same as the Discontinuous Lagrange elements on L.

In 2D, let T be a triangle in $\mathbb{R}^2$. The shape functions for $\mathrm{REG}^r(T)$ consist of symmetric 2-by-2 matrix-valued functions whose 3 components are in $\mathscr{P}^r(T)$. Let $\{E_1,E_2,E_3\}$ be the three edges of T. Since each E_i is a line segment, using the notation before, each u_{E_i} is a well-defined scalar-valued function on the entire triangle. The degrees of freedom are: for any symmetric 2-by-2 matrix-valued function u,

$$\begin{cases} \text{1D degrees of freedom on the restriction of } u \text{ to each } E_i, \\ \text{for } i = 1,2,3, \text{ integral of } u_{E_i} \text{ against } \mathscr{P}^{r-1}(T) \text{ on } T. \end{cases}$$

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 $r = 1$.

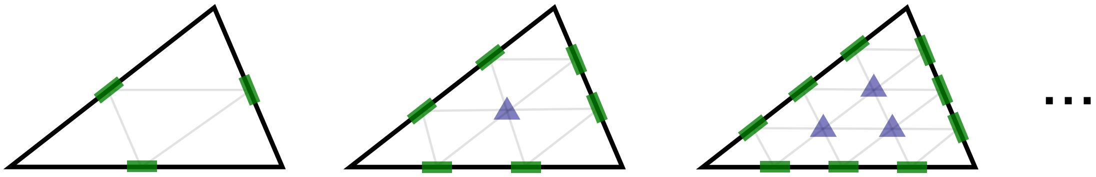


Figure 2.2: Degrees of freedom in 2D for $r = 0,1,2,\ldots$.

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

$$\begin{cases} \text{1D degrees of freedom on the restriction of } u \text{ to each } E_i, \\ \text{2D degrees of freedom on the restriction of } u \text{ to each } T_j, \\ \text{for } i = 1,\ldots,6, \text{ integral of } u_{E_i} \text{ against } \mathscr{P}^{r-2}(H) \text{ on } H. \end{cases}$$

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.

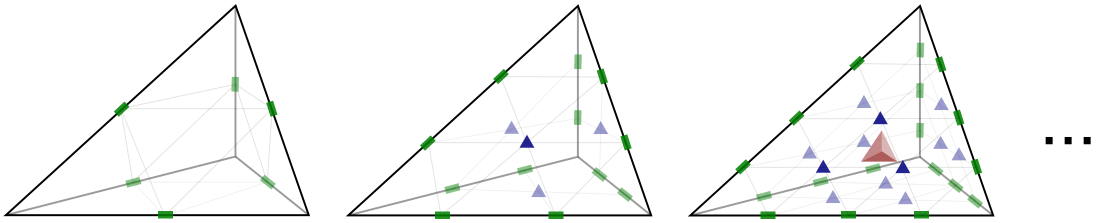


Figure 2.3: Degrees of freedom in 3D for $r = 0, 1, 2, \ldots$.

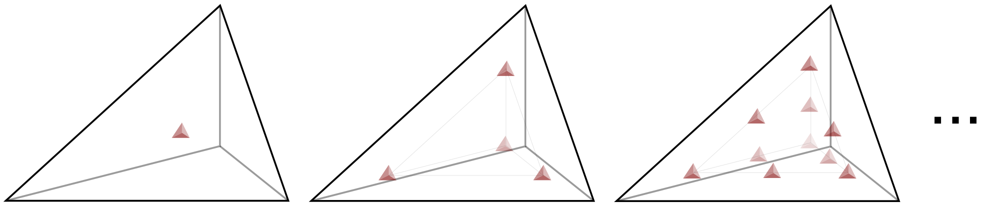


Figure 2.4: Interior degrees of freedom in 3D for $r = 2, 3, 4, \ldots$.

The general pattern for REG^r in dimension $n \geq 4$ is clear. A detailed description of this set of degrees of freedom can be found in Section 2.4.

The space REG^r unifies and generalizes several discrete structures previously known in a wide variety of fields. The lowest degree element REG^0 in all dimensions $n \geq 1$, called *Regge finite elements*, are well-known in the relativity, geometry, and finite element literature [24, 25, 28, 29, 96]. In dimension $n = 2$, one can consistently rotate all the edge tangent vectors to normal vectors of the triangle. Under this, 2D REG^r becomes the well-known Hellan-Herrmann-Johnson finite element [13, 20] for the biharmonic equation. In 2D, REG^1 is also equivalent to Pechstein-Schöberl's lowest degree normal-normal stress finite element [89, 100, 101] for the linear elasticity equation. These connections will be further studied and generalized later in the chapter on applications to solid mechanics.

We highlight several fundamental results concerning properties of REG^r proved in this chapter:

- (Theorem 2.1) The set of degrees of freedom for REG^r is unisolvent.
- (Theorem 2.3) REG^r on a mesh is characterized by *tangential-tangential continuity*: for $u \in \mathrm{REG}^r$ of a mesh, for any simplex f of dimension ≥ 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 point $x \in f$. Moreover a piecewise polynomial symmetric covariant 2-tensor field has tangential-tangential continuity if and only if it belongs to REG^r. This turns out to be the key property for its use in applications.
- (Theorem 2.2 and Theorem 2.4) REG^r is both *local* and *affine*. This makes it the *canon-*

ical finite element with respect to tangential-tangential continuity in the language of [52]. These two properties make REG^r easy to implement and analyze.

- (Theorem 2.5 and Theorem 2.6) The canonical interpolant of REG^r, induced by the degrees of freedom, has the optimal approximation properties.

This chapter is organized as follows. First, in Section 2.1, we give a coordinate-free definition of REG^r. In Section 2.2, we give a constructive proof of the characterization and unisolvency of REG^r. In Section 2.3, we study the affine properties of REG^r and prove the optimal approximation theorems. We end the chapter with Section 2.4 describing two sets of concrete degrees of freedom for REG^r. They form the basis of author's implementation of REG^r in the open source software FEniCS as part of this thesis.

2.1 Definition of generalized Regge family

In this section, we define the generalized Regge family precisely.

First, we clarify what is a mesh. In $\mathbb{R}^m$, the convex hull of $(n+1)$ points $\{v_0, \dots, v_n\}$ of general position is called an *n-simplex* $c = [v_0, \dots, v_n]$. 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* $\mathscr{T}$ is a finite collection of simplices in $\mathbb{R}^m$ satisfying:

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

The integer m is called the *geometric dimension* of $\mathscr{T}$ while n is called the *topological dimension* of $\mathscr{T}$. 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 $\mathscr{T}$ 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 [84, Section 7]. The manifold determined by $\mathscr{T}$, called the *carrier* of $\mathscr{T}$, is denoted by $|\mathscr{T}|$. Let M be any manifold (possibly with boundary). If $\mathscr{T}$ is a mesh with $|\mathscr{T}|$ diffeomorphic to M, then $\mathscr{T}$ is called a *triangulation* of M.

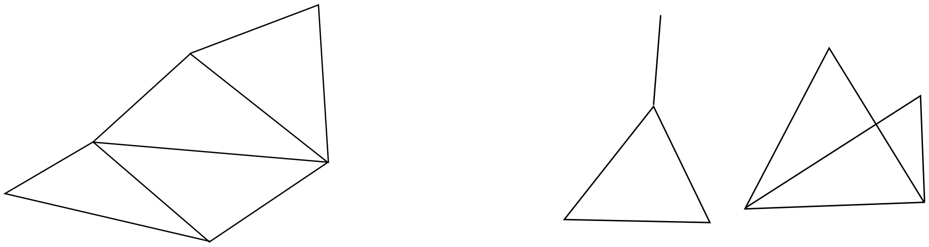


Figure 2.5: Mesh and non-meshes in 2D

Second, we review the notion of tensor fields in differential geometry [113, Chapter 2]. Let M be a smooth manifold. 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 smooth symmetric covariant 2-tensor fields on M is especially important and is denoted by $\mathscr{S}(M)$. 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,

$$[(\phi^*)_p g](u_1, \ldots, u_k) := g(d\phi_p(u_1), \ldots, d\phi_p(u_k)). \tag{2.1}$$

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 ϕ. In particular, $\phi^* : \mathscr{S}(N) \to \mathscr{S}(M)$ for any smooth ϕ. In this thesis, the manifolds M and N are frequently simplices, which have boundary. In this case, the functions are only required to be smooth in the interior and continuous up to the boundary. 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 ι_f. By definition, for any $g \in \mathscr{S}(c)$, its pullback $\iota_f^* g \in \mathscr{S}(f)$ 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 [28, 29]. The term tangential-tangential part is preferred in this thesis to single out the pullback for covariant 2-tensors.

Third, we develop some notations for polynomial spaces on simplices. For a n-simplex c, let $\mathscr{P}^r(c)$ be the space of polynomials of degree r or less on c as before. It is well-known [30,

Equation (2.2.2)] that

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

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 [113]). 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

$$\mathscr{P}^r\mathscr{S}(c) := \mathscr{P}^r(c) \otimes \mathbb{S}^n. \tag{2.2}$$

Equivalently, $\mathscr{P}^r\mathscr{S}(c)$ 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 $\mathbb{S}^n$ is isomorphic to the space of symmetric n-by-n matrices. Hence,

$$\dim \mathbb{S}^n = \binom{n+1}{2}, \qquad \dim \mathscr{P}^r\mathscr{S}(c) = \binom{n+r}{n}\binom{n+1}{2}. \tag{2.3}$$

Let $\mathscr{T}$ be any mesh of topological dimension n. A *piecewise polynomial symmetric covariant* 2*-tensor fields* of degree r or less is a function assigning to each cell c of $\mathscr{T}$ an element of $\mathscr{P}^r\mathscr{S}(c)$. These can be combined linearly in the obvious fashion. We denote the vector space of all piecewise polynomial symmetric covariant 2-tensor fields by $\mathscr{P}^r\mathscr{S}(\mathscr{T})$. Elements of $\mathscr{P}^r\mathscr{S}(\mathscr{T})$ can be interpreted as symmetric covariant 2-tensor fields on the carrier $|\mathscr{T}|$ of the mesh, which might be multi-valued on cell boundaries. It should be noted that one cannot define $\mathscr{P}^r\mathscr{S}(M)$ on a general smooth manifold M. However, for any triangulation $\mathscr{T}$ of M, $\mathscr{P}^r\mathscr{S}(\mathscr{T})$ is still well-defined.

Fourth, we review the definition of finite elements. A *simplicial finite element* is a triple (c, V, Σ). 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 Σ_f is a subspace of the dual space V_f'. This Σ 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, Σ), the simplex c is called the *domain*, elements of V are called *shape functions*, and elements of Σ 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 [30, 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 [8, Section 4]. For a particular software implementation, a basis B_f can be fixed for each Σ_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 [10, 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 $\mathscr{T}$, a finite element family F is called *assemblable* on $\mathscr{T}$ if all the cells of $\mathscr{T}$ 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 Σ_f on f. In such a situation, a *finite element space* on $\mathscr{T}$, denoted by $F(\mathscr{T})$, can be obtained through the *finite element assembly process*: $F(\mathscr{T})$ is the collection of functions u on $\mathscr{T}$ 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 $l \in \Sigma_f$.

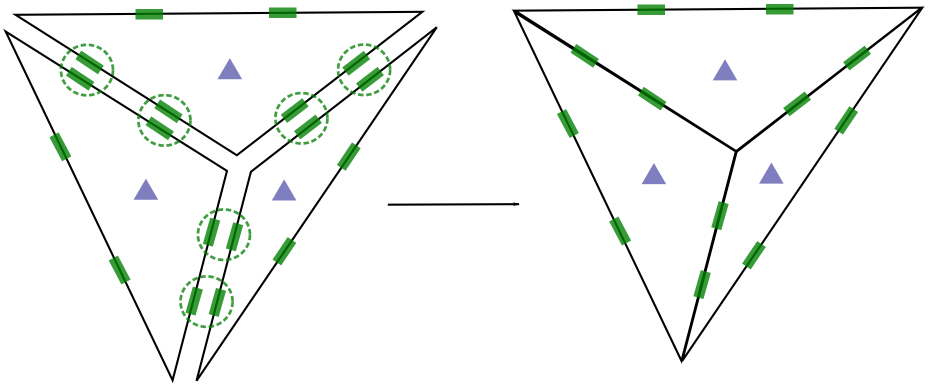


Figure 2.6: REG^1 assembly on a 3-triangle mesh.

A pictorial depiction of the assembly of REG^1 is given in Figure 2.6.

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:

$$\mathscr{P}^r \mathscr{S}(c) \tag{2.4a}$$

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

$$r_f := \iota_f^* : \mathscr{P}^r\mathscr{S}(c) \to \mathscr{P}^r\mathscr{S}(f), \qquad \Sigma_f := \{u \mapsto \int_f u : q \mid q \in \mathscr{P}^{r-k+1}\mathscr{S}(f), \tag{2.4b}$$

where the colon : denotes the Frobenius inner product on $\mathbb{S}^k$. It will be proven in Theorem 2.1 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.

We count the dimensions of the space and the degrees of freedom. The dimensions of $\mathscr{P}^r$ and $\mathscr{P}^r\mathscr{S}$ are already computed in (2.3). On one hand,

$$\dim V = \dim \mathscr{P}^r\mathscr{S}(c) = \binom{n+r}{r}\binom{n+1}{2}.$$

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:

$$\sum_{k=1}^{n} (\#\{k\text{-faces of } c\})(\dim \mathscr{P}^{r-k+1}\mathscr{S}(f)) = \sum_{k=1}^{n} \binom{n+1}{k+1}\binom{r+1}{k}\binom{k+1}{2}.$$

As a consequence of the unisolvency (Theorem 2.1), the following identity must hold:

$$\binom{n+r}{r}\binom{n+1}{2} = \sum_{k=1}^{n} \binom{n+1}{k+1}\binom{k+1}{2}\binom{r+1}{k}. \tag{2.5}$$

This identity can be verified independently as well. First, clearly

$$\binom{n+1}{k+1}\binom{k+1}{2} = \frac{(n+1)!}{(k+1)!(n-k)!}\frac{(k+1)!}{2!(k-1)!} = \frac{(n+1)!}{2!(n-1)!}\frac{(n-1)!}{(n-k)!(k-1)!} = \binom{n+1}{2}\binom{n-1}{n-k}.$$

Then,

$$\sum_{k=1}^{n} \binom{n-1}{n-k}\binom{r+1}{k} = \binom{n+r}{n}$$

can be derived from comparing the coefficients of x in the identity

$$(1+x)^{n-1}(1+x)^{r+1} = (1+x)^{n+r}.$$

We finally define REG^r, the generalized Regge finite element family of degree r. For any $r \geq 0$, it is a function space defined on all simplices c of dimension $n \geq 1$ which assigns the Regge finite element of degree r on c to each c.

Proposition 2.1. *For any fixed $r \geq 0$, 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 $\mathscr{T}$ is of dimension n so it is in the domain of REG^r. Suppose two cells c_1 and c_2 intersect at a k-face f. Then,

$$\iota^*_{f\to c_1}\mathscr{P}^r\mathscr{S}(c_1) = \mathscr{P}^r\mathscr{S}(f) = \iota^*_{f\to c_2}\mathscr{P}^r\mathscr{S}(c_2).$$

Finally, it is clear from the definition (2.4b) that Σ_f only depends on f and hence is the same in both $\mathrm{REG}^r(c_1)$ and $\mathrm{REG}^r(c_2)$. Hence REG^r is assemblable on $\mathscr{T}$. □

The resulting assembled space is denoted by $\mathrm{REG}^r(\mathscr{T})$.

2.2 Basic properties

We establish the basic properties of generalized Regge elements in this section. We outline the main results below and give the detailed proofs in the subsections.

The first result shows REG^r is well-defined.

Theorem 2.1 (Unisolvency)**.** *The set of degrees of freedom* (2.4b) *is unisolvent.*

The second result shows that REG^r satisfies the *locality property* defined in [52].

Theorem 2.2 (Locality)**.** *Let f be a k-face in an n-simplex c with $k \geq 1$ and $u \in \mathscr{P}^r\mathscr{S}(c)$. Then $\iota^*_f u$ is completely determined by the subset of degrees of freedom associated with f and its faces.*

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

Theorem 2.3 (Characterization)**.** *Let $\mathscr{T}$ be a mesh of topological dimension $n \geq 1$. Suppose $u \in \mathscr{P}^r\mathscr{S}(\mathscr{T})$ is a piecewise polynomial covariant symmetric 2-tensor. Then $u \in \mathrm{REG}^r(\mathscr{T})$ if and only if $\iota^*_f u$ is single-valued at each interior k-face f of the mesh with $k \geq 1$.*

On an simplex c, we call the subspace of $\mathscr{P}^r\mathscr{S}(c)$ with vanishing tangential-tangential parts *tangential-tangential bubble functions* and denote it by $\mathring{\mathscr{P}}^r\mathscr{S}(c)$. The fourth result concerns the structure of these bubble functions.

Proposition 2.2. *Let c be an n-simplex. The dual space to the space of tangential-tangential bubble functions is:*

$$[\mathring{\mathscr{P}}^r\mathscr{S}(c)]' = \left\{u \mapsto \int_c u : q \mid q \in \mathscr{P}^{r-n+1}\mathscr{S}(c)\right\}.$$

In particular,

$$\dim \mathring{\mathscr{P}}^r\mathscr{S}(c) = \dim \mathscr{P}^{r-n+1}\mathscr{S}(c) = \binom{n+1}{2}\binom{r+1}{n}.$$

In the rest of the section, we prove these four theorems using a construction known as the *geometric decomposition*. It gives an explicit basis for REG^r on a simplex indexed by its faces. Similar decompositions are useful both theoretically [9] and for software implementations [63].

In order to state the geometric decomposition, we need some notations. Fix an arbitrary n-simplex $c = [v_0 \dots v_n]$. Let $\{\lambda_j\}_{j=0}^n$ be the barycentric coordinates on c: λ_i are linear functions determined by $\lambda_i(v_j) = \delta_{ij}$. A *multi-index* $\alpha \in \mathbb{N}^{0:n}$ is an array $\alpha = (\alpha_0, \dots, \alpha_n)$ of $(n+1)$ non-negative integers $\alpha_i \geq 0$. Set

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

The *support* $[\![\alpha]\!]$ of α is defined to be

$$[\![\alpha]\!] := \{i \in \{0, \dots, n\} \mid \alpha_i \geq 1\}, \tag{2.6}$$

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

$$I(f) := \{f_0, \dots, f_k\}. \tag{2.7}$$

Under this notation, the (unnormalized) *Bernstein basis* $B^r(c)$ for $\mathscr{P}^r(c)$ is given by [76]:

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

Given two 1-forms l_1, l_2 on $c = [v_0 \dots v_n]$, their *symmetric tensor product* $l_1 \odot l_2$ is a symmetric covariant 2-tensor given by

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

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:

$$d\lambda_i \odot d\lambda_j \in \mathbb{S}^n.$$

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 $\mathscr{P}^r\mathscr{S}(c)$, for any $p \in \mathscr{P}^r(c)$ and any edge $[v_i v_j]$ of c, $p\, d\lambda_i \odot d\lambda_j \in \mathscr{P}^r\mathscr{S}(c)$.

We prove the following bases for $\mathscr{P}^r\mathscr{S}(c)$ and $\mathring{\mathscr{P}}^r\mathscr{S}(c)$.

Proposition 2.3 (Basis)**.** *Let $c = [v_0 \dots v_n]$ 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,*

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

forms a basis for $\mathscr{P}^r\mathscr{S}(c)$ and

$$\mathring{B}^r_{\mathscr{S}}(c) := \{\lambda^\alpha d\lambda_i \odot d\lambda_j \mid [v_iv_j] \in e(c), \alpha \in \mathbb{N}^{0:n}, |\alpha| = r, [\![\alpha]\!] \cup \{i,j\} = I(c)\}$$

forms a basis for $\mathring{\mathscr{P}}^r\mathscr{S}(c)$.

We need an extension operator to take polynomials on a face of a simplex to the simplex. Let $f = [v_{f_0} \dots v_{f_k}]$ be any k-face of c. This f is a k-simplex on its own and has its own barycentric coordinates $\{\lambda^f_{f_i}\}_{i=0}^k$. Further, $B^r(f)$ defined using these $\{\lambda^f_{f_i}\}$ forms a basis for $\mathscr{P}^r(f)$. There is a canonical map $E^r_{f\to c} : \mathscr{P}^r(f) \to \mathscr{P}^r(c)$, called the *barycentric extension* [9, Section 2.2], which simply replaces any appearance of $\lambda^f_{f_j}$ with $\lambda^c_{f_j}$ in the expansion of any $\mathscr{P}^r(f)$ in basis $B^r(f)$. For example, if $f = [v_0v_1v_2]$, then

$$E^2_{f\to c}(\lambda^f_1\lambda^f_2) = \lambda^c_1\lambda^c_2.$$

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

$$E^3_{f\to c}(\lambda^f_1\lambda^f_2) = E^3_{f\to c}(\lambda^f_1\lambda^f_2(\lambda^f_0 + \lambda^f_1 + \lambda^f_2)) = \lambda^c_0\lambda^c_1\lambda^c_2 + (\lambda^c_1)^2\lambda^c_2 + \lambda^c_1(\lambda^c_2)^2,$$

which is a cubic polynomial on c.

For any face $f = [v_{f_0} \dots v_{f_k}]$ of an n-simplex $c = [v_0 \dots v_n]$, the barycentric extension $E^r_{f\to c}$ can be extended to a map from $\mathscr{P}^r\mathscr{S}(f)$ to $\mathscr{P}^r\mathscr{S}(c)$ naturally via the basis given in Proposition 2.3. For example, if $f = [v_0v_1v_2]$ then for any edge $[v_iv_j]$ of f,

$$E^r_{f\to c}[\lambda^f_0(\lambda^f_1)^2\lambda^f_2 d\lambda^f_i \odot d\lambda^f_j] := \lambda^c_0(\lambda^c_1)^2\lambda^c_2 d\lambda^c_i \odot d\lambda^c_j.$$

For a cell c, the space $\mathscr{P}^r\mathscr{S}(c)$ can be decomposed as the extensions of bubble functions on faces of c.

Proposition 2.4 (Geometric decomposition)**.** *Let c be a simplex. Then,*

$$\mathscr{P}^r\mathscr{S}(c) = \bigoplus_{\dim f \geq 1} E^r_{f\to c}\mathring{\mathscr{P}}^r\mathscr{S}(f).$$

The basis decomposes accordingly:

$$B^r_{\mathscr{S}}(c) = \bigcup_{\dim f \geq 1} E^r_{f\to c}\mathring{B}^r_{\mathscr{S}}(f)$$

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

$$[\mathscr{P}^r\mathscr{S}(c)]' = \bigoplus_{\dim f \geq 1} \left\{u \mapsto \int_f (\iota^*_f u) : q \mid q \in \mathscr{P}^{r-\dim f+1}\mathscr{S}(f)\right\}.$$

For example, the unisolvency (Theorem 2.1) 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.

r	$[v_i v_j]$	$[v_i v_j v_k]$
0	$d\lambda_i \odot d\lambda_j$	
1	$\lambda_i d\lambda_i \odot d\lambda_j$, $\lambda_j d\lambda_i \odot d\lambda_j$	$\lambda_i d\lambda_j \odot d\lambda_k$, $\lambda_j d\lambda_i \odot d\lambda_k$, $\lambda_k d\lambda_i \odot d\lambda_j$
2	$\lambda_i^2 d\lambda_i \odot d\lambda_j$, $\lambda_j^2 d\lambda_i \odot d\lambda_j$, $\lambda_i \lambda_j d\lambda_i \odot d\lambda_j$	$\lambda_i^2 d\lambda_j \odot d\lambda_k$, $\lambda_j^2 d\lambda_i \odot d\lambda_k$, $\lambda_k^2 d\lambda_i \odot d\lambda_j$, $\lambda_i \lambda_k d\lambda_i \odot d\lambda_j$, $\lambda_j \lambda_k d\lambda_i \odot d\lambda_j$, $\lambda_i \lambda_j d\lambda_i \odot d\lambda_k$, $\lambda_k \lambda_j d\lambda_i \odot d\lambda_k$, $\lambda_j \lambda_i d\lambda_j \odot d\lambda_k$, $\lambda_k \lambda_i d\lambda_j \odot d\lambda_k$

Table 2.1: Bernstein-style Basis in 2D

r	$[v_i v_j]$	$[v_i v_j v_k]$	$[v_i v_j v_k v_l]$
0	Same as 2D, 1 per edge		
1	Same as 2D, 2 per edge	Same as 2D, 3 per triangle	
2	Same as 2D, 3 per edge	Same as 2D, 9 per triangle	$\lambda_i \lambda_j d\lambda_k \odot d\lambda_l$, $\lambda_j \lambda_k d\lambda_l \odot d\lambda_i$, $\lambda_k \lambda_l d\lambda_i \odot d\lambda_j$, $\lambda_l \lambda_i d\lambda_j \odot d\lambda_k$, $\lambda_i \lambda_k d\lambda_j \odot d\lambda_l$, $\lambda_j \lambda_l d\lambda_i \odot d\lambda_k$

Table 2.2: Bernstein-style Basis in 3D

2.2.1 Bernstein decomposition for Lagrange elements

Here we review the geometric decomposition of the scalar polynomial space $\mathscr{P}^r(c)$. This serves as a model and the basis for the more complicated decomposition of $\mathscr{P}^r \mathscr{S}^n(c)$.

As mentioned in the introduction of this section, on an n-simplex c, the unnormalized *Bernstein basis* $B^r(c)$ for $\mathscr{P}^r(c)$ is given by [76]:

$$B^r(c) := \{\lambda^\alpha \mid \alpha \in \mathbb{N}^{0:n}, |\alpha| = r\}, \tag{2.8}$$

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 implementations [74, Chapter 4]. This basis has many advantages numerically, for example see [63]. The Bernstein basis has the following elementary property:

Lemma 2.1. *Suppose $p \in \mathscr{P}^r(c)$ is divisible by λ_k for some k, then in the expansion of p in the Bernstein basis each summand contains a λ_k factor individually.*

Proof. Suppose the claim is false. Then necessarily, a linear combination of

$$\{\lambda_0^{\alpha_0}\cdots\lambda_n^{\alpha_n} \mid \alpha \in \mathbb{N}^{0:n}, |\alpha| = r, \alpha_k = 0\}$$

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 $B^{r-1}(c)$ again. In particular, each term of the thus expanded product in $\lambda_k q$ is a basis element in $B^r(c)$ and contains a λ_k factor. Thus a linear combination of elements in $B^r(c)$ which do not contain the factor λ_k equals a linear combination of elements in $B^r(c)$ which do contain the factor λ_k. This contradicts the fact that elements of $B^r(c)$ are linearly independent. □

Let the support $[\![\cdot]\!]$ and index set $I(\cdot)$ be defined as in (2.6) and (2.7) respectively. For any face f of an n-simplex c, set

$$B_c^r(f) := \{\lambda^\alpha \mid \alpha \in \mathbb{N}^{0:n}, |\alpha| = r, [\![\alpha]\!] = I(f)\} \tag{2.9}$$

which is the subset of the Bernstein basis $B^r(c)$ whose factors involve exactly λ_i associated with vertices of f. It is clear that every element of $B^r(c)$ is in a unique $B_c^r(f)$ for some face f of c. The trivial observation that the map $[v_{f_0} \dots v_{f_k}] \to \{v_{f_0} \dots v_{f_k}\}$ is a bijection between faces of c and subsets of vertices of c implies:

$$B^r(c) = \bigcup_{f \subset c} B_c^r(f),$$

where the union is disjoint and is taken over all faces f of c. Elements of $B_c^r(f)$ vanishes on the boundary of f.

On any simplex c, we call the subspace of $\mathscr{P}^r(c)$ which vanishes on the boundary the sapce of *bubble functions* and denote it by $\mathring{\mathscr{P}}^r(c)$. It turns out that following is a basis for $\mathring{\mathscr{P}}^r(c)$:

$$\mathring{B}^r(c) := \{\lambda^\alpha \mid \alpha \in \mathbb{N}^{0:n}, |\alpha| = r, [\![\alpha]\!] = I(c)\}, \tag{2.10}$$

Indeed, each λ_i vanishes on the facet opposite to vertex v_i and every facet is opposite to a vertex, every element of $\mathring{B}^r(c)$ vanishes on the boundary of c and is thus in $\mathring{\mathscr{P}}^r(c)$. On the other hand, if $p \in \mathring{\mathscr{P}}^r(c)$, then p is divisible by $\lambda_0 \cdots \lambda_n$. By Lemma 2.1, this implies that every term in the expansion of p in the Bernstein basis is in $\mathring{B}^r(c)$. Hence, $\mathring{B}^r(c)$ is a basis for $\mathring{\mathscr{P}}^r(c)$ (also derived in [9, Equation (2.4)]).

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

$$B^r_c(f) = E^r\mathring{B}^r(f).$$

Hence,

$$B^r(c) = \bigcup_{f\in c} E^r\mathring{B}^r(f), \qquad \mathscr{P}^r(c) = \bigoplus_{f\in c} E^r\mathring{\mathscr{P}}^r(f).$$

This is called the *Bernstein decomposition* in [9]. An example of this for $\mathscr{P}^3$ on a triangle is shown in Figure 2.7.

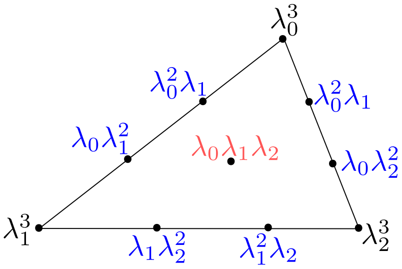


Figure 2.7: Geometric decomposition of $\mathscr{P}^3$ on a triangle.

This decomposition is useful for software implementation [44]. Moreover, it gives an elegant proof of unisolvency of degrees of freedoms for Lagrange elements. It is clear that the map from $\mathscr{P}^{r-n-1}(c)$ to $\mathring{\mathscr{P}}^r(c)$ given by

$$p \mapsto p\lambda_0\cdots\lambda_n$$

is an isomorphism. Thus the space of functionals

$$\left\{p \mapsto \int_c pq \mid q \in \mathscr{P}^{r-n-1}(c)\right\}$$

is isomorphic to the dual space $[\mathring{\mathscr{P}}^r(c)]'$. The Bernstein decomposition then implies that

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

This is known to be the degrees of freedom for Lagrange finite elements [9].

2.2.2 Geometric decomposition for Regge elements

This subsection is subtle. The main idea is to derive the Bernstein-style basis (Proposition 2.3) for REG^r first. This is used to give a constructive proof of the geometric decomposition (Proposition 2.4) and the bubble characterization (Proposition 2.2). 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 $\mathbb{S}^n$ in terms of barycentric coordinates (Proposition 2.5).
2. Derive the Bernstein-style basis $B^r_{\mathscr{S}}(c)$ for $\mathscr{P}^r\mathscr{S}(c)$.
3. Derive the action of pullback on basis elements (equation (2.16) and Proposition 2.6) and establish that $E^r_{f\to c}$ is the right inverse of the pullback.
4. Derive the Bernstein-style basis $\mathring{B}^r_{\mathscr{S}}(c)$ for $\mathring{\mathscr{P}}^r\mathscr{S}(c)$. 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 2.3 and Lemma 2.4). This combined with the second step proves Proposition 2.3.
5. Derive the geometric decomposition of $\mathscr{P}^r\mathscr{S}(c)$ (first part of Proposition 2.4). This is a constructive proof based on an edge-based Bernstein decomposition of $\mathscr{P}^r(c)$ (Lemma 2.6 and Lemma 2.7).
6. Prove the bubble characterization (Proposition 2.2) via an explicit bijection using the Bernstein-style basis. Then prove the dual geometric decomposition (Lemma 2.8). This along with the previous step proves Proposition 2.4.
7. Prove Theorem 2.1–2.3 as corollaries.

The first step is to recall the well-known connection [28, Proposition 3.2] between the barycentric coordinates and the space of piecewise constant symmetric covariant 2-tensors.

Proposition 2.5. *Let $c = [v_0 \dots v_n]$ 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*

$$\{d\lambda_i \odot d\lambda_j \mid [v_i v_j] \in e(c)\} \tag{2.11}$$

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

Proof. First, because each λ_i is linear, $d\lambda_i$ is constant. So is the symmetric tensor product $d\lambda_i \odot d\lambda_j$. Hence $d\lambda_i \odot d\lambda_j \in \mathscr{P}^0\mathscr{S}(c)$. Second, elementary dimension counts show that

$$\dim \mathscr{P}^0\mathscr{S}(c) = \binom{n+1}{2} = \#[e(c)]. \tag{2.12}$$

Thus the only thing left to show is that these $d\lambda_i \odot d\lambda_j$ 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 λ_i:

$$d\lambda_i(v_j - v_k) = (v_j - v_k)(\lambda_i) = \lambda_i(v_j) - \lambda_i(v_k) = \delta_{ij} - \delta_{ik}. \tag{2.13}$$

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

$$\begin{aligned}(d\lambda_i \odot d\lambda_j)(v_k - v_l, v_k - v_l) &= (\delta_{ik} - \delta_{il})(\delta_{jk} - \delta_{jl}) \\ &= \begin{cases} -1, & \text{if } (i = k \text{ and } j = l) \text{ or } (i = l \text{ and } j = k), \\ 0, & \text{otherwise.} \end{cases}\end{aligned} \tag{2.14}$$

Thus the span of the given set has dimension at least $\#[e(c)]$. This proves the linear independency 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

$$\begin{aligned}B^r_{\mathscr{S}}(c) &:= \{p\, d\lambda_i \odot d\lambda_j \mid p \in B^r(c), [v_i v_j] \in e(c)\} \\ &= \{\lambda^\alpha d\lambda_i \odot d\lambda_j \mid \alpha \in \mathbb{N}^{0:n}, |\alpha| = r, [v_i v_j] \in e(c)\}\end{aligned} \tag{2.15}$$

forms a basis for $\mathscr{P}^r\mathscr{S}(c)$. For any face f of c, the pullback ι_f^* is linear. Its action on a basis element is just:

$$\iota_f^*(p\, d\lambda_i \odot d\lambda_j) = (p \circ \iota_f)\iota_f^*(d\lambda_i \odot d\lambda_j). \tag{2.16}$$

Hence, the tensor part $d\lambda_i \odot d\lambda_j$ and the polynomial part can be dealt with separately.

Lemma 2.2. *Let c be any simplex and f any face of c. For any $u \in \mathscr{P}^r\mathscr{S}(f)$,*

$$\iota_f^* E^r_{f\to c} u = u.$$

Proof. Suppose $c = [v_0 \dots v_n]$ and $f = [v_{f_0} \dots v_{f_k}]$. Let $\{\lambda_i^c\}$ and $\{\lambda_{f_i}^f\}$ 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,

$$\lambda_{f_i}^c \circ \iota_f = \lambda_{f_i}^f, \qquad \iota_f^*(d\lambda_{f_i}^c) = d\lambda_{f_i}^f.$$

Hence $\iota_f^* E^r_{f\to c}$ is the identity on the basis elements in $B^r_{\mathscr{S}}(f)$. Both maps are linear so this extends to $\mathscr{P}^r\mathscr{S}(f)$. This proves the claim. □

The following theorem collects the key properties of the pullback of $d\lambda_i \odot d\lambda_j$:

Proposition 2.6. *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 $[v_i v_j] \subset f$. 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 $\mathscr{P}^0\mathscr{S}(f) = \mathbb{S}^k$, where λ_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 $\mathscr{P}^0\mathscr{S}(f)$. □

Corollary 2.1. *Let c be an n-simplex and $p \in \mathscr{S}(c)$ 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 Proposition 2.6, these $\iota_f^*(d\lambda_i \odot d\lambda_j)$ forms a basis for $\mathscr{P}^0\mathscr{S}(f)$. 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. □

The next step is find a basis for the tangential-tangential bubble space $\mathring{\mathscr{P}}^r\mathscr{S}(f)$. Let $c = [v_0 \dots v_n]$ be an n-simplex and $[v_i v_j] \in e(c)$ an edge. In light of Proposition 2.6 and the pullback formula (2.16), a basis element $p d\lambda_i \odot d\lambda_j \in \mathring{\mathscr{P}}^r\mathscr{S}(f)$ if p vanishes on the facets 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 $B^r_{ij}(c)$ to be the collection of elements of the Bernstein basis $B^r(c)$ which vanish on all the $(n-1)$ facets which do contain $[v_i v_j]$.

Lemma 2.3.

$$B^r_{ij}(c) = \{\lambda^\alpha \mid \alpha \in \mathbb{N}^{0:n}, |\alpha| = r, [\![\alpha]\!] \cup \{i,j\} = I(c)\}.$$

Further, $B^r_{ij}(c)$ forms a basis for the subspace of $\mathscr{P}^r(c)$ containing polynomials vanishing on all the $(n-1)$ facets containing the edge $[v_i v_j]$.

Proof. For the first part, using the fact that λ_i vanishes on the facet opposite to vertex v_i, the definition of $B^r_{ij}(c)$ implies that all elements of it are divisible by λ_k for all $k = 0,\dots,n$, except $k = i$ and $k = j$, which is exactly what the formula says. For the second part, it is clear that $B^r_{ij}(c)$ as a subset of the Bernstein basis $B^r(c)$ is linearly independent. It is also obvious that each element of $B^r_{ij}(c)$ is in the space it is claimed to be a basis of. Suppose $q \in \mathscr{P}^r(c)$ vanishes on all the facets containing the edge $[v_i v_j]$. Expand q in Bernstein basis $B^r(c)$. By Lemma 2.1, the fact that q is divisible by λ_k for all $k = 0,\dots,n$, except $k = i$ and $k = j$ implies that the same holds for each summand in the expansion. Hence q is a linear combination of elements in $B^r_{ij}(c)$. □

The following result, then, comes at no surprise:

Lemma 2.4. *Let c be an n-simplex. Then*

$$\begin{aligned}\mathring{B}^r_{\mathscr{S}}(c) &:= \{p\, d\lambda_i \odot d\lambda_j \mid p \in B^r_{ij}(c), [v_i v_j] \in e(c)\} \\ &= \{\lambda^\alpha d\lambda_i \odot d\lambda_j \mid [v_i v_j] \in e(c), \alpha \in \mathbb{N}^{0:n}, |\alpha| = r, [\![\alpha]\!] \cup \{i,j\} = I(c)\}\end{aligned}$$

forms a basis for bubbles $\mathring{\mathscr{P}}^r\mathscr{S}(c)$.

Proof. First, the preceding discussion showed that every element of $\mathring{B}^r_{\mathscr{S}}(c)$ is in $\mathring{\mathscr{P}}^r\mathscr{S}(c)$. Second, because both $B^r_{ij}(c)$ and $\{d\lambda_i \odot d\lambda_j \mid [v_i v_j] \in e(c)\}$ are linearly independent sets, elements in $\mathring{B}^r_{\mathscr{S}}(c)$ as their product are also linearly independent by the tensor product structure (2.2). Finally, suppose $q \in \mathring{\mathscr{P}}^r\mathscr{S}(c)$. Expand q in basis $B^r_{\mathscr{S}}(c)$ defined in (2.15):

$$q = \sum_{[v_i v_j] \in e(c)} \sum_{p \in B^r(c)} q_{p,i,j}\, p\, d\lambda_i \odot d\lambda_j.$$

By Corollary 2.1, for any edge $[v_i v_j] \in e(c)$, each

$$\iota^*_{\partial c}\left(\sum_{p \in B^r(c)} q_{p,i,j}\, p\, d\lambda_i \odot d\lambda_j\right) = \iota^*_{\partial c}(d\lambda_i \odot d\lambda_j) \sum_{p \in B^r(c)} q_{p,i,j}(p \circ \iota_{\partial c}) = 0.$$

By Proposition 2.6, the polynomial

$$\sum_{p\in B^r(c)} q_{p,i,j}(p\circ\iota_{\partial c})$$

must vanish on the $(n-1)$ boundary facets containing edge $[v_iv_j]$. By Lemma 2.3, this is in the span of $B^r_{ij}(c)$. Thus the linear span of $\mathring{B}^r_{\mathscr{S}}(c)$ contains $\mathring{\mathscr{P}}^r\mathscr{S}(c)$. This proves the claim. □

The next step is to derive another geometric decomposition of the Bernstein basis $B^r(c)$ which are based on edges. Let $c=[v_0\dots v_n]$ be an n-simplex and $[v_iv_j]$ be an edge. For any k-face f of c with $k\geq 1$, let

$$B^r_{c,ij}(f) := \{\lambda^\alpha d\lambda_i\odot d\lambda_j \mid \alpha\in\mathbb{N}^{0:n}, |\alpha|=r, [\![\alpha]\!]\cap\{i,j\}=I(f)\}, \tag{2.17}$$

where the barycentric coordinates $\{\lambda_i\}$ are for the cell c.

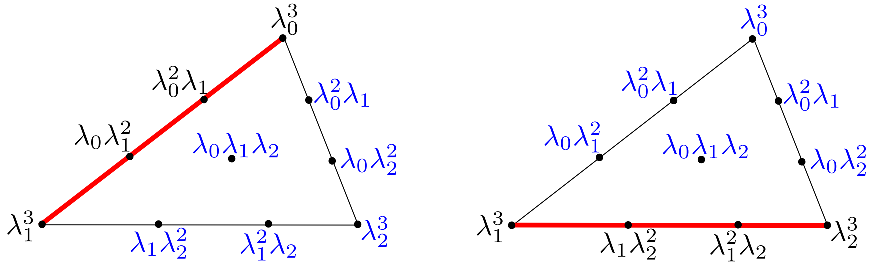


Figure 2.8: Edge-based Bernstein decomposition for $\mathscr{P}^3$ on a triangle. The chosen edge is in red. Basis associated with edges are in black while those associated with cells are in blue.

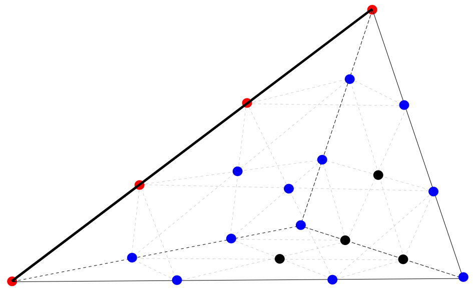


Figure 2.9: Edge-based Bernstein decomposition for $\mathscr{P}^3$ on a tetrahedron. The chosen edge is thickened. Basis associated with edges are in red, those associated with triangles are in blue, and those associated with cells are in black.

Lemma 2.5. *Let $c=[v_0\dots v_n]$ be an n-simplex. Then, for any fixed edge $[v_iv_j]$ of c,*

$$B^r(c) = \bigcup_{f\supset[v_iv_j]} B^r_{c,ij}(f),$$

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

Proof. It is clear that each $B^r_{c,ij}(f)$ is a subset of the Bernstein basis $B^r(c)$. 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 $p = \lambda^\alpha$ is any element of $B^r(c)$. 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 $B^r(c)$. This proves the claim. □

Let c be an n-simplex and f any k-face of c with $k \geq 1$. Further, let $E^r_{f\to c}$ be the barycentric extension defined before. Comparing the formula for $B^r_{ij}(f)$ in Lemma 2.3 with the definition of $B^r_{c,ij}(f)$ in equation (2.17), whenever f contains the edge $[v_i v_j]$, clearly,

$$B^r_{c,ij}(f) = E^r_{f\to c} B^r_{ij}(f).$$

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

Lemma 2.6. *Let $c = [v_0 \dots v_n]$ be an n-simplex. Then, for any edge $[v_i v_j]$ of c,*

$$B^r(c) = \bigcup_{f \supset [v_i v_j]} E^r_{f\to c} B^r_{ij}(f),$$

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 $B^r_{\mathscr{S}}(c)$.

Lemma 2.7. *Let c be a simplex. Then,*

$$B^r_{\mathscr{S}}(c) = \bigcup_{\dim f \geq 1} E^r_{f\to c} \mathring{B}^r_{\mathscr{S}}(f) \quad and \quad \mathscr{P}^r\mathscr{S}(c) = \bigoplus_{\dim f \geq 1} E^r_{f\to c} \mathring{\mathscr{P}}^r\mathscr{S}(f),$$

where the first union is disjoint.

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

$$B^r_{\mathscr{S}}(c) = \bigcup_{[v_i v_j] \in e(c)} \{p\, d\lambda_i \odot d\lambda_j \mid p \in B^r(c)\}.$$

Lemma 2.6 implies that

$$B^r_{\mathscr{S}}(c) = \bigcup_{[v_i v_j] \in e(c)} \bigcup_{f \supset [v_i v_j]} \{E^r_{f\to c} p\, d\lambda_i \odot d\lambda_j \mid p \in B^r_{ij}(f)\}.$$

Exchange the order of the two unions:

$$B^r_{\mathscr{S}}(c) = \bigcup_{\dim f \geq 1} \bigcup_{[v_i v_j] \in e(f)} \{E^r_{f\to c} p\, d\lambda_i \odot d\lambda_j \mid p \in B^r_{ij}(f)\}.$$

Finally, Lemma 2.4 says that the inner union is exactly $E^r_{f\to c}\mathring{B}^r_{\mathscr{S}}(f)$. Hence,

$$B^r_{\mathscr{S}}(c) = \bigcup_{\dim f \geq 1} E^r_{f\to c}\mathring{B}^r_{\mathscr{S}}(f)$$

proves the claim. □

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

Lemma 2.8. *For any n-simplex c,*

$$[\mathring{\mathscr{P}}^r\mathscr{S}(c)]' = \left\{u \mapsto \int_c u : q \mid q \in \mathscr{P}^{r-n+1}\mathscr{S}(c)\right\}.$$

Further, the geometric decomposition of $[\mathscr{P}^r\mathscr{S}(c)]'$ in Proposition 2.4 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 $\mathscr{P}^{r-n+1}(c)$ and the span of $B^r_{ij}(c)$. In particular, the dual relation holds:

$$[\operatorname{span} B^r_{ij}(c)]' = \left\{p \mapsto \int_c pq \mid q \in \mathscr{P}^{r-n+1}(c)\right\}.$$

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 Proposition 2.3 together implies the claim.

Finally, the geometric decomposition of $[\mathscr{P}^r\mathscr{S}(c)]'$ is derived. Dualize the geometric decomposition of $\mathscr{P}^r\mathscr{S}(c)'$ in Lemma 2.7 gives:

$$[\mathscr{P}^r\mathscr{S}(c)]' = \bigoplus_{\dim f \geq 1} [E^r_{f\to c}\mathring{\mathscr{P}}^r\mathscr{S}(f)]'.$$

By Lemma 2.2 $\iota^*_f E^r_{f\to c}$ is identity on the bigger space $\mathscr{P}^r\mathscr{S}(f)$. The first part of this lemma then implies for each k-face f,

$$[E^r_{f\to c}\mathring{\mathscr{P}}^r\mathscr{S}(f)]' = \left\{u \mapsto \int_f (\iota^*_f u) : q \mid q \in \mathscr{P}^{r-k+1}\mathscr{S}(f)\right\}.$$

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 2.1).

Lemma 2.2, the geometric decomposition (Lemma 2.7), and the characterization of bubbles (Lemma 2.8) combined implies the locality result (Theorem 2.2).

Finally, the characterization result (Theorem 2.3) is proved. Suppose $u \in \mathrm{REG}^r(\mathcal{T})$ for some mesh $\mathcal{T}$. By locality (Theorem 2.2), 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 $u \in \mathcal{P}^r\mathcal{S}(\mathcal{T})$ 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 $u' \in \mathrm{REG}^r(\mathcal{T})$. By unisolvency, the restrictions $u'|_c = u_c$ agree on each cell c of the mesh. Hence $u = u' \in \mathrm{REG}^r(\mathcal{T})$.

2.3 Affine and approximation properties

We prove two more important results on REG^r in this section. First, in Theorem 2.4, we show that REG^r forms an affine family of finite elements for any fixed dimension $n \geq 1$. Such affine families have many advantages [30, Section 2.3]. For example, for software implementations, this makes the assembly of bilinear forms involving such finite elements very efficient [75, Chapter 6]. Second, in Theorem 2.5 and Theorem 2.6, we prove the optimal approximation properties of the canonical interpolant induced by the degrees of freedom (2.4b), as a consequence of the affine property. Both results require some preparations to state precisely.

2.3.1 Affine property

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

$$\bar{r}_{\bar{f}}(\bar{V}) = \phi^*(r_f(V)) \qquad \text{and} \qquad \phi_*(\bar{\Sigma}_{\bar{f}}) = \Sigma_f,$$

where the ϕ_* is defined naturally: for any $\bar{l} \in \bar{\Sigma}_{\bar{f}}$ and $u \in r_f(V)$,

$$(\phi_* \bar{l})(u) = \bar{l}(\phi^* u).$$

This definition is adapted from the classical definition of equivalence of finite elements [17, Section 3.4]. A finite element family F is an *affine family* [30, 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 $\bar{c} = [\bar{v}_0, \ldots, \bar{v}_n]$ 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 $\{\bar{v}_1 - \bar{v}_0, \ldots, \bar{v}_n - \bar{v}_0\}$ 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 $\{\bar{v}_1 - \bar{v}_0, \ldots, \bar{v}_n - \bar{v}_0\}$ and $\phi : \mathbb{R}^n \to \mathbb{R}^n$ an affine map given by

$$\phi(x) := A(x - v_0) + \bar{v}_0. \tag{2.18}$$

It is clear that ϕ maps c to $\bar{c}$ 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

$$\hat{v}_0 = [0, \ldots, 0], \quad \hat{v}_1 = [1, 0, \ldots, 0], \quad \hat{v}_2 = [0, 1, 0 \ldots, 0], \quad \ldots, \quad \hat{v}_n = [0, \ldots, 0, 1], \tag{2.19}$$

is chosen to be the representative. This $\hat{c}$ is referred to as the *reference n-simplex.*

Given these definitions, we state the affine property of REG^r.

Theorem 2.4. *Fix any $r \geq 0$ and $n \geq 1$. Let $\hat{c}$ be the reference n-simplex. For any n-simplex c, $\mathrm{REG}^r(c)$ is affine equivalent to $\mathrm{REG}^r(\hat{c})$. Thus the restriction of generalized Regge family to simplices of the same dimension forms an affine family.*

To prove this theorem, we need the following lemma:

Lemma 2.9. *Let $\bar{c}$ and c be two n-simplices and ϕ the linear isomorphism mapping $\bar{c}$ to c defined in equation (2.18). For any face $\bar{f}$ of $\bar{c}$, $f := \phi(\bar{f})$ is a face of c. For any $u \in \mathscr{S}(c)$,*

$$\phi^* \iota_f^* u = \iota_{\bar{f}}^* \phi^* u.$$

Moreover, for any $u \in \mathscr{S}(c)$ and $v \in \mathscr{S}(\bar{c})$,

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

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

Proof. First since ϕ maps vertices to vertices, the image $f = \phi(\bar{f})$ is indeed a face of c. Identify the tangent space to $\bar{f}$ (or f) as a subspace of that of $\bar{c}$ (or c). Then $\phi^* \iota_f^* = \iota_{\bar{f}}^* \phi^*$ on $\mathscr{S}(c)$. 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 2.4. Let ϕ 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 $\phi^*\mathscr{P}^r\mathscr{S}(c)$ are still polynomials of degree r. The invertibility of $d\phi$ then implies that $\phi^*\mathscr{P}^r\mathscr{S}(c) = \mathscr{P}^r\mathscr{S}(\hat{c})$. Finally, the conditions on degrees of freedom have to be checked. First, this ϕ 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 $(\hat{r}_{\hat{f}}, \hat{\Sigma}_{\hat{f}}) \in \hat{\Sigma}$ and $(r_f, \Sigma_f) \in \Sigma$ satisfies

$$\hat{r}_{\hat{f}}(\hat{V}) = \mathscr{P}^r\mathscr{S}(\hat{f}) = \phi^*[\mathscr{P}^r\mathscr{S}(f)] = \phi^*[r_f(V)].$$

Note that in Lemma 2.9, $(\phi^*)^T$ 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 $\mathscr{P}^s\mathscr{S}(\hat{f})$ and $\mathscr{P}^s\mathscr{S}(f)$ for any integer s. Thus by definition of the degrees of freedom (2.4b), $\phi_*\hat{\Sigma}_{\hat{f}} = \Sigma_f$ as required. This proves the claim. □

2.3.2 Approximation properties of the canonical interpolant

We prove the optimal error rates for the canonical interpolant for REG^r.

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

In order to define the Sobolev spaces, a background Riemannian manifold is needed [12]. 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 $\mathrm{REG}^r(\mathscr{T})$, can serve as Riemannian metrics if it is everywhere positive definite. In the geometry literature (for example [25]), 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, $\mathrm{REG}^r(\mathscr{T})$ have many applications where it is not used as a discrete metric. So the error measured in the Euclidean background is always available.

Let Ω be a Lipschitz polytope in $\mathbb{R}^n$. We denote the standard Sobolev spaces [2] for scalar-valued functions on Ω by $W^{s,p}(\Omega)$. Under the background Euclidean coordinates, tensor fields acquire explicit components. For example, $\mathscr{S}(\Omega)$ becomes the space of $\mathbb{S}^n$-valued functions. We define the Sobolev space of tensor fields in a componentwise manner. In particular, the space of $W^{s,p}$-symmetric tensor fields on Ω, denoted by $W^{s,p}\mathscr{S}(\Omega)$, is identified with the Bochner space of $\mathbb{S}^n$-valued $W^{s,p}$-functions. More explicitly, for $g \in \mathscr{S}(\Omega)$, let $\{g_i\}_{i=1}^{n(n+1)/2}$ be its components in the background Euclidean space. The $W^{s,p}\mathscr{S}$-semi-norm of g is

$$|g|^p_{W^{s,p}\mathscr{S}} = \sum_{i=1}^{n(n+1)/2} |g_i|^p_{W^{s,p}},$$

where each component is treated as a scalar function on Ω. In the literature, the $W^{s,p}\mathscr{S}$-semi-norm is sometimes equivalently defined via the sum of the pointwise norms of the derivatives of g (for example, in [26, Appendix 1]). The componentwise approach is taken here because it is more convenient to apply theorems concerning scalar-valued Sobolev spaces.

For $u \in \mathscr{S}(\Omega)$, from the definition, the canonical interpolant $I^r_{\mathcal{T}}u$ is a piecewise polynomial which is discontinuous across the interior facets. By the standard trace theorems [45, Theorem 1.5.1.2],

$$I^r_{\mathcal{T}}u \in W^{s,p}\mathscr{S}(\Omega), \qquad \forall s \in [0, 1/p).$$

On the other hand, from the definition of the degrees of freedom (2.4b), the interpolation operator $I^r_{\mathcal{T}}$ needs the integral of the function restricted to k-faces for $1 \le k \le n$. By the trace theorems again, $I^r_{\mathcal{T}}$ can be extended from $\mathscr{S}(\Omega)$ boundedly to

$$I^r_{\mathcal{T}} : W^{s,p}\mathscr{S}(\Omega) \to \mathrm{REG}^r(\mathcal{T}), \qquad \forall s \in ((n-1)/p, \infty]. \tag{2.20}$$

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 c, the *inradius* ρ_c is the Euclidean diameter of the inscribing sphere of c, and the *shape constant* σ_c is defined to be the ratio h_c/ρ_c. The shape constant quantifies how far away the simplex c is from being degenerate (with vertices no longer of general position).

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

Lemma 2.10. *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*

$$\|d\phi\| \le \frac{\sqrt{n}h_{c'}}{\rho_c}, \qquad \|d\phi^{-1}\| \le \frac{\sqrt{n}h_c}{\rho_{c'}},$$

where the norm is the Euclidean Frobenius norm.

Proof. This result for the Euclidean operator 2-norm is well-known [30, 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 ρ_c can be realized as the difference between two points in c. Their images are at most $h_{c'}$ apart under ϕ, which proves the claim in the operator norm. For the invertible $d\phi$, let $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_n > 0$ be its singular values. It is well-known [43, Corollary 2.4.3] that the operator norm of $d\phi$ is σ_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 $W^{s,p}\mathscr{S}(\Omega)$-norm. This will be a key step in the proof of the approximation theorem.

Lemma 2.11. *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 $W^{s,p}\mathscr{S}(c)$ and $\bar{g} := \phi^* g$. Then, there exists a constant $C = C(n,s)$ such that*

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

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 [43, 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 $W^{s,p}\mathscr{S}(\Omega)$-semi-norm is estimated by:

$$|\bar{g}|_{W^{s,p}\mathscr{S}(\bar{c})} \leq \|d\phi\|^2 |(g \circ \phi)|_{W^{s,p}\mathscr{S}(\bar{c})}.$$

Recall the classical scaling result [30, Theorem 3.1.2] for scalar-valued functions $u \in W^{s,p}(c)$:

$$|u \circ \phi|_{W^{s,p}(\bar{c})} \leq C\|d\phi\|^s |\det(d\phi)|^{-1/p} |u|_{W^{s,p}(c)},$$

where $C = C(s,n)$. Since each component of $u \circ \phi$ is just a scalar function,

$$|(g \circ \phi)|_{W^{s,p}\mathscr{S}(\bar{c})} \leq C\|d\phi\|^{s}|\det(d\phi)|^{-1/p}|g|_{W^{s,p}\mathscr{S}(c)},$$

where the constant C depends on s,n. This proves the claim for ϕ^*. The same result applied to the inverse gives the second estimate. □

Compared with the classical scaling result [30, 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 [105, Theorem 5], which was used in [54] to derive estimates for Finite Element Exterior Calculus. It should be noted that in [54, 105], 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, we get an estimate for the canonical interpolant.

Theorem 2.5. *Let c be any n-simplex and I_c^r the REG^r canonical interpolant for any $r \geq 0$. Suppose $p \in [1,\infty]$ and $s \in ((n-1)/p, r+1]$. Then, for any $t \in [0,s]$, there exists a constant $C = C(n,r,t,s) > 0$ such that*

$$|g - I_c^r g|_{W^{t,p}\mathscr{S}(c)} \leq C\sigma_c^{t+2}h_c^{s-t}|g|_{W^{s,p}\mathscr{S}(c)}, \qquad \forall g \in W^{s,p}\mathscr{S}(c).$$

Compared with classical results for scalar-valued functions [17, Theorem 4.4.4], the only difference is that for covariant 2-tensors the exponent for σ_c is $(t+2)$ while for scalar functions it is t. This means that the approximation properties of $\mathrm{REG}^r(c)$ are more sensitive to the shape of c. In particular, while for scalar functions, the L^p-estimates are independent of σ_c, for $\mathrm{REG}^r(c)$ the L^p-estimates are still degraded if σ_c is large.

Proof of Theorem 2.5. 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 [16] and [30, 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:

$$\inf_{p \in \mathscr{P}^r(\hat{c})} \|u - p\|_{W^{r+1,p}} \leq C|u|_{W^{r+1,p}}, \qquad \forall u \in W^{r+1,p}(\hat{c}).$$

This implies a similar result for our Bochner space:

$$\inf_{p \in \mathscr{P}^r(\hat{c}) \otimes \mathbb{S}^n} \|u - p\|_{W^{r+1,p}\mathscr{S}(\hat{c})} \leq C|u|_{W^{r+1,p}\mathscr{S}(\hat{c})}, \qquad \forall u \in W^{r+1,p}(\hat{c}, \mathbb{S}^n),$$

where C depends on r and n. It was already shown before in (2.20) that for $s > (n-1)/p$, the canonical interpolant $I^r_{\hat{c}} : W^{s,p}\mathscr{S}(\hat{c}) \to \mathscr{P}^r\mathscr{S}(\hat{c})$ is bounded. The quantity

$$\|I^r_{\hat{c}}\|_{W^{r+1,p}\mathscr{S}(\hat{c})\to W^{s,p}\mathscr{S}(\hat{c})}$$

is just a constant depending only on r, n, s. It is also clear that $I^r_{\hat{c}}$ is a projection which preserves $\mathscr{P}^r\mathscr{S}(\hat{c})$. Hence, whenever $s \in ((n-1)/p, r+1]$, for any $t \in [0,s]$, there exists a constant $C = C(r,n,s,t)$ such that,

$$\begin{aligned}\|u - I^r_{\hat{c}}u\|_{W^{t,p}\mathscr{S}(\hat{c})} &= \inf_{p\in\mathscr{P}^r\mathscr{S}(\hat{c})} \|(I - I^r_{\hat{c}})(u-p)\|_{W^{t,p}\mathscr{S}(\hat{c})} \\ &\le \|(I - I^r_{\hat{c}})\|_{W^{s,p}\mathscr{S}(\hat{c})\to W^{t,p}\mathscr{S}(\hat{c})} \inf_{p\in\mathscr{P}^r\mathscr{S}(\hat{c})} \|(u-p)\|_{W^{t,p}\mathscr{S}(\hat{c})} \\ &\le C|u|_{W^{s,p}\mathscr{S}(\hat{c})}, \qquad \forall u \in W^{s,p}\mathscr{S}(\hat{c}).\end{aligned}$$

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 $u \in W^{s,p}\mathscr{S}(c)$ and $t \in [0,s]$, the second estimate in Lemma 2.11 implies

$$\|u - I^r_c u\|_{W^{t,p}\mathscr{S}(c)} \le C_1 \|d\phi^{-1}\|^{t+2} |\det(d\phi)|^{1/p} |\phi^*(u - I^r_c u)|_{W^{t,p}\mathscr{S}(\hat{c})},$$

where $C_1 = C_1(n,t)$. Crucially, the affine property (Theorem 2.4) implies that the canonical interpolation operator commutes with pullbacks $I^r_{\hat{c}}\phi^* = \phi^* I^r_c$. Thus, using the estimate for $I^r_{\hat{c}}$ in the previous step,

$$\begin{aligned}\|u - I^r_c u\|_{W^{t,p}\mathscr{S}(c)} &\le C_1 \|d\phi^{-1}\|^{t+2} |\det(d\phi)|^{1/p} |\hat{u} - I^r_c \hat{u}|_{W^{t,p}\mathscr{S}(\hat{c})} \\ &\le C_1 C_2 \|d\phi^{-1}\|^{t+2} |\det(d\phi)|^{1/p} |\hat{u}|_{W^{s,p}\mathscr{S}(\hat{c})},\end{aligned}$$

where $C_2 = C_2(r,n,s,t)$. Applying the first estimate in Lemma 2.11,

$$\|u - I^r_c u\|_{W^{t,p}\mathscr{S}(c)} \le C_1 C_2 C_3 \|d\phi^{-1}\|^{t+2} \|d\phi\|^{s+2} |u|_{W^{s,p}\mathscr{S}(c)},$$

where C_3 depends on n and s. Finally, this, along with the estimates for $\|d\phi\|$ and $\|d\phi^{-1}\|$ in Lemma 2.10, leads to the estimate in the claim. □

Last, for a mesh $\mathscr{T}$, the *mesh size* $h(\mathscr{T})$ and the *shape constant* $\sigma(\mathscr{T})$ are defined as the maximum of the cell-wise h_c and σ_c over all cells c in $\mathscr{T}$. We have the following very useful approximation theorem for REG^r.

Theorem 2.6. *Let Ω be a bounded Lipschitz polytope in $\mathbb{R}^n$ and $\{\mathscr{T}_h\}$ be a sequence of triangulations of Ω indexed by mesh size h with uniformly bounded shape constants* $\sup_h \sigma(\mathscr{T}_h) =:$

$\sigma < \infty$. The canonical interpolant I_h^r for $\mathrm{REG}^r(\mathscr{T}_h)$ *defined on smooth $\mathscr{S}(\Omega)$ can be extended boundedly to $W^{s,p}\mathscr{S}(\Omega)$ 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*

$$|g - I_h^r g|_{W^{t,p}\mathscr{S}(\Omega)} \leq C h^{s-t} |g|_{W^{s,p}\mathscr{S}(\Omega)}, \qquad \forall g \in W^{s,p}\mathscr{S}(\Omega).$$

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 2.5 to each cell c in $\mathscr{T}_h$ where $\sigma_c \leq \sigma$ is absorbed into the constant. Sum over all the cells in $\mathscr{T}_h$ leads to the estimate in the claim:

$$|g - I_h^r g|^p_{W^{t,p}\mathscr{S}(\Omega)} = \sum_{c \in \mathscr{T}_h} |g - I_h^r g|^p_{W^{t,p}\mathscr{S}(c)} \leq C^p h^{p(s-t)} \sum_{c \in \mathscr{T}_h} |g|^p_{W^{s,p}\mathscr{S}(c)} = C^p h^{p(s-t)} |g|^p_{W^{s,p}\mathscr{S}(\Omega)}.$$

□

2.4 Coordinate representations and implementable degrees of freedom

The first part of this section describes REG^r in coordinates. The most important results being equation (2.21) for the pullback in coordinates and Proposition 2.7 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 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 REG^0.

2.4.1 Coordinate representations

So far, 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 REG^r for concrete symmetric-matrix-valued functions. In this section, formulae for the coordinate representation are derived.

First, the standard coordinate identification (see for example [113]) 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 $\{dx_i \otimes dx_j\}_{1\le i,j\le n}$. Under this, covariant 2-tensor fields are identified with n-by-n matrix-valued functions and elements of $\mathscr{S}(\mathbb{R}^n)$ are identified with symmetric-matrix-valued functions. The evaluation of $g \in \mathscr{S}(\mathbb{R}^n)$ 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 Ω or a mesh $\mathscr{T}$, all these identifications are inherited. For a mesh $\mathscr{T}$ in $\mathbb{R}^n$, this is global in the sense that the same basis are used for all the cells. Under this, $\mathrm{REG}^r(\mathscr{T})$ 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:

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

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

$$(\phi^* g)(x) = [d\phi(x)]^T [g \circ \phi(x)][d\phi(x)]. \tag{2.21}$$

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 $c = [v_0 \dots v_n]$ be an n-simplex in $\mathbb{R}^n$ and f be a k-face with $1 \le k < n$. 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 $\{e_i \odot e_j\}_{1\le i\le j\le n}$. 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:

Proposition 2.7. *Let $c = [v_0 \dots v_n]$ be an n-simplex and $\{\lambda_i\}_{i=0}^n$ its barycentric coordinates. For any edge $e = [v_i v_j]$ of c, let*

$$g_e := -d\lambda_i \odot d\lambda_j, \qquad v_e := v_j - v_i.$$

Then,

$$\{g_e \mid e \subset c\} \textit{ is a basis for } \mathbb{S}^n, \qquad \{v_e \odot v_e \mid e \subset c\} \textit{ is a basis for } \mathbb{V}^n. \tag{2.22}$$

Further these two basis are dual to each other:

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

Proof. The fact the $\{g_e\}$ forms a basis for $\mathbb{S}^n$ has already been proved in Proposition 2.5. For $\{v_e \odot v_e\}$, it is enough to show that for any $g \in \mathbb{S}^n$, knowing the values

$$\{g(v_e, v_e) \mid e \subset c\}.$$

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 Proposition 2.5. □

Under the barycentric basis, the tangential-tangential pullback has a canonical representation. Indeed, each function $g \in \mathscr{S}(c)$ is an $\mathbb{R}^{n(n+1)/2}$-valued function where the components are indexed by edges of c. By the dual structure and Proposition 2.6, 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:

$$\iota_f^*(\sum_{e \subset c} a_e g_e) = \sum_{e \subset f} a_e g_e,$$

The classical REG^0 used in Regge Calculus is also given in this basis.

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.

2.4.2 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 [75] 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

$$\left\{u \mapsto \int_f (\iota_f^* u) : q \mid q \in \mathscr{P}^{r-k+1}\mathscr{S}(f)\right\}$$

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 $\iota_f^* u$ 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 $(\mathbb{S}^k)' = \mathbb{V}^k$,

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

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

$$(\iota_f^* u) \cdot (v_e \odot v_e) = u(v_e, v_e) = v_e^T u v_e,$$

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:

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

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 $\mathscr{P}^{r-k+1}(f)$. 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:

$$\{u \mapsto (v_e^T u v_e)(x) \mid e \subset f, x \in X^f_{r-k+1}\}. \tag{2.23}$$

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. 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* $X_r^{\hat{f}}$ is given by

$$X_r^{\hat{f}} := \left\{ \left(\frac{m_1}{r+2}, \dots, \frac{m_k}{r+2}\right) \middle| \, m_j \in \mathbb{Z} \text{ and } m_j \geq 1 \text{ for } j = 1, \dots, k. \sum_{j=1}^{k} m_j \leq r+1 \right\}. \tag{2.24}$$

Pictorially, this for various values of k and r are depicted in Figure 2.10.

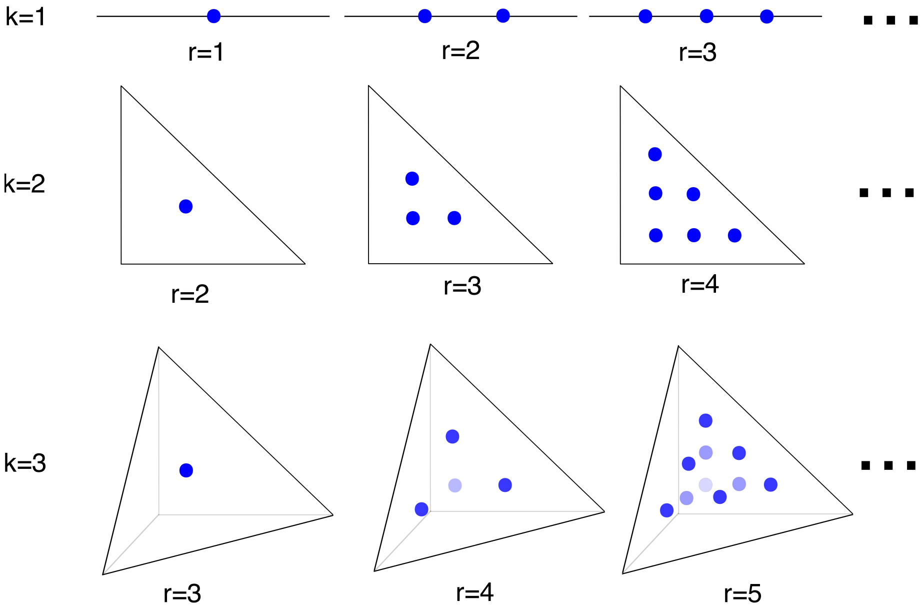


Figure 2.10: Pictures for $X_r^{\hat{f}}$.

On a general k-face f of an n-simplex c, the *equi-distance* X_r^f is defined as the image of $X_r^{\hat{f}}$ under the affine isomorphism ϕ mapping $\hat{f}$ to f. It is well-known that pointwise evaluation at points in $X_r^{\hat{f}}$ are linearly independent and forms a dual basis to $\mathscr{P}^r(f)$ [30].

There is another choice of X_r^f 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_r^f is best described with pictures. See examples for 2D in Figure 2.11 and for 3D in Figure 2.12.

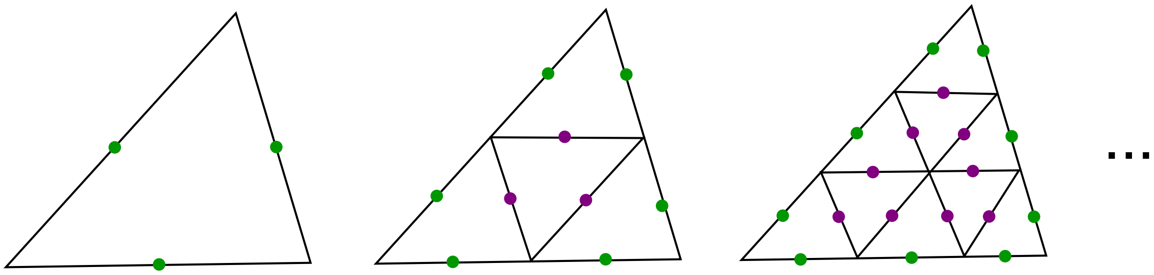


Figure 2.11: Subdivision-based degrees of freedom in 2D.

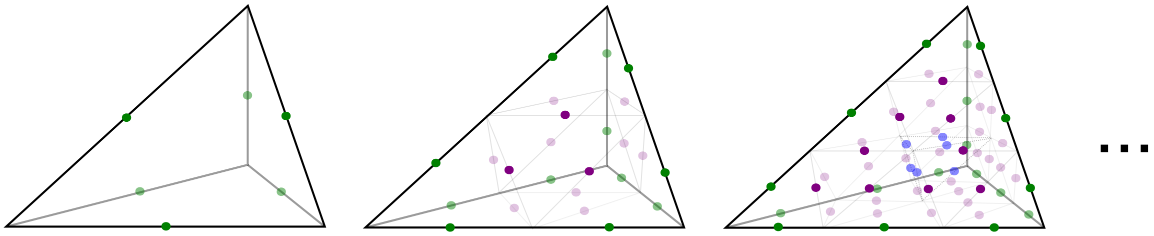


Figure 2.12: Subdivision-based degrees of freedom in 3D.

The pattern for higher dimensions is clear.

The subdivision-based X_r^f in fact leads to another set of degrees of freedom for REG^r. 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:

$$u \mapsto \int_0^1 (p_2 - p_1)^T u(p_1 + t(p_2 - p_1))(p_2 - p_1)\,dt. \tag{2.25}$$

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 $\mathrm{REG}^r(c)$. Indeed, the edge whose mid-point is inside a face f of c must be parallel to one of the undivided 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 $\mathscr{P}^{r-k+1}(f)$ as before.

This has a nice geometric interpretation: $\mathrm{REG}^r(c)$ 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 $\mathrm{REG}^r(c)$. This is the most geometric interpretation that clearly shows that $\mathrm{REG}^r(c)$ generalizes $\mathrm{REG}^0(c)$ 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 [110]. This generalization is much more natural and elegant.

It should be noted that both choices of X_r^f above are known to be not optimal [112]. The performance of choices of the degrees of freedom can be evaluated quantitatively by the *Lebesgue constant*. For a set of degrees of freedom Σ, let I_Σ be the induced interpolant. The Lebesgue constant Λ_Σ is the smallest constant such that the following holds for all smooth u:

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

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 $\mathscr{P}^r(c)$ with the smallest possible Lebesgue constant is known [112]. Further, in the same paper, it was shown that the first choice of X_f^r with the equi-distance lattice points (2.24) is not good for $r \geq 10$. In practice though, the optimal X_f^r is messy to implement and for real problems, degree r more than 3 is rarely used. So the easier equi-distance X_f^r was chosen for the software implementation of REG^r 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. We call everywhere positive definite functions in REG^r *generalized Regge metrics*, or simply REG^r *metrics*. In this chapter, we define and study geodesics on REG^r metrics.

The piecewise constant REG^0 has been studied extensively in the literature as a discrete model of geometry. Historically, Riemannian metrics in REG^0 are important in the mathematical study of Euclidean polyhedrons and are referred to as *polyhedral metrics* [6]. In General relativity, Lorentzian metrics, which are symmetric covariant 2-tensor fields similar to Riemannian metrics, play a central role. Tullio Regge used Lorentzian metrics in REG^0 to derive a geometric discretization of the Einstein field equation in his influential work [96].

In this chapter, we focus 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 the context of Riemannian geometry, geodesics are basic objects for quantifying and characterizing geometry. We examine various mathematical and computational aspects of geodesics on generalized Regge metrics in this chapter. Geodesics on discrete geometries are of considerable practical interest. Geodesics on 2D triangulations embedded in 3D Euclidean space are important in computer graphics [51] and computer-aided design [73]. In relativity, geodesics model the trajectories of light rays and free-falling test particles [111]. After Regge's initial work, physicists explored the interpretation and computation of geodesics on Regge metrics [18, 23, 118, 119]. Finally, in other parts of this thesis, REG^r 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 [55].

The theory of smooth geodesics on smooth Riemannian manifolds is well-understood [34]. We review this 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^r 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^r metrics. In Section 3.3, we study this in detail. The main result is that REG^r metrics have a well-behaved length structure, under which geodesics are well-defined. Further, REG^r metrics are the least smooth (thus the most general) piecewise polynomial Riemannian metric for which the usual sense of curve length is preserved (Theorem 3.1).

The local theory of geodesics turned out be subtle for REG^r 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^0, two seemingly unrelated ideas for resolving this are popular in the literature. The first idea [6] is of a geometrical nature. Take a 2D REG^0 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^0 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 [25]. In Section 3.5, this idea will be generalized to REG^r for all $r \geq 1$. The general case is quite subtle. For REG^r metrics, the mesh can be given a metric-dependent piecewise smooth globally C^1 atlas, 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.

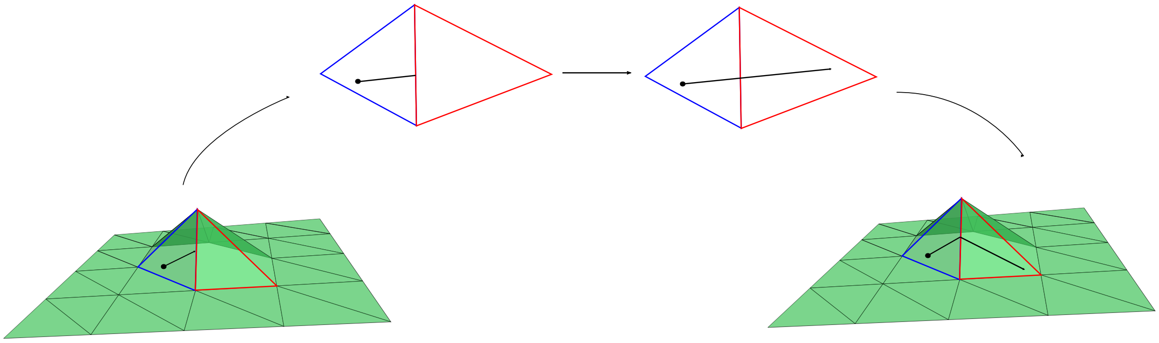


Figure 3.1: Illustration of the first geodesic idea.

The second idea [118] is motivated by physical interpretations. For 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. We show 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 REG^r in all dimensions and for all $r \geq 0$. We derive derive it rigorously in Section 3.4. Thus the geodesics are usual smooth geodesics inside each cell and rotates in this way when they cross interior facets.

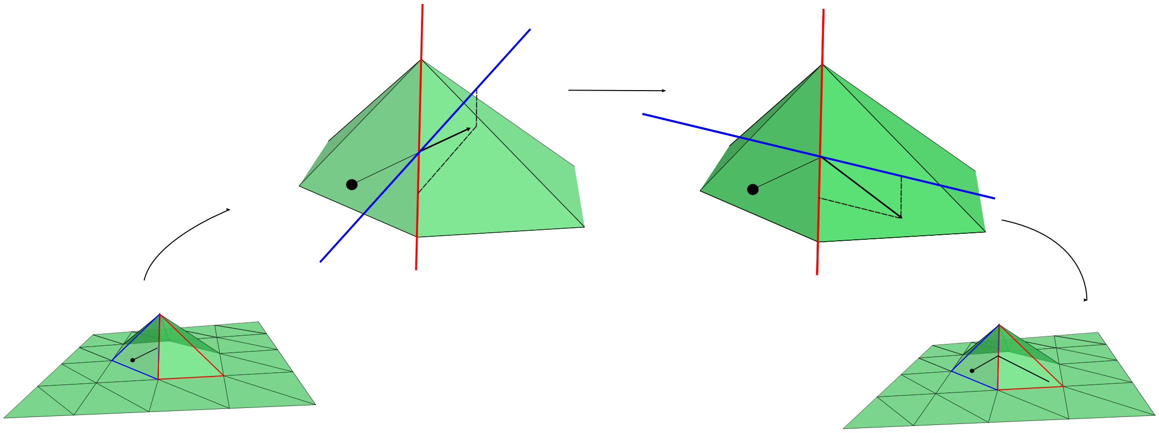


Figure 3.2: Illustration of the second geodesic idea. In the middle two figures, the red line indicates the tangential direction while the blue indicates the normal direction.

We prove prove that the geometric approach and the variational approach lead to the same definition for local geodesics on REG^r metrics (Theorem 3.5). While the variation approach is easy to understand and use, the more abstract geometric approach reveals some subtle structure of the geodesics. Moreover, we show in Section 3.6 that geodesics on REG^r metrics still have a symplectic structure in a subtle sense. This, for example, suggests that symplectic discretizations should be used to compute geodesics.

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 [6,93] 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. We discuss this in detail in Section 3.4.

In Section 3.7, we describe a robust algorithm to compute local geodesics on REG^r metrics. 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 negligible error. The author has implemented this algorithm is in Python using FEniCS as a part of this thesis.

In Section 3.8, we study the error between the smooth geodesic on a smooth Riemannian metric and the geodesics on a sequence of REG^r metrics approximating that metric. The main result is Theorem 3.6. 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 error estimates in Theorem 3.6 are effectively practical a priori error estimates between the smooth geodesic on the smooth metric and the computed numerical geodesic on the approximating REG^r 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 Kelperian orbit on the same mesh with REG^r for $r = 0, 1, 2, 3$. The exact orbit is periodic and follows an ellipse. The advantage of going to higher degrees is obvious.

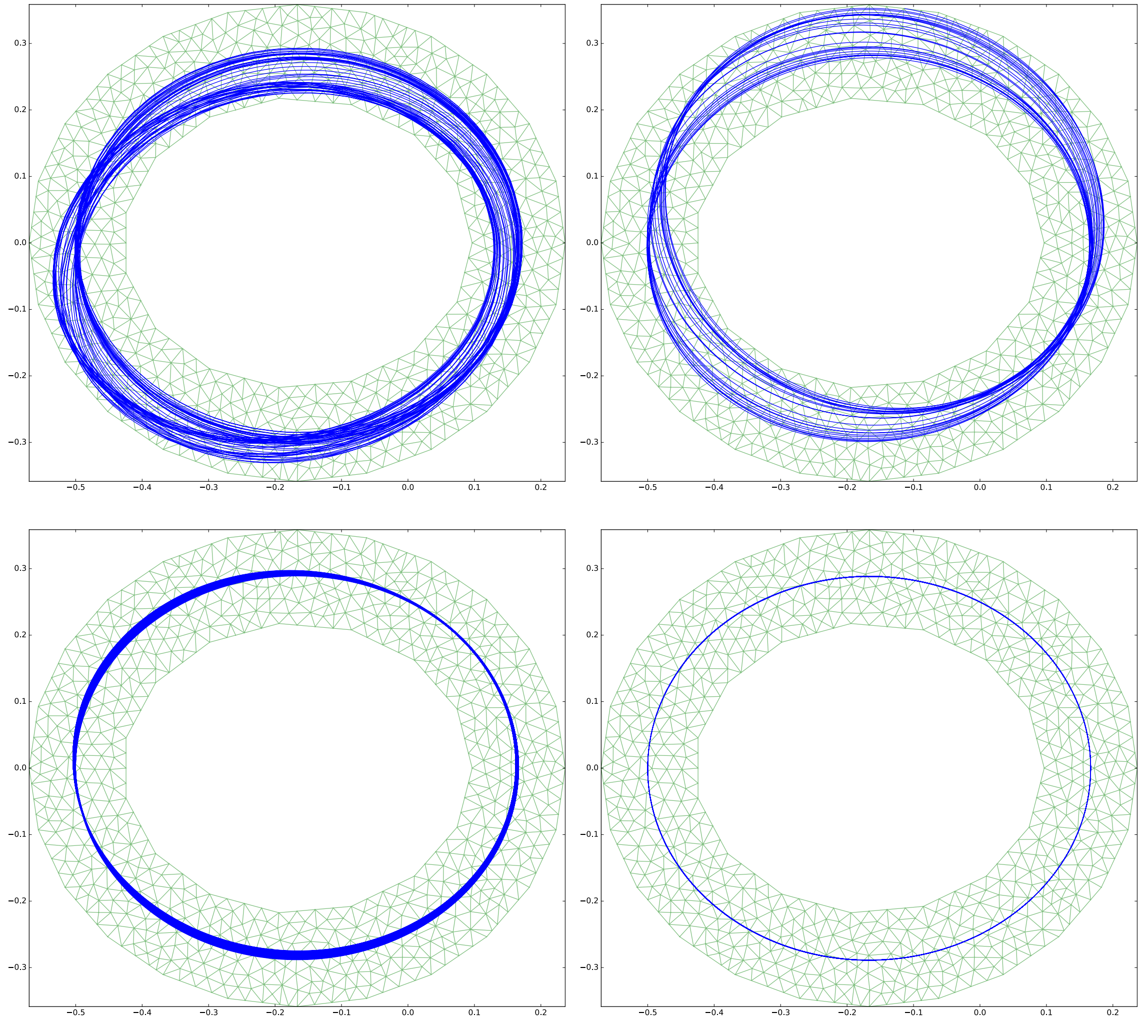


Figure 3.3: Keplerian orbits. Left to right, top to bottom $r = 0, 1, 2, 3$.

Figure 3.4 shows the result of the computation of 50 periods of a Schwarzschild orbit on the same mesh with 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 REG^r with higher degrees performs much better than REG^0. For REG^0, it is even difficult to tell if the orbit is Kelperian or Schwarzschild. This showcases the need for higher degree REG^r developed here for relativistic simulations.

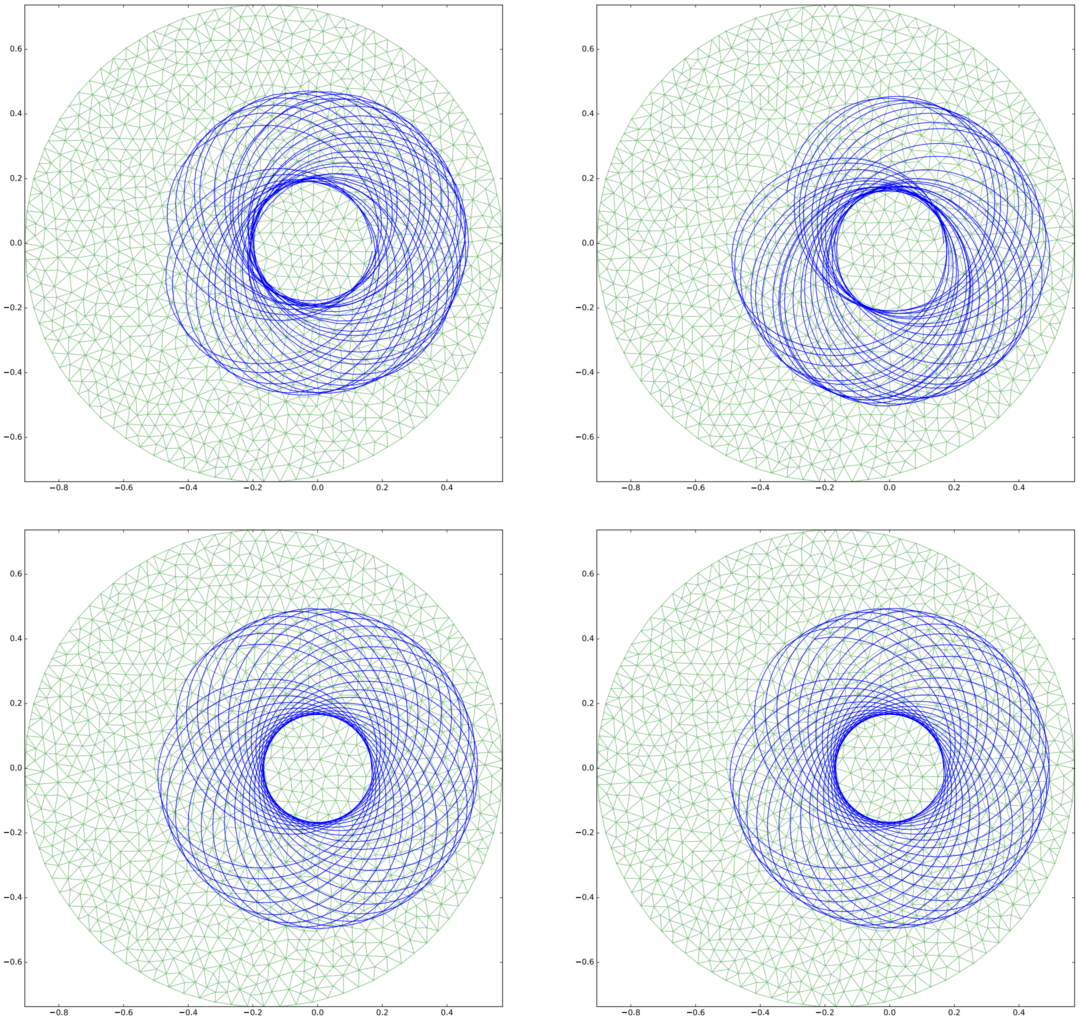


Figure 3.4: Schwarzschild orbits. Left to right, top to bottom, $r = 0,1,2,3$.

In the case of Kelperian 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 [48]. 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 an 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.

Method	Error in position	Error in energy	Error in momentum
Explicit Euler	$t^2 h_s$	th_s	th_s
Implicit Euler	$t^2 h_s$	th_s	th_s
Collocation at Gauss ($2r$)	th_s^{2r}	h_s^{2r}	0
Naive REG^r	$t^2 h^r$	th^r	th^r
REG^r	$(t+\epsilon t^2)h^{r+1}$	h^{r+1}	$(1+\epsilon t)h^{r+1}$

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 ϵ-terms for REG^r in the table. In practice, this effect is negligible, except for 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, we review basic facts of geodesics on smooth Riemannian manifolds.

Let (M,g) be a smooth Riemannian manifold. A *piecewise smooth curve* is a continuous function $\gamma : [a,b] \to M$ 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 γ is defined as

$$L(\gamma) := \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} \sqrt{g_{ij}(\gamma(t))\dot{\gamma}^i(t)\dot{\gamma}^j(t)}\,dt. \tag{3.1}$$

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

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

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

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

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. The parameterization is by arc length if and only 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

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

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

$$d(\gamma(t_1),\gamma(t_2)) = L(\gamma\,|_{[t_1,t_2]}) = c|t_1 - t_2|. \tag{3.2}$$

If γ 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 examples see [82]. 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*

$$E(\gamma) := \frac{1}{2}\sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} g_{ij}(\gamma(t))\dot{\gamma}^i(t)\dot{\gamma}^j(t)\,dt. \tag{3.3}$$

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 $\gamma : [a,b] \to M$,

$$L(\gamma)^2 \le 2(b-a)E(\gamma).$$

The equal sign holds if and only if the kinetic energy is constant. The following theorem is well-known ([34, Chpater 3 and 9]).

Proposition 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 γ is smooth and solves the* geodesic equation

$$\ddot{\gamma}^i + \Gamma^i_{kl}\dot{\gamma}^k\dot{\gamma}^l = 0, \tag{3.4}$$

where Γ^i_{jk} is the Christoffel symbol associated with g defined by

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

Moreover, such γ 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 initial-value problems given an initial position and velocity. Formally, given $p \in M$ and $v \in T_pM$, the *local geodesic problem* tries to find a smooth curve $\gamma(t)$ satisfying

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

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 [21, 46]. 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

$$L(\gamma) := \sup \sum d(\gamma(t_i), \gamma(t_{i+1})),$$

where the supremum is taken over all finite partitions $a = t_0 \le t_1 \le \cdots \le t_{n-1} \le t_n = b$ of $[a,b]$. When $L(\gamma) < \infty$, γ 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 *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 *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 *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:

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

Proof. From definition, $L(\gamma)$ is well-defined for a curve γ 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. By the characterization theorem of generalized Regge elements in Chapter 2, a piecewise polynomial covariant 2-tensor field on a mesh has single-valued pullbacks $\iota_f^* g$ for all faces of dimension ≥ 1 of the mesh if and only if $g \in \mathrm{REG}^r(\mathscr{T})$. This proves the first part. When $g \in \mathrm{REG}^r(\mathscr{T})$, 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 $(\mathscr{T}, d_g)$ is a length space by definition. □

For the lowest degree case REG^0, each simplex is flat. Riemannian metrics in REG^0 can be realized geometrically as triangulated polytopes embedded in some Euclidean space [25]. The global geodesics in this case have been studied extensively in mathematics [6] and in computational geometry [51, 73]. 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 2.

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 [62]. This is an extensively studied area in computational geometry. The generalized Regge case is future work.

3.4 Local geodesics on Regge metrics: variational approach

In this section, we define and study local geodesics on a generalized Regge metric. This is of particular interest in mathematical physics and numerical analysis.

Let $\mathscr{T}$ be a mesh and g a generalized Regge metric on $\mathscr{T}$. As described in the previous section, $(\mathscr{T}, g)$ is 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 $\mathscr{T}$ 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 $(\mathscr{T}, g)$. By Cauchy-Schwarz inequality again, it is clear that the above definition of a local geodesic is equivalent to requiring γ to be a critical point of the energy functional locally. This will be used to derive a local condition for local geodesics in Theorem 3.2. 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 nonsmooth 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. First we give some examples of the pathology. Then we give a more refined definition of a generalized local geodesic initial-value problem and describe its solution strategy.

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 $\mathscr{T}$, the state of the system is specified by a cell c of $\mathscr{T}$, a point $p \in c$, and a velocity vector $v \in T_p c$.

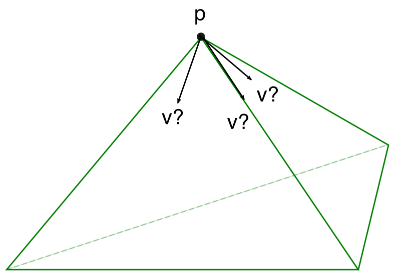


Figure 3.5: Failure of having a well-defined tangent space at a point.

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 REG^0, this is known in the computer graphics literature [93,94], but does not seem to be known in the physics literature. An example of this is illustrated below.

Proposition 3.2. *For a* REG^0 *metric on a 2D mesh, a curve passing through a vertex of positive angle deficit (the sum of angles around the vertex is smaller than* 2π*) cannot be a local geodesic.*

Proof. We call the vertex of positive angle deficit S and focus on the star of S (that is the union of all triangles intersect S). We call the star of S the tent. For example, the left panel of Figure 3.6 depicts a tent where S is surrounded by 4 triangles. Take any curve passing through S. If the curve lies entirely in one triangle, then it cannot be a geodesic, because a triangle is flat and geodesics are straight lines. Thus, the curve passes through from one triangle to another one. By the flatness of triangles again, within each triangle in order to be a local geodesic, the curve has to be a straight line. Hence we only need to consider the case where the curve is a piecewise straight line from one triangle to another turning at S. The left panel of Figure 3.6 shows such a generic situation. Take two points P and Q on the curve from the interior of the two triangles, say path $\overrightarrow{PSQ}$ connects $P \in \triangle ABS$ and $Q \in \triangle DCS$. We show that the curve $\overrightarrow{PSQ}$ cannot be a geodesic.

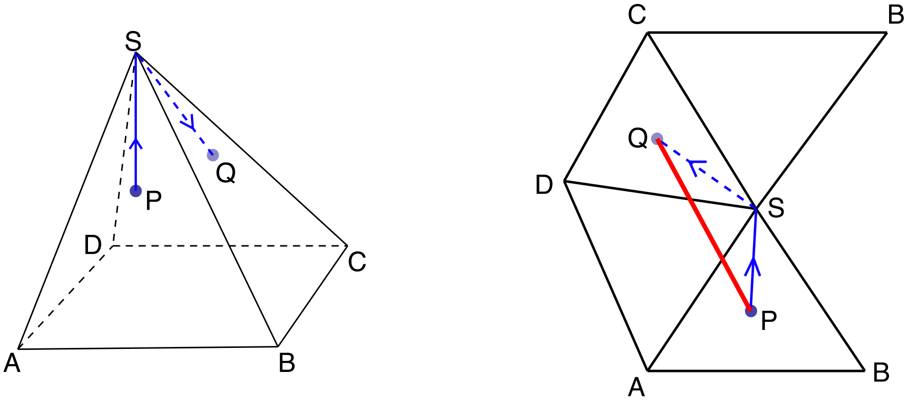


Figure 3.6: Pathology of generalized geodesic

Because of the positive angle deficit, we can always find an edge at S, say $\overline{SB}$ here, such that if we cut along that edge and flatten the tent then the line segment $\overline{PQ}$ lies completely inside the flattened triangles. The right panel of Figure 3.6 depicts such a flattened tent, where $\overline{SB}$ on the left is cut and becomes $\overline{SB}$ and $\overline{SB'}$ on the right. Note that this cut-and-flatten operation is an isometry. Using the triangle inequality in the flattened tent, it is clear that the length of $\overrightarrow{PQ}$ is shorter than the path $\overrightarrow{PSQ}$. Hence the original path $\overrightarrow{PSQ}$ cannot be locally distance minimizing and therefore not a local geodesic. □

Given the above proposition, if a local geodesic hits a vertex of positve angle deficit, then it cannot be extended further. A similar argument shows that if the angle deficit is negative, then a local geodesic has an infinite family of extensions. For the two-dimensional REG^0 case, various generalizations of the notion of local geodesics were proposed in the literature [6, 93], 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 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, we will show 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 REG^r are similar to their smooth counterparts. The next step is to prove an analog of Proposition 3.1 describing a local condition for local geodesics. In this case, there is nothing special about $\mathrm{REG}^r(\mathcal{T})$, 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. We consider this slightly more general case 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 γ 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 γ at p. Quantities like g_{ij}^-, n^-, and $\dot{\gamma}_-^i$ are similarly defined in c^-.

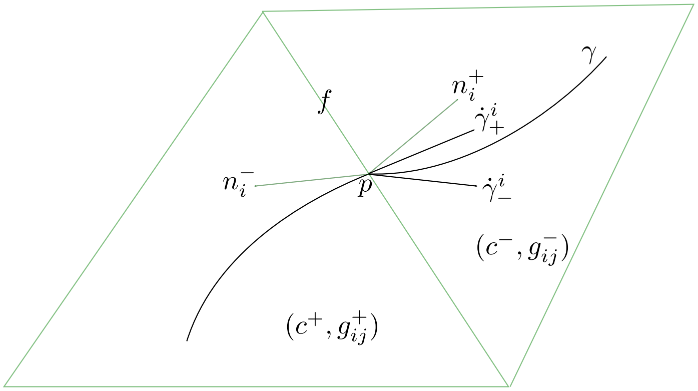


Figure 3.7: Definitions of quantities when a curve crosses an interior facet

Theorem 3.2. *Let $\mathcal{T}$ be a mesh of dimension n and g a piecewise smooth Riemannian metric with tangential-tangential continuity. A piecewise smooth curve $\gamma : [a,b] \to \mathcal{T}$ 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 γ intersects a facet f, the* *tangential projection of $\dot\gamma$ is the same on both sides: for all vectors $t^j \in T_p f$,*

$$g^+_{ij}\dot\gamma^i_+ t^j = g^-_{ij}\dot\gamma^i_- t^j$$

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, γ is $C^{0,1}$ globally. If g happens to be in C^k globally, $k \geq 0$, then γ is in $C^{k+1,1}$ globally. If g happens to be smooth, then γ is smooth and solves the usual smooth geodesic equation everywhere.

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

Corollary 3.1. *Suppose g is piecewise smooth with tangential-tangential continuity and γ 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:*

$$\dot\gamma^i_- = \dot\gamma^i_+ - (g^+_{jk}\dot\gamma^j_+ n^k_+)(n^i_+ + n^i_-), \tag{3.6}$$

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

$$t^i_+ = t^i_-, \qquad a_+ + a_- = 0.$$

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 *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 3.2 for non-smooth metrics are derived through variational methods [50,77], or through Filippov's theory for differential inclusions [104], or through the regularization of the metric [78, 79]. In another direction, similar results for curved interface were derived in [41]. The REG^0 case was derived in [118]. The case considered here has a simple proof and much stronger conclusions (namely uniqueness and regularity). To prove Theorem 3.2, the following lemma on the variation of the energy functional is needed.

Lemma 3.1. *Let $\mathscr{T}$ be a mesh of dimension n and g a piecewise smooth Riemannian metric. Suppose $\gamma : [a,b] \to \mathscr{T}$ is a piecewise smooth curve which does not cross interior faces of dimension $\leq (n-2)$ in $\mathscr{T}$. Let $\gamma_s(t) : (-\epsilon,\epsilon)\times[a,b] \to \mathscr{T}$ be a smooth family of variations:*

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

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 γ_s relative to γ:*

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

Then the variation of the energy functional is

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

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

Proof. This is just a direct computation from the definition

$$E(\gamma_s) = \sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}} g_{ij}\dot\gamma^i_s\dot\gamma^j_s\,dt,$$

and integration by parts. □

Then the main theorem of this section follows:

Proof of Theorem 3.2. Lemma 3.1 gives the condition for γ 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 γ to be smooth and solves the geodesic equation in the interior of each cell. At a point t_i where γ crosses an interior facet f at $p = \gamma(t_i)$, the conditions for critical points require

$$g^+_{ij}\dot\gamma^i_+\dot\gamma^j_+ = g^-_{ij}\dot\gamma^i_-\dot\gamma^j_-,$$

$$g^+_{ij}\dot\gamma^i_+ w^j = g^-_{ij}\dot\gamma^i_- w^j, \quad \text{for all } w^j \in T_p f.$$

The first equation implies the first condition in Theorem 3.2 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 γ_s crosses that facet. So the corresponding variational vector field must be tangential to f.

Now, set $a_\pm := g^\pm_{ij}\dot\gamma^i_\pm n^j_\pm$ and $t^i_\pm := \dot\gamma^i_\pm - a_\pm n^i_\pm$. The second condition and tangential-tangential continuity together imply that $t^i_+ = t^i_-$. Then the first condition implies

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

But γ 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, γ 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 REG^0. In the previous section, the variational approach was generalized to handle higher degree REG^r cases. In this section, the cut-flatten-glue approach is generalized to piecewise smooth metrics with tangential-tangential continuity, which include 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 [25]. 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 $\mathrm{REG}^0(\mathscr{T})$ 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 $\mathrm{REG}^0(\mathscr{T})$ of dimension n, let $\mathring{\mathscr{T}}$ be the manifold obtained by removing faces of dimension $\leq (n-2)$ from the mesh $\mathscr{T}$. Then an abstract smooth manifold $\mathring{\mathscr{T}}$ with the Euclidean metric can be obtained from $\mathrm{REG}^0(\mathscr{T})$ using a similar construction.

For the general case, we have the following.

Theorem 3.3. *Let $\mathscr{T}$ be a mesh of dimension n and g a piecewise smooth Riemannian metric with tangential-tangential continuity on $\mathscr{T}$. There exists an atlas depending on g for $\mathscr{T}$ which is piecewise smooth, globally C^1 on $\mathring{\mathscr{T}}$, and singular at $\mathscr{T} - \mathring{\mathscr{T}}$, under which the piecewise smooth metric g can be extended to a globally C^0-Riemannian metric on $\mathring{\mathscr{T}}$. Let $\mathring{\mathscr{T}}^g$ denote the C^1-manifold obtained from the topological manifold $\mathring{\mathscr{T}}$ with the aforementioned atlas. Then g is a C^0-Riemannian metric on $\mathring{\mathscr{T}}^g$ satisfying the condition that each cell in $\mathring{\mathscr{T}}^g$ is isometric to its corresponding cell in $(\mathscr{T}, g)$ via a smooth map whose differential is identity on vectors tangential to the boundary facets of each interior cell. Further such $(\mathring{\mathscr{T}}^g, g)$ is unique up to isometry.*

This theorem is a direct consequence of the gluing lemma [31] below.

Lemma 3.2. *Let $(M^\pm, g^\pm)$ be two smooth compact Riemannian manifolds with boundary, having smooth submanifolds $\Sigma^\pm$ of the boundaries $\partial M^\pm$ isometric to each other. Let M be the disjoint union of $M^\pm$ with $\Sigma^\pm$ identified via the isometry. Identify $M^\pm$ as subsets of M and let g be a piecewise function on M with $g = g^\pm$ depending on where g is evaluated. Then there exists a unique C^1 altas on M, which is compatible with the smooth atlas on $M^\pm$ and under which g can be extended to a C^0 Riemannian metric on M by continuity.*

Notice that for REG^0, as described before, $(\mathring{\mathscr{T}}^g, g)$ is a smooth manifold with a smooth (globally constant) Euclidean metric. For generalized Regge metrics 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 [71, 104]. In general, it was proved in [104] 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 [37]. 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 3.4. *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 $\gamma : [0,T] \to M$ by alternating between solving the smooth geodesic equation inside a cell and move to the next cell by continuity of γ and $\dot{\gamma}$. This process can go on as long as γ 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, γ 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. □

Local geodesics defined in this way agree with the local geodesics defined variationally in the previous section:

Theorem 3.5. *Let $\mathscr{T}$ be a mesh of dimension n and g a piecewise smooth Riemannian metric with tangential-tangential continuity on $\mathscr{T}$. Let $(\mathring{\mathscr{T}}^g, g)$ be the induced abstract Riemannian manifold and $\Phi : \mathring{\mathscr{T}}^g \to \mathscr{T}$ be the piecewise smooth isometry in Theorem 3.3. Take any cell c in $\mathscr{T}$, any point $p \in c \cap \mathring{\mathscr{T}}^g$, and any vector $v \in T_p c$. Let γ in $\mathring{\mathscr{T}}^g$ be the curve defined in Theorem 3.4 for $(\mathring{\mathscr{T}}^g, g)$ with initial data (q, v) and γ' in $\mathscr{T}$ be the local geodesic constructed using Theorem 3.2 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, Φ is a smooth isometry. By definition, γ and γ' are solutions to the same smooth geodesic equation with the same initial data. By standard ODE theory, γ and γ' 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 $(\mathring{\mathcal{T}}^g, g)$, because g is C^0, the geodesic equation (3.4) implies that $\dot{\gamma}$ is at least C^0 and therefore the solution γ is at least C^1. Hence, necessarily the kinetic energy and the facet tangential part of γ are preserved crossing f. These two conditions determines the velocity $\dot{\gamma}$ on the other side of the facet uniquely in $(\mathring{\mathcal{T}}^g, g)$. Both conditions are invariant under Φ. The preservation of these two are exactly the conditions for local geodesics in $(\mathcal{T}, g)$ in Theorem 3.2. 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 [11]. It is well-known that smooth local geodesics can also be formulated in the Hamiltonian framework [35, Section 28.3]. In numerical analysis, it is also well-known 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 [48].

In this section, we show 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.

First, we review the smooth case. 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

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

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

$$\begin{aligned} \dot{q}^i &= \frac{\partial H}{\partial p^i} = g^{ij} p_j, \\ \dot{p}_i &= -\frac{\partial H}{\partial q^i} = -\frac{1}{2} p_j p_k \partial_i g^{jk}. \end{aligned} \tag{3.7}$$

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 [48]. 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 $g_{ij}\dot{\gamma}^i\dot{\gamma}^j$ (Proposition 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 [48, 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)* ω is a 2-form on T^*M given by:

$$\omega := \sum_{i=1}^{n} dq^i \wedge dp_i. \tag{3.8}$$

It is easy to verify that ω is closed $d\omega = 0$ and *non-degenerate*: $\omega(u,v) = 0$ for all v if and only if $u = 0$ [11, 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 ω: for $u \in T(T^*M)$,

$$[J_\omega(u)](v) := \omega(u,v), \qquad \forall v \in T(T^*M).$$

Due to the non-degeneracy of ω, J_ω 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 ω is conserved along the flows of X_H:

$$\mathscr{L}_{X_H}\omega = \iota_{X_H} d\omega + d\iota_{X_H}\omega = d\iota_{X_H}\omega = d[J_\omega(X_H)] = ddH = 0, \tag{3.9}$$

where $\mathscr{L}_{X_H}$ is the Lie derivative, the first step uses Cartan's magic formula, the second step uses the fact that ω is closed, the third and fourth step use the definition of contraction and J_ω. The relevance of this to the current discussion is clear with a computation in coordinates. In the coordinates of (p_i, q_j), J_ω becomes a $2n$-by-$2n$ block matrix [11, 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 [48, 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 [48, Chapter X].

Let $\mathscr{T}$ 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 $(\mathscr{T}, g)$ 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 [78, 79]. 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 $(\mathring{\mathscr{T}}^g, g)$ be the abstract Riemannian manifold constructed in Theorem 3.3. Since $\mathring{\mathscr{T}}^g$ is only C^1 globally, its cotangent bundle $T^*\mathring{\mathscr{T}}^g$ is a C^0 manifold of dimension $2n$. In particular, it does not make sense to talk about vector fields and differential forms on $T^*\mathring{\mathscr{T}}^g$ directly. Let (q_i, p_j) be a local coordinate patch for $T^*\mathring{\mathscr{T}}^g$. 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 $T^*\mathring{\mathscr{T}}^g$ are only piecewise smooth. Because g is piecewise smooth and globally C^0, using a similar argument to the one used in Theorem 3.4, 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 ω be defined on $\mathring{\mathscr{T}}^g$ 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 $\mathrm{REG}^r(\mathscr{T})$ have a (metric dependent) symplectic structure.

Because the symplectic structure is defined with respect to the abstract metric dependent manifold $(\mathring{\mathscr{T}}^g, g)$, it is not immediately clear its explicit corresponding structure in the original computational coordinates $(\mathscr{T}, g)$. But this does show that on $(\mathscr{T}, g)$ it is possible to define a generalized Hamiltonian system using certain non-smooth theory for ODEs. This will not be pursued here. Given Theorem 3.5, 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 3.2 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 $\le (n-2)$. In practice, however, there are many problems due to numerical issues and practical concerns. In what follows, we describe and implement a robust method 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 $\le (n-2)$.

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 $\le (n-1)$, 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 ϵ (for example $\epsilon = 0.01$), $q' := q + \epsilon h v$ is computed. The cell with the minimum distance to its q' is chosen. If there are tied minimizers, a random choice is made.

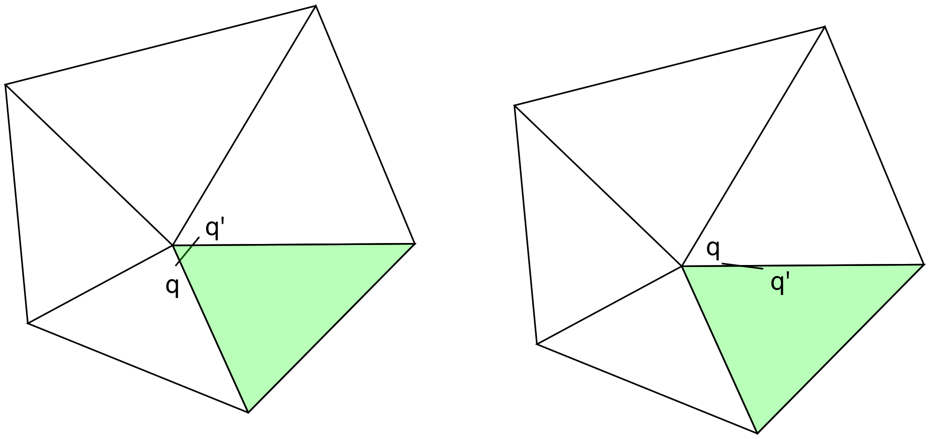


Figure 3.8: Possible bad initial conditions. The green cell is chosen.

For the next step, the geodesic ODE needs to be solved in the interior of each cell. For 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 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 [48]. 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) cannot be represented accurately in a finite element space for symbolic derivative computation. Instead, the gradient of the inverse metric is evaluated exactly via:

$$\partial_i g^{jk} = -g^{jm} g^{kn} \partial_i g_{mn}. \tag{3.10}$$

Second, due to performance concern, the collocation method is implemented via its equivalent Runge-Kutta method [48, Chapter II Theorem 1.4]. In practice, REG^r 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 [48, 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 [14] 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 [88]. 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 $\le (n-2)$, 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*.

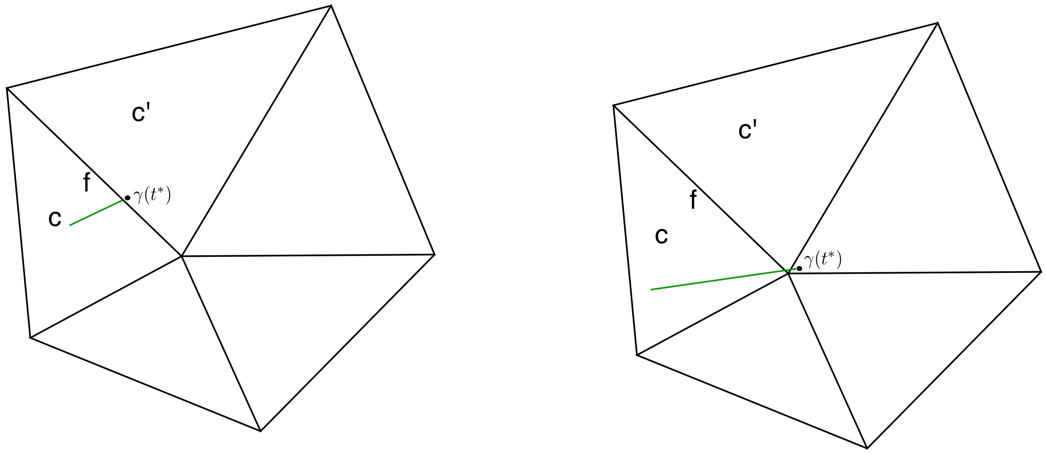


Figure 3.9: Left: a good crossing. Right: a bad crossing that needs a restart.

Finally, the momentum is rotated using the following formula derived from equation (3.6):

$$p_i^- = g_{ik}^-(g_+^{kj} - n_+^k n_+^j - n_-^k n_+^j)p_j^+. \tag{3.11}$$

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 [48, 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 `geodesics/regge_geodesics.py` in the companion code repository to this thesis. All the numerical examples later in this chapter are computed using this library.

3.8 Error analysis

Let (M,g) be a smooth Riemannian manifold and $\gamma : [a,b] \to M$ a smooth geodesic. Suppose $\{\mathscr{T}_h\}$ is a sequence of triangulations of M, on which $g_h \in \mathrm{REG}^r(\mathscr{T}_h)$ are Riemannian metrics and γ_h geodesics to $(\mathscr{T}_h, g_h)$ with the same initial data as γ. We study when γ_h is close to γ 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 $\mathscr{T}$ is given as an embedded manifold in some $\mathbb{R}^n$, the mesh size of $\mathscr{T}$ 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 $\mathscr{T}$, notations like $\|u\|_{W^{s,p}(\mathscr{T})}$ mean the piecewise $W^{s,p}$-norm on each cells in $\mathscr{T}$ combined using the scaling of p-norms in the obvious way. For example, $\|g\|_{W^{1,\infty}}(\mathscr{T})$ 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 $\mathscr{T}$ in the notation will be suppressed. For example, for a smooth metric g on $\mathscr{T}$, $\|g\|_{W^{2,\infty}}$ 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 3.6. *Let M be a domain in $\mathbb{R}^n$ and $\mathscr{T}_h$ a family of triangulations of M parameterized by the mesh size h. Suppose g is a smooth Riemannian metric on M and $g_h \in \mathrm{REG}^r(\mathscr{T}_h)$ a family of Riemannian metrics satisfying $\|g-g_h\|_{L^\infty} \le \frac{1}{2}\|g^{-1}\|_{L^\infty}^{-1}$ uniformly in h. Suppose $\gamma : [0,T] \to M$ is a smooth geodesic under g and γ_h a family of geodesics under g_h with the same initial conditions as γ. Moreover, assume the "no-stuck" condition: there exists a constant $V > 0$ such that the time γ_h takes to traverse through a single mesh cell is bounded above by h/V uniformly for all cells of $\mathscr{T}_h$ and all h. Then, there exists a constant C depending only on $\|g\|_{W^{2,\infty}}$, $\|g^{-1}\|_{L^\infty}$, V, T, and $|\dot\gamma(0)|$, such that*

$$|\dot\gamma(t) - \dot\gamma_h(t)| \le C(\|g-g_h\|_{W^{1,\infty}(\mathscr{T}_h)} + h^{-1}\|g-g_h\|_{L^\infty}),$$
$$|\gamma(t) - \gamma_h(t)| \le C(h\|g-g_h\|_{W^{1,\infty}(\mathscr{T}_h)} + \|g-g_h\|_{L^\infty}).$$

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^∞-norm with no control over the derivatives.

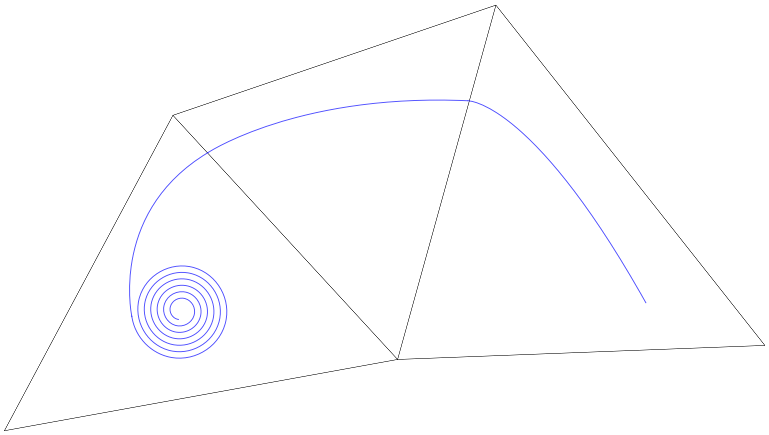


Figure 3.10: A geodesic is stuck in a cell.

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

Corollary 3.2. *Under the assumption of Theorem 3.6, suppose the shape constants of the meshes are bounded uniformly, g is known, and g_h are the Regge canonical interpolants. Then,*

$$|\dot{\gamma}(t) - \dot{\gamma}_h(t)| \leq Ch^r, \qquad |\gamma(t) - \gamma_h(t)| \leq Ch^{r+1},$$

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 in Chapter 2. □

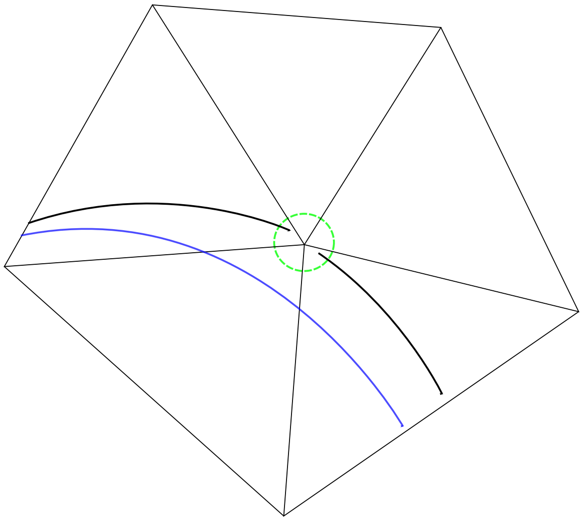


Figure 3.11: Restart a geodesic.

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 3.3. *Under the assumption of Theorem 3.6, suppose γ_h comes to a distance ϵh, $\epsilon < 1$, to a face of dimension $\le (n-2)$ at time t_*. Restart the extension of γ_h by keeping the Euclidean velocity vector $\dot{\gamma}_h(t^*)$ while moving its position to a point in another cell within the ϵ-sphere. Still call this (discontinuous) curve γ_h after t^*, and extend it as usual. Then, the error estimates still holds with an additional ϵh error in both the position and velocity estimates.*

Proof. It is clear that an extra error of ϵh is incurred for the position at time $t = t^*$. 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^*, the original estimate applies to the restarted geodesic approximation problem to the smooth geodesic under g with the same initial condition as $\gamma_h(t_+^*)$. 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 3.7 later). This proves the claim. □

In practice, ϵh is close to machine precision. So this is negligible.

The proof of Theorem 3.6 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.

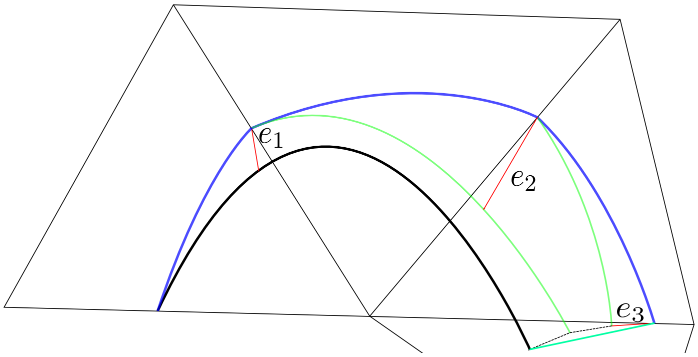


Figure 3.12: The black curve is the smooth geodesic. The blue curve is the geodesic on g_h.

In each cell, consider auxiliary smooth geodesics under the smooth metric using the position and velocity of γ_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 γ_h is further rotated. At this time, the difference between γ_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 3.6. Fix a particular h. Let $t_1, t_2, \dots, t_n$ be the time γ_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 λ_k which morphs from γ to γ_h as depicted in Figure 3.12: for $k = 0, \dots, (n+1)$,

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

where $f_k : [t_k, T] \to M$ is the geodesic under g with the initial condition

$$f_k(t_k) = \gamma_h(t_k), \qquad \dot{f}_k(t_k) = \dot{\gamma}_h(t_k+).$$

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

$$\lambda_0(t) = \gamma(t), \qquad \lambda_{m+1}(t) = \gamma_h(t).$$

Hence,

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

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

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

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 3.5, 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

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

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

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

The right-hand side can be estimated by using Lemma 3.6 for the two geodesics with the same initial condition but different metric and then applying Lemma 3.7 for the rotation of the velocity at the interior facet. The result is:

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

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

$$|\gamma(t) - \gamma_h(t)| + |\dot{\gamma}(t) - \dot{\gamma}_h(t+)| \le C_3 \sum_{k=0}^{m} [h \|g - g_h\|_{W^{1,\infty}(c_{k+1})} + \|g - g_h\|_{L^\infty}].$$

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

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

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

$$\begin{aligned}|\gamma(t)-\gamma_h(t)| \le C_3 \sum_{k=0}^{m} (t_{k+1}-t_k)(h\|g-g_h\|_{W^{1,\infty}(c_{k+1})} + \|g-g_h\|_{L^\infty}) \\ \le C_3 T(h\|g-g_h\|_{W^{1,\infty}(\mathscr{T}_h)} + \|g-g_h\|_{L^\infty}),\end{aligned}$$

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

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

Proof. By Theorem (3.2), the speed of γ measured in g is constant along γ:

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

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 [47, Corollary I.14.6]:

Theorem 3.7. *Let $y(t,t_0,y_0)$ be the solution to*

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

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

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

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

$$z(t) - y(t) = \int_0^1 \frac{\partial y}{\partial y_0}(t, t_0, y_0 + s(z_0 - y_0))(z_0 - y_0)\,ds + \int_{t_0}^t \frac{\partial y}{\partial y_0}(t, s, z(s))\delta(s, z(s))\,ds.$$

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 $v^j = \dot\gamma^j$.

$$\begin{aligned}\dot q^i &= v^i, \\ \dot v^j &= -\Gamma^j_{kl} v^k v^l,\end{aligned} \qquad \Leftrightarrow \qquad \dot y(t) = F(y(t)), \tag{3.12}$$

where the Christoffel symbol Γ^j_{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 3.4. *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,

$$\begin{gathered}
\|\Gamma\|_{L^\infty} \leq \|g^{-1}\|_{L^\infty} |g|_{W^{1,\infty}}, \\
\|\Gamma\|_{W^{1,\infty}} \leq \|g^{-1}\|_{L^\infty}^2 |g|_{W^{1,\infty}}^2 + \|g^{-1}\|_{L^\infty} |g|_{W^{2,\infty}}, \\
\cdots
\end{gathered}$$

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 (∂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 [60, I.4.24] implies that

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

This also shows that $\|g_1^{-1}\|_{L^\infty} \leq 2\|g_2^{-1}\|_{L^\infty}$. This proves the second estimate. □

This lemma gives a crude stability bound for the smooth geodesic equation:

Lemma 3.5. *Let $y(t, t_0, y_0)$ be the solution to equation (3.12) with initial data $y(t_0) = y_0$. 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 [47, Theorem I.14.3], Φ solves the linear ODE:

$$\dot{\Phi}(t) = \frac{\partial F}{\partial y}(t, y(t, t_0, y_0))\Phi(t), \qquad \Phi(t_0) = I,$$

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

$$\frac{\partial F}{\partial y} = \begin{bmatrix} 0 & I \\ -\partial_i \Gamma^j_{kl} v^k v^l & -2\Gamma^j_{il} v^l \end{bmatrix}.$$

Then Lemma 3.4 applies to the Γ-terms and Lemma 3.3 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. □

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 3.6. *Let c be an n-simplex in $\mathbb{R}^n$ of Euclidean diameter h. Suppose $\bar{g}$ is a Riemannian metric on c and $\bar{\gamma}$ a geodesic with initial condition $\gamma(t_0) = q_0 \in c$ and $\dot{\gamma}(t_0) = v_0$. Suppose g is any Riemannian metric on c with $\|g - \bar{g}\|_{L^\infty} \leq \frac{1}{2}\|\bar{g}^{-1}\|^{-1}_{L^\infty}$. Let γ be the geodesic under g with the same initial data (q_0, v_0). Set $y := [q^i, v^j]$ for γ as before and define $\bar{y}$ similarly. Then, before γ exits c, there exist constants C and M depending only on $\|\bar{g}\|_{W^{2,\infty}}$, $\|\bar{g}^{-1}\|_{L^\infty}$, and $|v_0|$ such that*

$$|y(t) - \bar{y}(t)| \leq C e^{M(t-t_0)} h \|g - \bar{g}\|_{W^{1,\infty}}.$$

Proof. By Theorem 3.7 and the definition of $F(y)$ in equation (3.12),

$$|y(t) - \bar{y}(t)| \leq \left| \int_{t_0}^t \frac{\partial \bar{y}}{\partial y_0} (\Gamma^i_{jk} - \bar{\Gamma}^i_{jk}) v^j v^k \, ds \right| \leq (\sup_t |v(t)|_{l^\infty}) \left\| \frac{\partial \bar{y}}{\partial y_0} \right\| \|\Gamma - \bar{\Gamma}\|_{L^\infty} \left| \int_{t_0}^t v \, ds \right|.$$

Because γ cannot exit c,

$$\left| \int_{t_0}^t v \, ds \right| = |q(t) - q(t_0)| \leq h.$$

By Lemma 3.4, 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 3.5, there is a constant C_2 depending on $\|\bar{g}\|_{W^{2,\infty}}$, $\|\bar{g}^{-1}\|_{L^\infty}$, and $|v_0|$, such that

$$\left\| \frac{\partial \bar{y}}{\partial y_0} \right\| \leq e^{C_2(t-t_0)}.$$

Moreover, on finite dimensional spaces, the $|\cdot|$-norm controls the l^∞-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 3.3. Combining all these estimates, one obtains

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

□

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 3.7. *Fix an $(k-1)$-dimensional hyperplane H in $\mathbb{R}^k$ and any basis $\{t_1,\dots,t_{k-1}\}$ 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 $|g-\bar{g}| \le \frac{1}{2}|\bar{g}^{-1}|^{-1}$. 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*

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

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 3.4, it can be shown that

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

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

$$T^T u n = 0, \qquad n^T u n = 1.$$

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,

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

This is a linear system. Solve for n',

$$n' = -\begin{bmatrix} T \\ n \end{bmatrix}^{-T} u^{-1} \begin{bmatrix} T \\ n/2 \end{bmatrix}^{T} u' n.$$

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,

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

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

3.9 Numerical examples: Kepler and Schwarzchild systems

In this section, we give two interesting numerical examples for the geodesic algorithm: the Keplerian orbits and the Schwarzschildian orbits. All the Python scripts used in this section can be found in the directory `geodesics` in the companion code repository to this thesis.

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

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

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 [48, Equation (2.5)]:

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

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

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

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

Lemma 3.8 (Equation (2.10) of [48]). *Let $e = \sqrt{1+2HL^2}$. Then, r and θ 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 θ. 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

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

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:

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

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 $\overline{RQ}$ with the circle of radius a centered at the origin. The new angle variable $E := \angle QOR$ is called the *eccentric anomaly*.

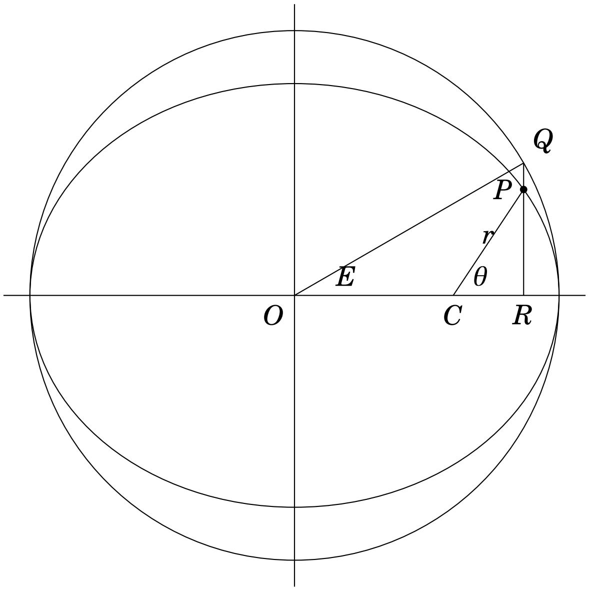


Figure 3.13: Definition of the eccentric anomaly

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

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

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* [38, 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]:

Proposition 3.3. *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 $\mathscr{E}$ is a geodesic $\gamma(s)$ of the Riemannian manifold $(M, \bar{g})$ with the Jacobi metric *$\bar{g} := 2(\mathscr{E} - V)g$ under the reparameterization*

$$s(\tau) = 2\int_0^\tau \mathscr{E} - V(\gamma(t))\,dt.$$

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

$$g_{ij} = 2(\mathscr{E} + |q|^{-1})\delta_{ij}. \tag{3.16}$$

Here the potential V is always negative and is normalized so that $V \to 0$ at infinity. Therefore, when $\mathscr{E} \geq 0$, the trajectories are unbounded. When $\mathscr{E} < 0$, the trajectories are trapped inside the region where $\mathscr{E} - V$ 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:

$$\dot{q}^i = \frac{p^i}{2(\mathscr{E} + |q|^{-1})}, \qquad \dot{p}_i = -\frac{|p|^2 q_i}{4(\mathscr{E} + |q|^{-1})^2|q|^3}.$$

The solution $q(s)$ to this system is related to the exact solution $q(t)$ before via the reparameterization:

$$s(\tau) = \int_0^\tau 2(\mathscr{E} + |q(t)|^{-1})\,dt.$$

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 `mshr`. A visualization of the discrete Kepler metric is given in Figure 3.14.

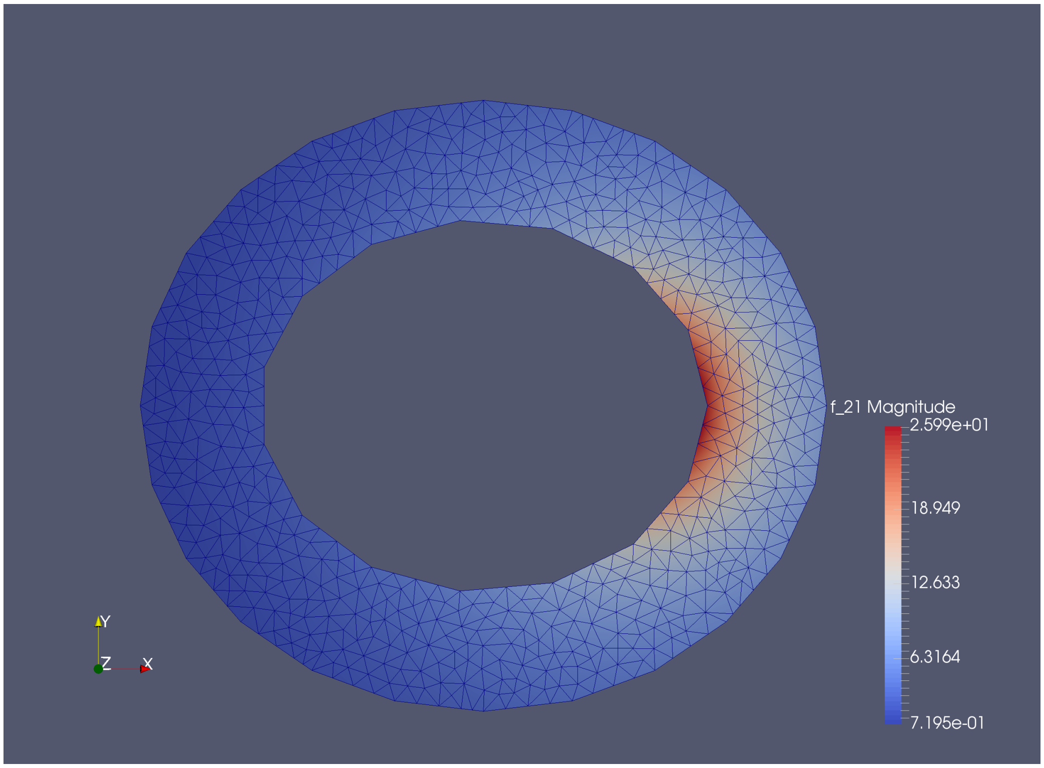


Figure 3.14: Plot of a discrete Kepler metric. The color indicates the pointwise Euclidean norm of the metric.

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 2×10^{-5}, which is smaller than the smallest mesh size $h_m \sim 7 \times 10^{-5}$. This ensures the error convergence rate is due to the better approximation of the metric. After each computation, the L^∞-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.

Metric	mesh sizes	max error in position	max error in H	max error in L
REG^0	$[64,128,256,512,1024]$	h_m^1 (1)	h_m^1 (1)	h_m^1 (1)
REG^1	$[64,128,256,512,1024]$	h_m^2 (h_m^2)	h_m^2 (h_m^2)	h_m^2 (h_m^2)
REG^2	$[32,64,128,256,512]$	h_m^3 (h_m^2)	h_m^3 (h_m^2)	h_m^3 (h_m^2)
REG^3	$[16,32,64,128,256]$	h_m^4 $(h_m^{3.5})$	h_m^4 $(h_m^{3.5})$	h_m^4 $(h_m^{3.5})$

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.

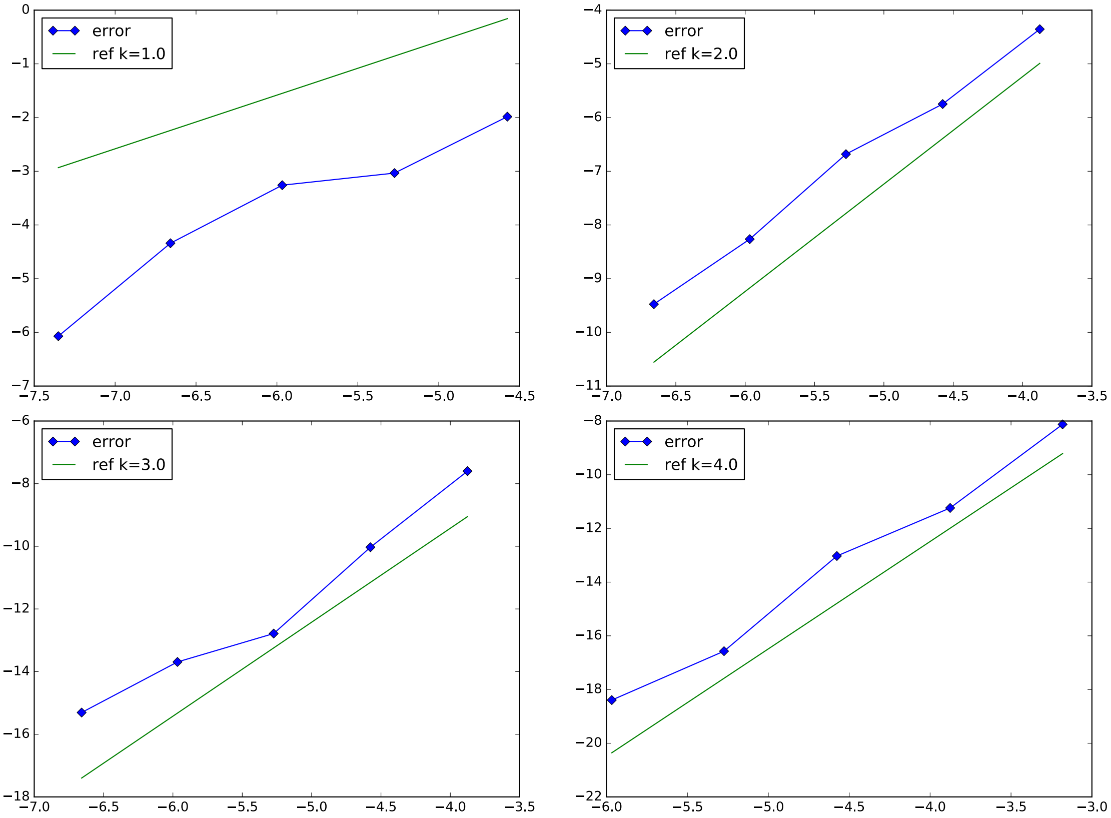


Figure 3.15: Blue: log-log plot of mesh size against position error for degree 0, 1, 2, 3. Red: reference slope for convergence of order 1, 2, 3, 4.

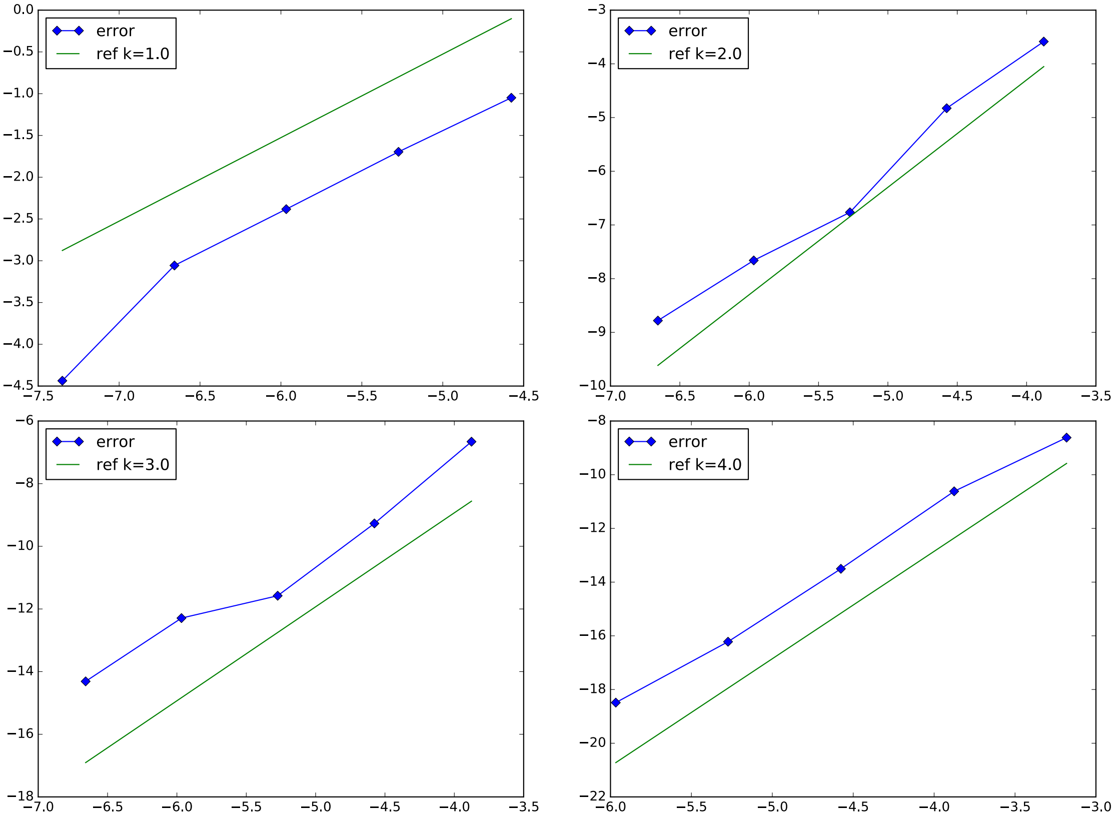


Figure 3.16: Blue: log-log plot of mesh size against error in the energy for degree $0,1,2,3$. Red: reference slope for convergence of order $1,2,3,4$.

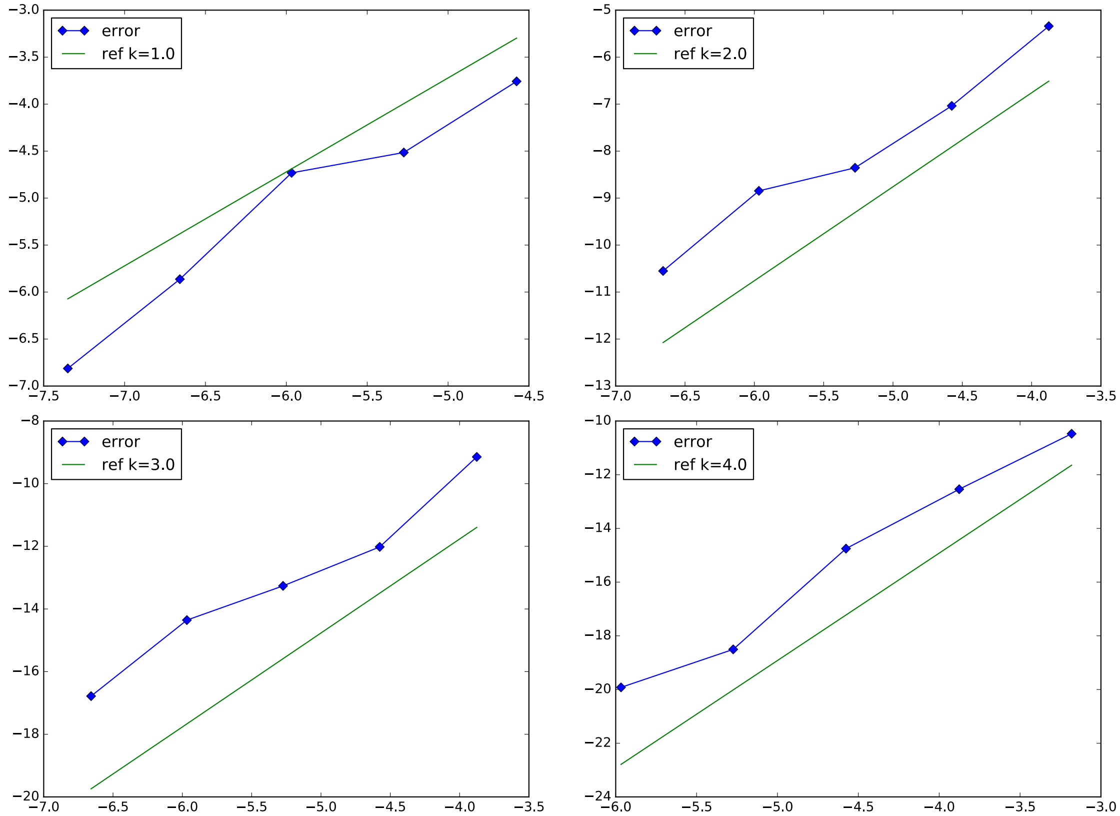


Figure 3.17: Blue: log-log plot of mesh size against momentum error for degree $0,1,2,3$. Red: reference slope for convergence of order $1,2,3,4$.

In the second sets of numerical experiments, the long time behavior of the error is assessed. The Kepler Jacobi metric is interpolated into 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.

Metric	mesh size	error in position	error in H	error in L
REG^0	160	t^1 (t^2)	1 (t^1)	1 (t^1)
REG^1	96	t^1 (t^2)	1 (t^1)	1 (t^1)
REG^2	96	t^1 (t^2)	1 (t^1)	1 (t^1)
REG^3	48	t^1 (t^2)	1 (t^1)	ϵt^1 (t^1)

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 REG^0 is somewhat sporadic.

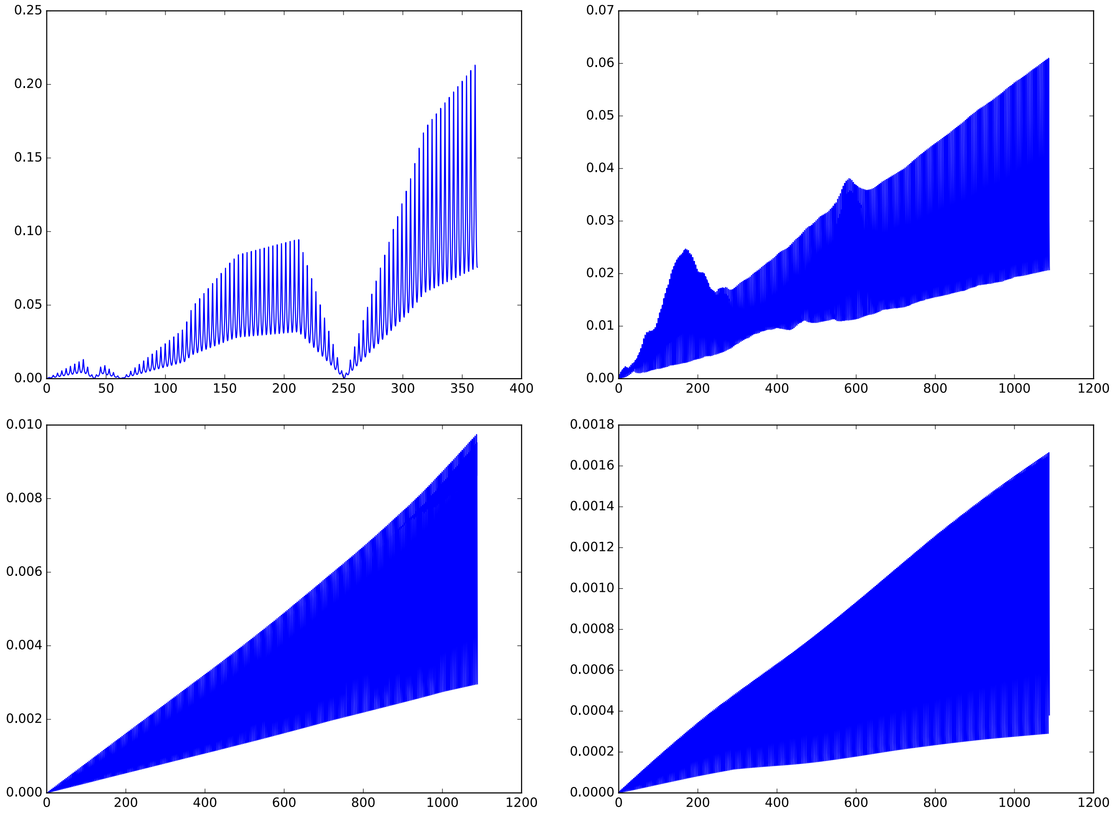


Figure 3.18: Plot of time against the error in position for degree $0, 1, 2, 3$.

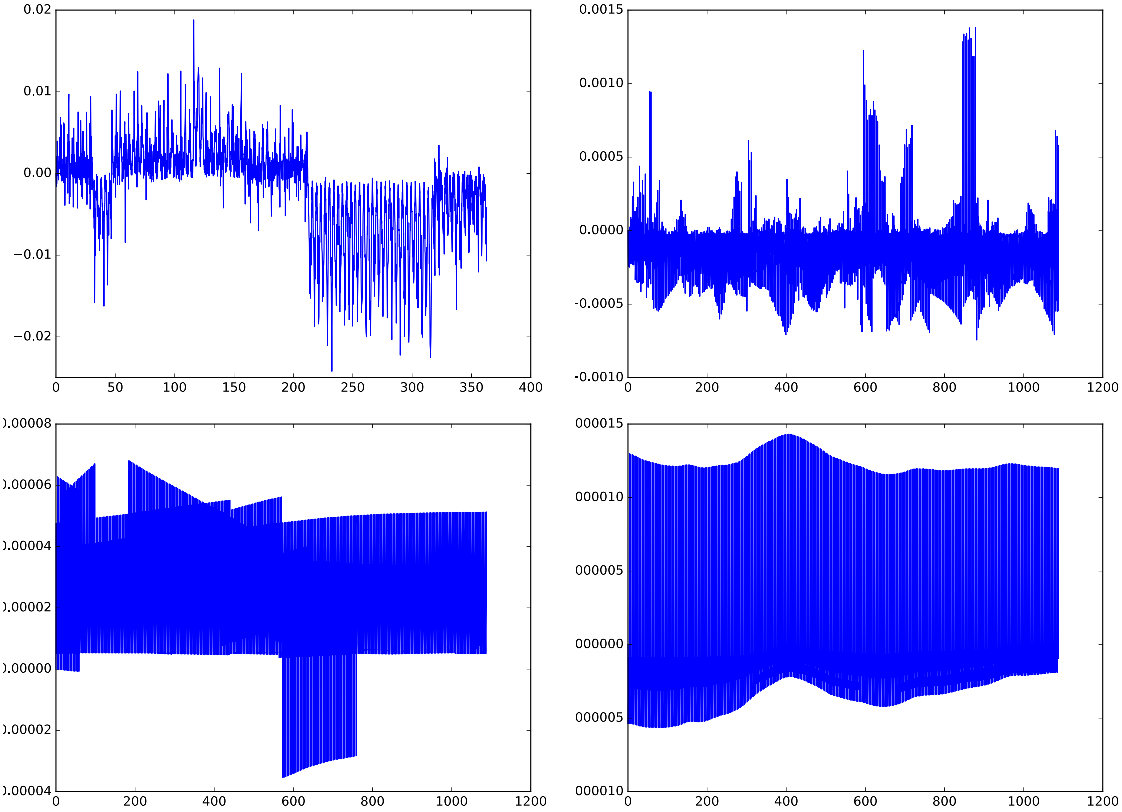


Figure 3.19: Plot of time against the relative error in energy for degree $0, 1, 2, 3$.

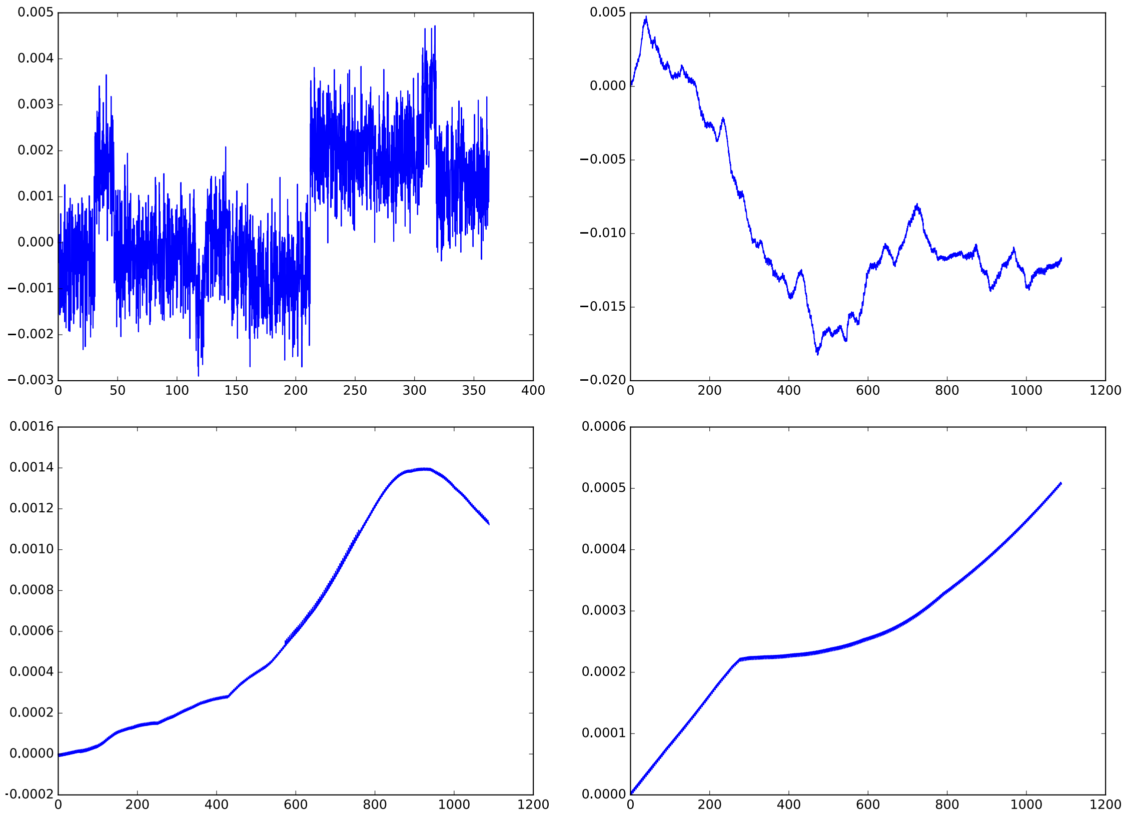


Figure 3.20: Plot of time against the relative error in momentum for degree $0,1,2,3$.

3.9.4 Schwarzschild system

The Schwarzschild metric is the most general static spherically symmetric solution to the Einstein field equation in general relativity [15]. 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 [111, Chapter 6]. In spherical coordinates, the metric for a star of mass M in natural units has the form [111, 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 [42, 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:

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

It is known that the orbits are almost ellipses with precession (that is, the major axis of the ellipsis rotates). A plot of the metric is shown in Figure 3.21. The mesh is obtained from `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.

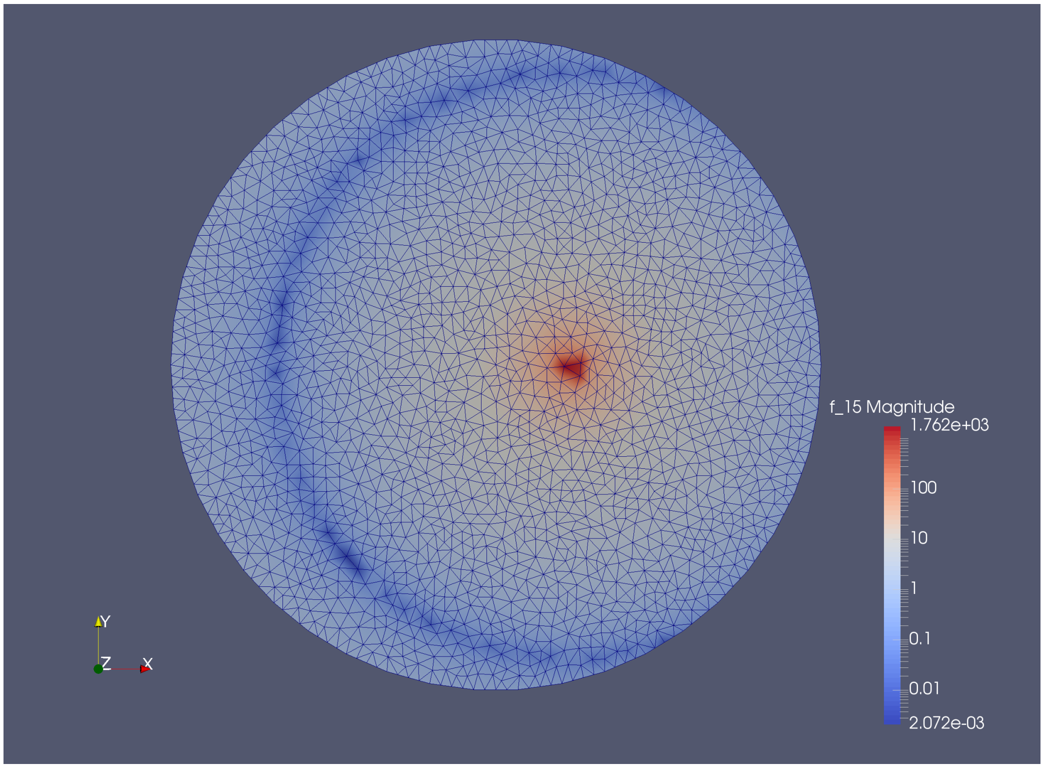


Figure 3.21: Plot of the discrete Schwarzschild metric. The color indicates the pointwise Euclidean norm of the metric.

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

Chapter 4

Rotated generalized Regge finite elements with applications in solid mechanics

In this chapter, we use generalized Regge finite elements, REG^r, to solve problems in solid mechanics. In particular, in Section 4.2, we propose a mixed method for the biharmonic equation in dimension $n \geq 2$, where we use REG^r to discretize the $\operatorname{div}\operatorname{div}$ operator on symmetric matrix fields. Moreover, in Section 4.3, we propose another mixed method for the elasticity equation in dimension $n \geq 2$, where we use REG^r to discretize div on symmetric matrix fields. We demonstrate the effectiveness and convergence properties of both methods via numerical examples.

In both methods, symmetric matrix-valued finite elements with normal-normal continuity are needed. The key idea, is to use the trace shifting map

$$Su := u - I \operatorname{tr} u,$$

to transform tangential-tangential continuous REG^r to normal-normal continuous finite elements. We study the properties of this transformation and its geometric interpretations in Section 4.1. We call $S(\mathrm{REG}^r)$ *rotated generalized Regge finite elements*.

This study also reveals connections of REG^r to previously known finite elements for symmetric tensor fields. In particular, in 2D, $S(\mathrm{REG}^r)$ is equivalent to the well-known Hellan-Herrmann-Johnson (HHJ) elements [13, 20] for the bending moment tensor in plate models. In 3D, $S(\mathrm{REG}^r)$ forms a strict subspace of the TDNNS stress elements for elasticity by Pechstein-Schöberl [89–91].

On one hand, we can view the two proposed mixed methods as using $S(\mathrm{REG}^r)$ to discretize $\operatorname{div}\operatorname{div}$ and div on symmetric matrix fields for applications in solid mechanics. This is the main view of this chapter. On the other hand, however, we can equivalent consider these two methods as using REG^r to discretize $\operatorname{div}\operatorname{div}S$ and $\operatorname{div}S$ on symmetric matrix fields. These two operators play important roles in the discretization of linearized relativity in later chapters. We end this chapter with discussions of this connection and some other potential applications in Section 4.4.

4.1 Rotated generalized Regge finite element

A vector-valued finite element with tangential continuity can clearly be transformed into one with normal continuity via a simple rotation by 90° in dimension 2, as shown in Figure 4.1. This, for example, relates Nédélec edge elements of the first kind [85] to Raviart-Thomas elements [95].

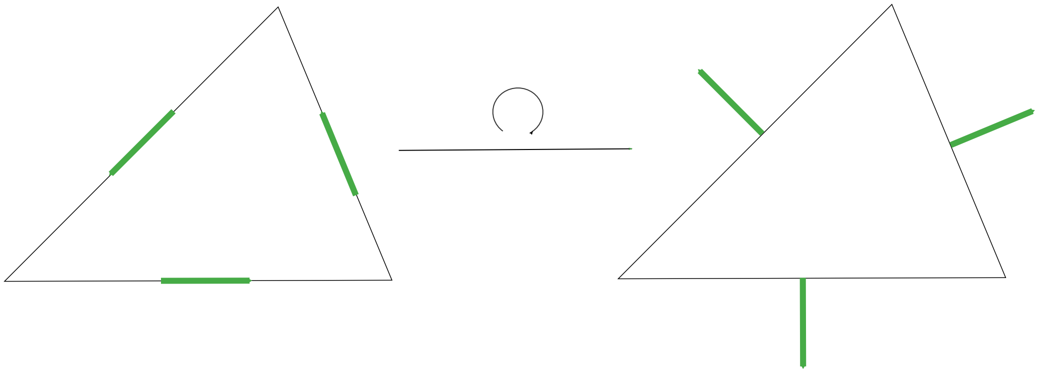


Figure 4.1: Rotation of 2D Nédélec edge elements to Raviart-Thomas elements

However, such linear algebraic map between tangential continuous and normal continuous vector finite elements can only exist in 2D.

Proposition 4.1. *For $m \geq 3$, suppose $A \in \mathbb{R}^{m\times m}$ has the property that for any $(m-1)$-plane P in $\mathbb{R}^m$, A maps tangential vectors to P to normal vectors to P. Then $A = 0$. In particular, there is no nonzero linear map taking piecewise smooth tangential continuous vector fields on a mesh to piecewise smooth normal continuous vector fields.*

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

For symmetric matrix fields, however, the situation is quite different. Let P be an $(n-1)$-plane in $\mathbb{R}^n$ and $\{t_1, \ldots, t_{n-1}, n\}$ 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 *tangential-tangential 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 tangential-tangential continuous or normal-normal continuous if the tangential-tangential parts or normal-normal parts are single-valued at all interior facets.

Theorem 4.1. *Suppose $m \geq 2$. Let S be the linear map on symmetric matrix fields:*

$$Su := u - I \operatorname{tr} u.$$

The map S takes tangential-tangential 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 $\mathbb{R}^m$ adapted to f. By definition,

$$n^T (Su) n = n^T u n - \operatorname{tr} u.$$

The trace can be computed via:

$$\operatorname{tr} u = n^T u n + \sum_{i=1}^{m-1} t_i^T u t_i.$$

Hence,

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

is just the trace of the tangential-tangential part of u to f, which is continuous across interior facets by assumption. □

The definition of S is most intuitive in 2D. Let R be the clockwise 90°-rotation matrix

$$R := \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{4.1}$$

It is clear that this matrix rotates tangential vectors to every 1-plane in $\mathbb{R}^2$ to normal vectors. Thus if u is tangential-tangential continuous, $R^T uR$ is normal-normal continuous. A direct computation shows that

$$R^T uR = -(Su)^T. \tag{4.2}$$

In dimension $m \geq 3$, there is no such R as we show in Proposition 4.1.

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

Theorem 4.2. *In dimension $m \geq 2$,*

$$S(\mathrm{REG}^r(\mathscr{T})) \subset \mathrm{NN}^r(\mathscr{T}).$$

In 2D, the two spaces are equal. For $m \geq 3$, the inclusion is proper.

Proof. Theorem 4.1 implies the inclusion. In dimension 2, the rotation R is clearly invertible. Equation (4.2) is the isomorphism between the two spaces. For $m \geq 2$, S is invertible

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

Normal-normal continuity entails only 1 condition at every interior facet, while tangential-tangential continuity entails $(n-1)$ conditions at every interior facet. Thus when $m \geq 3$, for $v \in \mathrm{NN}^r(\mathscr{T})$, the inverse image $S^{-1}v$ is not necessarily in $\mathrm{REG}^r(\mathscr{T})$ in general. Hence $S(\mathrm{REG}^r(\mathscr{T}))$ is only a strict subspace. □

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

Let Ω 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

$$\begin{aligned} \Delta\Delta u &= f, \quad \text{in } \Omega, \\ u = \partial_n u &= 0, \quad \text{on } \partial\Omega. \end{aligned} \tag{4.3}$$

This is a classic model problem with many applications including 2D Kirchhoff-Love plate models [108], potential formulations in 3D elasticity [99], and stationary Cahn-Hilliard phase separation models [114].

The continuous theory for the 2D biharmonic equation is well-established. In particular, it can be shown that given $f \in H^{-2}$ there is a unique solution $u \in \mathring{H}^2$. An exposition of the existence and regularity theory on Lipschitz domains can be found in [45, Chapter 7]. Its finite element discretization is also very mature. A survey can be found in [30, 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 [7, 13, 20] of equation (4.3). We then show that $S(\mathrm{REG}^r)$ is equivalent to the HHJ element for the symmetric matrix field variable. After that we propose a mixed method for the biharmonic equation in dimension $m \geq 2$ using REG^r and study the convergence properties numerically.

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-by-n matrices and

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

where div div means first taking the divergence of a matrix field row by row and then take the divergence again of the resulting vector field. This space caries the graph norm.

The mixed formulation of equation (4.3) is, given $f \in H^{-1}$, find $(\sigma, u) \in H(\operatorname{div}\operatorname{div}) \times \mathring{H}^1$ such that

$$\begin{aligned} (\sigma, \tau) - \langle u, \operatorname{div}\operatorname{div}\tau\rangle &= 0, \quad \forall \tau \in H(\operatorname{div}\operatorname{div}), \\ \langle \operatorname{div}\operatorname{div}\sigma, v\rangle &= \langle f, v\rangle, \quad \forall v \in \mathring{H}^1, \end{aligned} \tag{4.4}$$

where $\langle \cdot, \cdot \rangle$ is the duality pairing between H^{-1} and $\mathring{H}^1$.

The following two theorems are well-known in the literature [69] (see also [13, 20]). 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 (4.4) itself is well-posed:

Theorem 4.3 (Theorem 2.2 of [69]). *Given $f \in H^{-1}$, there exists a unique pair (σ, u) solving system (4.4). Further there exists a constant $C > 0$ depending only on Ω such that*

$$\|\sigma\|_{H(\operatorname{div}\operatorname{div})} + \|v\|_{H^1} \leq C\|f\|_{H^{-1}}.$$

Proof. This follows from Brezzi's theorem [19]. It is clear that (σ, τ) is coercive over the kernel of div div. It remains to show the inf-sup condition for the bilinear form $\langle v, \operatorname{div}\operatorname{div}\tau\rangle$. Note

$$\operatorname{div}\operatorname{div} I v = \operatorname{div}\nabla v = \Delta v.$$

For any $v \in \mathring{H}^1$, let $\tau = -Iv \in H^1 \otimes \mathbb{S}^n \subset H(\operatorname{div}\operatorname{div})$. Then,

$$\langle v, \operatorname{div}\operatorname{div}\tau\rangle = \langle v, -\Delta v\rangle = (\nabla v, \nabla v) \geq c\|v\|_{H^1}^2,$$

where c depends on the Poincaré constant for Ω and

$$\|\tau\|_{H(\operatorname{div}\operatorname{div})}^2 = \|\tau\|_{L^2}^2 + \|\operatorname{div}\operatorname{div}\tau\|_{H^{-1}}^2 = \|Iv\|_{L^2}^2 + \|\Delta v\|_{H^{-1}}^2 \leq (n^2+1)\|v\|_{H^1}^2,$$

where n is the dimension of Ω. Thus the inf-sup constant is bounded below by a constant depending only on the domain. □

Second, the mixed system (4.4) can be used to solve the biharmonic equation (4.3).

Theorem 4.4 (Corollary 2.3 of [69]). *Given $f \in H^{-1}$, suppose (σ, u) is a solution to system (4.4), then $\sigma = \nabla\nabla u$ and $u \in \mathring{H}^2$ solves the biharmonic equation (4.3) as a distribution.*

Proof. The biharmonic equation (4.3) has a unique solution, say $w \in \mathring{H}^2$ with $\Delta\Delta w = f$. It is clear that $\nabla\nabla w \in H(\operatorname{div}\operatorname{div})$. Once we show that $(\sigma, u) := (\nabla\nabla w, w)$ solves system (4.4), the theorem is then proved by the well-posedness of system (4.4). First, for test functions y,

$$\langle y, \operatorname{div}\operatorname{div}\tau\rangle = (\nabla\nabla y, \tau).$$

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

$$\langle w, \operatorname{div}\operatorname{div}\tau\rangle = (\nabla\nabla w, \tau) = (\sigma, \tau),$$

which shows that the first equation of system (4.4) holds. Similarly, by definition, $\operatorname{div}\operatorname{div}\sigma = \operatorname{div}\operatorname{div}\nabla\nabla w = \Delta\Delta w = f$ as a distribution. Since the set of test functions is also dense in $\mathring{H}^1$ the second equation of system (4.4) also holds. □

4.2.2 Hellan-Herrmann-Johnson discretization

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

$$V := \{\sigma \in L^2 \otimes \mathbb{S}^n \text{ is piecewise } H^1 \text{ with normal-normal continuity}\},$$
$$W := \{u \in \mathring{H}^1 \text{ is piecewise } H^2\}.$$

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

$$\sigma_{nn} := n^T \sigma n, \qquad \sigma_{n\tau} := \sigma n - n\sigma_{nn}.$$

Similarly, for $u \in W$, let

$$\partial_n u := n \cdot \nabla u, \qquad \partial_\tau u := \nabla u - n\partial_n u.$$

We give V and W the following mesh dependent norms:

$$\|\sigma\|_V^2 := \sum_c \|\sigma\|_{L^2(c)}^2 + h\|\sigma_{nn}\|_{L^2(\partial c)}^2,$$
$$\|u\|_W^2 := \sum_c \|u\|_{H^2(c)}^2 + h^{-1}\|\partial_n u\|_{L^2(\partial c)}^2,$$

where both sum over all the cells c in the mesh $\mathscr{T}_h$. Using the same notation, we define a *mesh-dependent* $\operatorname{div}\operatorname{div}_h$ by:

$$\langle \operatorname{div}\operatorname{div}_h \sigma, v\rangle := \sum_c \left(\int_c \sigma : \nabla\nabla v - \int_{\partial c} \sigma_{nn}\partial_n v \right) = \sum_c \int_c \sigma : \nabla\nabla v - \sum_f \int_f \sigma_{nn} [\![\partial_n v]\!], \tag{4.5}$$

where $\sum_f$ means sum over all facets f of $\mathscr{T}_h$ and the *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

$$|\langle \operatorname{div}\operatorname{div}_h \sigma, v\rangle| \leq C\|\sigma\|_V \|u\|_W.$$

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

$$V_h := \mathrm{NN}^r(\mathscr{T}_h), \qquad W_h := \mathrm{CG}^{r+1}(\mathscr{T}_h) \cap \mathring{H}^1, \qquad r \geq 0.$$

The 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 $\mathrm{NN}^r(T)$ is defined by the shape functions

$$\mathscr{P}^r(T) \otimes \mathbb{S}^2$$

and the degrees of freedom

$$\sigma \mapsto \int_e (n^T \sigma n) q, \qquad \forall q \in \mathscr{P}_r(e) \text{ and all edges } e \text{ of } T,$$
$$\sigma \mapsto \int_T \sigma : \tau \qquad \forall \tau \in \mathscr{P}_{r-1}(T) \otimes \mathbb{S}^2,$$

where n is the outward unit normal vector to T.

Let R be the 90°-rotation matrix defined in equation (4.1) and $t := Rn$ the unit tangent vector to the edges of T. The generalized Regge element $\mathrm{REG}^r(T)$ in 2D is given by the same shape functions

$$\mathscr{P}^r(T) \otimes \mathbb{S}^2$$

but tangential degrees of freedom

$$\sigma \mapsto \int_e (t^T \sigma t) q, \qquad \forall q \in \mathscr{P}_r(e) \text{ and all edges } e \text{ of } T,$$
$$\sigma \mapsto \int_T \sigma : \tau \qquad \forall \tau \in \mathscr{P}_{r-1}(T) \otimes \mathbb{S}^2.$$

Note that in 2D, we have

$$S^{-1} = S.$$

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

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

we conclude that $S : \mathrm{REG}^r(T) \to \mathrm{NN}^r(T)$ is an isomorphism of finite elements. Since we proved in early chapters of this thesis that REG^r is unisolvent, $\mathrm{NN}^r(T)$ 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:

$$\begin{aligned} (\sigma, \tau) - \langle u, \operatorname{div} \operatorname{div}_h \tau \rangle &= 0, \quad \forall \tau \in V_h, \\ \langle \operatorname{div} \operatorname{div}_h \sigma, v \rangle &= \langle f, v \rangle, \quad \forall v \in W_h, \end{aligned} \tag{4.6}$$

First, we have consistency.

Theorem 4.5. *Suppose $u \in H^3 \cap \mathring{H}^2$ solves the biharmonic equation (4.3). Let $\sigma := \nabla\nabla u$. Then (σ, u) satisfies the discrete system (4.6).*

Proof. First, $u \in W$ so $\langle u, \operatorname{div} \operatorname{div}_h \tau \rangle$ makes sense. Because $u \in \mathring{H}^2$, $[\![\partial_n u]\!] = 0$ at all facets of the mesh. Hence, the first equation of (4.3) reads:

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

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 $\langle \operatorname{div} \operatorname{div}_h \sigma, v \rangle$ still makes sense. Then, for an interior facet f, $[\![\sigma_{n\tau}]\!] = 0$ at f because σ is continuous across facets. On the other hand, for boundary facets f, $\partial_\tau v = 0$ because $\tau \in \mathring{H}^1$. Hence,

$$\sum_c \int_{\partial c} \sigma_{n\tau} \partial_\tau v = \sum_f \int_f [\![\sigma_{n\tau}]\!] \partial_\tau v = 0.$$

Thus, by the identity $\sigma n \cdot \nabla v = \sigma_{nn} \partial_n v + \sigma_{n\tau} \partial_\tau v$ and integration by parts:

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

Sum over the cells and integrate by parts again:

$$\langle \operatorname{div}\operatorname{div}_h \sigma, v\rangle = \int_\Omega -\operatorname{div}\sigma\cdot\nabla v = \int_\Omega v\operatorname{div}\operatorname{div}\sigma - \int_{\partial\Omega} nv\cdot\operatorname{div}\sigma = \int_\Omega v\operatorname{div}\operatorname{div}\sigma,$$

where the last equality follows from the fact that $v = 0$ on the boundary. By definition, $\operatorname{div}\operatorname{div}\sigma = \Delta\Delta u = f$. So the second equation of (4.6) is also satisfied. This proves the claim. □

In [13], the following stability and convergence theorem was proved:

Theorem 4.6. *Suppose the domain is a convex polygon. Let $u \in H^3$ be a solution to the biharmonic equation (4.3) and $\sigma = \nabla\nabla u$. The discrete system (4.6) has a unique solution $(\sigma_h, u_h) \in \mathrm{NN}^r \times \mathrm{CG}^{r+1} \cap \mathring{H}^1$. This pair satisfies:*

$$\|\sigma - \sigma_h\|_{L^2} + \|u - u_h\|_{H^1} \le Ch\|u\|_{H^3}.$$

Moreover, if u is smooth, then

$$\|\sigma - \sigma_h\|_{L^2} \le Ch^{r+1}\|u\|_{H^{r+3}},$$

and for $r = 0$,

$$\|u - u_h\|_{H^1} \le Ch\|u\|_{H^3}, \qquad \|u - u_h\|_{L^2} \le Ch^2\|u\|_{H^4},$$

while for $r \ge 1$,

$$\|u - u_h\|_{H^1} \le Ch^{r+1}\|u\|_{H^{r+2}}, \qquad \|u - u_h\|_{L^2} \le Ch^{r+2}\|u\|_{H^{r+3}}.$$

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

In dimension m, $m \ge 3$, the form of the continuous biharmonic equation (4.3), the continuous mixed formulation (4.4), and the mesh-dependent divdiv (4.5) remain the same as those in 2D. We noted that $S(\mathrm{REG}^r)$, which is defined in dimension m for all $m \ge 2$, is a discrete subspace of the infinite-dimensional mesh-dependent space V. This opens up the possibility of using the pair

$$V_h = S(\mathrm{REG}^r), \qquad W_h = \mathrm{CG}^{r+1} \cap \mathring{H}^1, \qquad r \ge 0, \tag{4.7}$$

in higher dimensions for the discretization (4.6) to solve the biharmonic equation. In this subsection, we first validate that in 2D, the space $S(\mathrm{REG}^r)$ can be used to solve the biharmonic equation in place of HHJ^r in practical implementations. Then we study the convergence of the discrete space choice (4.7) for solving the 3D biharmonic equation.

The finite element pair (4.7) can be implemented practically by: given f, find $(\mu, u) \in \mathrm{REG}^r \times W_h$ satisfying:

$$\begin{aligned} (S\mu, S\rho) - \langle u, \operatorname{div}\operatorname{div}_h S\rho\rangle &= 0, && \forall \rho \in \mathrm{REG}^r, \\ \langle \operatorname{div}\operatorname{div}_h S\mu, v\rangle &= \langle f, v\rangle, && \forall v \in W_h. \end{aligned}$$

Direct computation shows that:

$$(S\mu, S\rho) = (\mu, \rho) + (m-2)(\operatorname{tr}\mu, \operatorname{tr}\rho),$$

which is coercive over the L^2 norm for all $m \geq 2$. The operator $\operatorname{div}\operatorname{div} S$ 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 HHJ^{r+1} as part of this thesis in FEniCS. To make this numerical test more realistic and interesting, the biharmonic equation:

$$\Delta\Delta u = f,$$

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.

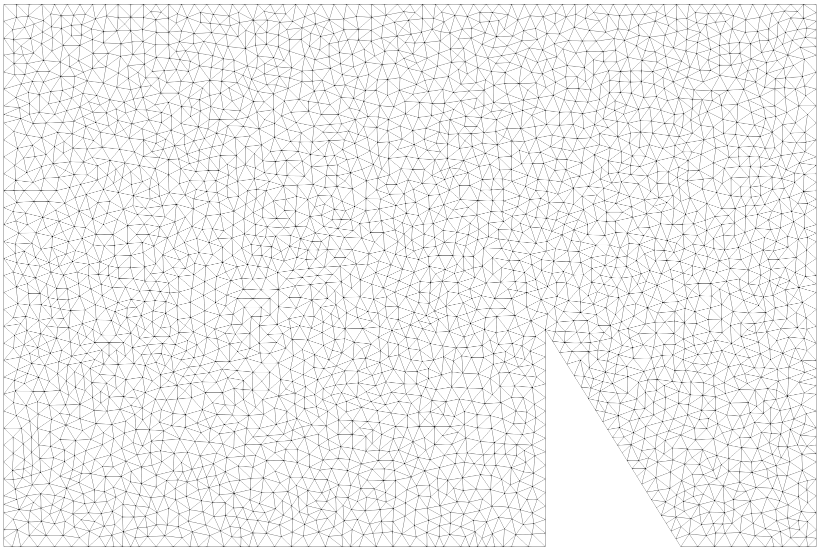


Figure 4.2: Domain and mesh of the comparison test

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 $n^T(\nabla\nabla u)n = 0$. 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 $S(\mathrm{REG}^2)$ to solve this problem. A plot of the solution is given in Figure 4.3.

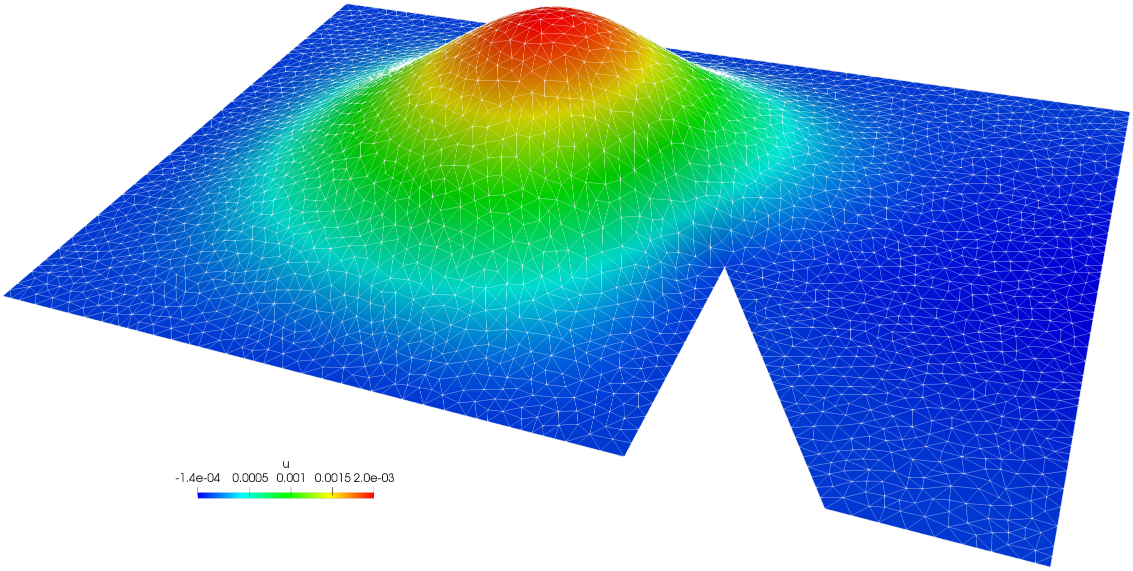


Figure 4.3: 2D biharmonic equation demo

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 $S(\mathrm{REG}^2)$ 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 $S(\mathrm{REG}^r)$ 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

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

A sequence of unstructured meshes are generated using the FEniCS package mshr, which in turn internally uses CGAL [106] 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.

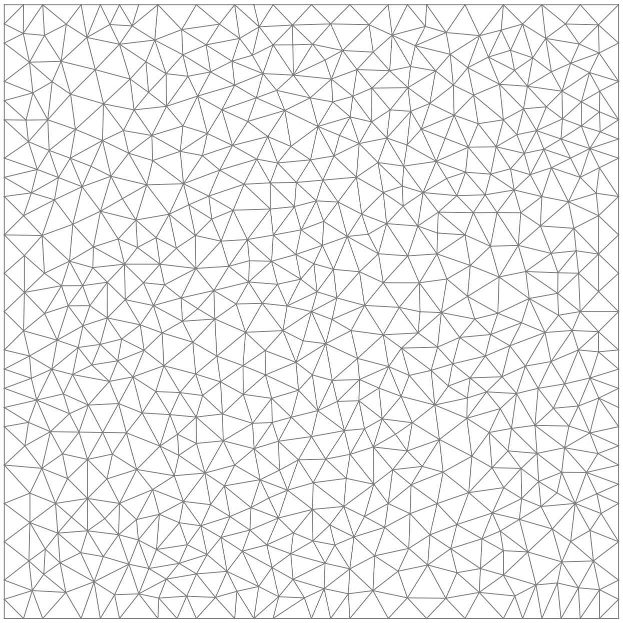


Figure 4.4: Example of an unstructured 2D mesh for the convergence test

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 σ and u are observed.

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

Table 4.1: 2D biharmonic degree 0

Mesh size	$\|u-u_h\|$	Rate	$\|\nabla(u-u_h)\|$	Rate	$\|\sigma-\sigma_h\|$	Rate
4	5.389649e-03		1.667406e-01		1.578812e+00	
8	6.811778e-04	2.97	4.421494e-02	1.91	4.364165e-01	1.85
16	8.296795e-05	3.12	1.074281e-02	2.09	1.096000e-01	2.05
32	1.053586e-05	2.89	2.754467e-03	1.91	2.859750e-02	1.88
64	1.356525e-06	2.94	6.987778e-04	1.97	7.258006e-03	1.97

Table 4.2: 2D biharmonic degree 1

Mesh size	$\|u-u_h\|$	Rate	$\|\nabla(u-u_h)\|$	Rate	$\|\sigma-\sigma_h\|$	Rate
2	8.358579e-03		2.291764e-01		2.037017e+00	
4	5.105807e-04	4.42	2.396781e-02	3.57	2.367370e-01	3.40
8	3.798470e-05	3.74	3.408628e-03	2.80	3.306554e-02	2.83
16	2.197489e-06	4.22	4.115211e-04	3.13	4.095130e-03	3.09
32	1.457209e-07	3.81	5.348559e-05	2.86	5.273608e-04	2.87

Table 4.3: 2D biharmonic degree 2

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

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

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.

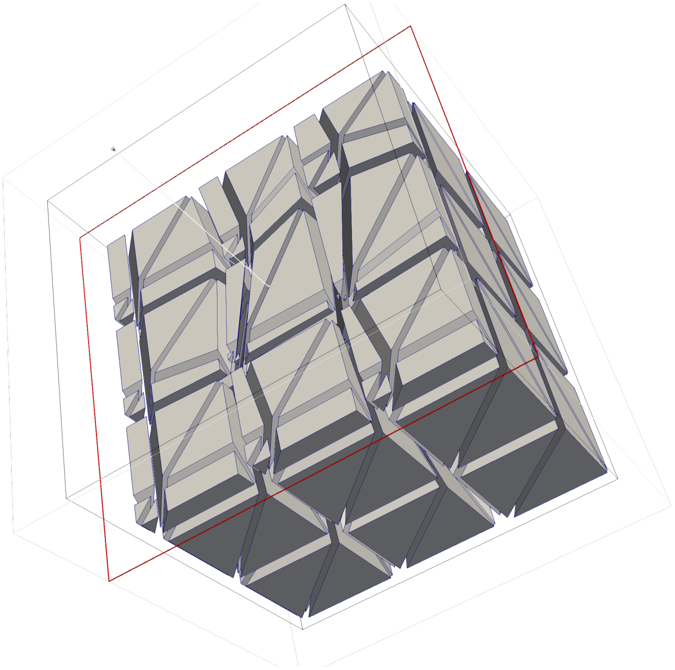


Figure 4.5: Example of a randomly perturbed 3D mesh for the convergence test

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.

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

Table 4.4: 3D biharmonic degree 0

Mesh size	$\|u-u_h\|$	Rate	$\|\nabla(u-u_h)\|$	Rate	$\|\sigma-\sigma_h\|$	Rate
3	3.958634e-02		4.178022e-01		4.115063e+00	
6	1.139772e-02	1.83	1.636156e-01	1.38	2.451652e+00	0.76
9	4.954564e-03	2.36	7.946402e-02	2.04	1.654588e+00	1.11
12	2.639824e-03	1.97	4.685069e-02	1.65	1.213616e+00	0.97

Table 4.5: 3D biharmonic degree 1

Mesh size	$\|u-u_h\|$	Rate	$\|\nabla(u-u_h)\|$	Rate	$\|\sigma-\sigma_h\|$	Rate
2	3.470993e-02		3.689862e-01		3.453816e+00	
4	3.516079e-03	3.34	7.192239e-02	2.39	1.221303e+00	1.52
6	8.712791e-04	3.68	2.456078e-02	2.83	6.156819e-01	1.81
8	3.382587e-04	3.25	1.129189e-02	2.67	3.872619e-01	1.59

Table 4.6: 3D biharmonic degree 2

4.3 Solving the elasticity equation via the Pechstein-Schöberl mixed formulation

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

$$\begin{aligned} \operatorname{div} C\epsilon u &= -f, \qquad \text{in } \Omega, \\ u &= 0, \qquad \text{on } \partial\Omega, \end{aligned} \tag{4.8}$$

where the *compliance tensor* C is given such that $(C\cdot,\cdot)$ 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 [80].

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 $u \in \mathring{H}^1 \otimes \mathbb{R}^n$ solving the problem. An exposition of the regularity theory on various types of domains can be found in [45].

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 [89,90,100, 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 $H(\operatorname{div}\operatorname{div})$ for the stress variable $\sigma = C\epsilon u$. This is where rotated Regge elements would fit. We still need another space to pair with $\operatorname{div}\sigma$. 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:

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

This is a Hilbert space under the graph inner product

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

In 2D, $H\Lambda^1$ is the space $H(\mathrm{rot})$. In 3D, $H\Lambda^1$ is the space $H(\mathrm{curl})$. In general, this is the space of L^2-differential 1-forms [8, 10]. It can be shown using integration by parts and a density argument [8, page 19] that elements of $H\Lambda^1$ has a well-defined tangential trace to the boundary. More precisely, there is a bounded linear map $H\Lambda^1 \to H^{-1/2}(\partial\Omega) \otimes \mathbb{R}^n$. Let $\mathring{H}\Lambda^1$ be the subspace of $H\Lambda^1$ with vanishing tangential trace. We will show that there is a duality pairing $\langle \operatorname{div}\tau, v\rangle$ for $\tau \in H(\operatorname{div}\operatorname{div})$ and $v \in \mathring{H}\Lambda^1$. This requires several steps.

First, we recall the following the regular decomposition result (Lemma 5 in [33] with $k = 1$):

Proposition 4.2. *On a bounded Lipschitz domain Ω, for all $u \in \mathring{H}\Lambda^1$, there exists $\phi \in \mathring{H}^1$, $z \in \mathring{H}^1 \otimes \mathbb{R}^n$ such that $u = \nabla\phi + z$ with $\|\phi\|_{H^1} + \|z\|_{H^1} \le M\|u\|_{H\Lambda}$ for some constant M depending only on Ω.*

Second, using a similar argument to Lemma 2.1 of [89], we show the following duality result:

Proposition 4.3. *On a bounded Lipschitz domain, let*

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

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

$$(H^{-1}(\mathrm{div}))' = \mathring{H}\Lambda^1.$$

In particular, $\operatorname{div} H(\operatorname{div}\operatorname{div}) \subset H^{-1}(\mathrm{div})$, therefore the pairing $\langle \operatorname{div}\tau, v\rangle$ makes sense for $\tau \in H(\operatorname{div}\operatorname{div})$ and $v \in \mathring{H}\Lambda^1$ and leads to a bounded bilinear form on this pair of spaces.

Proof. By Proposition 4.2, we have the following equivalence of norms on distributions:

$$\|f\|_{(\mathring{H}\Lambda^1)'} = \sup_{u\in\mathring{H}\Lambda^1} \frac{\langle f,u\rangle}{\|u\|_{H\Lambda^1}} \sim \sup_{\phi,z} \frac{\langle f,\nabla\phi+z\rangle}{\|\phi\|_1+\|z\|_1} \sim \sup_{\phi} \frac{\langle f,\nabla\phi\rangle}{\|\phi\|_1} + \sup_{z} \frac{\langle f,z\rangle}{\|z\|_1} = \|\operatorname{div} f\|_{-1} + \|f\|_{-1},$$

where $a \sim b$ means that there exist constants c and C depending only on the domain such that $ca \le b \le Ca$. By definition, $(\mathring{H}\Lambda^1)'$ is the space of distributions with bounded dual norm. This implies the first claim. Finally, it is clear that for $\sigma \in H(\operatorname{div}\operatorname{div})$, $\operatorname{div}\sigma \in H^{-1}\otimes\mathbb{R}^n$ and $\operatorname{div}\operatorname{div}\sigma \in H^{-1}$. This proves the last claim. □

Let $A := C^{-1}$ be the *compliance tensor*. The TDNNS continuous formulation is: given $f \in H^{-1}(\operatorname{div})$, find $\sigma \in H(\operatorname{div}\operatorname{div})$, $u \in \mathring{H}\Lambda^1$ such that

$$\begin{aligned} (A\sigma,\tau) + \langle u,\operatorname{div}\tau\rangle &= 0, \quad && \forall \tau \in H(\operatorname{div}\operatorname{div}),\\ \langle \operatorname{div}\sigma, v\rangle &= -\langle f,v\rangle, && \forall v \in \mathring{H}\Lambda^1. \end{aligned} \tag{4.9}$$

First we show that this system is well-posed. The theorem below largely follows the arguments in Theorem 2.3 of [89]. The proof there has a gap where they only proved $\langle \operatorname{div}\sigma, v\rangle = (w,v)_{H(\mathrm{curl})}$ for $v \in \mathring{H}^1\otimes\mathbb{R}^3$ but in the end took $v = w$ where $w \in H(\mathrm{curl})$ is from a bigger space. Here we give the correct proof with more details and greater generality.

Theorem 4.7. *On a bounded Lipschitz domain Ω in $\mathbb{R}^n$, there exists a unique solution (σ,u) to system (4.9). Further there exists a constant M depending only on Ω and the coefficient C such that*

$$\|\sigma\|_{H(\operatorname{div}\operatorname{div})} + \|u\|_{H\Lambda^1} \le M\|f\|_{(\mathring{H}\Lambda^1)'}.$$

Proof. This follows from Brezzi's theorem [19]. We only need to show the inf-sup condition for $\langle \operatorname{div}\tau, v\rangle$. Fix any $v \in \mathring{H}\Lambda^1$. Let $w \in \mathring{H}^1\otimes\mathbb{R}^n$ be the solution to

$$(C\epsilon w, \epsilon y) = (v,y)_{H\Lambda^1}, \qquad \forall y \in \mathring{H}^1\otimes\mathbb{R}^n.$$

The left-hand side is a bounded coercive bilinear form. Hence such unique solution w exists. Define $\tau := C\epsilon w$. Then $\tau \in L^2\otimes\mathbb{S}^n$ with

$$\|\tau\|_{L^2} = \|C\epsilon w\|_{L^2} \le M_1\|w\|_{H^1},$$

where M_1 is a constant depending on Ω and C. By Korn's inequality and the equation defining w,

$$\|w\|_{H^1}^2 \le M_2(C\epsilon w,\epsilon w) = M_2(v,w)_{H\Lambda^1} \le M_2\|v\|_{H\Lambda^1}\|w\|_{H\Lambda^1} \le M_2\|v\|_{H\Lambda^1}\|w\|_{H^1},$$

for some constant M_2 depending on Ω and C. Hence,

$$\|w\|_{H^1} \le M_2\|v\|_{H\Lambda^1}, \qquad \|\tau\|_{L^2} \le M_1M_2\|v\|_{H\Lambda^1}.$$

Moreover, by definition

$$(\tau, \epsilon y) = (v, y)_{H\Lambda^1}, \qquad \forall y \in \mathring{H}^1 \otimes \mathbb{R}^n. \tag{4.10}$$

Take any test function ρ. 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:

$$(\tau, \epsilon\nabla\rho) = (v, \nabla\rho)_{H\Lambda^1} = (v, \nabla\rho) \leq \|v\|_{L^2} \|\rho\|_{H^1}. \tag{4.11}$$

Notice $\epsilon\nabla$ is just the Hessian. Thus by definition, $\operatorname{div}\operatorname{div}\tau$ is a distribution in H^{-1} with:

$$\|\operatorname{div}\operatorname{div}\tau\|_{H^{-1}} \leq \|v\|_{L^2}. \tag{4.12}$$

Hence $\tau \in H(\operatorname{div}\operatorname{div})$ with

$$\|\tau\|_{H(\operatorname{div}\operatorname{div})} \leq M_3 \|v\|_{H\Lambda^1}.$$

Now take the regular decomposition of $v =: \nabla\phi + z$ for some $\phi \in \mathring{H}^1$ and $z \in \mathring{H}^1 \otimes \mathbb{R}^3$ with $\|\phi\|_{H^1} + \|z\|_{H^1} \leq M_4 \|v\|_{H\Lambda^1}$. The z-part can be plugged into equation (4.10) by choosing $y = z$,

$$\langle -\operatorname{div}\tau, z\rangle := (\tau, \epsilon z) = (v, z)_{H\Lambda^1}.$$

Because test functions are dense in $\mathring{H}^1$ and both sides of equation (4.11) are continuous in ρ under the H^1-norm on ρ. By continuity, equation (4.11) holds for $\rho \in \mathring{H}^1$. Choose $\rho = \phi$,

$$\langle -\operatorname{div}\tau, \nabla\phi\rangle := (\tau, \epsilon\nabla\phi) = (v, \nabla\phi)_{H\Lambda^1}.$$

Adding two preceding equations up, we get

$$\langle -\operatorname{div}\tau, v\rangle = \|v\|^2_{H\Lambda^1}.$$

This equation and estimate (4.12) together imply the inf-sup condition for $\langle \operatorname{div}\tau, v\rangle$. This proves the theorem. □

The mixed system (4.9) solves the linear elasticity equation (4.8) when the body force f is in $H^{-1}(\operatorname{div})$.

Theorem 4.8. *On a bounded Lipschitz domain, suppose $f \in H^{-1}(\operatorname{div})$. Let (σ, u) be the unique solution to the mixed system (4.9). Then $u \in \mathring{H}^1 \otimes \mathbb{R}^n$ and its solves the elasticity equation (4.8).*

Proof. The elasticity equation has a unique solution in $\mathring{H}^1 \otimes \mathbb{R}^n$ when f is from a bigger space $H^{-1} \otimes \mathbb{R}^n$. Given this special f, let w be that unique solution in $\mathring{H}^1 \otimes \mathbb{R}^n$. Since the mixed system has a unique solution, we have proven the theorem if we can show that $(C\epsilon w, w)$ 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 $\operatorname{div} \sigma = -f \in H^{-1}(\operatorname{div})$, that is, $\operatorname{div}\operatorname{div} \sigma \in H^{-1}$. Hence $\sigma \in H(\operatorname{div}\operatorname{div})$ and satisfies the second equation of system (4.9). For vector-valued test functions y we have,

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

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

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

This shows that the first equation of system (4.9) is satisfied as well. □

4.3.2 Rotated Regge element discretization

In this subsection, we show how to discretize the mixed formulation (4.9) 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 [13]. A different analysis approach for a different finite element discretization of the same mixed formulation (4.9) is given in [90].

The relationship between $H(\operatorname{div}\operatorname{div})$ and piecewise normal-normal continuous finite elements were already studied in the biharmonic section of this chapter. We still use $S(\mathrm{REG}^r)$ to discretize $H(\operatorname{div}\operatorname{div})$. The finite element theory for the space $H\Lambda^1$ is well-understood [8, 10]. We use the FEEC element $\mathscr{P}^r\Lambda^1$ to discretize $H\Lambda^1$. In dimension 2 and 3, $\mathscr{P}^r\Lambda^1$ is the space of Nédéléc edge elements of the second kind, which is widely used. The only thing that remains here is to derive the formula for the pairing $\langle \operatorname{div} \tau, v \rangle$. It is more natural to define this in mesh dependent spaces, in a fashion very similar to the that of the biharmonic case.

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

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

$$W := \{\text{piecewise } H^1 \text{ vector fields in } \mathring{H}\Lambda^1 \text{ with tangential continuity}\}.$$

Note that piecewise H^1 vector fields with tangential continuity already forms a subspace of $H\Lambda^1$. Hence the condition $\mathring{H}\Lambda^1$ 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:

$$\|\sigma\|_V^2 = \sum_c \|\sigma\|_{L^2(c)}^2 + h\|\sigma_{nn}\|_{L^2(\partial c)}^2,$$

$$\|u\|_W^2 = \sum_c \|u\|_{H^1(c)}^2 + h^{-1}\|u\|_{L^2(\partial c)}^2,$$

where both sums are over all the cells c in mesh $\mathscr{T}_h$. We define a mesh-dependent div_h operator: for any $(\tau, v) \in V \times W$,

$$\langle \operatorname{div}_h \tau, v\rangle := \sum_c \int_c -\tau : \epsilon v + \int_{\partial c} \tau_{nn} v_n = \sum_c \int_c -\tau : \epsilon v + \sum_f \int_f \tau_{nn} [\![v_n]\!], \tag{4.13}$$

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

$$|\langle \operatorname{div}_h \tau, v\rangle| \le M\|\tau\|_V \|v\|_W.$$

We now introduce our finite element choices as subspaces of V and W. For $r \ge 1$, let

$$V_h := S(\mathrm{REG}^r), \qquad W_h := \mathscr{P}^r\Lambda^1 \cap \mathring{H}\Lambda^1.$$

The discrete problem corresponding to mixed system (4.9) is: given $f \in H^{-1}(\operatorname{div})$, find $(\sigma, u) \in V_h \times W_h$, such that

$$\begin{aligned} (A\sigma, \tau) + \langle u, \operatorname{div}_h \tau\rangle &= 0, \quad && \forall \tau \in V_h, \\ \langle \operatorname{div}_h \sigma, v\rangle &= -\langle f, v\rangle, && \forall v \in W_h. \end{aligned} \tag{4.14}$$

An obvious question is, given that div_h is not really div and $V_h \times W_h$ does not have an apparent relationship to $H(\operatorname{div}\operatorname{div}) \times \mathring{H}\Lambda^1$, how is system (4.14) a discretization of the mixed system (4.9) 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 (4.14) in some sense, and discrete stability, which means that (4.14) 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 (4.8) holds for smooth C on convex polyhedral or C^2 domains [45] for $f \in L^2 \otimes \mathbb{R}^n$.

Theorem 4.9. *Suppose $u \in (\mathring{H}^1 \cap H^2) \otimes \mathbb{R}^n$ solves the elasticity equation (4.8). Let $\sigma := C\epsilon u$. Then (σ, u) satisfies system (4.14).*

Proof. First, it is clear that $u \in W$ and $\sigma \in H^1 \otimes \mathbb{S}^n \subset V$. Hence the equations in system (4.14) still make sense for this continuous (σ, u). Since $u \in \mathring{H}^1$ globally, $[\![u_n]\!] = 0$ (this includes the condition that u vanishes on the domain boundary). Thus,

$$\langle u, \operatorname{div}_h \tau \rangle = \int_\Omega -\tau : \epsilon u = \int_\Omega -A\tau : C\epsilon u = \int_\Omega -A\tau : \sigma = -(A\sigma, \tau).$$

This proves that the first equation is satisfied. Second, because $\sigma \in H^1 \otimes \mathbb{S}^n$, we have $[\![\sigma_{n\tau}]\!] = 0$ at all interior facets. On the other hand, $v_\tau = 0$ at all boundary facets. So overall,

$$\sum_c \int_{\partial_c} \sigma_{n\tau} v_\tau = \sum_f \int_f [\![\sigma_{n\tau}]\!] v_\tau = 0.$$

Because $u \in H^2 \otimes \mathbb{R}^n$, we have $f \in L^2 \otimes \mathbb{R}^n$. Thus,

$$\begin{aligned}(-f, v) = (\operatorname{div}\sigma, v) = \sum_c \int_c (\operatorname{div}\sigma) v = \sum_c \int_c \sigma : (-\epsilon v) + \int_{\partial c} \sigma_n \cdot v \\ = \sum_c \int_c \sigma : (-\epsilon v) + \int_{\partial c} \sigma_{nn} v_n = \langle \operatorname{div}_h \sigma, v \rangle,\end{aligned}$$

where the second to last equation used the decomposition $\sigma_n \cdot v = \sigma_{nn} v_n + \sigma_{n\tau} v_\tau$ and the previous identity. This proves that the second equation is satisfied as well. □

For software implementation of system (4.14) in an environment where REG^r is already implemented, as is the case for FEniCS, the following equivalent formulation should be used: find $(\rho, u) \in \mathrm{REG}^r \times W_h$ such that

$$\begin{aligned}(AS\rho, S\tau) - \langle u, \operatorname{div}_h S\tau \rangle &= 0, \quad \forall \tau \in \mathrm{REG}^r, \\ \langle \operatorname{div}_h S\rho, v \rangle &= -\langle f, v \rangle, \quad \forall v \in W_h.\end{aligned} \tag{4.15}$$

4.3.3 Numerical experiments

In this subsection, we show through numerical experiments that discretization (4.15) is stable but converges with suboptimal L^2 error rates in the σ variable. We further show that this method can be implemented to handle different boundary conditions in both 2D and 3D and does not suffer from locking.

First, we look at the convergence test for 2D. The elasticity equation (4.8) is solved on the unit square with the following exact solution:

$$u = \begin{bmatrix} \sin(\pi x)\sin(\pi y) \\ 15x(1-x)y(1-y) \end{bmatrix} \in C^\infty \cap \mathring{H}^1.$$

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/tdnns_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 σ.

We also tested the same formulation with Nédéléc edge elements of the first kind of the same degree, a smaller space, in place of Nédéléc edge elements of the second kind. The resulting scheme is 1 order suboptimal in both the stress and displacement variables. In particular, it is unstable for degree 1.

Mesh size	$\|u - u_h\|$	Rate	$\|\sigma - \sigma_h\|$	Rate
8	7.094244e-03		1.687453e-01	
16	1.742923e-03	2.08	8.163971e-02	1.08
32	4.548902e-04	1.88	4.239306e-02	0.92
64	1.130114e-04	2.00	2.114090e-02	1.00
128	2.834921e-05	1.98	1.062154e-02	0.99

Table 4.7: 2D elasticity degree 1

Mesh size	$\|u - u_h\|$	Rate	$\|\sigma - \sigma_h\|$	Rate
4	1.945019e-03		3.559110e-02	
8	2.633687e-04	2.87	8.278951e-03	2.10
16	3.096338e-05	3.17	1.910408e-03	2.17
32	3.931342e-06	2.89	4.794675e-04	1.94
64	5.015333e-07	2.95	1.224961e-04	1.96

Table 4.8: 2D elasticity degree 2

Mesh size	$\|u-u_h\|$	Rate	$\|\sigma-\sigma_h\|$	Rate
2	1.346217e-03		1.646650e-02	
4	1.031960e-04	4.06	2.660197e-03	2.88
8	7.203260e-06	3.83	3.054329e-04	3.11
16	4.397091e-07	4.14	3.678944e-05	3.13
32	2.857008e-08	3.83	4.766975e-06	2.87
64	1.779997e-09	3.98	5.977013e-07	2.98

Table 4.9: 2D elasticity degree 3

We then study the convergence rates in 3D. The linear elasticity equation is solved on the unit cube with the following exact solution:

$$u = \begin{bmatrix} \sin(\pi x)\sin(\pi y)\sin(\pi z) \\ 15x(1-x)y(1-y)z(1-z) \\ 7x(1-x)\sin(\pi y)\sin(\pi z) \end{bmatrix} \in C^\infty \cap \mathring{H}^1.$$

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 σ.

Mesh size	$\|u-u_h\|$	Rate	$\|\sigma-\sigma_h\|$	Rate
2	3.391169e-01		1.832970e+00	
4	1.049142e-01	1.86	1.048096e+00	0.89
6	4.901894e-02	1.54	7.042752e-01	0.81
8	2.892313e-02	2.01	5.382773e-01	1.02
10	1.908383e-02	2.07	4.314971e-01	1.10

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

$$\epsilon u = \frac{(1+\nu)\sigma - \nu I \operatorname{tr}\sigma}{E},$$

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.

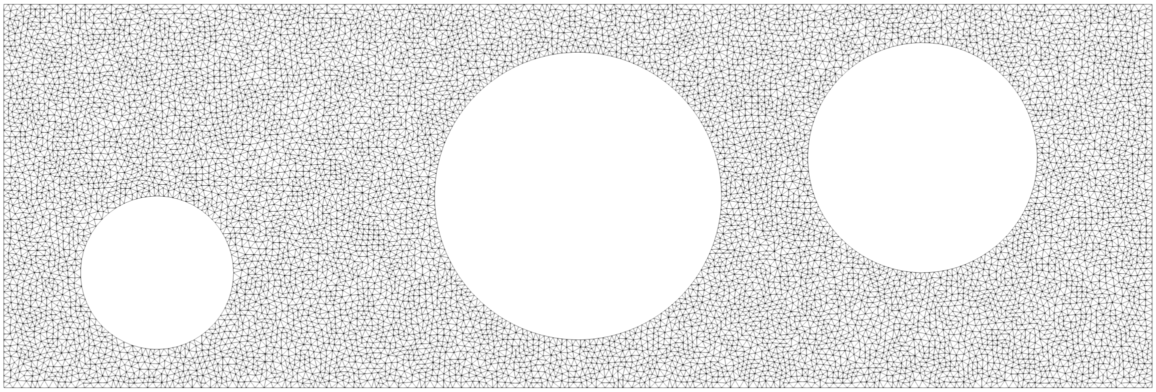


Figure 4.6: Domain and mesh for the 2D elasticity example

We note that in the TDNNS formulation, the tangential part of the displacement u_τ is an essential boundary condition, while the normal part is a natural boundary condition. Suppose $un = gn$ on the part of the boundary Γ_N. Then we get

$$\int_{\Gamma_N} \tau_{nn} g_n$$

in the right-hand side of the first equation of (4.14). 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 (4.14).

This problem is solved with degree $r = 1$. A plot of the solution is shown in Figuer 4.7. 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 $\sigma_d := \sigma - \frac{1}{2} I \operatorname{tr}\sigma$.

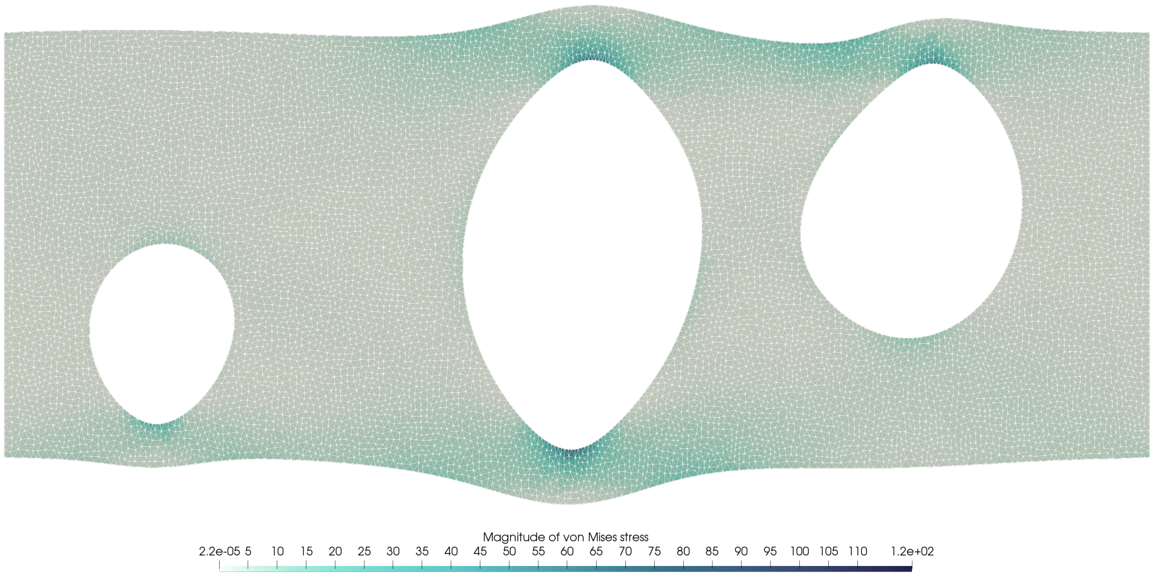


Figure 4.7: Visualization of the 2D solution with Poisson ratio $\nu = 0.2$

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 ration is close to 0.5, the quality of the numerical solution degrades substantially. Mixed methods should not suffer from this. In Figure 4.8, 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.

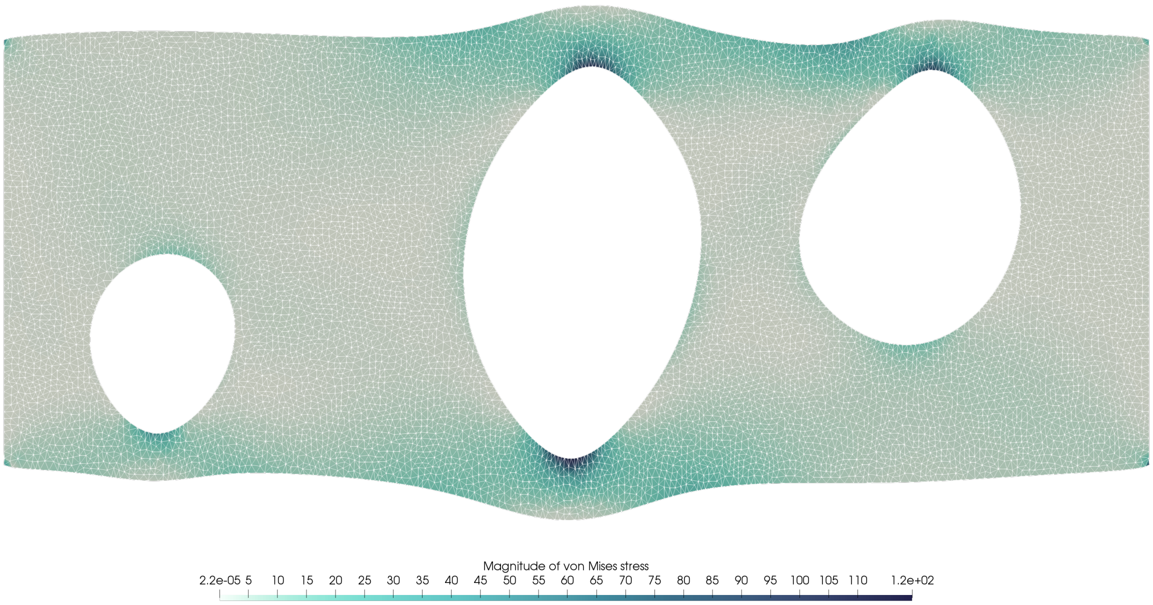


Figure 4.8: Visualization of the 2D solution with Poisson ratio $\nu = 0.499999$

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 4.9 and Figure 4.10.

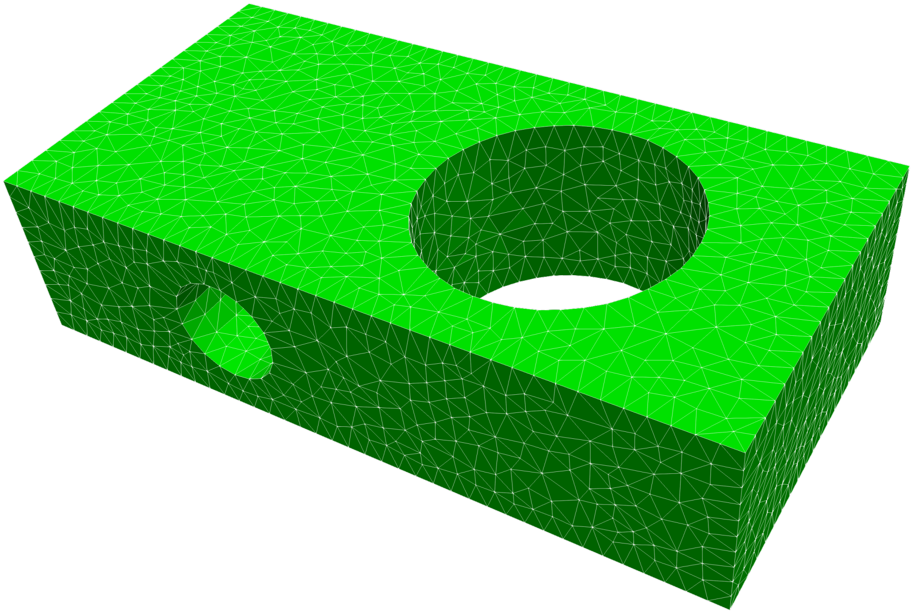


Figure 4.9: 3D problem domain

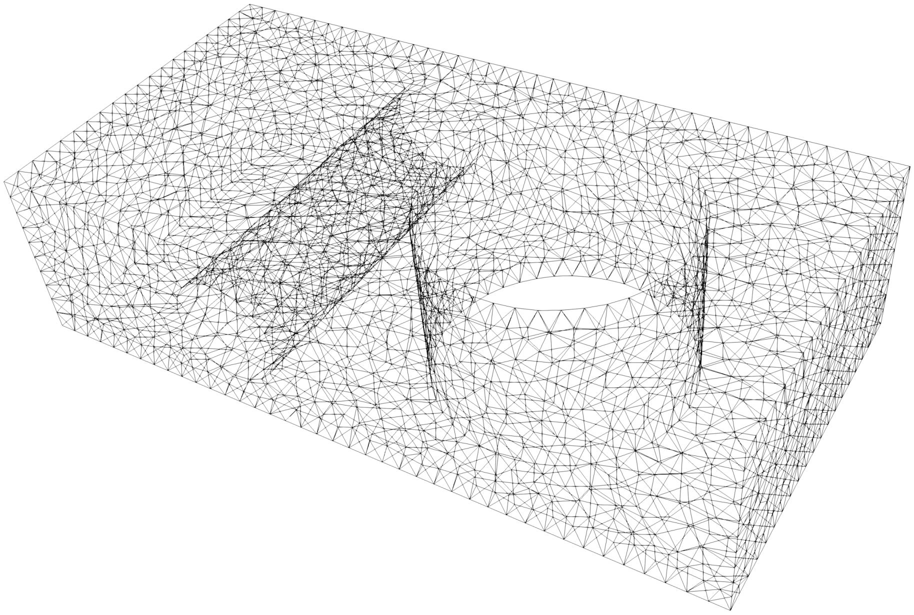


Figure 4.10: 3D problem mesh

The material is again isotropic and homogeneous with Young's modulus $E = 1.0$ and Poisson'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 4.11 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 $\sigma_d := \sigma - \frac{1}{3} I tr \sigma$.

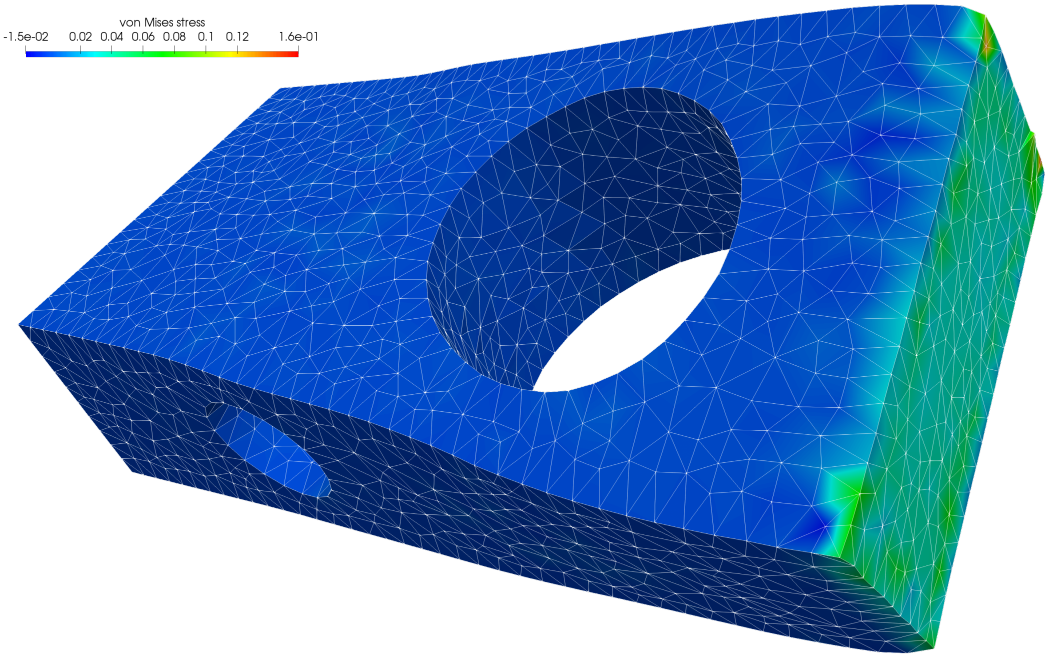


Figure 4.11: Visualization of the 3D solution with Poisson ratio $\nu = 0.2$

Figure 4.12 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.

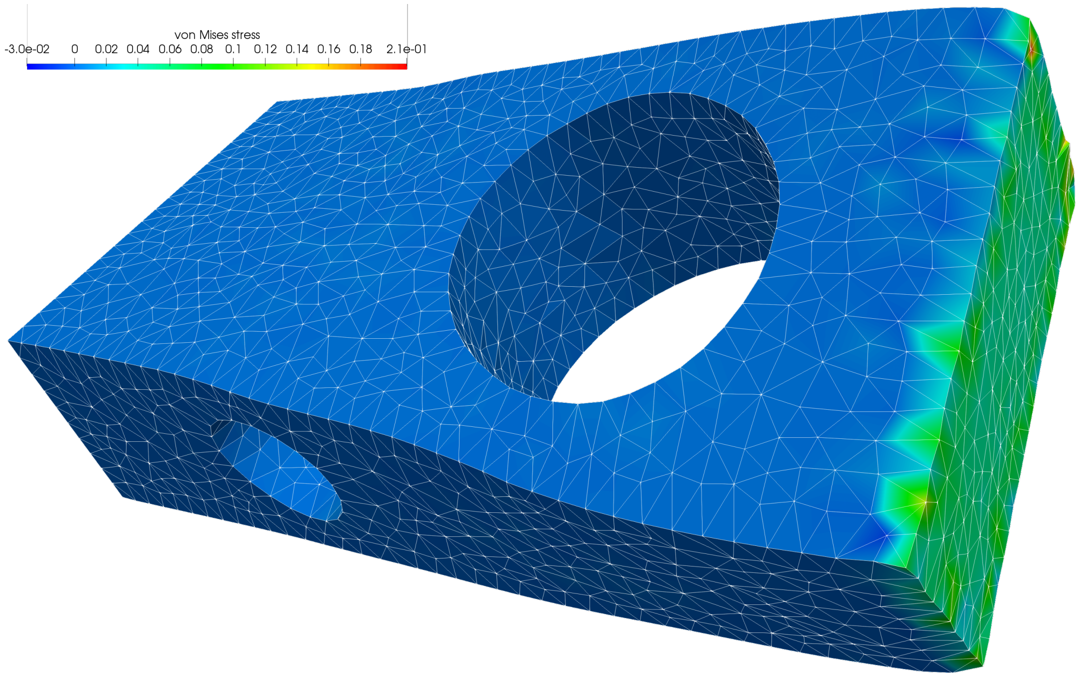


Figure 4.12: Visualization of the 3D solution with Poisson ratio $\nu = 0.499999$

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:

$$\begin{aligned} &\operatorname{div}\operatorname{div} S\gamma = 0,\\ &\operatorname{div} S\gamma' + \operatorname{curl}\operatorname{curl}\beta = 0,\\ &S\gamma'' + 2\operatorname{ein}\gamma + S\nabla\nabla\alpha - 2S\epsilon\beta' = 0. \end{aligned}$$

where α is scalar field, β is a vector field, γ 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 γ. For example, we would at least want to know to what extent the constraints are satisfied. We then need to look at $\operatorname{div}\operatorname{div} S$ and $\operatorname{div} S$ of functions in REG^r. Notices 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.

In some sense, these two are the most natural equations derived from the operators $\operatorname{div}\operatorname{div}S$ and $\operatorname{div}S$. 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:

$$\min \|\sigma\|, \qquad \text{subject to } L\sigma = f.$$

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 $\operatorname{div}\operatorname{div}S$ and $\operatorname{div}S$ 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 $\operatorname{div}\operatorname{div}S\gamma$ is likely to hold in the discrete sense, when tested against CG^{r+1}, while the constraint involving $\operatorname{div}S\gamma'$ is likely to hold when tested against $\mathscr{P}^r\Lambda^1$.

Chapter 5

Model problems in relativity for discretization

In this chapter, we first identify two linear problems below which are of key importance for developing and analyzing the Galerkin discretization of the fully nonlinear space-time Einstein equation. They are the Cauchy Problem 5.1 and the Source Problem 5.2. Both are quoted below for the convenience of reader. Here S and J are two algebraic operators on symmetric matrix-valued functions:

$$Su = u - I\,\mathrm{tr}\,u, \qquad Ju = u - \frac{1}{2} I\,\mathrm{tr}\,u,$$

and ein is a second-order linear differential operator derived from linearizing the Einstein tensor at the Euclidean metric (defined in equation (5.5)).

Cauchy Problem. Given two smooth symmetric matrix fields γ_0, γ_1 satisfying the *compatibility conditions*:

$$\mathrm{div}\,\mathrm{div}\, S\gamma_0 = 0, \qquad \mathrm{div}\, S\gamma_1 = 0,$$

find a symmetric matrix field, such that $\gamma(0) = \gamma_0$, $\gamma'(0) = \gamma_1$, and for all $t > 0$:

$$S\gamma'' + 2\,\mathrm{ein}\,\gamma = 0.$$

Source Problem. Given a smooth symmetric matrix field f in the range of ein, find a symmetric matrix field u such that

$$2\,\mathrm{ein}\,u = f, \qquad \mathrm{div}\, Ju = 0.$$

The derivation of the two problems from the fully nonlinear space-time Einstein equation is cut into 5 parts, from Section 5.1 to Section 5.5. We also discuss why they are important and other applications of these two problems.

In Section 5.6, we prove that both model problems are well-posed on the flat torus (a cube with periodic boundary conditions): Theorem 5.2 for the Cauchy Problem and Theorem 5.1 for the Source Problem. However, we point out that both problems are flawed in some sense and need regularization. In particular, the Source Problem is only solvable for special divergence-free data. With numerical errors, it is likely that any discretization will lead to an inconsistent linear system. There is also no efficient solvers for such potentially inconsistent linear systems. Thus it is not suitable for direct discretization. The Cauchy Problem is only weakly hyperbolic but not strongly hyperbolic. As will be shown, this means that though the equation itself is well-posed, it can become ill-posed with either lower-order terms or variable coefficients. Since these two are inevitable eventually for solving the nonlinear problem, the Cauchy Problem is not suitable for direct discretization either.

In Section 5.7, we introduce the regularized versions of the two model problems: Regularized Source Problem 5.3 and Regularized Cauchy Problem 5.4. These two are also quoted below for the convenience of the reader. Let Ω be a bounded smooth contractible domains in $\mathbb{R}^3$ and n its unit outward normal vector on the boundary. Then $\tau := I - nn^T$ is the projection to the tangential space of the boundary. For a symmetric matrix-valued function u on Ω, on the boundary we define

$$u_{nn} := n^T un, \qquad u_{n\tau} := n^T u\tau, \qquad u_{\tau n} := \tau^T un = u_{n\tau}^T, \qquad u_{\tau\tau} := \tau^T u\tau \tag{5.1}$$

to be the normal-normal, normal-tangential, tangential-normal, and tangential-tangential part of u respectively. Set

$$V := \{u \in H^1 \otimes \mathbb{S}^3 \,|\, u_{\tau\tau} = 0,\, u_{nn} = 0 \text{ on } \partial\Omega\}, \qquad Y := V \cap \{u \in H^2 \otimes \mathbb{S}^3 \,|\, \partial_n u_{n\tau} = 0 \text{ on } \partial\Omega\}.$$

Regularized Source Problem. Given $f \in L^2 \otimes \mathbb{S}^3$, find $u \in Y$ such that

$$-\Delta u = f.$$

To simplify the notation, time dependent function spaces like $C^0([0,T], H^1)$ is shortened to C^0H^1.

Regularized Hyperbolic Problem. Given $u_0, u_1 \in Y$ and $f \in C^0(L^2 \otimes \mathbb{S}^3)$, find $u \in C^0Y \cap C^1V \cap C^2(L^2 \otimes \mathbb{S}^3)$ such that $u(0) = u_0$, $u'(0) = u_1$, and for all $t > 0$,

$$u'' - \Delta u = f.$$

Theorem 5.3 shows that the Regularized Source Problem is a well-posed elliptic problem. Theorem 5.4 proves that the Source Problem can be solved using the Regularized Source Problem with the right-hand side Sf. Theorem 5.5 shows that the Regularized Hyperbolic

Problem is also well-posed. In fact it is strongly hyperbolic. Theorem 5.6 proves that we can use the Regularized Hyperbolic Problem to solve the Cauchy Problem. The key conclusion is that the two regularized problems are suitable for discretization and can be used to solve problems in relativity.

Finally in the last Section 5.8, we hint at how generalized Regge elements can be used to solve these two regularized problems. This study is still in its very early stages and will be future work.

5.1 The fully nonlinear space-time Einstein equation

The Einstein field equation [36] is a well-established model for large-scale structures of the universe. It is a nonlinear second-order partial differential equation for symmetric 2-tensor fields. Relevant facts are recalled here for the convenience of the reader. Further details can be found in many textbooks, for example [49,81,111].

We will write down the Einstein equation in coordinates following the notation in [5]. 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 of the metric containing its first-order derivatives [5, Equation (1.8.12)]:

$$\Gamma^{\alpha}_{\beta\gamma} := \frac{1}{2} g^{\alpha\mu}(\partial_\gamma g_{\beta\mu} + \partial_\beta g_{\gamma\mu} - \partial_\mu g_{\beta\gamma}),$$

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 [5, Equation (1.9.2)]:

$$R^{\alpha}_{\beta\mu\nu} := \partial_\mu \Gamma^{\alpha}_{\beta\nu} - \partial_\nu \Gamma^{\alpha}_{\beta\mu} + \Gamma^{\alpha}_{\rho\mu}\Gamma^{\rho}_{\beta\nu} - \Gamma^{\alpha}_{\rho\nu}\Gamma^{\rho}_{\beta\mu}.$$

The *Einstein tensor* is then defined in terms of the Riemann tensor [5, Equation (1.10.4)]:

$$G_{\mu\nu} := R^{\alpha}_{\mu\alpha\nu} - \frac{1}{2} g_{\mu\nu} g^{\beta\gamma} R^{\alpha}_{\beta\alpha\gamma},$$

Our main equation, the vacuum *Einstein field equation* is:

$$G_{\mu\nu} = 0. \tag{5.2}$$

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 $u_{\alpha\beta}g^{\mu\alpha}$.

Proposition 5.1. *The Einstein tensor as a function of the metric splits as a second-order principal term and lower-order terms:*

$$G_{\mu\nu} = \frac{1}{2}\left(-\partial^{\lambda}\partial_{\lambda}g_{\mu\nu} + \partial_{\mu}\partial^{\lambda}g_{\lambda\nu} + \partial_{\nu}\partial^{\lambda}g_{\lambda\mu} - g^{\alpha\beta}\partial_{\mu}\partial_{\nu}g_{\alpha\beta} - g_{\mu\nu}\partial^{\alpha}\partial^{\beta}g_{\alpha\beta} + g_{\mu\nu}g^{\alpha\beta}\partial^{\lambda}\partial_{\lambda}g_{\alpha\beta}\right) + Q_{\mu\nu}, \tag{5.3}$$

where each component of $Q_{\mu\nu}$ is a polynomial in $g_{\alpha\beta}$, $g^{\alpha\beta}$, and $\partial_{\lambda}g_{\alpha\beta}$. In particular, each term of each component of $Q_{\mu\nu}$ is exactly quadratic in $\partial_{\lambda}g_{\alpha\beta}$.

Proof. Notes that derivatives on the inverse metric can be moved to the metric via:

$$0 = \partial_{\nu}(\delta^{\alpha}_{\mu}) = \partial_{\nu}(g^{\alpha\beta}g_{\beta\mu}) = g_{\beta\mu}\partial_{\nu}g^{\alpha\beta} + g^{\alpha\beta}\partial_{\nu}g_{\beta\mu}.$$

Then the definitions of the Christoffel symbol and the Riemann tensor imply:

$$R^{\alpha}_{\beta\mu\nu} = \frac{1}{2}g^{\alpha\lambda}[\partial_{\mu}(\partial_{\beta}g_{\lambda\nu} - \partial_{\lambda}g_{\beta\nu}) - \partial_{\nu}(\partial_{\beta}g_{\lambda\mu} - \partial_{\lambda}g_{\beta\mu})] + P^{\alpha}_{\beta\mu\nu},$$

where $P^{\alpha}_{\beta\mu\nu}$ is a polynomial in $g^{\alpha\beta}$ and $\partial_{\lambda}g_{\alpha\beta}$, each term of each component of which is exactly quadratic in the latter. Taking the $\alpha\mu$ trace, we get,

$$R^{\alpha}_{\beta\alpha\nu} = -\frac{1}{2}\partial^{\lambda}\partial_{\lambda}g_{\beta\nu} + \frac{1}{2}(\partial_{\beta}\partial^{\lambda}g_{\lambda\nu} + \partial_{\nu}\partial^{\lambda}g_{\lambda\beta}) - \frac{1}{2}g^{\alpha\lambda}\partial_{\beta}\partial_{\nu}g_{\alpha\lambda} + S_{\beta\nu},$$

where $S_{\beta\nu}$ is a quadratic polynomial in $g^{\alpha\beta}$ and $\partial_{\lambda}g_{\alpha\beta}$, each term of each component of which is exactly quadratic in the latter. Taking another trace using the metric:

$$g^{\beta\nu}R^{\alpha}_{\beta\alpha\nu} = \partial^{\nu}\partial^{\lambda}g_{\lambda\nu} - g^{\beta\nu}\partial^{\lambda}\partial_{\lambda}g_{\beta\nu} + g^{\beta\nu}S_{\beta\nu},$$

Plugging these into the definition of the Einstein tensor proves the claim. □

5.2 Linearized space-time Einstein equation

From the perspective of numerical analysis of Galerkin methods, broadly speaking, the following diagram commutes:

$$\begin{tikzcd}
\text{Continuous nonlinear problem} \arrow[r, "\text{Galerkin discretization}"] \arrow[d, "\text{Iterative linearization}"] & \text{Discrete nonlinear problem} \arrow[d, "\text{Iterative linearization}"] \\
\text{Continuous linear problem} \arrow[r, "\text{Galerkin discretization}"] & \text{Discrete linear problem}
\end{tikzcd}$$

where iterative linearization means solving nonlinear problems via solving a sequence of linearized problems (for example, via Newton's method). Hence for sufficiently regular nonlinear problems, a good discretization scheme for the linearized problem directly gives a method to solve the nonlinear problem.

It is clear that the Minkowski metric $\eta_{\mu\nu} = \mathrm{diag}[-1,1,1,1]$ on $\mathbb{R}^4$ satisfies the Einstein equation. As a first step, we linearize the Einstein equation around that. Physically, this leads to models of gravitational waves passing through the empty space. The theorem below holds for $g_{\mu\nu} = \bar{g}_{\mu\nu} + sh_{\mu\nu}$ with any constant background $\bar{g}_{\mu\nu}$ in $\mathbb{R}^m$ with the same proof. But we only state it for the Minkowski metric $\eta_{\mu\nu}$ to simplify the notation.

Proposition 5.2. *Let $g_{\mu\nu} = \eta_{\mu\nu} + sh_{\mu\nu}$ for some symmetric 2-tensor field $h_{\mu\nu}$ with $s \in \mathbb{R}$ and $G_{\mu\nu}$ the Einstein tensor of $g_{\mu\nu}$. We define a linear operator* ein *on symmetric 2-tensor fields by*

$$(\mathrm{ein}\, h)_{\mu\nu} := \frac{d}{ds} G_{\mu\nu}\Big|_{s=0}.$$

Then,

$$(2\,\mathrm{ein}\, h)_{\mu\nu} = -\partial^\lambda \partial_\lambda h_{\mu\nu} + \partial_\mu \partial^\lambda h_{\lambda\nu} + \partial_\nu \partial^\lambda h_{\lambda\mu} - \partial_\mu \partial_\nu h^\alpha_\alpha - \eta_{\mu\nu} \partial^\alpha \partial^\beta h_{\alpha\beta} + \eta_{\mu\nu} \partial^\lambda \partial_\lambda h^\alpha_\alpha, \tag{5.4}$$

where the background metric $\eta_{\mu\nu}$ is used to raise and lower indices.

Proof. Plug $g_{\mu\nu} := \eta_{\mu\nu} + sh_{\mu\nu}$ into equation (5.3) of Proposition 5.1 and compute to the first-order in s. Because $\partial_\lambda g_{\alpha\beta} = \epsilon \partial_\lambda h_{\alpha\beta}$, the $Q_{\mu\nu}$ part, having each term of each component exactly quadratic in $\partial_\lambda g_{\alpha\beta}$, is of the order $Q_{\mu\nu} \sim O(s^2)$. Further, the inverse metric of $g_{\mu\nu}$ is:

$$g^{\mu\nu} = \eta^{\mu\nu} - sh^{\mu\nu} + O(s^2).$$

Then the first three principal terms of the Einstein tensor gives:

$$\frac{s}{2}\left(-\partial^\lambda \partial_\lambda h_{\mu\nu} + \partial_\mu \partial^\lambda h_{\lambda\nu} + \partial_\nu \partial^\lambda h_{\lambda\mu}\right) + O(s^2),$$

where the s^2 term comes from the fact that in the nonlinear formula $g^{\mu\nu}$ is used to raise the index in ∂^λ. The computation for the rest three principal terms is tedious. For example: the $g_{\mu\nu} g^{\alpha\beta} \partial^\lambda \partial_\lambda g_{\alpha\beta}$ term becomes:

$$\begin{aligned}(\eta_{\mu\nu} + sh_{\mu\nu})(\eta^{\alpha\beta} - sh_{\alpha\beta})(\eta^{\lambda\tau} - sh_{\lambda\tau})\partial_\tau \partial_\lambda(\eta_{\alpha\beta} + sh_{\alpha\beta})& \\ = s\eta_{\mu\nu}\eta^{\alpha\beta}\eta^{\lambda\tau}\partial_\tau \partial_\lambda h_{\alpha_\beta} + o(s^2) = s\eta_{\mu\nu}\partial^\lambda \partial_\lambda h^\alpha_\alpha + O(s^2).&\end{aligned}$$

The computation for the other two are similar. In all last three principal terms contribute:

$$\frac{s}{2}\left(-\partial_\mu \partial_\nu h^\alpha_\alpha - \eta_{\mu\nu}\partial^\alpha \partial^\beta h_{\alpha\beta} + \eta_{\mu\nu}\partial^\lambda \partial_\lambda h^\alpha_\alpha\right) + O(s^2).$$

Combining these, we get the claim. □

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 Einstein tensor would be the exactly same as $(\mathrm{ein}\, h)_{\mu\nu}$, except that the indices are raise by a different background metric. Hence, up to low lower-order terms and variable coefficients in the principal part, the linearized Einstein tensor at any solutions is of the form

$$(\mathrm{ein}\, h)_{\mu\nu} = 0.$$

We therefore, reduce the problem of developing and analysing numerical methods for the full nonlinear Einstein equation to the same problem for the much simpler linear equation above, requiring the numerical methods to be robust against variable coefficients and lower-order terms.

5.3 Matrix calculus notation

In this section, we switch from the index notation to matrix calculus notation, which is more familiar. The background Minkowski metric on $\mathbb{R}^4$ establishes a canonical Euclidean coordinate system, under which symmetric 2-tensor fields are identified with symmetric matrix fields and ein becomes a matrix of familiar differential operators in calculus.

First, we recall some basic operators. For any scalar field u and vector field v, we have the *gradient*, *hessian*, and the *symmetric gradient*:

$$(\nabla u)_\alpha := \partial_\alpha u, \qquad (\nabla\nabla u)_{\alpha\beta} := \partial_\alpha \partial_\beta u, \qquad (\epsilon v)_{\alpha\beta} := \frac{1}{2}(\partial_\alpha v_\beta + \partial_\beta v_\alpha).$$

The next batch of the differential operators depend on the metric. For us, this metric is either Euclidean $I := \mathrm{diag}[1,\dots,1]$ or Minkowskian $\eta := \mathrm{diag}[-1,1,\dots,1]$. The *divergence* and the *Laplacian* under a metric g is:

$$(\mathrm{div}_g\, u) := g^{\alpha\beta}\partial_\alpha u_\beta, \qquad (\mathrm{div}_g\, v)_\beta := g^{\alpha\lambda}\partial_\lambda v_{\alpha\beta}, \qquad \Delta_g := g^{\alpha\beta}\partial_\alpha\partial_\beta.$$

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,

$$(\mathrm{tr}_g\, u) := g^{\alpha\beta} u_{\alpha\beta}, \qquad J_g u := u - \frac{1}{2} g(\mathrm{tr}_g\, u), \qquad S_g u := u - g(\mathrm{tr}_g\, u).$$

To further simply the notation, the subscript for the metric is omitted when it is clear from the context.

Under Minkowski metric η, the linearized Einstein operator (5.4) becomes:

$$2\,\mathrm{ein}\,u = -\Delta u + 2\epsilon\,\mathrm{div}\,u - \nabla\nabla\,\mathrm{tr}\,u - \eta\,\mathrm{div}\,\mathrm{div}\,u + \eta\Delta\,\mathrm{tr}\,u.$$

We immediately see that if u happens to be divergence-free and trace-free, then $2\,\mathrm{ein}\,u$ is the d'Alembertian of u. In that case, the linearized Einstein equation is just a component-wise wave equation.

Several calculus identities which will be used very frequently are collected here:

Lemma 5.1. *In dimension m, under any constant background pseudo-Riemannian metric g, the following identities hold:*

$$J^{-1}u = u - \frac{1}{m-2}g(\mathrm{tr}\,u), \qquad S^{-1}u = u - \frac{1}{m-1}g(\mathrm{tr}\,u),$$
$$\mathrm{div}\,\epsilon = \frac{1}{2}\Delta + \frac{1}{2}\nabla\,\mathrm{div}, \qquad \mathrm{div}\,J\epsilon = \frac{1}{2}\Delta,$$
$$\mathrm{div}\,S\nabla\nabla = 0, \qquad \mathrm{div}\,\mathrm{div}\,S\epsilon = 0$$

For example, we have a more compact formula for the linearized Einstein:

$$2\,\mathrm{ein} = -J\Delta + 2J\epsilon\,\mathrm{div}\,J. \tag{5.5}$$

Proposition 5.3.

$$\mathrm{div}\,\mathrm{ein} = 0, \qquad \mathrm{ein}\,\epsilon = 0.$$

Proof. This is a direct consequence of $\mathrm{div}\,J\epsilon = \frac{1}{2}\Delta$ in Lemma 5.1:

$$\mathrm{div}\,\mathrm{ein} = -\frac{1}{2}\,\mathrm{div}\,J\Delta + (\mathrm{div}\,J\epsilon)\,\mathrm{div}\,J = -\frac{1}{2}\,\mathrm{div}\,J\Delta + \frac{1}{2}\Delta\,\mathrm{div}\,J = 0.$$

$$\mathrm{ein}\,\epsilon = -\frac{1}{2}\Delta J\epsilon + J\epsilon(\mathrm{div}\,J\epsilon) = -\frac{1}{2}\Delta J\epsilon + \frac{1}{2}J\epsilon\Delta = 0.$$

□

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 $\mathrm{ein}\,u = 0$, then $u + \epsilon\phi$ is also a solution for any vector field ϕ. This is an important feature of the Einstein equation.

5.4 Model linear Cauchy problem

As note before, the principal part of the Einstein equation can be understood as a d'Alembertian plus some unhelpful terms. To use the Einstein equation, it is reasonable 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 details, see monographs [26, 98]. Here we only need to set up an initial-value problem for the linearized Einstein equation. This is the topic of this section.

For the purpose of computation, we consider the background manifold $([0,T]\times\Omega,\eta)$ where Ω is a bounded smooth domain in $\mathbb{R}^3$ and η the Minkowski metric. We will defer the discussion on boundary conditions later when we discuss well-posedness.

We start by carrying out what is known as the $(1+3)$-decomposition: separating the temporal and spatial part of the matrix fields according to the natural factorization of $[0,T]\times\Omega$. 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. For symmetric matrix fields, we use the matrix notation:

$$h_{(4)} = \begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix},$$

where α is a scalar field for h_{00}, β is a 3D vector field for h_{0i}, and γ is a 3D symmetric matrix field for h_{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 decomposed as:

$$u_{(4)} = \begin{bmatrix} \phi \\ w \end{bmatrix},$$

where ϕ is a scalar field for u_0 and w is a 3D vector for u_i. The time derivative ∂_0 will be denoted by prime $'$ while spatial differential operators will continue be denoted using the matrix calculus notation. The following is proved by a direct computation:

Proposition 5.4.

$$\nabla_{(4)}u = \begin{bmatrix} u' \\ \nabla u \end{bmatrix}, \qquad \nabla\nabla_{(4)}u = \begin{bmatrix} u'' & (\nabla u')^T \\ \nabla u' & \nabla\nabla u \end{bmatrix}, \qquad \epsilon_{(4)}\begin{bmatrix} \phi \\ w \end{bmatrix} = \begin{bmatrix} \phi' & \frac{1}{2}(\nabla\phi + w')^T \\ \frac{1}{2}(\nabla\phi + w') & \epsilon w \end{bmatrix}$$

$$\operatorname{tr}_{(4)}\begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix} = -\alpha + \operatorname{tr}\gamma, \qquad \operatorname{div}_{(4)}\begin{bmatrix} \phi \\ w \end{bmatrix} = -\phi' + \operatorname{div} w, \qquad \operatorname{div}_{(4)}\begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix} = \begin{bmatrix} -\alpha' + \operatorname{div}\beta \\ -\beta' + \operatorname{div}\gamma \end{bmatrix},$$

$$J_{(4)}\begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(\alpha + \operatorname{tr}\gamma) & \beta^T \\ \beta & J\gamma + \frac{1}{2}I\alpha \end{bmatrix}, \qquad \Delta_{(4)}w = -w'' + \Delta w.$$

Given this, we can compute the $(1+3)$ decomposition of ein:

Proposition 5.5.

$$2\operatorname{ein}_{(4)}\begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix} = \begin{bmatrix} \operatorname{div}\operatorname{div} S\gamma & [\operatorname{div} S(\gamma' - 2\epsilon\beta)]^T \\ \operatorname{div} S(\gamma' - 2\epsilon\beta) & S\gamma'' + 2\operatorname{ein}\gamma + S\nabla\nabla\alpha - 2S\epsilon\beta' \end{bmatrix}. \tag{5.6}$$

Proof. This is a long computation using the previous theorem and formula (5.5). Since we know the matrices are symmetric, to simplify the notation, we omit the upper right corner. First,

$$-\frac{1}{2}J\Delta_{(4)}\begin{bmatrix}\alpha & \beta^T\\ \beta & \gamma\end{bmatrix} = \begin{bmatrix}\frac{1}{4}(\alpha''-\Delta\alpha+\operatorname{tr}\gamma''-\Delta\operatorname{tr}\gamma) & \dots\\ \frac{1}{2}(\beta''-\Delta\beta) & \frac{1}{2}(J\gamma''-\Delta J\gamma)+\frac{1}{4}I(\alpha''-\Delta\alpha)\end{bmatrix}.$$

Second,

$$\begin{aligned}J\epsilon\operatorname{div}J_{(4)}\begin{bmatrix}\alpha & \beta^T\\ \beta & \gamma\end{bmatrix} &= J\epsilon\operatorname{div}_{(4)}\begin{bmatrix}\frac{1}{2}(\alpha+\operatorname{tr}\gamma) & \dots\\ \beta & J\gamma+\frac{1}{2}I\alpha\end{bmatrix} = J\epsilon_{(4)}\begin{bmatrix}-\frac{1}{2}(\alpha'+\operatorname{tr}\gamma')+\operatorname{div}\beta\\ -\beta'+\operatorname{div}J\gamma+\frac{1}{2}\nabla\alpha\end{bmatrix}\\ &= J_{(4)}\begin{bmatrix}-\frac{1}{2}(\alpha''+\operatorname{tr}\gamma'')+\operatorname{div}\beta' & \dots\\ \frac{1}{2}(-\beta''+\operatorname{div}\gamma'-\nabla\operatorname{tr}\gamma'+\nabla\operatorname{div}\beta) & -\epsilon\beta'+\epsilon\operatorname{div}J\gamma+\frac{1}{2}\nabla\nabla\alpha\end{bmatrix}\\ &= \begin{bmatrix}\frac{1}{4}(-\alpha''-\operatorname{tr}\gamma''+\Delta\alpha)+\frac{1}{2}\operatorname{div}\operatorname{div}J\gamma & \dots\\ \frac{1}{2}(-\beta''+\operatorname{div}\gamma'-\nabla\operatorname{tr}\gamma'+\nabla\operatorname{div}\beta) & J\epsilon\operatorname{div}J\gamma+\frac{1}{2}J\nabla\nabla\alpha-\frac{1}{4}I(\alpha''+\operatorname{tr}\gamma'')-S\epsilon\beta'\end{bmatrix}.\end{aligned}$$

Combining these two parts, we get the claim. □

Identity (5.6) is valid in $(1+m)$-dimension for all $m \geq 1$. We have some extra identities which are only true for the case we care about $m = 3$ and are convenient:

Lemma 5.2. *In dimension* 3 *under the Euclidean metric, for a vector field u and a symmetric matrix field w,*

$$J^{-1} = S, \qquad \nabla\operatorname{div}u-\Delta u = \operatorname{curl}\operatorname{curl}u, \qquad \operatorname{div}S\epsilon u = -\frac{1}{2}\operatorname{curl}\operatorname{curl}u, \qquad 2\operatorname{ein}w = \operatorname{curl}(\operatorname{curl}w)^T,$$

where curl *of a matrix is defined row by row.*

The proof is a direct computation. In 3D, $2\operatorname{ein}$ is also known as the *Saint-Venant's operator* or the *incompatibility operator* in the solid mechanics literature.

We thus arrived at the $(1+3)$ linearized Einstein equation: a triple (α,β,γ) of time-dependent scalar, vector, symmetric matrix fields on Ω are components of a solution to the linearized Einstein equation if and only if:

$$\begin{aligned}&\operatorname{div}\operatorname{div}S\gamma = 0,\\ &\operatorname{div}S(\gamma'-2\epsilon\beta) = 0,\\ &S\gamma''+2\operatorname{ein}\gamma+S\nabla\nabla\alpha-2S\epsilon\beta' = 0.\end{aligned} \tag{5.7}$$

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:

Proposition 5.6. *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. By Lemma 5.1 and Proposition 5.3, $\operatorname{div} S \nabla\nabla = 0$ and $\operatorname{div} \operatorname{ein} = 0$. Take the divergence of the evolution equation:

$$\operatorname{div} S\gamma'' - 2 \operatorname{div} S\epsilon\beta' = 0.$$

This is the time derivative of the second constraint equation. 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. By the third identity in Lemma 5.2,

$$\operatorname{div} \operatorname{div} S\gamma' = 0.$$

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. □

Before we state the initial-value problem, we interpret the decomposition of the 4-metric:

$$h_{(4)} = \begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix}.$$

Consider another coordinate system $(\hat{t}, \hat{x})$ related to the Euclidean (t, x) on $[0, T] \times \Omega$ via a linear reparameterization of time:

$$t = H\hat{t} + F^T\hat{x}, \qquad x = \hat{x},$$

for some $H \in \mathbb{R}$ and $F \in \mathbb{R}^3$. Then the pullback metric in the hat coordinates is:

$$\begin{bmatrix} H & F^T \\ 0 & I \end{bmatrix} \begin{bmatrix} \alpha & \beta^T \\ \beta & \gamma \end{bmatrix} \begin{bmatrix} H & 0 \\ F & I \end{bmatrix} = \begin{bmatrix} \alpha H^2 + 2H\beta^T F + F^T\gamma F & (H\beta + \gamma F)^T \\ H\beta + \gamma F & \gamma \end{bmatrix}.$$

This shows that α and β component of the 4-metric can be interpreted as a choice for the linear parameterization of the t-coordinates. For the formulation of initial-value problems, we can therefore consider α and β as given data and consider the evolution of γ alone.

We then derive the $(1+3)$ initial-value model problem for linearized relativity from system (5.7). Given a smooth scalar field $\alpha(t)$, a smooth vector field $\beta(t)$, and two smooth symmetric matrix fields γ_0, γ_1 satisfying:

$$\operatorname{div} \operatorname{div} S\gamma_0 = 0, \qquad \operatorname{div} S(\gamma_1 - 2\epsilon\beta(0)) = 0,$$

we find a symmetric matrix field $\gamma(t)$ on Ω, such that $\gamma(0) = \gamma_0$, $\gamma'(0) = \gamma_1$, and for all $t > 0$:

$$S(\gamma'' + \nabla\nabla\alpha - 2\epsilon\beta') + 2\,\mathrm{ein}\,\gamma = 0.$$

Note that $\nabla\nabla\alpha - 2\epsilon\beta' = \epsilon(\nabla\alpha - 2\beta')$ is in the image of ϵ, which is in turn in the kernel of ein. This means that α and β terms do not contribute to the evolution equation. Indeed, define:

$$\hat{\gamma}(t) = \gamma(t) + \int_0^t \int_0^s \nabla\nabla\alpha(v)\,dv\,ds - 2\int_0^t \epsilon\beta(s)\,ds.$$

Then $\gamma(t)$ solves the problem with the given α and β if and only if $\hat{\gamma}(t)$ solves the same system with $\alpha = 0$ and $\beta = 0$. Thus without loss of generality, we can set $\alpha = 0$ and $\beta = 0$.

We are ready to state the model problem for linearized Einstein equation:

Problem 5.1 (Linearized Cauchy problem)**.** *Given two smooth symmetric matrix fields* γ_0, γ_1 *satisfying the* compatibility conditions*:*

$$\mathrm{div}\,\mathrm{div}\,S\gamma_0 = 0, \qquad \mathrm{div}\,S\gamma_1 = 0,$$

find a symmetric matrix field $\gamma(t)$*, such that* $\gamma(0) = \gamma_0$*,* $\gamma'(0) = \gamma_1$*, and for all* $t > 0$*:*

$$S\gamma'' + 2\,\mathrm{ein}\,\gamma = 0.$$

This should be compared with the nonlinear Cauchy problem for the Einstein equation [64, Definition 3].

5.5 Model linear source problem

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. Moreover, it is well-known in the finite element literature that the understanding of the corresponding source problem is the first step in analyzing the discretization of time-dependent problems (for example see [107]). 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 (5.7) is:

$$2\,\mathrm{ein}\,\gamma = 0.$$

The source problem is thus, given a symmetric matrix field f, find a symmetric matrix field u such that

$$2\,\mathrm{ein}\,u = f.$$

It is clear that this problem cannot be well-posed because $\operatorname{div}\operatorname{ein} = 0$ is an obstruction to existence while $\operatorname{ein}\epsilon = 0$ is an obstruction to uniqueness. From these considerations, we formulate the following problem:

Problem 5.2. *Given a symmetric matrix field f in the range of* ein*, find a symmetric matrix field u such that*

$$2\operatorname{ein} u = f, \qquad \operatorname{div} Ju = 0.$$

The reason for the choice of $\operatorname{div} Ju = 0$ for removing the kernel of ein will become clear later.

Problem 5.2 also shows up in the geometric theory for defects and plasticity. This has a long history in mathematics, solid mechanics, and physics. It started with Volterra's paper [109] on plasticity, where curvature is used to model certain types of disclinations. This was subsequently picked up and developed further by engineers and physicists with a substantial literature. For good surveys, see [59,65–67,70,87,120]. Of these, Kröner [70] studied this problem explicitly with a different constraint $\operatorname{div} u = 0$ instead of $\operatorname{div} Ju = 0$. He related this problem to the component-wise biharmonic equation on the whole space and solved it using the fundamental solution. On the more direct application side, the review [58] describes a model of growth in blood vessel walls using this problem, where f is related to the growth of the blood vessels and u is the residual stress caused by the growth. The author is not aware of any treatment of this problem in numerical analysis literature yet.

5.6 Fourier analysis: well-posedness and weak hyperbolicity

In this section, we analyze the Cauchy Problem 5.1 and the source Problem 5.2 on the flat torus $\mathbb{T}^3$ via Fourier analysis. The goals are two. First we prove that both problems are well-posed. Second, we show that the hyperbolic Problem 5.1 is not strongly hyperbolic. This means that although it is well-posedness in the sense of Hadamard, adding lower-order perturbations and variable coefficients can make it ill-posed. Since our goal eventually is to solve the nonlinear Einstein equation, where such lower-order perturbations and variable coefficients are inevitable, we have to regularize Problem 5.1. That will be the subject of the next section.

The flat torus, $\mathbb{T}^3$, is the cube of side length 2π with the periodic boundary conditions. It is very convenient mainly because linear differential calculus is reduced to algebra via Fourier series here. On $\mathbb{T}^3$, a scalar field u can be represented as a formal infinite sum:

$$u(x) = \sum_{k\in\mathbb{Z}^3} u_k e^{ik\cdot x},$$

where $u_k \in \mathbb{C}$ are constants. We need norms to make these sums well-defined. Let H^s be the usual Sobolev spaces. It can be characterized by the Fourier coefficients: for $s \geq 0$,

$$u = \sum_{k \in \mathbb{Z}^3} u_k e^{ik\cdot x} \in H^s \quad \text{if and only if} \quad \sum_{k \in \mathbb{Z}^3} (1+|k|^2)^{s/2} |u_k|^2 < \infty$$

For linear differential operators, it is sufficient to study their behavior for each k individually. Fix $k \in \mathbb{Z}^3$, $k \neq 0$. 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:

$$[a]_k := ae^{ik\cdot x}, \qquad \begin{bmatrix} a \\ b \\ c \end{bmatrix}_k := (am + bn + cl)e^{ik\cdot x},$$

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & j \end{bmatrix}_k := (amm^T + bmn^T + cml^T + dnm^T + enn^T + fnl^T + glm^T + hln^T + jll^T)e^{ik\cdot x}.$$

Because the basis is orthonormal, the trace and transpose work directly in the matrix notation:

$$\operatorname{tr} \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & j \end{bmatrix}_k = [a+e+j]_k, \qquad \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & j \end{bmatrix}_k^T = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & j \end{bmatrix}_k .$$

Matrix calculus is reduced to matrix algebra in this notation:

Proposition 5.7. *The following holds:*

$$\nabla [a]_k = i|k| \begin{bmatrix} a \\ 0 \\ 0 \end{bmatrix}_k, \qquad \operatorname{curl} \begin{bmatrix} a \\ b \\ c \end{bmatrix}_k = i|k| \begin{bmatrix} 0 \\ -c \\ b \end{bmatrix}_k, \qquad \operatorname{div} \begin{bmatrix} a \\ b \\ c \end{bmatrix}_k = i|k| [a]_k,$$

$$\epsilon \begin{bmatrix} a \\ b \\ c \end{bmatrix}_k = i|k| \begin{bmatrix} a & b/2 & c/2 \\ b/2 & 0 & 0 \\ c/2 & 0 & 0 \end{bmatrix}_k, \qquad \operatorname{div} \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & j \end{bmatrix}_k = i|k| \begin{bmatrix} a \\ d \\ g \end{bmatrix}_k,$$

$$2\operatorname{ein} \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}_k = |k|^2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & -f & e \\ 0 & e & -d \end{bmatrix}_k .$$

The proof is just a direct computation. Given this, for example, it is obvious that:

$$\operatorname{curl}\nabla = 0, \qquad \operatorname{div}\operatorname{curl} = 0, \qquad \operatorname{ein}\epsilon = 0, \qquad \operatorname{div}\operatorname{ein} = 0.$$

We first show the source problem is well-posed:

Theorem 5.1. *Problem 5.2 is elliptic on $\mathbb{T}^3$. Given $f \in H^{-1}(\mathbb{T}^3)$ satisfying $\operatorname{div} f = 0$ and $\int f = 0$, there exists a unique $u \in H^1 \otimes \mathbb{S}^3$ satisfying*

$$2\,\mathrm{ein}\, u = f, \qquad \operatorname{div} Ju = 0, \qquad \int u = 0.$$

Further, there exists a constant $C > 0$ independent of f such that

$$\|u\|_{H^1} \le C\|f\|_{H^{-1}}.$$

Proof. By linearity, we examine the system for each k independently. The zero integral condition means that all components of the coefficient for $k = 0$ vanishes. For a fixed $k \neq 0$, let

$$u = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}_k .$$

Because $\operatorname{div} f = 0$, the k component of f can be written as

$$f = \begin{bmatrix} 0 & 0 & 0 \\ 0 & g & h \\ 0 & h & l \end{bmatrix}_k .$$

The equation system is thus

$$|k|^2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & -f & e \\ 0 & e & -d \end{bmatrix}_k = \begin{bmatrix} 0 & 0 & 0 \\ 0 & g & h \\ 0 & h & l \end{bmatrix}_k , \qquad ik \begin{bmatrix} (a-d-f)/2 \\ b \\ c \end{bmatrix}_k = 0.$$

All the claims of the theorem are then clear. □

The linearized Einstein equation is more interesting. Let γ be a symmetric matrix field for a fixed nonzero $k \in \mathbb{Z}^3$ of the form:

$$\gamma = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}_k .$$

The $S\gamma'' + 2\,\mathrm{ein}\,\gamma = 0$ reads:

$$\begin{bmatrix} -d-f & b & c \\ b & -a-f & e \\ c & e & -a-d \end{bmatrix}_k'' + |k|^2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & -f & e \\ 0 & e & -d \end{bmatrix}_k = 0.$$

Equivalently, it is an ODE system:

$$
\begin{gathered}
b'' = 0, \qquad c'' = 0, \qquad (d+f)'' = 0, \\
e'' + |k|^2 e = 0, \\
(a+f)'' + |k|^2 f = 0, \\
(a+d)'' + |k|^2 d = 0.
\end{gathered}
$$

Define $g = \frac{d+f}{2}$ and $h = \frac{d-f}{2}$. Eliminate d and f from the above using g and h, we get:

$$
\begin{gathered}
b'' = 0, \qquad c'' = 0, \qquad g'' = 0, \\
e'' + |k|^2 e = 0, \\
h'' + |k|^2 h = 0, \\
a'' + |k|^2 g = 0.
\end{gathered}
\tag{5.8}
$$

The constraint $\operatorname{div}\operatorname{div} S\gamma(0) = 0$ reads:

$$g(0) = 0,$$

and $\operatorname{div} S\gamma' = 0$ reads:

$$g'(0) = 0, \qquad b'(0) = 0, \qquad c'(0) = 0.$$

The compatible initial data thus reads:

$$
\begin{aligned}
a(0) &= a_0, \quad a'(0) = a_1, \qquad b(0) = b_0, \quad b'(0) = 0, \\
c(0) &= c_0, \quad c'(0) = 0, \qquad e(0) = e_0, \quad e'(0) = e_1, \\
g(0) &= 0, \quad g'(0) = 0, \qquad h(0) = h_0, \quad h'(0) = h_1.
\end{aligned}
$$

We see that $g \equiv 0$, which implies $a'' = 0$. Thus $a(t)$ is linear in time, $b(t)$ and $c(t)$ are constant in time. Then there are two oscillatory components e and h, both of which are sinusoidal with frequency $|k|$. In the physics literature, h is called the + polarization while e is called × polarization of the gravitational wave.

For $k = 0$, the equation simply reads $\gamma'' = 0$. Hence overall, all components of $\gamma(t)$ can grow at most linear in t independent of k. This proves the following theorem:

Theorem 5.2. *The linear hyperbolic Problem 5.1 is well-posed. In particular, for the unique solution $\gamma(t)$, there exist constants C and D independent of γ_0 and γ_1 such that*

$$\|\gamma(t)\|_{H^1} + \|\gamma'(t)\|_{L^2} \le (Ct + D)(\|\gamma_0\|_{H^1} + \|\gamma_1\|_{L^2})$$

for all $t \ge 0$.

For the purpose as a model problem for the nonlinear Einstein equation, Theorem 5.2 is not enough. This relates to the well-known problem of weak hyperbolicity, which we will demonstrate here. For general theory on the well-posedness hyperbolic problems in the presence of lower-order terms and variable coefficients, see [68].

First, the study of hyperbolicity looks at the equation with arbitrary initial data, including incompatible ones. This is realistic because in the discretization, it is usually impossible to impose the compatibility conditions exactly. The ODE system (5.8) decouples. We only need to look at the offending subsystem:

$$g'' = 0, \qquad a'' + |k|^2 g = 0.$$

Introduce two auxiliary variables: the scalar $l := g'$ and the vector $m := ikg$. We rewrite the above into a first-order system:

$$\begin{bmatrix} a \\ m \\ g \\ l \end{bmatrix}' = \begin{bmatrix} 0 & ik^T & 0 & 0 \\ 0 & 0 & ik & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ m \\ g \\ l \end{bmatrix}.$$

For a first-order system like the one above, it is called *weakly hyperbolic* if all the eigenvalues of the matrix are real. It is easy to see that a weakly hyperbolic system is well-posed in the sense of Hadamard in the L^2-norm.

It is *strongly hyperbolic* if the matrix is further diagonalizable. It can be proven that strongly hyperbolic systems are still well-posed with variable coefficients and additional lower-order terms [68]. The system above is only weakly hyperbolic but not strongly hyperbolic because the geometric multiplicity of the only eigenvalue 0, which is the dimension of its kernel, is just 1, instead of 4.

The lack of robustness of the above first order system against lower-order perturbation is easy to demonstrate. Suppose the linearized Einstein equation is perturbed by a zero-th order term such that instead of $g'' = 0$, we have $g'' = 4a$. The system above becomes:

$$\begin{bmatrix} a \\ m \\ g \\ l \end{bmatrix}' = \begin{bmatrix} 0 & ik^T & 0 & 0 \\ 0 & 0 & ik & 0 \\ 0 & 0 & 0 & 1 \\ 4 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ m \\ g \\ l \end{bmatrix}.$$

The eigenvalues of the above matrix matrix are:

$$\lambda = \pm 1 \pm i\sqrt{|k|}.$$

For $k \neq 0$, both a and g can grow at least as fast as $e^{\sqrt{|k|}t}$. This is clearly ill-posed because there does not exist any $C > 0$ and $M > 0$ independent of k, such that the solution in L^2 norm is bounded by Ce^{Mt}. It was shown in [68] that this is in fact the typical behavior of such weakly but not strongly hyperbolic equations.

5.7 Regularized well-posed model problems on bounded smooth domains

Although the Cauchy Problem 5.1 and Source Problem 5.2 are well-posed. Both have fatal flaws making them unsuitable for discretization. The source problem is only solvable for compatible data. It can be difficult to ensure the discrete data is compatible. Thus its discretization can lead to inconsistent linear systems. The Cauchy problem is only weakly hyperbolic, so both numerical error and potential lower-order terms and variable coefficients can make it ill-posed. In this section, we regularize these problems to symmetric matrix-valued Poisson equation and wave equation with special boundary conditions, which are suitable for discretization. Moreover, we show that they can be used to solve the unregularized problems.

The first goal is to prove the well-posedness of the symmetric matrix-valued Poisson problem

$$-\Delta u = f$$

and the symmetric matrix-valued wave equation

$$u'' - \Delta u = f,$$

subject to boundary conditions:

$$u_{\tau\tau} = 0, \qquad u_{nn} = 0, \qquad \partial_n u_{n\tau} = 0,$$

where these boundary components are defined as in equation (5.1).

Note that there are exactly $3 + 1 + 2 = 6$ boundary conditions for the 6 components of u. We will show that the Poisson problem is well-posed for all right-hand sides and the wave equation is strongly hyperbolic.

We start from the weak form of the elliptic problem. Define the following subspace of $H^1 \otimes \mathbb{S}^3$:

$$V := \{u \in H^1 \otimes \mathbb{S}^3 \,|\, \text{on } \partial\Omega\text{: } u_{\tau\tau} = 0,\ u_{nn} = 0\} \tag{5.9}$$

and a symmetric bilinear form $B : V \times V \to \mathbb{R}$:

$$B(u, v) = \int_\Omega \nabla u : \nabla v. \tag{5.10}$$

Lemma 5.3. *On a bounded connected smooth domain Ω in $\mathbb{R}^3$, if u is a constant symmetric matrix-valued function satisfying $u_{\tau\tau} = 0$, then $u = 0$.*

Proof. Since Ω is connected, u can be identified with its value, a symmetric matrix, C. Since Ω is bounded, there is at least one point p in Ω with the largest x-coordinate. Clearly $p \in \partial\Omega$. Because Ω is smooth, the normal vector at p is just $[1,0,0]$ and the yz-coordinate plane is parallel to the tangent space at p. The boundary conditions then imply that

$$C = \begin{bmatrix} a & b & c \\ b & 0 & 0 \\ c & 0 & 0 \end{bmatrix}$$

for some $a, b, c \in \mathbb{R}$. Repeat the same argument with a point having the largest y-coordinate, we get $a = c = 0$. Repeat again the same argument with a point having the largest z-coordinate, we get $b = a = 0$. Hence $u = 0$. □

The dimension and the smoothness requirement can be substantially weakened. In fact, as long as a connected domain has three linearly independent normal vector, then the above holds by a similar argument.

Proposition 5.8. *On a bounded connected smooth domain Ω in $\mathbb{R}^3$, the symmetric bilinear form B in equation (5.10) is symmetric, bounded, and coercive. Moreover, for any $f \in L^2 \otimes \mathbb{S}^3$, there exists a unique $u \in V$ such that*

$$B(u,v) = \langle f, v\rangle, \qquad \forall v \in V,$$

and there exists a constant $C > 0$ only depending on the domain such that

$$\|u\|_{H^1} \le C\|f\|_{L^2}.$$

Proof. From the definition, it is clear that B is symmetric and bounded. Suppose B fails to be coercive. Then there exists a sequence $u_n \in V$ such that $\|u_n\|_{H^1} = 1$ but $B(u_n, u_n) \to 0$ as $n \to \infty$. Since $V \subset H^1 \otimes \mathbb{S}^3$ is compact in $L^2 \otimes \mathbb{S}^3$, we can find a subsequence u_{n_k} such that $u_{n_k} \to u$ in L^2 for some $u \in L^2$ and $B(u_{n_k}, u_{n_k}) \to 0$. The last fact implies that $\nabla u = 0$ in L^2. The boundary conditions are clearly preserved by taking the limit. Hence u is a constant function in V. By Lemma 5.3, $u = 0$. This is a contradiction because u is the L^2-limit of u_{n_k} with $\|u_{n_k}\|_{L^2} = 1$. Hence B is coercive. Then the final claim follows from Lax-Milgram Theorem. □

We note that in this proposition, f can be taken from the dual space V' in general and the norm on f can be weakened. But we do not need that for the purpose here.

We then proceed to study the strong solutions to the elliptic problem. Define the following subspace of V:

$$Y := \{u \in V \,|\, u \in H^2 \otimes \mathbb{S}^3 \text{ and on } \partial\Omega\colon \partial_n u_{n\tau} = 0\}. \tag{5.11}$$

We state the full source problem first:

Problem 5.3 (Regularized Elliptic Problem)**.** *On a bounded connected smooth domain Ω in $\mathbb{R}^3$, let Y be defined as in equation (5.11). Given $f \in L^2 \otimes \mathbb{S}^3$, find $u \in Y$ such that*

$$-\Delta u = f.$$

Theorem 5.3. *Problem 5.3 has a unique solution $u \in Y$. Moreover, there exists a constant C depending only on the domain such that*

$$\|u\|_{H^2} \le C\|f\|_{L^2}.$$

Proof. First, by Proposition 5.8 there exists a unique $u \in V$ satisfying

$$B(u,v) = (f,v), \qquad \forall v \in V.$$

Second, we need to show $u \in Y$. This uses the standard Agmon-Douglis-Nirenberg theory of elliptic systems [3,4]. In particular, the version stated as Theorem 9.31 of [97] is applied. The fact that $-\Delta$ has an elliptic symbol is clear. The only additional thing necessary is to show that the boundary conditions on Y are complementary to $-\Delta u = 0$ as defined in Definition 9.28 of [97]. We notice that all the operators are invariant under rotations. So without loss of generality, we only need to prove this for the upper half plane. Suppose $u(x_1,x_2,z) = e^{i\xi\cdot x}v(z) \in V$, $z \ge 0$, solves $-\Delta u = 0$, where $v \to 0$ as $z \to \infty$. We need to show that $u = 0$. First, the equation $-\Delta u = 0$ implies that:

$$v'' - |\xi|^2 v = 0.$$

For $\xi \neq 0$, the decay condition on v implies that

$$v(z) = Ce^{-|\xi|z}, \qquad z \ge 0$$

for some symmetric matrix C. Then $u_{\tau\tau} = 0$ and $u_{nn} = 0$ at $z = 0$ implies that C can be written in the form:

$$C = \begin{bmatrix} 0 & 0 & c \\ 0 & 0 & e \\ c & e & 0 \end{bmatrix}.$$

Then $\partial_n u_{n\tau} = 0$ at $z = 0$ implies:

$$-|\xi| c = 0, \qquad -|\xi| e = 0.$$

Hence $c = e = 0$. Thus $C = 0$ which implies $v = 0$ and in turn $u = 0$. On bounded domains, there is no nontrivial constant function in V by assumption. This shows that the boundary conditions are complementary. □

The boundary conditions on Y implies more useful identities:

Lemma 5.4. *Suppose $u \in Y$, then on the boundary, we further have*

$$\operatorname{tr} u = 0, \qquad (\operatorname{div} u) \times n = 0, \qquad \operatorname{div}\operatorname{div} u = \partial_n \partial_n u_{nn} = \Delta u_{nn}.$$

Proof. For $u \in Y$, we have $u_{\tau\tau} = 0$, $u_{nn} = 0$, and $\partial_n u_{n\tau} = 0$ on the boundary. First $\operatorname{tr} u = \operatorname{tr} u_{\tau\tau} + u_{nn} = 0$. Second, we use div_τ to denote the divergence on $\partial\Omega$. Under this, in the coordinate system straightening out the boundary,

$$\operatorname{div} u = \operatorname{div} \begin{bmatrix} u_{\tau\tau} & u_{n\tau} \\ u_{n\tau}^T & u_{nn} \end{bmatrix} = \begin{bmatrix} \partial_n u_{n\tau} \\ \operatorname{div}_\tau u_{n\tau} + \partial_n u_{nn} \end{bmatrix}.$$

In particular, the tangential part of $\operatorname{div} u$ is $\partial_n u_{n\tau}$ which vanishes. Equivalently, this means $(\operatorname{div} u) \times n = 0$. Third, taking the divergence of the above, we further get,

$$\operatorname{div}\operatorname{div} u = \partial_n(\operatorname{div}_\tau u_{n\tau} + \partial_n u_{nn}) = \operatorname{div}_\tau \partial_n u_{n\tau} + \partial_n \partial_n u_{nn} = \partial_n \partial_n u_{nn},$$

where the first term vanishes because $\partial_n u_{n\tau} = 0$ is constant on the boundary. Let Δ_τ be the Laplacian on the boundary. Then,

$$-\Delta u_{nn} = -\Delta_\tau u_{nn} - \partial_n \partial_n u_{nn} = -\partial_n \partial_n u_{nn},$$

where the first term vanishes because $u_{nn} = 0$ is constant on the boundary. This proves the claim. □

We then state in what sense Regularized Source Problem 5.3 solves Source Problem 5.2:

Theorem 5.4. *Suppose f is a smooth symmetric matrix-valued function in the image of* ein *satisfying $f_{nn} = 0$ on the boundary. Let ϕ be the solution to Regularized Source Problem 5.3 with the right-hand side f. Then $u := S\phi \in Y$ solves the corresponding Source Problem 5.2 in the sense that* $\operatorname{div} Ju = 0$ *and* $2 \operatorname{ein} u = f$.

Proof. By Theorem 5.3, we can find a unique $\phi \in Y$ solving $-\Delta\phi = f$. In particular, $\operatorname{div}\phi \in H^1 \otimes \mathbb{R}^3$ satisfies a vector Laplace equation:

$$-\Delta \operatorname{div}\phi = \operatorname{div} f = 0.$$

On the boundary, by Lemma 5.4, we have

$$(\operatorname{div}\phi) \times n = 0, \qquad \operatorname{div}(\operatorname{div}\phi) = \Delta\phi_{nn} = -f_{nn} = 0.$$

This is a well-known set of complementary boundary conditions for the vector Laplacian (see, for example equation (64) of [10]). Hence $\operatorname{div}\phi = 0$ on the whole domain. Finally, recall the identity (5.5)

$$2\,\mathrm{ein} = -J\Delta + 2J\epsilon \operatorname{div} J$$

and the fact that in 3D, $J = S^{-1}$. First, combining the two, we get

$$2S\,\mathrm{ein} - 2\epsilon \operatorname{div} J = -\Delta. \tag{5.12}$$

Let $u := S\phi$. On the boundary, ϕ is trace-free. Further, S does not change the tangential-normal and the normal-tangential components. Hence $u \in Y$ as well. Clearly $\operatorname{div} Ju = \operatorname{div}\phi = 0$. Now,

$$Sf = -S\Delta\phi = -\Delta S\phi = -\Delta u = 2S\,\mathrm{ein}\,u - 2\epsilon \operatorname{div} Ju = 2S\,\mathrm{ein}\,u.$$

Applying J to both sides, we get $2\,\mathrm{ein}\,u = f$ as claimed. □

We then look at the corresponding wave equation. In what follows, the time domain is always assumed to be the interval $[0,T]$ for some $T > 0$. For linear hyperbolic equations, solutions are always global in time so T is not interesting. To simplify the notation, we write time-dependent function spaces like $C^0([0,T],V)$ simply as C^0V.

Problem 5.4 (Regularized Hyperbolic Problem)**.** *On a bounded connected smooth domain, let V and Y be defined as in equation (5.9) and (5.11) respectively. Given $u_0, u_1 \in Y$ and $f \in C^0(L^2 \otimes \mathbb{S}^3)$, find $u \in C^0Y \cap C^1V \cap C^2(L^2 \otimes \mathbb{S}^3)$ such that $u(0) = u_0$, $u'(0) = u_1$, and for all $t > 0$,*

$$u'' - \Delta u = f.$$

Theorem 5.5. *Problem 5.4 is well-posed. It has a unique solution u satisfying the energy estimate: for some absolute constant $C > 0$,*

$$|u(t)|_{H^1} + \|u'(t)\|_{L^2} \le C(|u_0|_{H^1} + \|u_1\|_{L^2} + \int_0^t \|f\|_{L^2}).$$

Proof. For this, we use the standard semi-group theory [97, Chapter 12]. We write the wave equation as a first-order in time system by introducing $v = u'$. Define an operator:

$$A = \begin{bmatrix} 0 & I \\ \Delta & 0 \end{bmatrix} : Y \times V \subset (V \times (L^2 \otimes \mathbb{S}^3)) \to (V \times (L^2 \otimes \mathbb{S}^3)).$$

By Proposition 5.8, we can use the H^1-seminorm as the norm on V. If A generates a continuous semi-group, then the following abstract ODE has a unique solution:

$$\begin{bmatrix} u \\ v \end{bmatrix}' = \begin{bmatrix} 0 & I \\ \Delta & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ f \end{bmatrix}.$$

This u then solves the wave equation as required. We prove A indeed generates a contraction semi-group using Lumer-Phillips Theorem [97, Theorem 12.22]. In particular, we need to check that A is densely defined, $(x, Ax) \leq 0$ for x in the domain of A, and there exists $\lambda > 0$ such that $A - \lambda I$ is onto.

First, clearly, A is densely defined because test functions are dense in L^2. Second, we note for $u \in Y$ and $v \in V$,

$$\int_\Omega -\Delta u : v = \int_\Omega \nabla u : \nabla v - \int_{\partial\Omega} v : \partial_n u = \int_\Omega \nabla u : \nabla v.$$

The boundary term vanishes because $v_{\tau\tau} = 0$, $v_{nn} = 0$ while by symmetry $\partial_n u_{n\tau} = \partial_n u^T_{\tau n} = 0$. Thus, A is dissipative (as defined in Definition 12.25 of [97]):

$$\left(\begin{bmatrix} u \\ v \end{bmatrix}, A \begin{bmatrix} u \\ v \end{bmatrix} \right)_{V \times L^2} = (\nabla u, \nabla v) + (\Delta u, v) = 0.$$

Finally, given $(g, h) \in V \times L^2 \otimes \mathbb{R}^3$, we need one $\lambda > 0$ such that we can solve

$$(A - \lambda I) \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} g \\ h \end{bmatrix}.$$

Equivalently,

$$(-\Delta + \lambda^2) u = -h - \lambda g.$$

$-\Delta$ alone is already coercive over V. So a solution $u \in V$ exists. Now apply Theorem 5.3 to $-\Delta u = -\lambda^2 u - h - \lambda g \in L^2 \otimes \mathbb{S}^3$. We get $u \in Y$ as needed. Thus by Lumer-Phillips theorem, A generates a contraction semi-group, which gives us a unique solution u to the wave equation.

Finally, multiplying u' on both sides of the equation and integration by parts:

$$\frac{1}{2}[(u', u')' + (\nabla u, \nabla u)] = (f, u).$$

Integrate this in time gives the energy estimate. □

At last, we show how to use this wave equation to solve the Cauchy Problem 5.1. This requires several lemmas. The first lemma concerns the boundary condition:

Lemma 5.5. *For a scalar field w and a vector field u, on the boundary*

$$(\nabla\nabla w)_{\tau\tau} = (\nabla\nabla)_{\partial\Omega} w, \qquad (\epsilon u)_{\tau\tau} = \epsilon_{\partial\Omega} u_\tau,$$

where u_τ is the tangential part of u, $(\nabla\nabla)_{\partial\Omega}$ is the tangential Hessian on the boundary, and $\epsilon_{\partial\Omega}$ is the tangential symmetric gradient on the boundary. In particular, if $w = 0$ on the boundary, then $(\nabla\nabla w)_{\tau\tau} = 0$. Similarly, if $u \times n = 0$, then $(\epsilon u)_{\tau\tau} = 0$.

Proof. Due to rotational invariance of ϵ, $\nabla\nabla$, and the tangential trace, it is enough to show this for the upper half-plane, where this is obvious. □

The second lemma concerns the structure of the compatibility condition for the Cauchy problem (5.1).

Lemma 5.6. *Let Ω be a bounded smooth contractible domain in $\mathbb{R}^3$. Suppose $u \in Y$ satisfies* $\operatorname{div}\operatorname{div} Su = 0$. *Then there exists a vector field $\xi \in H^3 \otimes \mathbb{R}^3$ such that $\bar{u} := u + \epsilon\xi \in Y$ satisfies:*

$$\operatorname{tr}\bar{u} = 0, \qquad \operatorname{div}\bar{u} = 0, \qquad \textit{in } \Omega.$$

Proof. For this, we need the standard well-posedness and elliptic regularity of scalar and vector Poisson problems. First, define a vector field $v \in H^1 \otimes \mathbb{R}^3$ as the unique solution to the vector Poisson problem:

$$-\Delta v = \operatorname{div} Su,$$

with the boundary condition $v \times n = 0$ and $\operatorname{div} v = 0$ (this is the well-posed 1-form Hodge Laplacian problem in 3D [10]). Using elliptic regularity, $v \in H^3$ because $\operatorname{div} Su \in H^1$. Further $\operatorname{div} v \in H^2$ satisfies the homogeneous scalar Laplace equation with homogeneous Dirichlet boundary conditions because $\operatorname{div}\operatorname{div} Su = 0$. Hence in Ω,

$$\operatorname{div} v = 0.$$

We know on vector fields:

$$-\Delta = -\nabla \operatorname{div} + \operatorname{curl}\operatorname{curl}.$$

Hence, $-\Delta v = \operatorname{div} Su$ implies

$$\operatorname{curl}\operatorname{curl} v = \operatorname{div} Su.$$

Second, define $\phi \in H^1$ as the unique solution to the scalar Poisson problem:

$$-\Delta\phi = \frac{1}{2}\operatorname{tr} u,$$

with the boundary condition $\phi = 0$. This is well-known to be well-posed. By elliptic regularity, $\phi \in H^4$ because $\operatorname{tr} u \in H^2$.

Now we define $\xi := v + \nabla\phi$ and $\bar{u} := u + 2\epsilon\xi$. First, it is clear that $\bar{u} \in H^2$.

Second, in Ω, using the identities $\operatorname{tr}\epsilon = \operatorname{div}$ and $\operatorname{tr}\nabla\nabla = \Delta$, by the definition of v and ϕ, we have,

$$\operatorname{tr}\bar{u} = \operatorname{tr}(u + 2\epsilon v + 2\nabla\nabla\phi) = \operatorname{tr} u + 2\operatorname{div} v + 2\Delta\phi = \operatorname{tr} u + 0 - \operatorname{tr} u = 0.$$

Moreover, recall from Lemma 5.2, $2\operatorname{div} S\epsilon = -\operatorname{curl}\operatorname{curl}$ and from Lemma 5.1, $\operatorname{div} S\nabla\nabla = 0$. Hence,

$$\operatorname{div} S\bar{u} = \operatorname{div} S(u + 2\epsilon v + 2\nabla\nabla\phi) = \operatorname{div} Su - \operatorname{curl}\operatorname{curl}(v + \nabla\phi) = \operatorname{div} Su - \operatorname{curl}\operatorname{curl} v = 0.$$

This further implies that $\operatorname{div}\bar{u} = 0$. Third, by Lemma 5.5, $\bar{u}_{\tau\tau} = 0$. But because $\operatorname{tr}\bar{u} = 0$, $\bar{u}_{nn} = 0$ on the boundary as well. Finally, $\operatorname{div}\bar{u} = 0$ and $\bar{u}_{\tau\tau} = 0$ implies that $\partial_n \bar{u}_{n\tau} = 0$ on the boundary because it is exactly the tangential part of $\operatorname{div} u$ there. Hence $\bar{u} \in Y$. □

Finally, we can state how to use the Regularized Hyperbolic Problem 5.4 to solve the Cauchy Problem 5.1.

Theorem 5.6. *Given $\gamma_0, \gamma_1 \in Y$ satisfying the compatibility conditions*

$$\operatorname{div}\operatorname{div} S\gamma_0 = 0, \qquad \operatorname{div} S\gamma_1 = 0.$$

Let $\xi_0, \xi_1 \in H^3 \otimes \mathbb{R}^3$ be the vector fields for γ_0 and γ_1 respectively, as defined in Lemma 5.6, such that both $\bar{\gamma}_0 := \gamma_0 + \epsilon\xi_0$ and $\bar{\gamma}_1 := \gamma_1 + \epsilon\xi_1$ are divergence-free, trace-free, and in Y. Let $\bar{\gamma}(t)$ be the solution to Problem 5.4 with zero right-hand side and initial data $\bar{\gamma}(0) = \bar{\gamma}_0$, $\bar{\gamma}'(0) = \bar{\gamma}_1$. Then

$$\gamma(t) := \bar{\gamma}(t) - \epsilon\xi_0 - t\epsilon\xi_1$$

solves Problem 5.1.

Proof. First, it is clear that $\gamma(0) = \gamma_0$ and $\gamma'(0) = \gamma_1$. It remains to check that $S\gamma'' + 2\operatorname{ein}\gamma = 0$.

Second, we prove that the boundary conditions on Y propagates the conditions that $\operatorname{tr}\bar{\gamma} = 0$ and $\operatorname{div}\bar{\gamma} = 0$. First, it is clear that $\operatorname{tr}\bar{\gamma}$ satisfies a homogeneous scalar wave equation and by Lemma 5.4 $\operatorname{tr}\bar{\gamma} = 0$ on the boundary. Hence $\operatorname{tr}\bar{\gamma} = 0$ for all time. Moreover, $\operatorname{div}\bar{\gamma}$ satisfies a vector wave equation. By Lemma 5.4 again, we have on the boundary,

$$(\operatorname{div}\bar{\gamma}) \times n = 0, \qquad \operatorname{div}(\operatorname{div}\bar{\gamma}) = \Delta\bar{\gamma}_{nn} = \bar{\gamma}''_{nn} = 0.$$

These ensure that $\operatorname{div}\bar{\gamma} = 0$ for all time. Finally, we again use the identity (5.12). The fact that $\bar{\gamma}'' - \Delta\bar{\gamma} = 0$ implies

$$\bar{\gamma}'' + 2S\operatorname{ein}\bar{\gamma} - 2\epsilon\operatorname{div} J\bar{\gamma} = 0.$$

Because $\operatorname{div}\bar{\gamma} = 0$ and $\operatorname{tr}\bar{\gamma} = 0$, $\operatorname{div} J\bar{\gamma} = 0$. Further, $\operatorname{tr}\bar{\gamma} = 0$ also implies that $S\bar{\gamma} = S^{-1}\bar{\gamma}$. Hence the wave equation for $\bar{\gamma}$ implies:

$$S\bar{\gamma}'' + 2\operatorname{ein}\bar{\gamma} = 0.$$

Finally, the difference between γ and $\bar{\gamma}$ is linear in time and a symmetric gradient. Since $\operatorname{ein}\epsilon = 0$, γ satisfies the equation above as well. This proves the claim. □

We have shown that with proper boundary conditions and transformations, the well-posed Poisson equation and the wave equation on symmetric matrix-valued functions can be used to solve the elliptic and hyperbolic model problems from relativity. These clarify in a Hilbert space context exactly what it means to solve the model problems. Second, the two regularized continuous problems can then be used as a starting point for discretization. This approach, for example, is very different from and more likely to be successful than the Regge Calculus approach, which tries to discretize a weakly hyperbolic continuous system.

5.8 Regge elements discretization

In this final section, we hint at how Regge elements can be used to solve the regularized elliptic Problem 5.3 and regularized hyperbolic Problem 5.4. This is largely inspired by the success of Finite Element Exterior Calculus for the Hodge Laplacian problems [8, 10]. There we try to solve the Poisson problems on anti-symmetric tensor-valued functions with their appropriate boundary conditions. The strategy was to break the Laplacian into different terms using the boundary conditions as a hint and construct a mixed formulation. The detailed study of the methods mentioned here will be future work.

In the previous chapters, we showed that REG^r paired with CG^{r+1} can be used to discretize the bilinear form $\langle \operatorname{div}\operatorname{div} Su, \eta\rangle$ and REG^r paired with NED^r can be used to discretize the bilinear form $\langle \operatorname{div} Su, p\rangle$. In Christiansen's work [29], we saw that at least REG^0 paired with itself can be used to discretize the bilinear form $\langle \operatorname{ein} u, v\rangle$. Here we propose a way to connect all these together to use REG^r to solve Problem 5.3 and Problem 5.4.

The key is the following identity:

Proposition 5.9.

$$-S\Delta = 2\operatorname{ein} - 2S\epsilon \operatorname{div} S - S\nabla\nabla\operatorname{tr} - I\operatorname{div}\operatorname{div} S. \tag{5.13}$$

Proof. We derive it from identity (5.12):

$$-\Delta = 2S\operatorname{ein} - 2\epsilon\operatorname{div} J.$$

Using $\operatorname{tr} S = -2\operatorname{tr}$, the above implies

$$2\operatorname{tr}\operatorname{ein} = -\operatorname{div}\operatorname{div} S.$$

Use this to expand $2S\operatorname{ein}$. Note $2\epsilon\operatorname{div} J = 2\epsilon\operatorname{div} S + \nabla\nabla\operatorname{tr}$. Identity (5.12) becomes:

$$-\Delta = 2\operatorname{ein} - I\operatorname{div}\operatorname{div} S - 2\epsilon\operatorname{div} S - \nabla\nabla\operatorname{tr}.$$

Apply S to both sides. Use $SI = -2I$ and the trace ein formula again, we get the claim. □

The crucial unintuitive idea is to solve problems associated with $-S\Delta$ instead of the Laplacian itself. The operator $-S\Delta$ is not elliptic. Nevertheless, the associated source and Cauchy problems are still well-posed because it differs from $-\Delta$ by just a point-wise linear algebraic operation.

For the regularized source problem 5.3, $-\Delta u = f$, we use a mixed formulation for $-S\Delta u = Sf$. Let

$$p = \operatorname{div} Su, \qquad \eta = \operatorname{tr} u, \qquad \theta = \operatorname{div}\operatorname{div} Su.$$

We have a system:

$$\begin{gathered}\eta - \operatorname{tr} u = 0,\\ p - \operatorname{div} Su = 0,\\ \theta - \operatorname{div}\operatorname{div} Su = 0,\\ 2\operatorname{ein} u - 2S\epsilon p - S\nabla\nabla\eta - I\theta = Sf.\end{gathered}$$

The boundary conditions for Y defined in equation 5.11 becomes:

$$\eta = 0, \qquad p_\tau = 0, \qquad \theta = 0, \qquad u_{\tau\tau} = 0,$$

by Lemma 5.4. These exactly matches the boundary conditions where the discretizations of $\operatorname{div}\operatorname{div} S$, $\operatorname{div} S$, and ein using REG^r are valid. For example, we can write down an implementable mixed method via the Lagrange multipliers. On a triangulation of Ω, define

$$V_h = (\mathrm{CG}^{r+1} \cap \mathring{H}^1) \times (\mathrm{NED}^r \cap \mathring{H}(\mathrm{curl})) \times (\mathrm{CG}^{r+1} \cap \mathring{H}^1) \times \{u \in \mathrm{REG}^r,\ u_{\tau\tau} = 0\}.$$

Given $f \in L^2 \otimes \mathbb{S}^3$, we find $(\eta, p, \theta, u) \in V_h$ such that for all $(\lambda, q, \xi, v) \in V_h$,

$$\begin{gathered}(\eta, \lambda) - (\operatorname{tr} u, \lambda) = 0,\\ (p, q) - \langle \operatorname{div} Su, q\rangle = 0,\\ (\theta, \xi) - \langle \operatorname{div}\operatorname{div} Su, \xi\rangle = 0,\\ 2(\operatorname{ein} u, v) + 2\langle p, \operatorname{div} Sv\rangle - \langle \eta, \operatorname{div}\operatorname{div} Sv\rangle - (\theta, \operatorname{tr} v) = (Sf, v).\end{gathered}$$

All the integration by parts are allowed by the boundary conditions. It remains to be seen if this is indeed a stable mixed method.

For the regularized hyperbolic Problem 5.4, we again solve the trace-shifted version:

$$Su'' - S\Delta u = 0.$$

Using identity (5.13), the above is equivalent to:

$$Su'' + 2\operatorname{ein} u - 2S\epsilon \operatorname{div} Su - S\nabla\nabla \operatorname{tr} u - I \operatorname{div}\operatorname{div} Su = 0.$$

This is basically the evolution equation in the linearized Einstein equation system (5.7) with some additional terms. Two of the terms are related to the constraints. The $-S\nabla\nabla\operatorname{tr} = 2\operatorname{ein} I$ term brings control of the trace of the metric to the evolution equation. Recall that for $S\gamma'' + 2\operatorname{ein}\gamma = 0$, we have instead, $-2\operatorname{tr}\gamma'' + 2\operatorname{tr}\operatorname{ein}\gamma = 0$ implies $\operatorname{tr}\gamma'' = 0$ by the constraint $0 = \operatorname{div}\operatorname{div} S\gamma = -2\operatorname{tr}\operatorname{ein}\gamma$. This was exactly the cause of weak hyperbolicity for constraint-violating solutions.

For discretization using REG^r, we again introduce auxiliary variables to get a first-order in time system. Rewrite the wave equation as:

$$\begin{aligned}
\eta' &= \operatorname{tr} u,\\
p' &= \operatorname{div} Su,\\
\theta' &= \operatorname{div}\operatorname{div} Su,\\
w' &= 2\operatorname{ein} u,\\
Su' + w - 2S\varepsilon p - S\nabla\nabla\eta - I\theta &= 0.
\end{aligned}$$

We again have an implementable mixed discretization. On a triangulation of Ω, define

$$Q_h = (\mathrm{CG}^{r+1} \cap \mathring{H}^1) \times (\mathrm{NED}^r \cap \mathring{H}(\mathrm{curl})) \times (\mathrm{CG}^{r+1} \cap \mathring{H}^1) \times \{u \in \mathrm{REG}^r,\ u_{\tau\tau} = 0\}^2.$$

Given initial data, we find $(\eta(t), p(t), \theta(t), w(t), u(t)) \in V_h$ such that for all $(\lambda, q, \xi, y, v) \in V_h$ at each time t,

$$\begin{aligned}
(\eta', \lambda) - (\operatorname{tr} u, \lambda) &= 0,\\
(p', q) - \langle \operatorname{div} Su, q\rangle &= 0,\\
(\theta', \xi) - \langle \operatorname{div}\operatorname{div} Su, \xi\rangle &= 0,\\
(w', y) - 2\langle \operatorname{ein} u, y\rangle &= 0,\\
(u', Sv) + (w, v) + 2\langle p, \operatorname{div} Sv\rangle - \langle \eta, \operatorname{div}\operatorname{div} Sv\rangle - (\theta, \operatorname{tr} v) &= 0.
\end{aligned}$$

It remains to be seen if this is stable.

Chapter 6

Two failure modes of Regge Calculus

In 1961, Regge [96] proposed Regge Calculus as a space-time geometric discretization of the Einstein field equation. Only a decade later, Sorkin [103], 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 [40, 117]. 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 [92, 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 [25, 29], 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.

In particular, the methods proposed at the end of Chapter 5 are examples of such methods.

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:

$$\begin{aligned}
\operatorname{div} E &= \rho,\\
\operatorname{div} B &= 0,\\
\operatorname{curl} E + B' &= 0,\\
\operatorname{curl} B - E' &= j,
\end{aligned}$$

where E is the electric field, B is the magnetic field, ρ is the charge density, and j is the current. The right-hand side is required to satisfy the conservation law:

$$\rho' + \operatorname{div} j = 0.$$

We look at a simple case where $\rho = 0$. This implies $\operatorname{div} j = 0$. Taking the time derivative of the fourth equation and eliminating B using the third equation, we get a constrained evolution system:

$$\begin{aligned}
\operatorname{div} E &= 0,\\
E'' + \operatorname{curl}\operatorname{curl} E &= -j'.
\end{aligned}$$

This is the vector analog of the linearized Einstein equation from the previous chapter:

$$\operatorname{div}\gamma = \operatorname{div}\gamma' = 0, \qquad \operatorname{tr}\gamma = \operatorname{tr}\gamma' = 0,$$
$$S\gamma'' + 2\operatorname{ein}\gamma = 0.$$

The key point of this analogy is that both $\operatorname{curl}\operatorname{curl}$ 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 2π 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*:

$$\begin{aligned} \int_{\mathbb{T}^3} a = \int_{\mathbb{T}^3} b = 0, &\qquad \int_{\mathbb{T}^3} f(t) = 0,\ \forall t \in [0,T], \\ \operatorname{div} a = \operatorname{div} b = 0, &\qquad \operatorname{div} f = 0,\ \forall t \in [0,T], \end{aligned}$$

find $u : [0,T] \times \mathbb{T}^3 \to \mathbb{R}^3$ such that $u(0) = a$, $u'(0) = b$, and

$$u'' + \operatorname{curl}\operatorname{curl} u = f. \tag{6.1}$$

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

$$\operatorname{div} u'' = 0.$$

By assumption $\operatorname{div} u(0) = \operatorname{div} u'(0) = 0$. Hence any solution u also satisfies $\operatorname{div} u \equiv 0$. 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:

$$\operatorname{curl}\operatorname{curl} - \nabla \operatorname{div} = -\Delta, \tag{6.2}$$

where Δ is the component-wise Laplace operator. Thus the solution u in the above theorem will also satisfy the component-wise wave equation:

$$u'' - \Delta u = f. \tag{6.3}$$

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:

$$C^\infty \otimes \mathbb{R}^3 = \operatorname{curl}(C^\infty \otimes \mathbb{R}^3) \oplus \nabla(C^\infty) \oplus \mathbb{R}^3,$$

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):

$$w'' = 0, \qquad w(0) = w'(0) = 0,$$

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 Nédéléc edge elements [86], which is known to be a good spatial discretization for problems involving the curl operator [10, 53, 83]. Let NED^1 be the finite element space of Nédéléc edge elements of degree 1 on a uniform mesh of $\mathbb{T}^3$. We solve the problem: given $a_h, b_h \in \mathrm{NED}^1$ and $f : [0,T] \times \mathbb{T}^3 \to \mathbb{R}^3$, find $u_h : [0,T] \to \mathrm{NED}^1$ satisfying $u_h(0) = a_h$, $u_h'(0) = b_h$, and

$$(u_h'', w) + (\operatorname{curl} u_h, \operatorname{curl} w) = (f, w), \qquad \forall w \in \mathrm{NED}^1, \tag{6.4}$$

where $(\cdot,\cdot)$ denotes the L^2-inner product.

Then the temporal part is discretized using the Crank-Nicolson scheme [32], 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: $u_h(0) = a_h$, $v_h(0) = b_h$,

$$\begin{aligned} &(u_h', y) - (v_h, y) = 0, && \forall y \in \mathrm{NED}^1, \\ &(v_h', w) + (\operatorname{curl} u_h, \operatorname{curl} w) = (f, w), && \forall w \in \mathrm{NED}^1. \end{aligned}$$

Finally, the Crank-Nicolson scheme is applied to this system: for time step size k,

$$\begin{aligned} &\left(\frac{u_h^{n+1} - u_h^n}{k}, y\right) - \frac{(v_h^{n+1}, y) + (v_h^n, y)}{2} = 0, && \forall y \in \mathrm{NED}^1, \\ &\left(\frac{v_h^{n+1} - v_h^n}{k}, w\right) + \frac{(\operatorname{curl} u_h^{n+1}, \operatorname{curl} w) + (\operatorname{curl} u_h^n, w)}{2} = \frac{(f(n+1)k, w) + (f(nk), w)}{2}, && \forall w \in \mathrm{NED}^1, \end{aligned}$$

with initial data $u_h^0 = a_h$ and $v_h^0 = b_h$.

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,

$$(u, v) = V^T M U, \qquad (\operatorname{curl} u, \operatorname{curl} v) = V^T A U.$$

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}:

$$(4+k^2A)U^{n+1}=4(U^n+kV^n)+k^2(F^{n+1}+F^n-AU^n).$$

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 $(4+k^2A)$ 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 $8\times8\times8$ 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 $(0,0,0)$.

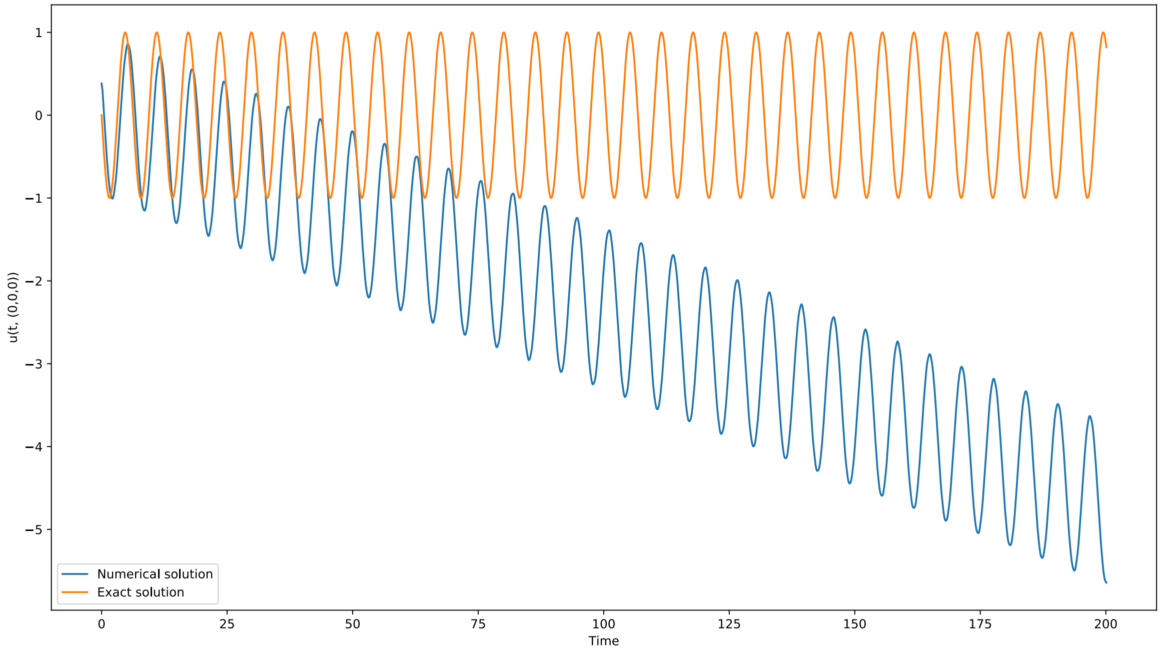


Figure 6.1: Plot of numerical and exact solution

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 [92]). The reason for this is easy to explain. It is known [8, 10] that our discrete space Nédéléc also has a discrete Hodge decomposition:

$$\mathrm{NED}^1 = \nabla\,\mathrm{CG}^1 \oplus \mathbb{R}^3 \oplus V_h,$$

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 $\nabla\,\mathrm{CG}^1$ 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 $\nabla\,\mathrm{CG}^1$ and $\mathbb{R}^3$ component of the numerical solution in time. In particular, the $\nabla\,\mathrm{CG}^1$ 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.

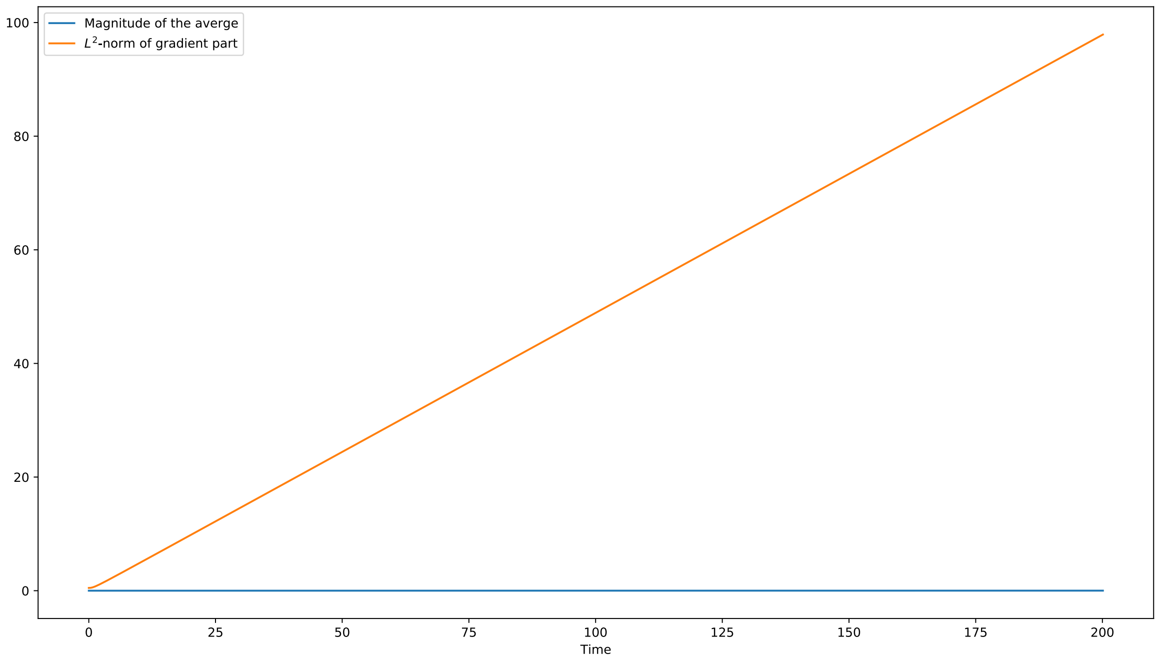


Figure 6.2: Plot of the growth of the gradient and harmonic part of the numerical solution

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 $u \in \mathrm{NED}^1$, find $\phi \in \mathrm{CG}^1$ satisfying:

$$(\nabla\phi, \nabla\psi) = (u, \nabla\psi), \qquad \forall \psi \in \mathrm{CG}^1.$$

Then set:

$$Qu := u - \nabla\phi - \int_{\mathbb{T}^3} u.$$

It is clear that $Q : \mathrm{NED}^1 \to V_h$. After we interpolate the initial data into 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 $\nabla(\mathrm{CG}^1)$ 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.

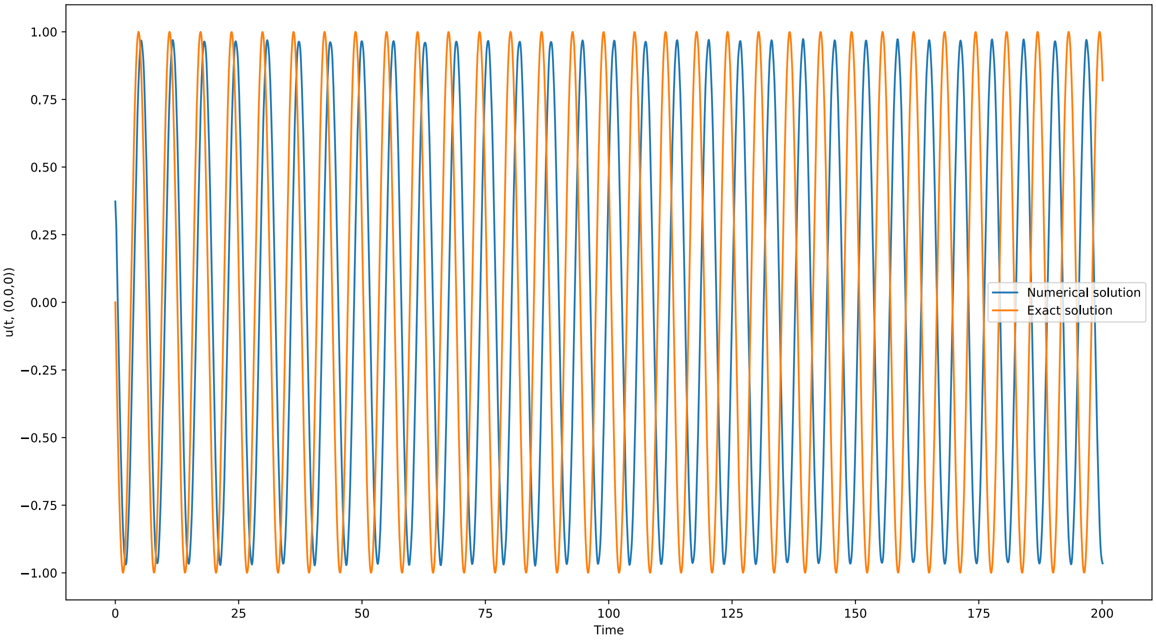


Figure 6.3: Plot of numerical and exact solution with projected initial data

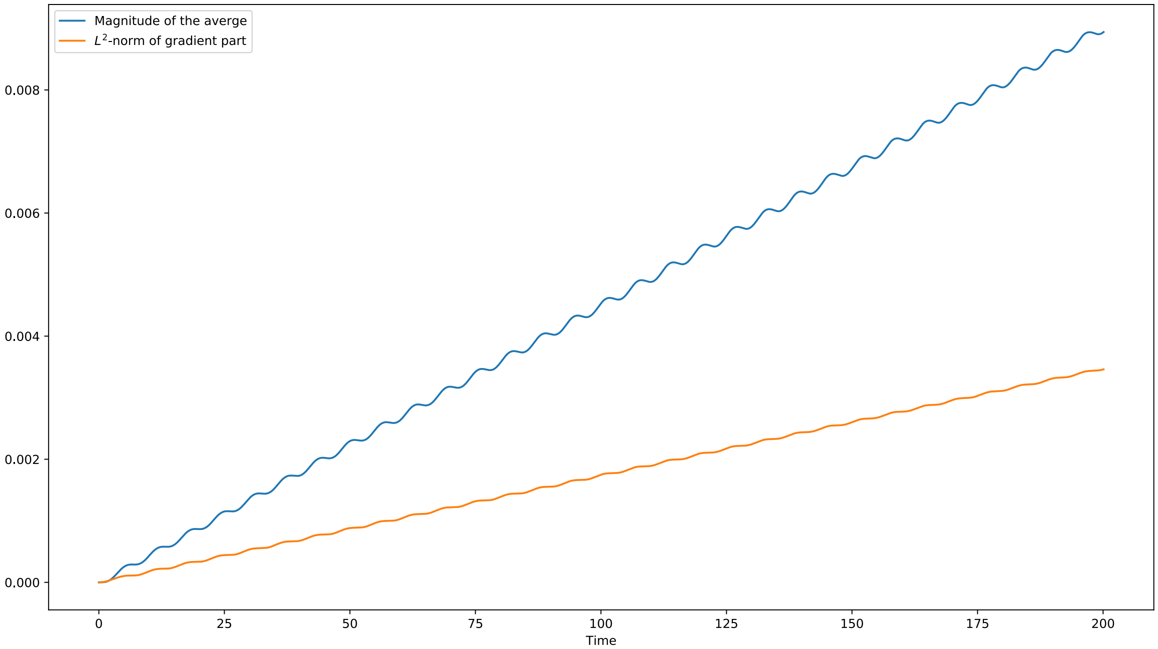


Figure 6.4: Plot of the growth of the gradient and harmonic part of the numerical solution with projected initial data

This problem becomes much more severe for nonlinear problems. Next, we look at a nonlinear perturbed problem:

$$u'' + \mathrm{curl}(1 + \epsilon u \times)\mathrm{curl} = f,$$

where ϵ 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 $\nabla(\mathrm{CG}^1)$ 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 [92] showing the blow-up of the Sorkin-style space-time Regge calculus.

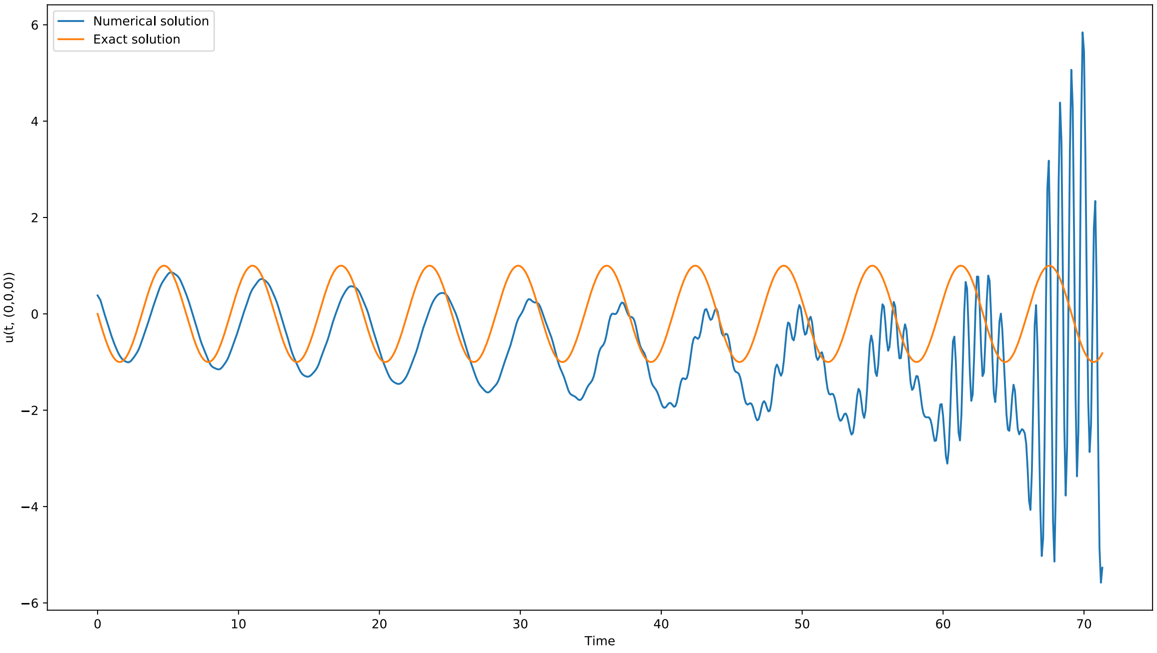


Figure 6.5: Plot of numerical and exact solution for the nonlinear problem

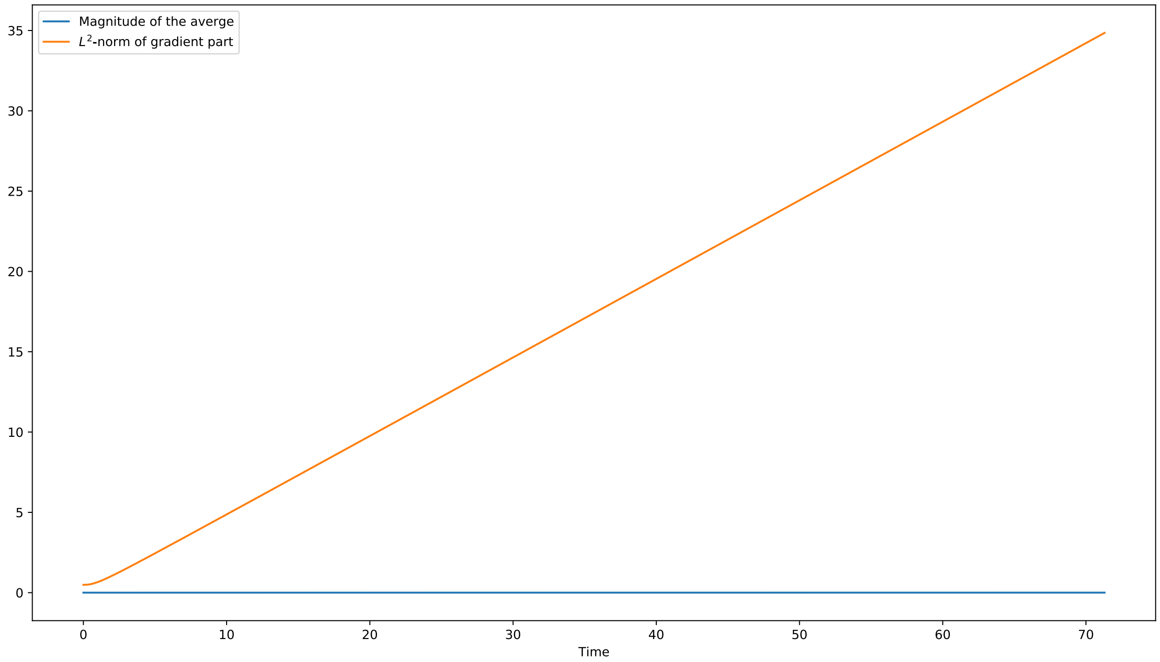


Figure 6.6: Plot of the growth of the gradient and harmonic part of the numerical solution to the nonlinear problem

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 $\nabla(\mathrm{CG}^1)$ 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.

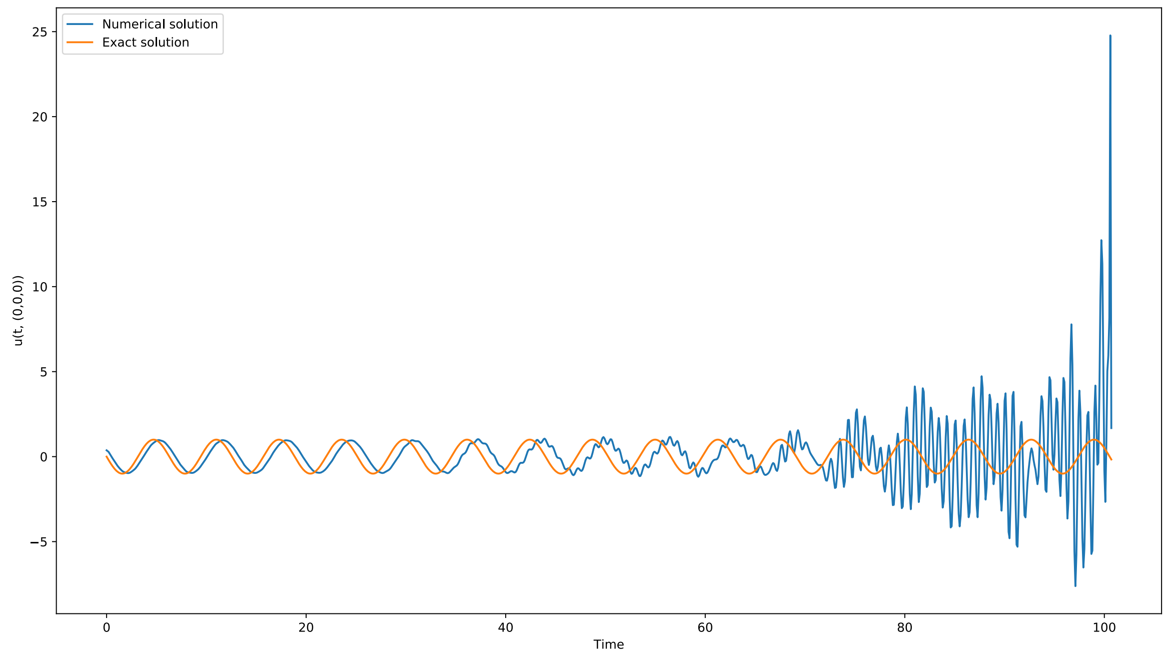


Figure 6.7: Plot of numerical and exact solution with projected initial data

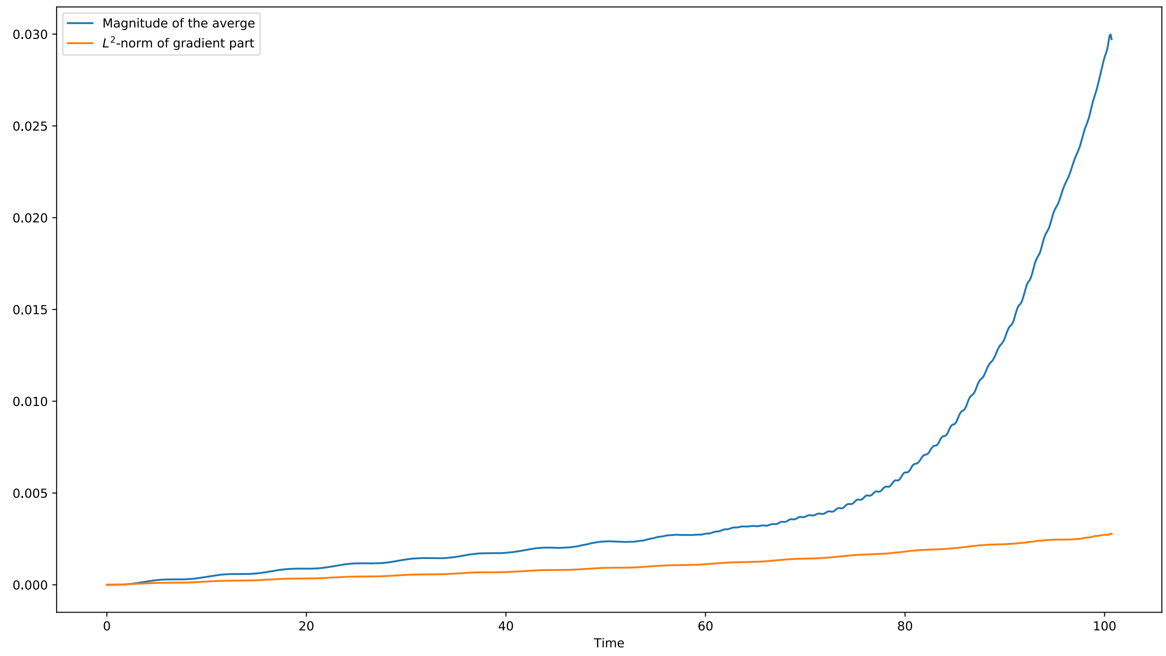


Figure 6.8: Plot of the growth of the gradient and harmonic part of the numerical solution with projected initial data

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 alraedy shown by numerical experiments in the literature [92], 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 [27]. 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:

$$\begin{aligned}
\sigma' &= -\operatorname{div} u,\\
u' &= v,\\
v' &= -\operatorname{curl}\operatorname{curl} u + \nabla\sigma + f.
\end{aligned}$$

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:

$$\Box u = \operatorname{div}(\eta \nabla u) = 0, \tag{6.5}$$

where $\Box$ is the d'Lambertian, η 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:

$$\operatorname{ein} g = 0.$$

As a companion to his space-time Regge Calculus paper [103], Sorkin [102] 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:

$$(u,v) := \int_\Omega uv, \qquad ((u,v)) := \int_{[0,T]\times\Omega} uv.$$

The action S for a scalar field u on the space-time is given by:

$$S(u) := \int_{[0,T]\times\Omega} \eta_{\alpha\mu}(\partial^\alpha u)(\partial^\beta u) = \frac{1}{2}[-((u', u')) + ((\nabla u, \nabla u))]. \tag{6.6}$$

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 a on Ω, find u on $[0,T]\times\Omega$ satisfying $u(0) = 0$, $u'(0) = a$, and

$$\begin{aligned} &u'' - \Delta u = 0, && \text{in } [0,T]\times\Omega, \\ &u = 0, && \text{on } [0,T]\times\partial\Omega, \end{aligned}$$

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 $\mathbb{T}^3$ 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

$$X := H^1L^2 \cap L^2\mathring{H}^1, \qquad X^0 = \{u \in X | u(0) = 0\}, \qquad X^T = \{u \in X | u(T) = 0\},$$

where $\mathring{H}^1$ 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

$$B(u, v) = (b, v(0)), \qquad \forall v \in X^T. \tag{6.7}$$

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 [22, 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 $\mathscr{T}$ 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 $\mathscr{T}$, 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,

$$S_h(u) := \int_{[0,T]\times\Omega} \eta_{\alpha\mu}(\partial^\alpha u)(\partial^\beta u) = \frac{1}{2}[-((u',u')) + ((\nabla u, \nabla u))].$$

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 $\mathscr{T}$. This is the Regge calculus-style discrete wave equation.

This is very natural from the finite element point of view. Let CG^1 be the space of Lagrange elements of degree 1 on $\mathscr{T}$. It is well-known that $\mathrm{CG}^1 \subset H^1H^1$. Define discrete subspaces with temporal and spatial boundary conditions:

$$X_h := \mathrm{CG}^1 \cap L^2\mathring{H}^1, \qquad X_h^0 = \{u \in X_h | u(0) = 0\}, \qquad X_h^T = \{u \in X_h | u(T) = 0\},$$

Clearly,

$$X_h \subset X, \qquad X_h^0 \subset X^0, \qquad X_h^T \subset X^T.$$

We thus get a conforming discretization of equation (6.7) via the Galerkin projection: find $u \in X_h^0$ such that

$$B(u,v) = (b, v(0)), \qquad \forall v \in X_h^T. \tag{6.8}$$

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 tent 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.

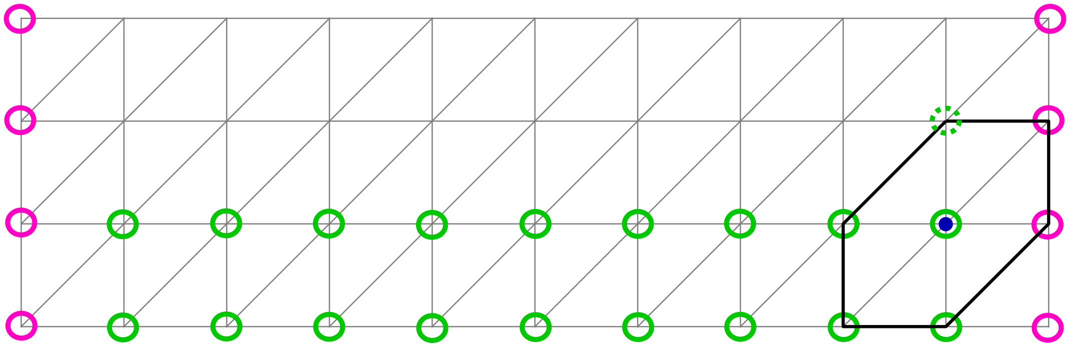


Figure 6.9: Illustration of the marching scheme

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.

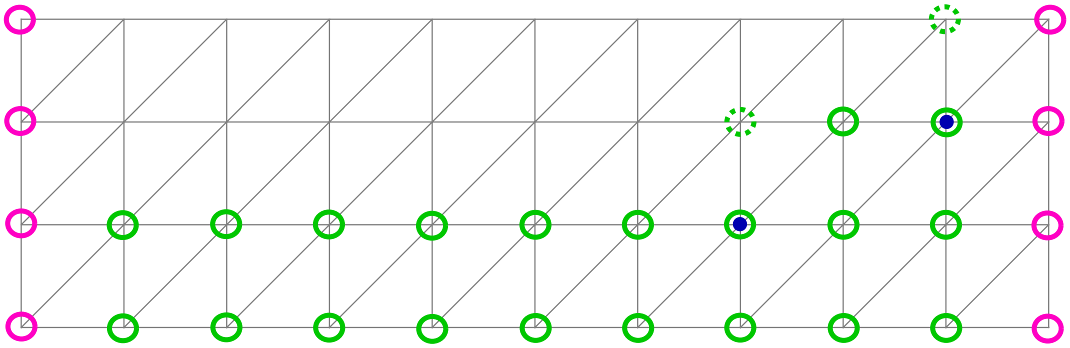


Figure 6.10: Parallel marching scheme

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.

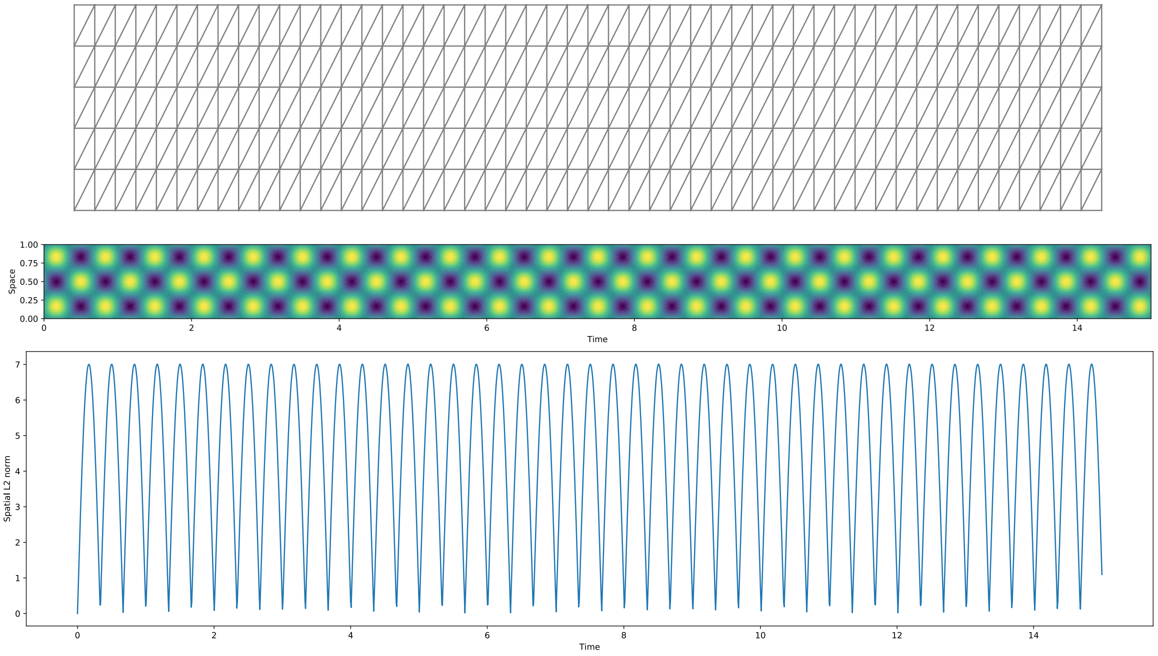


Figure 6.11: Uniform mesh

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.

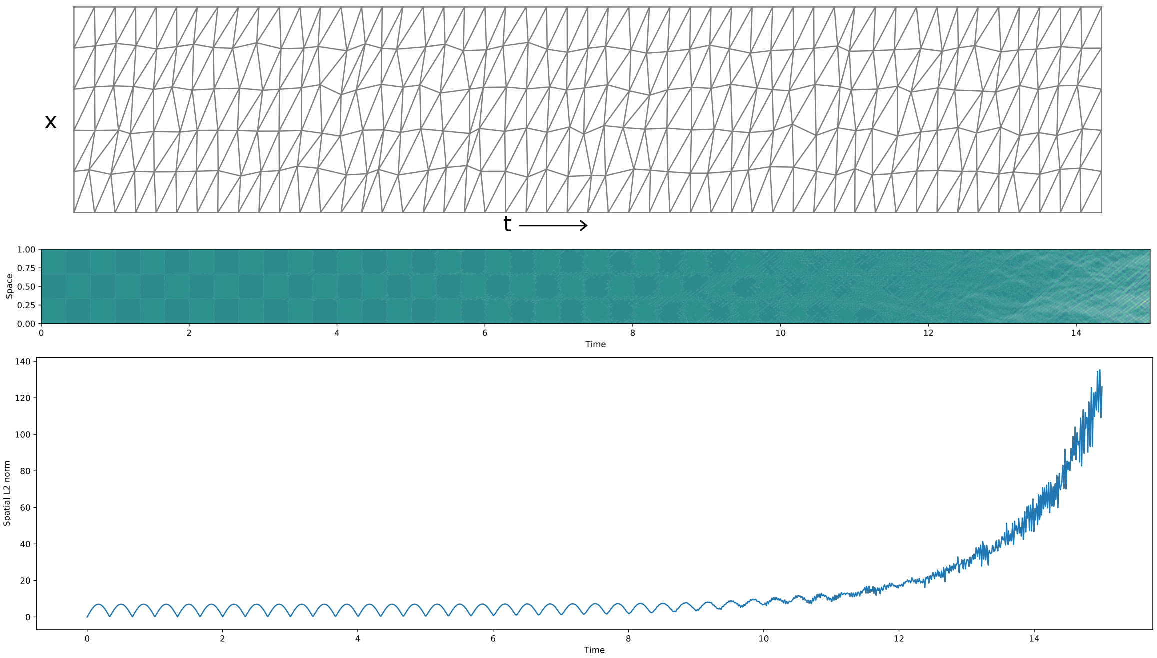


Figure 6.12: Randomly perturbed mesh

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.

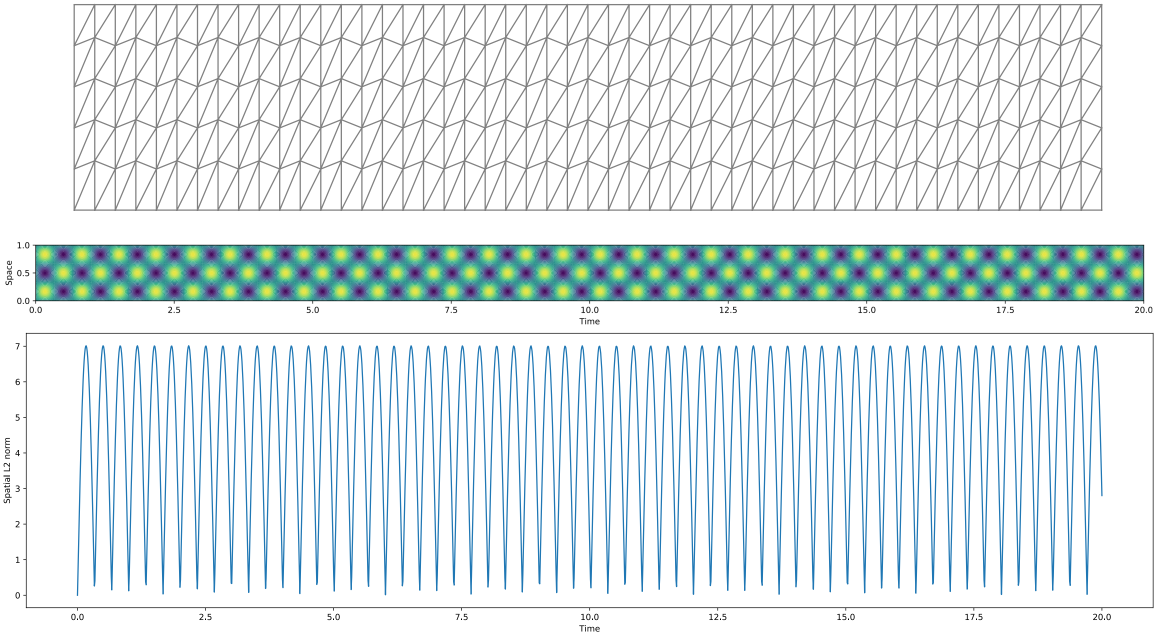


Figure 6.13: Perturb every second spatial slice

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 quicklly (notice the scale of the y-axis on the bottom panel).

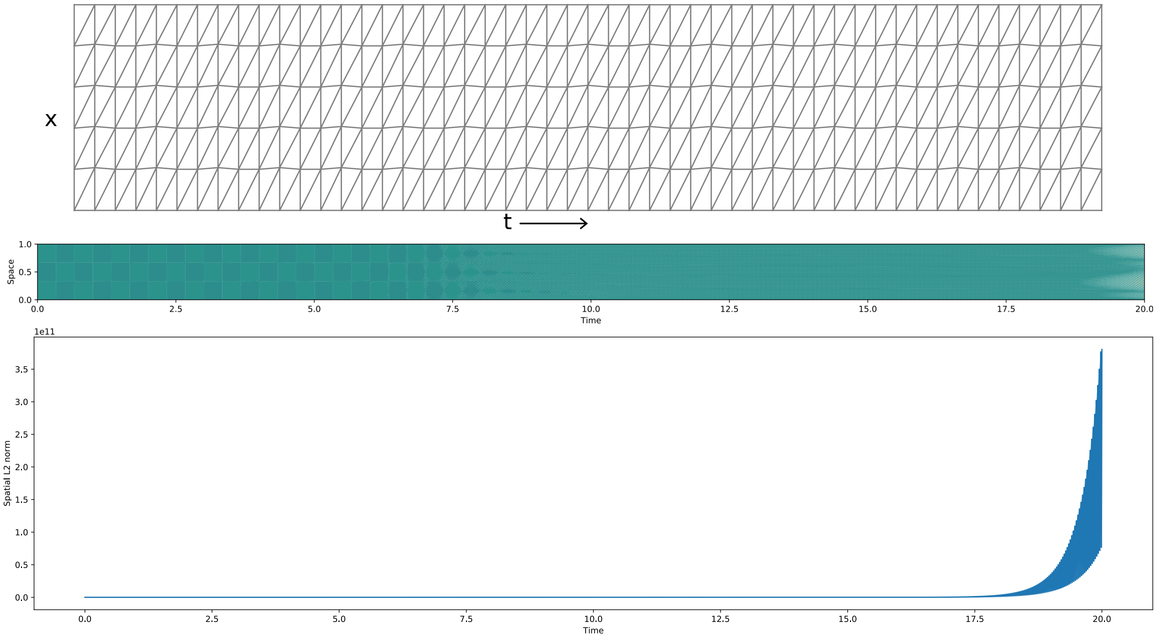
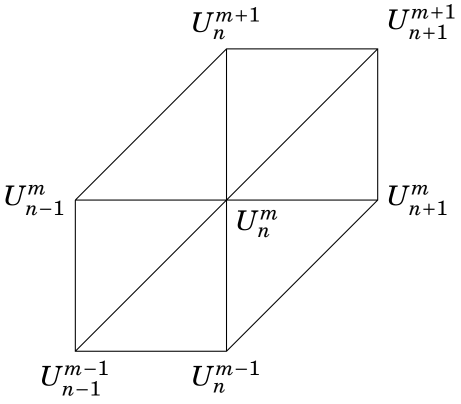


Figure 6.14: Perturb every third spatial slice

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:



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, 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:

$$k/h \le 1. \tag{6.10}$$

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 [72, Section 9.6]. The idea is simple: we check the behavior of the method by marching a discrete function of the form

$$U_n^m = r^m e^{in\theta}.$$

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 θ. Our method is stable if $|r| \le 1$ for all θ.

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.

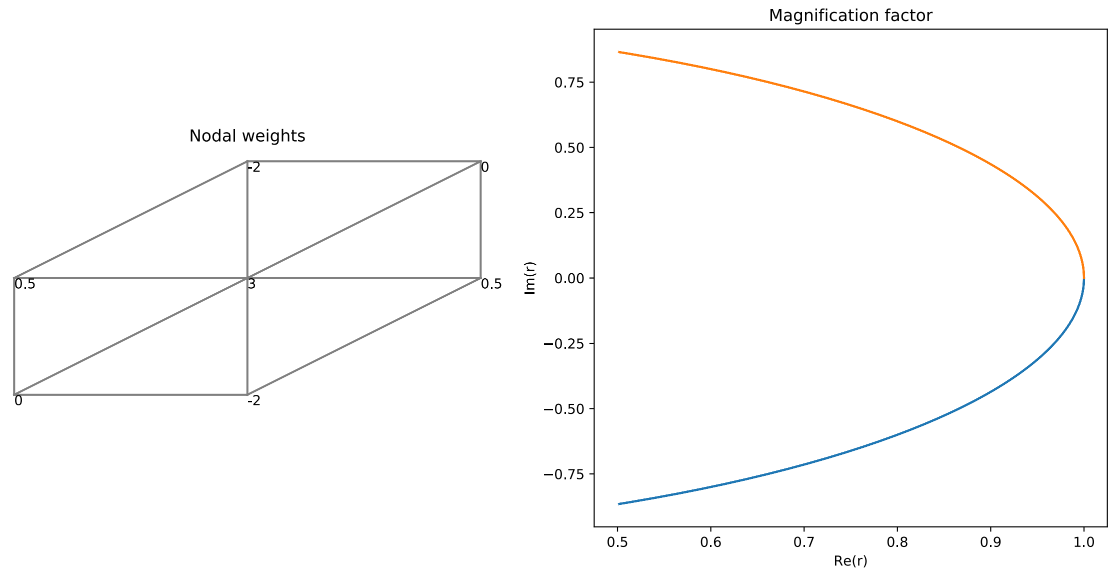


Figure 6.15: Uniform mesh $k = 0.5$ and $h = 1.0$

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.

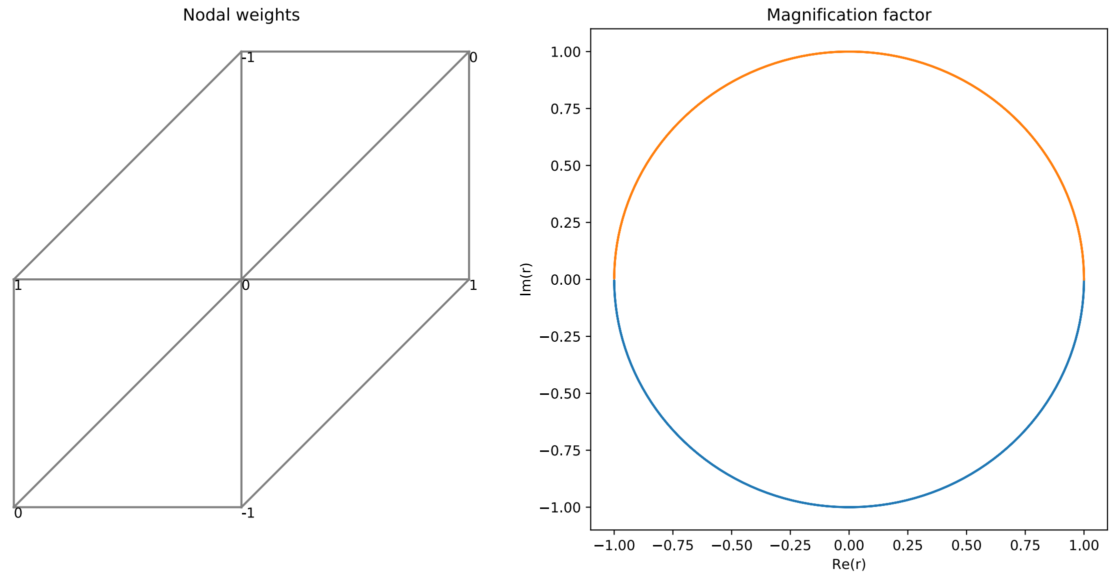


Figure 6.16: Uniform mesh $k = 1.0$ and $h = 1.0$

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.

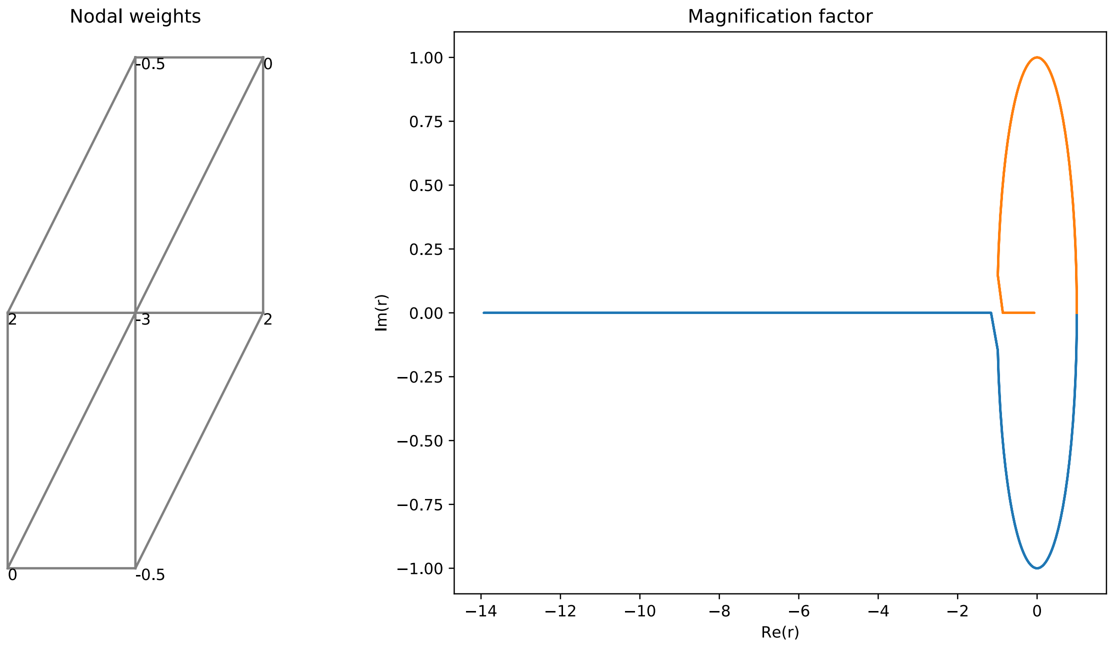


Figure 6.17: Uniform mesh $k = 1.0$ and $h = 0.5$

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.

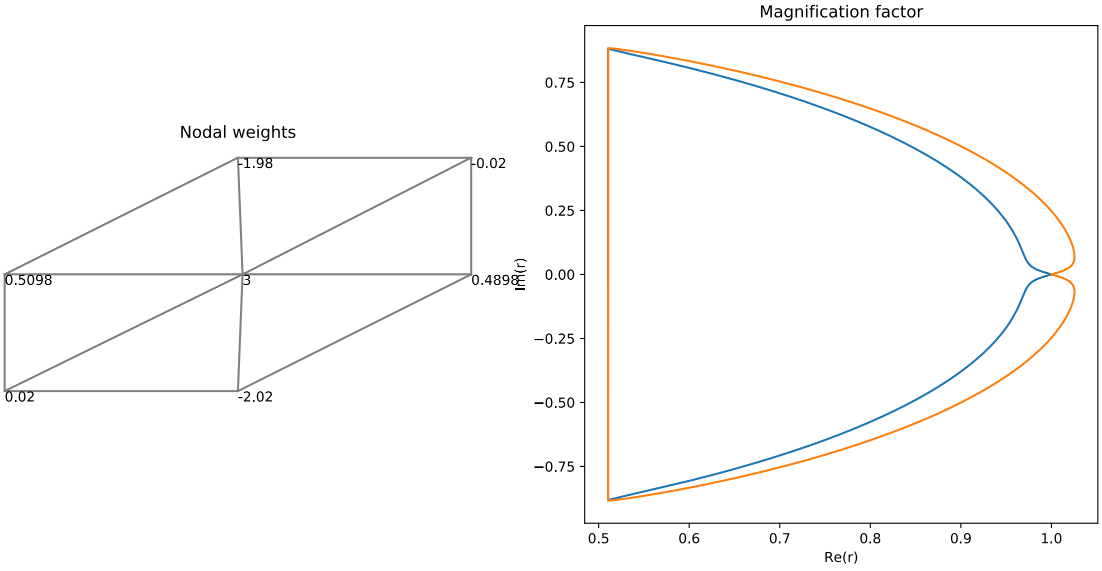


Figure 6.18: Perturb the middle node spatially

Figure 6.18 shows the situation where the very middle node is perturbed by a large amount in the temporal direction. This, however, is still stable.

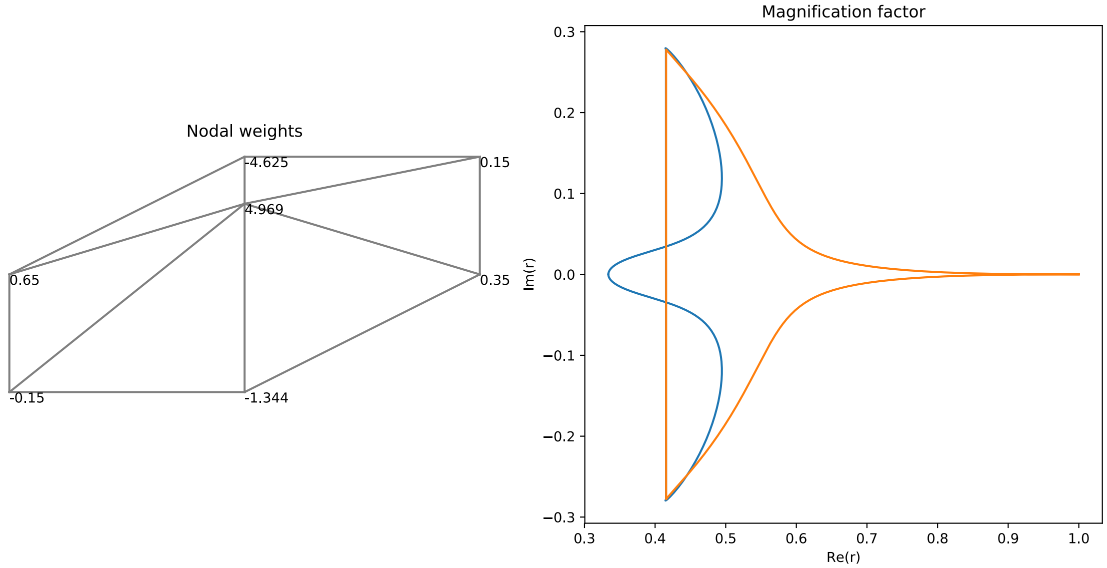


Figure 6.19: Perturb the middle node temporally

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.

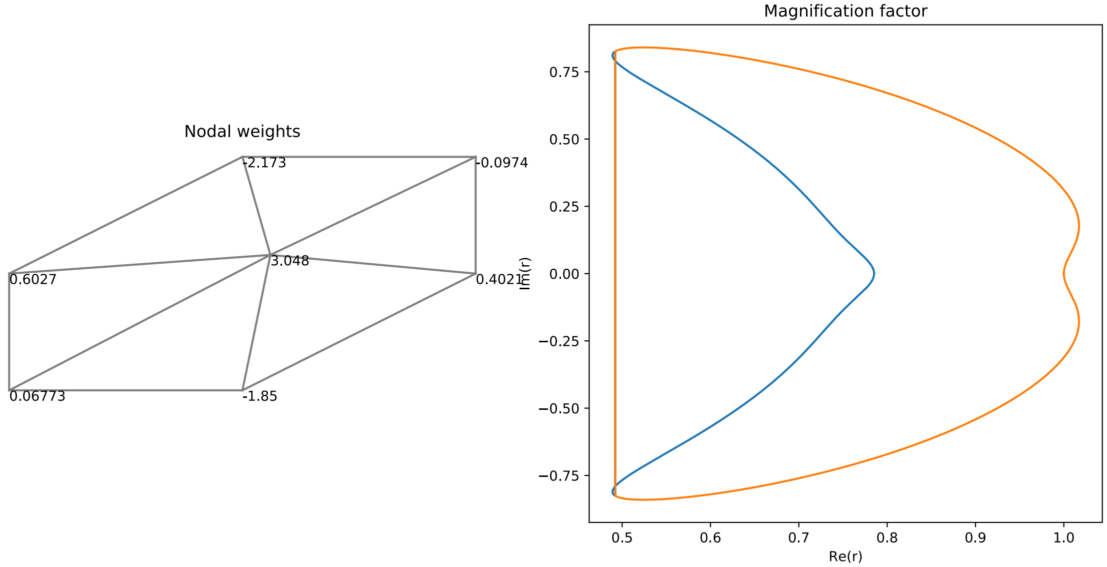


Figure 6.20: General perturbation of the middle node

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.

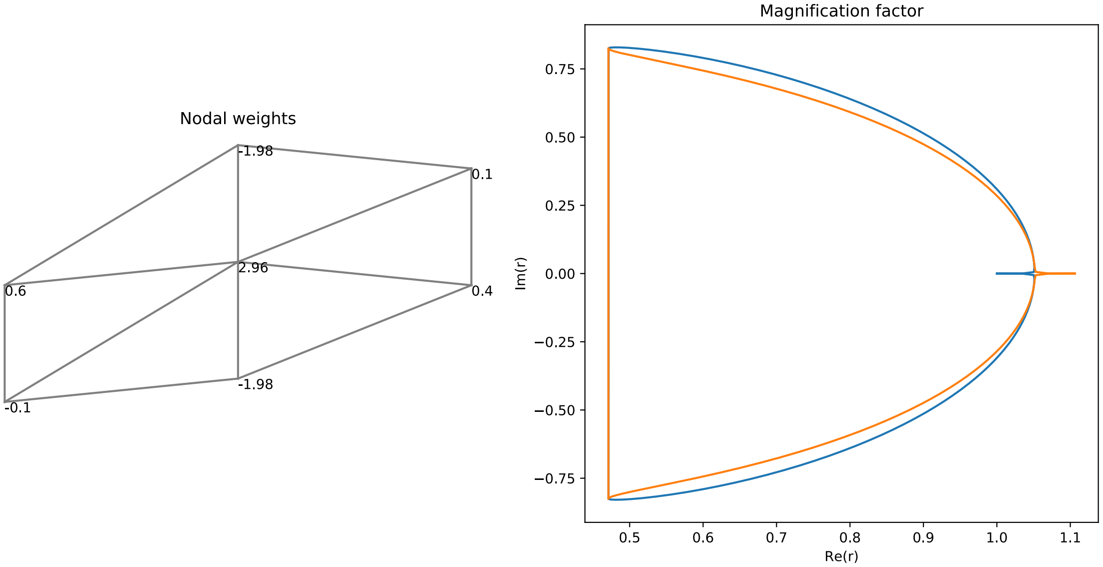


Figure 6.21: Move all middle nodes temporally

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.

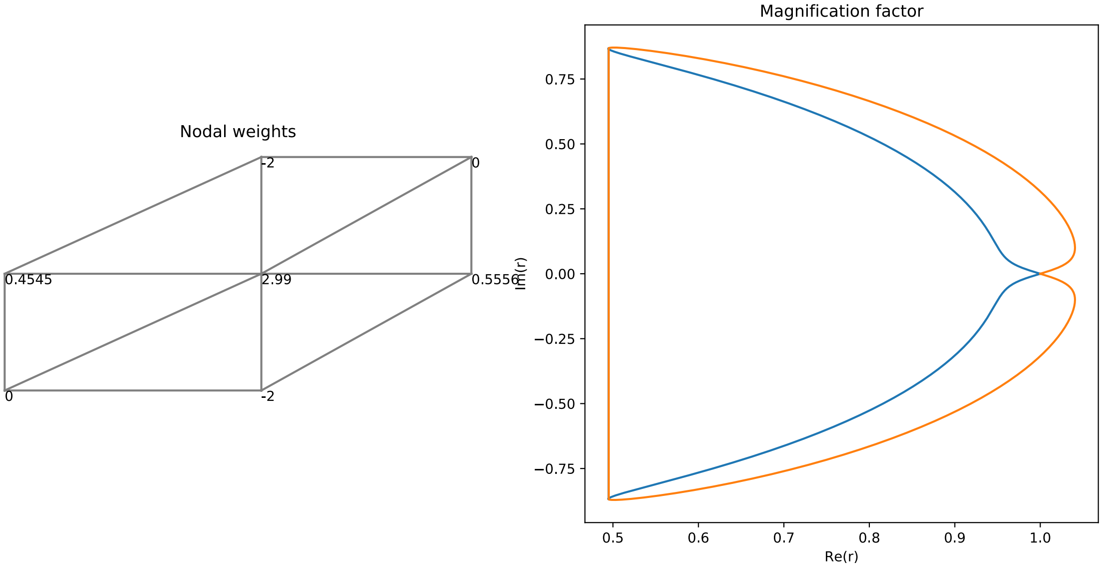


Figure 6.22: Move all middle nodes spatially

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.

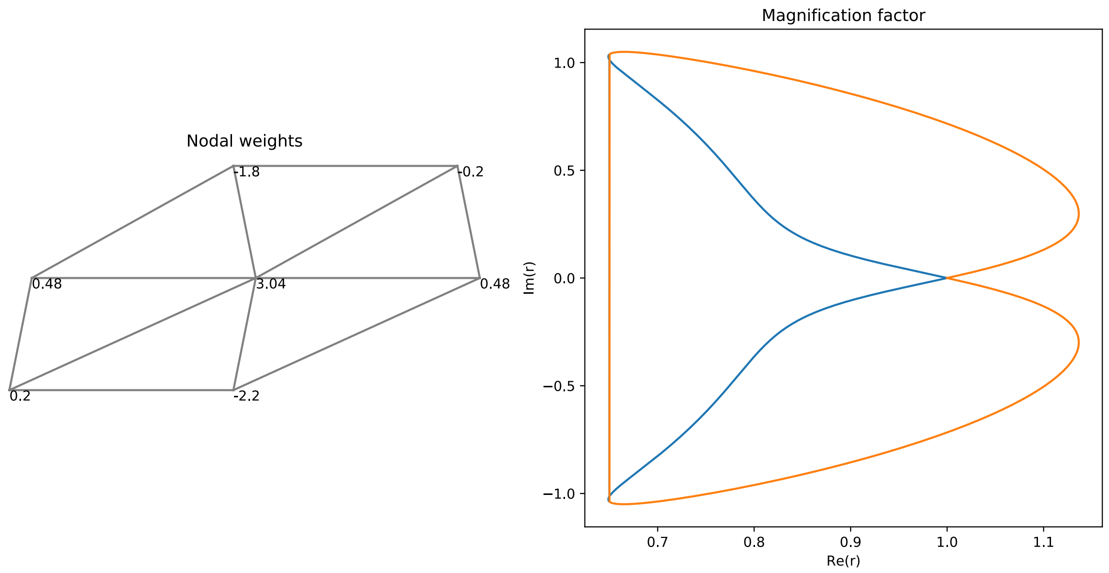


Figure 6.23: Move all middle nodes spatially

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 $|g| > 1$.

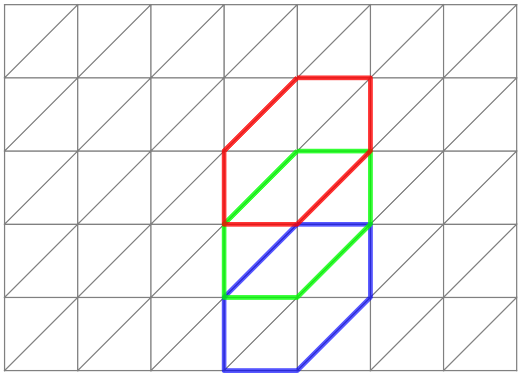


Figure 6.24: Uniform mesh $k = 0.5$ and $h = 1.0$

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 [39, 56, 57]. 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:

$$\begin{aligned} ((u',v)) - ((p,v)) &= 0, && \forall v \\ ((p',q)) + A(u,q) &= 0, && \forall q, \end{aligned}$$

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:

$$-((u',v')) + A(u,v) = 0, \qquad \forall v$$

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.

Bibliography

[1] Ralph Abraham and Jerrold E. Marsden. *Foundations of mechanics*. Benjamin/Cummings Publishing Co., Inc., Advanced Book Program, Reading, Mass., 1978. Second edition, revised and enlarged, With the assistance of Tudor Raţiu and Richard Cushman.

[2] Robert A. Adams and John J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[3] S. Agmon, A. Douglis, and L. Nirenberg. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. I. *Comm. Pure Appl. Math.*, 12:623–727, 1959.

[4] S. Agmon, A. Douglis, and L. Nirenberg. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. II. *Comm. Pure Appl. Math.*, 17:35–92, 1964.

[5] Miguel Alcubierre. *Introduction to* $3+1$ *numerical relativity*, volume 140 of *International Series of Monographs on Physics*. Oxford University Press, Oxford, 2008.

[6] A. D. Aleksandrov and V. A. Zalgaller. *Intrinsic geometry of surfaces*. Translated from the Russian by J. M. Danskin. Translations of Mathematical Monographs, Vol. 15. American Mathematical Society, Providence, R.I., 1967.

[7] D. N. Arnold and F. Brezzi. Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. *RAIRO Modél. Math. Anal. Numér.*, 19(1):7–32, 1985.

[8] Douglas N. Arnold, Richard S. Falk, and Ragnar Winther. Finite element exterior calculus, homological techniques, and applications. *Acta Numer.*, 15:1–155, 2006.

[9] Douglas N. Arnold, Richard S. Falk, and Ragnar Winther. Geometric decompositions and local bases for spaces of finite element differential forms. *Comput. Methods Appl. Mech. Engrg.*, 198:1660–1672, 2009.

[10] Douglas N. Arnold, Richard S. Falk, and Ragnar Winther. Finite element exterior calculus: from Hodge theory to numerical stability. *Bull. Amer. Math. Soc. (N.S.)*, 47:281–354, 2010.

[11] V. I. Arnol'd. *Mathematical methods of classical mechanics*, volume 60 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1989. Translated from the Russian by K. Vogtmann and A. Weinstein.

[12] Thierry Aubin. *Nonlinear analysis on manifolds. Monge-Ampère equations*, volume 252 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York, 1982.

[13] I. Babuška, J. Osborn, and J. Pitkäranta. Analysis of mixed methods using mesh dependent norms. *Math. Comp.*, 35(152):1039–1062, 1980.

[14] Jean-Paul Berrut and Lloyd N. Trefethen. Barycentric lagrange interpolation. *SIAM Review*, 46(3):501–517, 2004.

[15] George David Birkhoff. *Relativity and modern physics*. Harvard University Press, 1923.

[16] J. H. Bramble and S. R. Hilbert. Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation. *SIAM J. Numer. Anal.*, 7:112–124, 1970.

[17] Susanne C. Brenner and L. Ridgway Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[18] L. Brewin. Particle paths in a Schwarzschild spacetime via the Regge calculus. *Classical and Quantum Gravity*, 10(9):1803, 1993.

[19] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, 8(R-2):129–151, 1974.

[20] F. Brezzi and P.-A. Raviart. Mixed finite element methods for 4th order elliptic equations. In *Topics in numerical analysis, III (Proc. Roy. Irish Acad. Conf., Trinity Coll., Dublin, 1976)*, pages 33–56. Academic Press, London, 1977.

[21] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001.

[22] Thierry Cazenave and Alain Haraux. *An introduction to semilinear evolution equations*, volume 13 of *Oxford Lecture Series in Mathematics and its Applications*. The Clarendon Press, Oxford University Press, New York, 1998. Translated from the 1990 French original by Yvan Martel and revised by the authors.

[23] Sukanya Chakrabarti, Adrian P. Gentle, Arkady Kheyfets, and Warner A. Miller. Geodesic deviation in Regge calculus. *Classical and Quantum Gravity*, 16(7):2381, 1999.

[24] J. Cheeger, W. Müller, and R. Schrader. Kinematic and tube formulas for piecewise linear spaces. *Indiana Univ. Math. J.*, 35(4):737–754, 1986.

[25] Jeff Cheeger, Werner Müller, and Robert Schrader. On the curvature of piecewise flat spaces. *Comm. Math. Phys.*, 92(3):405–454, 1984.

[26] Yvonne Choquet-Bruhat. *General relativity and the Einstein equations*. Oxford Mathematical Monographs. Oxford University Press, Oxford, 2009.

[27] Alexandre Joel Chorin. Numerical solution of the Navier-Stokes equations. *Math. Comp.*, 22:745–762, 1968.

[28] Snorre H. Christiansen. A characterization of second-order differential operators on finite element spaces. *Mathematical Models and Methods in Applied Sciences*, 14(12):1881–1892, 2004.

[29] Snorre H. Christiansen. On the linearization of Regge calculus. *Numer. Math.*, 119(4):613–640, 2011.

[30] Philippe G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.

[31] C. J. S. Clarke and T. Dray. Junction conditions for null hypersurfaces. *Classical and Quantum Gravity*, 4(2):265, 1987.

[32] J. Crank and P. Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type [reprint of MR0019410 (8,409b)]. *Adv. Comput. Math.*, 6(3-4):207–226 (1997), 1996. John Crank 80th birthday special issue.

[33] Alan Demlow and Anil N. Hirani. A posteriori error estimates for finite element exterior calculus: the de Rham complex. *Found. Comput. Math.*, 14(6):1337–1371, 2014.

[34] Manfredo Perdigão do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.

[35] B. A. Dubrovin, A. T. Fomenko, and S. P. and Novikov. *Modern geometry—methods and applications. Part II*, volume 104 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1985. The geometry and topology of manifolds, Translated from the Russian by Robert G. Burns.

[36] Albert Einstein. Die Feldgleichungen der Gravitation. *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, 2:844–847, 1915.

[37] A. F. Filippov. *Differential equations with discontinuous righthand sides*, volume 18 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1988. Translated from the Russian.

[38] Richard Fitzpatrick. *An introduction to celestial mechanics*. Cambridge University Press, Cambridge, 2012.

[39] Donald A. French and Todd E. Peterson. A continuous space-time finite element method for the wave equation. *Math. Comp.*, 65(214):491–506, 1996.

[40] Adrian P. Gentle and Warner A. Miller. A brief review of Regge calculus in classical numerical relativity. In *The Ninth Marcel Grossmann Meeting*, pages 1467–1468. World Scientific Publishing Company, 2012.

[41] Roberto Giambò and Fabio Giannoni. Minimal geodesics on manifolds with discountinuous metrics. *Journal of the London Mathematical Society*, 67:527–544, April 2003.

[42] G.W. Gibbons. The Jacobi metric for timelike geodesics in static spacetimes. *Classical and Quantum Gravity*, 33(2):025004, 2016.

[43] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[44] J. Gopalakrishnan, L. E. García-Castillo, and L. F. Demkowicz. Nédélec spaces in affine coordinates. *Comput. Math. Appl.*, 49(7-8):1285–1294, 2005.

[45] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985.

[46] Misha Gromov. *Metric structures for Riemannian and non-Riemannian spaces*. Modern Birkhäuser Classics. Birkhäuser Boston, Inc., Boston, MA, english edition, 2007. Based on the 1981 French original, With appendices by M. Katz, P. Pansu and S. Semmes, Translated from the French by Sean Michael Bates.

[47] E. Hairer, S. P. Nø rsett, and G. Wanner. *Solving ordinary differential equations. I*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 1993. Nonstiff problems.

[48] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006.

[49] S. W. Hawking and G. F. R. Ellis. *The large scale structure of space-time*. Cambridge University Press, London, 1973. Cambridge Monographs on Mathematical Physics, No. 1.

[50] David Hilbert. Über das Dirichletsche Prinzip. *Math. Ann.*, 59(1-2):161–186, 1904.

[51] Klaus Hildebrandt, Konrad Polthier, and Max Wardetzky. On the convergence of metric and geometric properties of polyhedral surfaces. *Geometriae Dedicata*, 123(1):89–112, 2007.

[52] R. Hiptmair. Canonical construction of finite elements. *Math. Comp.*, 68(228):1325–1346, 1999.

[53] R. Hiptmair. Finite elements in computational electromagnetism. *Acta Numerica*, 11:237–339, 2002.

[54] Michael Holst and Ari Stern. Geometric variational crimes: Hilbert complexes, finite element exterior calculus, and problems on hypersurfaces. *Found. Comput. Math.*, 12(3):263–293, 2012.

[55] Ingrid Hotz, Louis Feng, Hans Hagen, Bernd Hamann, and Kenneth Joy. *Visualization and processing of tensor fields*, pages 269–281. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[56] Thomas J. R. Hughes and Gregory M. Hulbert. Space-time finite element methods for elastodynamics: formulations and error estimates. *Comput. Methods Appl. Mech. Engrg.*, 66(3):339–363, 1988.

[57] Gregory M. Hulbert and Thomas J. R. Hughes. Space-time finite element methods for second-order hyperbolic equations. *Comput. Methods Appl. Mech. Engrg.*, 84(3):327–348, 1990.

[58] Gareth Wyn Jones and S. Jonathan Chapman. Modeling growth in biological materials. *SIAM Rev.*, 54(1):52–118, 2012.

[59] Mikhail O Katanaev. Geometric theory of defects. *Physics-Uspekhi*, 48(7):675, 2005.

[60] Tosio Kato. *Perturbation theory for linear operators*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.

[61] Parandis Khavari. *Regge calculus as a numerical approach to General relativity*. PhD thesis, Univeristy of Toronto, 2009.

[62] R. Kimmel and J. A. Sethian. Computing geodesic paths on manifolds. *Proceedings of the National Academy of Sciences*, 95(15):8431–8435, 1998.

[63] Robert C. Kirby. Fast simplicial finite element algorithms using Bernstein polynomials. *Numer. Math.*, 117(4):631–652, 2011.

[64] Sergiu Klainerman. Mathematical challenges of general relativity. *Rend. Mat. Appl. (7)*, 27(2):105–122, 2007.

[65] H. Kleinert. Towards a unified field theory of defects and stresses. *Lett. Nuovo Cimento (2)*, 35(2):41–45, 1982.

[66] Hagen Kleinert. *Multivalued fields in condensed matter, electromagnetism, and gravitation*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2008.

[67] Kazuo Kondo. On the geometrical and physical foundations of the theory of yielding. In *Proceedings of the Second Japan National Congress for Applied Mechanics, 1952*, pages 41–47. Science Council of Japan, Tokyo, 1953.

[68] Heinz-Otto Kreiss and Jens Lorenz. *Initial-boundary value problems and the Navier-Stokes equations*, volume 47 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2004. Reprint of the 1989 edition.

[69] Wolfgang Krendl, Katharina Rafetseder, and Walter Zulehner. A decomposition result for biharmonic problems and the Hellan-Herrmann-Johnson method. *Electron. Trans. Numer. Anal.*, 45:257–282, 2016.

[70] Ekkehart Kröner. Continuum theory of defects. In Roger Balian, Maurice Kléman, and Jean-Paul Poirier, editors, *Physics of defects, Les Houches 1980*, 1980.

[71] Alexander Lecke, Roland Steinbauer, and Robert Švarc. The regularity of geodesics in impulsive pp-waves. *General Relativity and Gravitation*, 46(1):1648, 2013.

[72] Randall J. LeVeque. *Finite difference methods for ordinary and partial differential equations*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007. Steady-state and time-dependent problems.

[73] André Lieutier and Boris Thibert. Convergence of geodesics on triangulations. *Computer Aided Geometric Design*, 26(4):412 – 424, 2009. Geometric Modeling and Processing 2008, 5th International Conference on Geometric Modeling and Processing.

[74] A. Logg. Efficient representation of computational meshes. *International Journal of Computational Science and Engineering*, 4(4):283–295, 2009.

[75] Anders Logg, Kent-Andre Mardal, and Garth Wells. *Automated solution of differential equations by the finite element method*. Springer Berlin Heidelberg, 2012.

[76] Tom Lyche and Karl Scherer. On the p-norm condition number of the multivariate triangular Bernstein basis. *J. Comput. Appl. Math.*, 119(1-2):259–273, 2000. Dedicated to Professor Larry L. Schumaker on the occasion of his 60th birthday.

[77] Alexander Lytchak and Asli Yaman. On Hölder continuous Riemannian and Finsler metrics. *Trans. Amer. Math. Soc.*, 358(7):2917–2926 (electronic), 2006.

[78] J. E. Marsden. Generalized Hamiltonian mechanics: A mathematical exposition of non-smooth dynamical systems and classical Hamiltonian mechanics. *Arch. Rational Mech. Anal.*, 28:323–361, 1967/1968.

[79] J. E. Marsden. Non-smooth geodesic flows and classical mechanics. *Canad. Math. Bull.*, 12:209–212, 1969.

[80] Jerrold E. Marsden and Thomas J. R. Hughes. *Mathematical foundations of elasticity*. Dover Publications, Inc., New York, 1994. Corrected reprint of the 1983 original.

[81] Charles W. Misner, Kip S. Thorne, and John Archibald Wheeler. *Gravitation*. W. H. Freeman and Co., San Francisco, Calif., 1973.

[82] Joseph S. B. Mitchell and Christos H. Papadimitriou. The weighted region problem: finding shortest paths through a weighted planar subdivision. *J. ACM*, 38(1):18–73, January 1991.

[83] Peter Monk. *Finite element methods for Maxwell's equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2003.

[84] James R. Munkres. *Elementary differential topology*, volume 1961 of *Lectures given at Massachusetts Institute of Technology, Fall*. Princeton University Press, Princeton, N.J., 1966.

[85] J.-C. Nédélec. Mixed finite elements in $\mathbf{R}^3$. *Numer. Math.*, 35(3):315–341, 1980.

[86] J.-C. Nédélec. A new family of mixed finite elements in $\mathbf{R}^3$. *Numer. Math.*, 50(1):57–81, 1986.

[87] Walter Noll. Materially uniform simple bodies with inhomogeneities. *Arch. Rational Mech. Anal.*, 27:1–32, 1967.

[88] Joseph O'Rourke. *Computational geometry in C*. Cambridge university press, 1998.

[89] Astrid Pechstein and Joachim Schöberl. Tangential-displacement and normal-normal-stress continuous mixed finite elements for elasticity. *Math. Models Methods Appl. Sci.*, 21(8):1761–1782, 2011.

[90] Astrid Pechstein and Joachim Schöberl. An analysis of the TDNNS method using natural norms. *arXiv:1606.06853 [math.NA]*, 2016.

[91] Astrid Pechstein and Joachim Schöberl. The TDNNS method for Reissner-Mindlin plates. *arXiv:1704.03649 [math.NA]*, 2017.

[92] Frank Peuker. *Simplicial method for solving selected problems in General relativity numerically*. PhD thesis, Friedrich-Schiller-Universität Jena, 2009.

[93] Konrad Polthier and Markus Schmies. *Straightest geodesics on polyhedral surfaces*, pages 135–150. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[94] Konrad Polthier and Markus Schmies. Geodesic flow on polyhedral surfaces. In *Data Visualization 99: Proceedings of the Joint EUROGRAPHICS and IEEE TCVG Symposium on Visualization in Vienna*, pages 179–188, Vienna, 1999. Springer Vienna.

[95] P. A. Raviart and J. M. Thomas. A mixed finite element method for 2nd order elliptic problems. In *Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975)*, pages 292–315. Lecture Notes in Math., Vol. 606. Springer, Berlin, 1977.

[96] T. Regge. General relativity without coordinates. *Il Nuovo Cimento*, 19(3):558–571, 1961.

[97] Michael Renardy and Robert C. Rogers. *An introduction to partial differential equations*, volume 13 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2004.

[98] Hans Ringström. *The Cauchy problem in general relativity*. ESI Lectures in Mathematics and Physics. European Mathematical Society (EMS), Zürich, 2009.

[99] A. P. S. Selvadurai. *Partial differential equations in mechanics. 2*. Springer-Verlag, Berlin, 2000. The biharmonic equation, Poisson's equation.

[100] Astrid Sinwel and Joachim Schöberl. Tangential-displacement and normal-normal-stress continuous mixed finite elements for elasticity. Technical report, Johann Radon Institute for Computational and Applied Mathematics Austrian Academy of Sciences, 2007.

[101] Astrid Sabine Sinwel. *A new family of mixed finite elements for elasticity*. PhD thesis, Johannes Kepler Universität, 2009.

[102] Rafael Sorkin. The electromagnetic field on a simplicial net. *Journal of Mathematical Physics*, 16(12):2432–2440, 1975.

[103] Rafael Sorkin. Time-evolution problem in regge calculus. *Phys. Rev. D*, 12:385–396, Jul 1975.

[104] Roland Steinbauer. Every Lipschitz metric has C^1-geodesics. *Classical Quantum Gravity*, 31(5):057001, 3, 2014.

[105] Ari Stern. L^p change of variables inequalities on manifolds. *Math. Inequal. Appl.*, 16(1):55–67, 2013.

[106] The CGAL Project. *CGAL User and reference manual*. CGAL Editorial Board, 4.11 edition, 2017.

[107] Vidar Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006.

[108] Stephen P Timoshenko and Sergius Woinowsky-Krieger. *Theory of plates and shells*. McGraw-hill, 1959.

[109] Vito Volterra. Sur l'équilibre des corps élastiques multiplement connexes. *Ann. Sci. École Norm. Sup. (3)*, 24:401–517, 1907.

[110] Chris Wainwright and Ruth M Williams. Area Regge calculus and discontinuous metrics. *Classical and Quantum Gravity*, 21(21):4865, 2004.

[111] Robert M. Wald. *General relativity*. University of Chicago Press, Chicago, IL, 1984.

[112] T. Warburton. An explicit construction of interpolation nodes on the simplex. *J. Engrg. Math.*, 56(3):247–262, 2006.

[113] Frank W. Warner. *Foundations of differentiable manifolds and Lie groups*, volume 94 of *Graduate Texts in Mathematics*. Springer-Verlag, New York-Berlin, 1983. Corrected reprint of the 1971 edition.

[114] Juncheng Wei and Matthias Winter. Stationary solutions for the Cahn-Hilliard equation. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 15(4):459–492, 1998.

[115] Hassler Whitney. *Geometric integration theory*. Princeton University Press, Princeton, N. J., 1957.

[116] R. Williams. Quantum Regge calculus. In Daniele Oriti, editor, *Approaches to Quantum Gravity*, pages 360–377. Cambridge University Press, 2009. Cambridge Books Online.

[117] R. M. Williams and P. A. Tuckey. Regge calculus: a brief review and bibliography. *Classical and Quantum Gravity*, 9(5):1409, 1992.

[118] Ruth M. Williams and G. F. R. Ellis. Regge calculus and observations. I. Formalism and applications to radial motion and circular orbits. *General Relativity and Gravitation*, 13(4):361–395, 1981.

[119] Ruth M. Williams and G. F. R. Ellis. Regge calculus and observations. II. Further applications. *General Relativity and Gravitation*, 16(11):1003–1021, 1984.

[120] Arash Yavari and Alain Goriely. Riemann-Cartan geometry of nonlinear dislocation mechanics. *Arch. Ration. Mech. Anal.*, 205(1):59–118, 2012.