# A Mathematical Primer for Computational Data Sciences

C. Bajaj

September 2, 2018

# Contents

# Preface

A mix of algebra and geometry, combinatorics, and topology,
A mathematical primer for those working in algorithmic and computational structural biology, i.e.modeling and discovering biological structure to function relationships using the computer
Lacks statistics
Thanks to members of ccv

# Introduction

intro to math (algebra, geometry, topology, statistics) for data sciences

applicable math: Algebra and Trigonometry: The ideas of linear algebra are used throughout . vectors, matrices, tensors

"Linear Algebra and Its Applications" Gilbert Strang Academic Press

Matrix Computations Gene Golub and Charles Van Loan Johns Hopkins University Press

Differential Geometry: Multivariable calculus is the prerequisite for this area.

Elementary Differential Geometry Barrett O'Neill Academic Press

These sub-areas include sampling theory, matrix equations, numerical solution of differential equations, and optimization. Book recommendation:

Numerical Recipes in C: The Art of Scientific Computing William Press, Saul Teukolsky, William Vetterling and Brian Flannery Cambridge University Press

Sampling Theory and Signal Processing

At the heart of sampling theory are concepts such as convolution, the Fourier transform, and spatial and frequency representations of functions. These ideas are also important in the fields of image and audio processing. Book recommendation:

The Fourier Transform and Its Applications Ronald N. Bracewell McGraw Hill

Computational Geometry in C Joseph O'Rourke Cambridge University Press

Computational Geometry: An Introduction Franco Preparata and Michael Shamos Springer-Verlag

Index to content of chapters:

chap 2 vector spaces, metric spaces, Hilbert spaces,

chap 2 describes triangulations, Delaunay, convex decompositions, tilings, packings, Voronoi diagrams

chap 3 describes polynomials, piecewise polynomials, algebraic functions, splines over triangulations, quads, convex decompositions

chap 4 describes differential geometry, inner products and discretization of differential operators used in finite difference and finite element solution of PDE's

chap 5 describes morse-smale complexes, critical point, integral manifold stratifications of shape and smooth and discrete functions

chap 6 describes exterior calculus, differential forms and homology of discrete function spaces used in solution of PDE's

chap 7 describes branching topology and geometry

# Chapter 1

# Graphs, Triangulations and Complexes

## Key Chapter Concepts

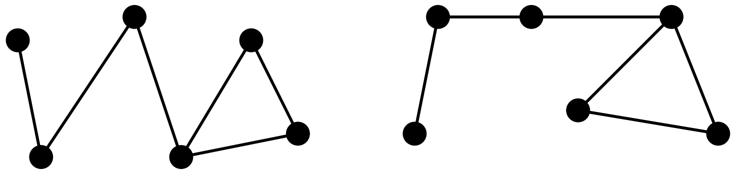
- Intertwined role of geometry, topology and combinatorics in domain definition.
- Unification of concepts for describing shape in any dimmension.
- Smooth shape description needed at all scales of biological modeling.

## 1.1 Graph Theory

## 1.2 Combinatorial vs. Embedded Graphs

A **graph** is a set of $V$ of vertices and a set $E$ of edges between vertices.
An **embedded graph** is a graph where $V \subset \mathbb{R}^n$. A **planar graph** is a graph where $V \subset \mathbb{R}^2$ such that no vertices are duplicates and no edges intersect. If a graph is not embedded, it is called a **combinatorial graph**.

**Embedded Graphs**



The two graphs shown are the same if interpreted as combinatorial graphs but different if interpreted as embedded graphs. This reflects the basic notion that topological properties (e.g. the adjacency relations between vertices) are more fundamental to a shape than geometrical properties (e.g. the particular location in space of each vertex). In biological modeling, topological properties are often well known while geometrical properties are more difficult to characterize. For instance, the sequence of amino acids along a particular protein stays fixed even while the actual shape of the protein undergoes rapid, frequent, and sometimes dramatic geometric changes.

The **genus of a graph** is defined to be the smallest value $g$ such that the graph can be embedded on a surface of genus $g$.
The **Euler characteristic** of a graph is given by

$$\chi := V - E + F$$

where $V, E, F$ are the number of vertices, edges and faces in the mesh, respectively. It is a theorem of algebraic topology that $\chi$ is an invariant of a domain, i.e. its value is independent of the mesh used to compute it so long as that mesh is homeomorphic to the domain.
The Euler characteristic is related to the genus of the domain by the relationship

$$\chi = 2 - 2g$$

### 1.2.1 Network Theory

A **direceted graph** is a graph whose edges have a specific orientation. A **flow network** is a directed graph where each edge also has a capacity.

**Max Cut Min Flow Theorem** A flow network can be thought of as a highway system with only one-way streets. Vertices represent locations and edges represent the streets between them. The direction of an edge indicates which way traffic is allowed to travel on that street. The capacity of an edge represents the maximum traffic that can flow down the street at any one moment (e.g. the number of lanes in the road).
The Max Cut Min Flow Theorem states that in a flow network, the maximum amount of flow passing from a source to a sink is equal to the minimum capacity which when removed in a specific way from the network causes the situation that no flow can pass from the source to the sink. It is a formalization and generalization of the familiar notion that a chain is only as strong as its weakest link.

### 1.2.2 Trees and Spanning Trees

Minimal spanning trees, etc.

## 1.3 Topological Complexes

### 1.3.1 Pointset Topology

Let $S$ be a set and let $T$ be a family of subsets of $S$. Then $T$ is called a **topology** on $S$ if

- Both the empty set and $S$ are elements of $T$.
- Any union of arbitrarily many elements of $T$ is an element of $T$.
- Any intersection of finitely many elements of $T$ is an element of $T$.

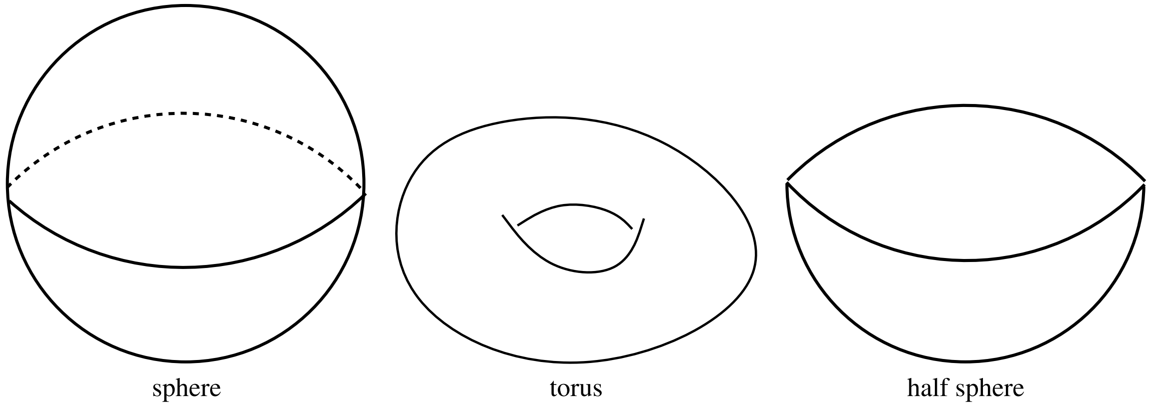
If $T$ is a topology on $S$, then the pair ($S$, $T$) is called a **topological space**. If the topology is implicit, the space $S$ is listed without mention of $T$.

The members of $T$ are called the **open sets** of $S$. A set is called **closed** if its complement is in $T$.

A **neighborhood** of a point **x** $\in S$ is any element of $T$ containing **x**.

We will often deal with subsets of $\mathbb{R}^n$ with the usual topology. This means the set $S$ is the points of $\mathbb{R}^n$ and the $T$ is formed from all open balls of any radius in $\mathbb{R}^n$.

**Manifolds**



A manifold is a special kind of topological space commonly used in domain modeling. At any point **x** in an $n$-manifold $M$, there exists a neighborhood of **x** which is homeomorphic to $\mathbb{R}^n$. Thus the surface of a sphere, the surface of a cube and the surface of a torus are all examples both 2-manifolds.
If the manifold has boundary, then its boundary is described by those points which only have neighborhoods homeomorphic to $\mathbb{R}^{n-1}$.
In the figure, the sphere and torus have genus 1 and 2, respectively. The half sphere is an example of a 2-manifold with boundary.

### 1.3.2 CW-complexes

A **Hausdorff space** is a topological space in which distinct points have disjoint neighborhoods.

**Definition 1.1.** A **CW-complex** is a Hausdorff space $X$ together with a partition of $X$ into open cells of varying dimension such that

1. For each $n$-dimensional open cell $C$ in the set $X$, there exists a continuous map $f$ from the $n$-dimensional closed ball to $X$ such that the restriction of $f$ onto the interior of the ball is a homeomorphism onto the cell $C$.

2. The image of the boundary of the open ball (i.e. the boundary of the open cell $C$) intersects only finitely many other cells.

A CW-complex can be presented as a sequence of spaces and maps

$$X_0 \hookrightarrow X_1 \hookrightarrow \ldots \hookrightarrow X_n \hookrightarrow \ldots$$

where each space $X_n$, called the $n$-dimensional skeleton of the presentation, is the result of attaching copies of the $n$-disk $D^n := \{x \in \mathbb{R}^n : ||x|| \leq 1\}$ along their boundaries $S^{n-1} := \partial D^n$ to $X_{n-1}$.
A Voronoi decomposition is a special kind of the more general class of CW-complexes.

**Computational Homology**
Triangulationas and CW-complexes can be used to compute the homology groups of a manifold, a topological invariant. The ranks of these groups are called the Betti numbers, a simple and geometrically meaninful topological invariant. We will discuss this further in Section 1.4.1.

### 1.3.3 Morse Functions and the Morse-Smale Complex

Morse theory(Adding a citation here) provides useful results not only for the construction of contour trees but also for an evaluation of the smooth distance function $h_\Sigma$ defined at the end of Section 1. As in the previous section, we consider a smooth function $f : M \to \mathbb{R}^1$, now adding the restriction that $M$ is a compact 3-manifold without boundary. Let $m \in M$ be a non-degenerate critical point of $f$, meaning the derivative map $df_m$ is the zero map and the Hessian at $m$ is non-singular. We note that a critical *point* $m$ lies in the domain $M$ of $f$ as opposed to a critical *value* $r$ which lies in the range $\mathbb{R}^1$ of $f$. The Morse Lemma [154] states that $f$ exhibits quadratic behavior in a small neighborhood $m$. That is, we may choose a coordinate chart about $m$ such that locally

$$f(x, y, z) = f(m) \pm x^2 \pm y^2 \pm z^2$$

We define the *index* of $m$ to be the number of minus signs in the equation above. It can be shown that the index is independent of the coordinate chart and that it is equal to the number of negative eigenvalues of the Hessian at $m$.

Thus, in three dimensions, a non-degenerate critical point can have index 0, 1, 2, or 3. These indices correspond to minima, 1-saddles, 2-saddles, and maxima of the function $f$, respectively

We can unambiguously connect these critical points into a meaningful structure known as the Morse complex. Away from critical points, the gradient vector $\nabla f$ is non-zero and points in the direction of maximum positive change. If we construct a maximal path whose velocity vectors coincide with the gradient vector at each point on the path, we will always connect two critical points. Such a path is called an *integral path* and necessarily terminates at a critical point of index 1, 2, or 3. We define the *stable manifold* of a non-degenerate critical point $m \in M$ to be the union of $m$ and the images of all integral paths on $M$ terminating at $m$. We note that an unstable manifold can be defined similarly, but we will not need it in this paper.(Critical inconsistency, changes shall be made here) For our purposes, the *Morse Complex* is defined to be the union of all maxima and their stable manifolds.

Previous work employing the Morse complex has dealt primarily with two types of input functions: grids and surfaces. The complex has been defined on two-dimensional grids and three-dimensional unstructured tetrahedral grids by Edelsbrunner, et al. [72, 71]. Cazals, Chazal and Lewiner used the Morse complex for molecular shape analysis in [43]. As described in Section 3, our work uses the Morse complex to aid in the curation of molecular surfaces.

Upper left: critical points (minima, saddle points and maxima pictured as blue, green and red disks), and three integral lines (pink curves) of a Morse function. Black arrows show the gradient of that function. Upper right: ascending 2-manifolds : the set of points belonging to integral lines whose destination is the same minimum (critical point of index 0). Lower left: descending 2-manifolds : the set of points belonging to integral lines whose origin is the same maximum (critical point of index 2). Lower right: the Morse-Smale complex : a natural tesselation of space into cells induced by the gradient fo the function. Each cell is the set of points belonging to integral lines whose origin and destination are identical (i.e. each cell is the intersection of an ascending and a descending manifold). The purple region is a 2-cell: intersection of an ascending and a descending 2-manifold (red and blue regions) where all field lines have the same orgin and destination (a minimum and a maxium). The yellow curve is a 1-cell (also called an arc): the intersection of and ascending 2-manifold (blue region) and a descending 1-manifolds (green+yellow curves, originating from the same saddle point).

### 1.3.4 Signed Distance Function and Critical Points of Discrete Distance Functions

It seems that this paragraph is cited from somewhere, what is the citation reference? (yiwang)

The key ingredient in ranking the topological features of the extracted level set is the *distance function* over $\mathbb{R}^3$. The distance function has been used earlier for reconstruction and image feature identification [16, 44, 69, 96, 233]. Chazal and Lieutier [47] have used it for stable medial axis construction. Dey, Giesen and Goswami have used distance function for object segmentation and matching [66]. Goswami, Dey and Bajaj have used it for detailed feature analysis of shape via an annotation of flat and tubular features in addition to shape segmentation [101]. Recently, Bajaj and Goswami have shown a novel use of distance function, induced by a molecular surface, in order to detect secondary structural motifs of a protein molecule [20]. The close connection between the critical point structure of the distance function and the topology of the surface and its complement is what we utilize to detect and remove small topological artifacts.

In this context, $\Sigma$ is the extracted isosurface. For the ease of computation, we approximate $h_\Sigma$ by $h_P$ which assigns to every point in $\mathbb{R}^3$, the distance to the nearest point from the set $P$ which finitely samples $\Sigma$.

$$h_P \; : \; \mathbb{R}^3 \to \mathbb{R}, \;\; x \mapsto \min_{p \in P} \|x - p\|$$

With our extracted isosurface, we make use of the distance function introduced above. Given a compact surface $\Sigma$ smoothly embedded in $\mathbb{R}^3$, a distance function $h_\Sigma$ can be designed over $\mathbb{R}^3$ that assigns to each point its distance to $\Sigma$.

$$h_\Sigma \; : \; \mathbb{R}^3 \to \mathbb{R}, \;\; x \mapsto \inf_{p \in \Sigma} \|x - p\|$$

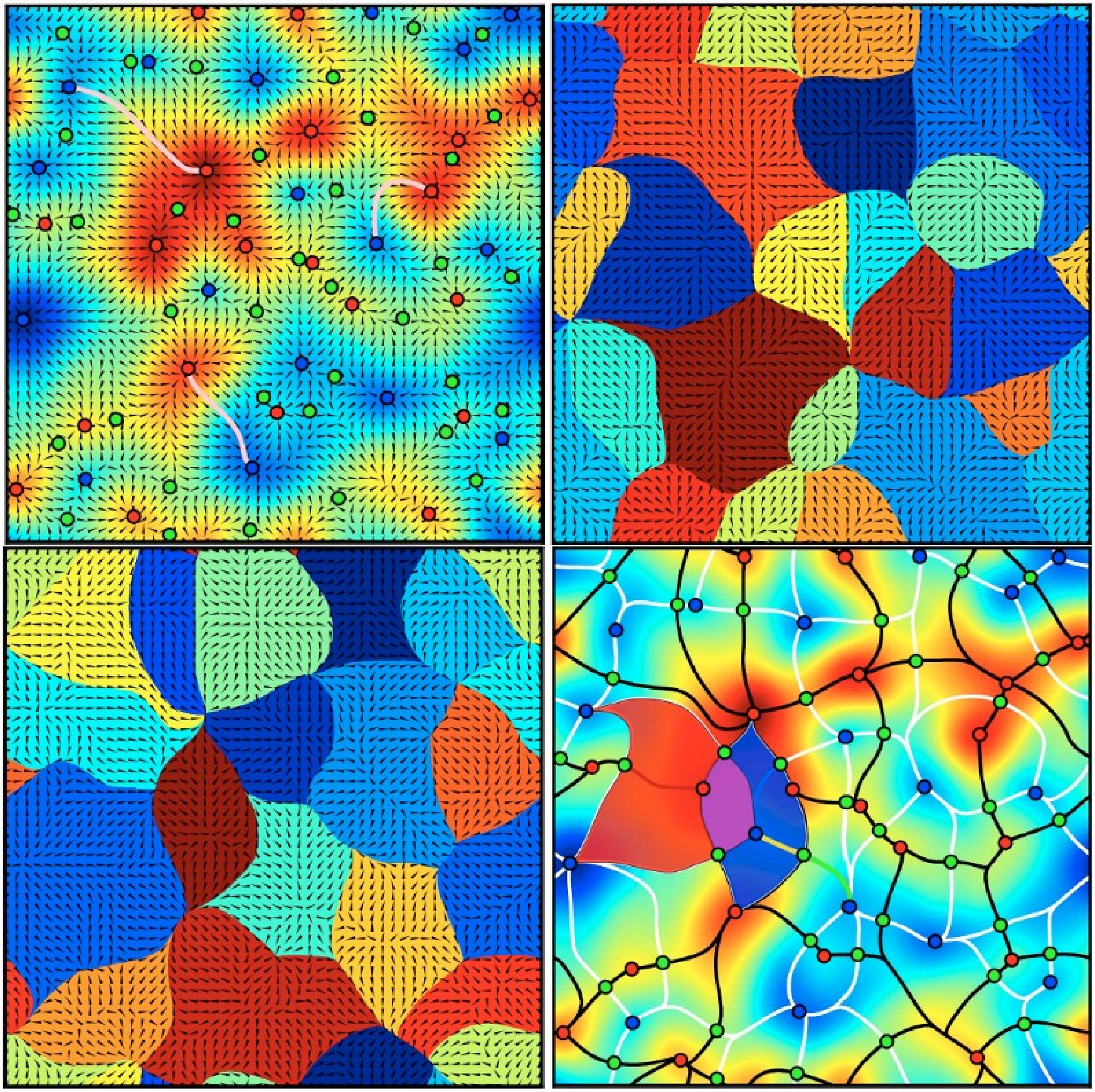


Figure 1.1: Upper left: critical points (minima, saddle points and maxima pictured as blue, green and red disks), and three integral lines (pink curves) of a Morse function. Black arrows show the gradient of that function. Upper right: ascending 2-manifolds : the set of points belonging to integral lines whose destination is the same minimum (critical point of index 0). Lower left: descending 2-manifolds : the set of points belonging to integral lines whose origin is the same maximum (critical point of index 2). Lower right: the Morse-Smale complex : a natural tesselation of space into cells induced by the gradient fo the function. Each cell is the set of points belonging to integral lines whose origin and destination are identical (i.e. each cell is the intersection of an ascending and a descending manifold). The purple region is a 2-cell: intersection of an ascending and a descending 2-manifold (red and blue regions) where all field lines have the same orgin and destination (a minimum and a maxium). The yellow curve is a 1-cell (also called an arc): the intersection of and ascending 2-manifold (blue region) and a descending 1-manifolds (green+yellow curves, originating from the same saddle point).

We identify the maxima and index 2 saddle points of $h_P$ which lie outside the level set. The stable manifolds of these critical points help detect the tunnels and the pockets of $\Sigma$. Additionally, these stable manifolds are used to compute geometric attributes of the detected topological features to which they correspond. In this way, we obtain a description of the isosurface, and its complement, in terms of its topological features. These features are quantified by their geometric properties and may be selectively removed.

The function $h_P$ induces a flow at every point $x \in \mathbb{R}^3$ and this flow has been characterized earlier [96, 101]. See also [69].

The critical points of $h_P$ are those points where $h_P$ has no non-zero gradient along any direction. These are the points in $\mathbb{R}^3$ which lie within the convex hull of its closest points from $P$. It was shown by Siersma [194] that the critical points of $h_P$ are the intersection points of the Voronoi objects with their dual Delaunay objects.

- *Maxima* are the Voronoi vertices contained in their dual tetrahedra,
- *Index 2 saddles* lie at the intersection of Voronoi edges with their dual Delaunay triangles,
- *Index 1 saddles* lie at the intersection of Voronoi facets with their dual Delaunay edges, and
- *Minima* are the sample points themselves as they are always contained in their Voronoi cells.

In this discrete setting, the index of a critical point is the dimension of the lowest dimensional Delaunay simplex that contains the critical point.

At every $x \in \mathbb{R}^3$, a unit vector can be assigned that is oriented in the direction of the steepest ascent of $h_P$. The critical points are assigned zero vectors. This vector field, which may not be continuous, nevertheless induces a flow in $\mathbb{R}^3$. This flow tells how a point $x$ moves in $\mathbb{R}^3$ along the steepest ascent of $h_P$ and the corresponding path is called the *orbit* of $x$.

For a critical point $c$ its stable manifold is the set of points whose orbits end at $c$. The stable manifold of a maximum is a three dimensional polytope whose boundary is composed of the stable manifolds of the index 2 saddle points which in turn are bounded by the stable manifolds of index 1 saddle points and minima. See [66, 96] for the detailed discussion on the structure and computation of the stable manifolds of the critical points of $h_P$.

### 1.3.5 Contouring Tree Representation

## 1.4 Complementary space topology and geometry

**Compactifying Complementary Space**

A molecular surface $S$ bounds a finte portion of $R^3$, viz. the interior volume $V$ of the molecule. The complementary space, defined to be $\mathbb{R}^3 - V$, contains useful geometric and topological information about the surface such as the number of connected components and number of tunnels passing through the surface. Since $\mathbb{R}^3 - V$ is unbounded, we compactify complementary space to get a handle on these features by using some results from Morse theory.

We construct an approximation of the Morse complex for $h_\Sigma$ described in Section 1.3.4 based on the critical points known for $h_P$. First we describe the classification of the requisite critical points and then describe how they are clustered together. The critical points of $h_P$ are detected by checking the intersection of the Voronoi and its dual Delaunay diagram of the point set $P$ sampled from $\Sigma$. The critical points are primarily of three types depending on if the Voronoi/Delaunay object involved lies interior to $\Sigma$, exterior to $\Sigma$, or if the Voronoi object crosses $\Sigma$. There are some exceptions: maxima can not lie on the surface and therefore come in only two types - interior and exterior. The minima are sample points themselves and therefore they are always on $\Sigma$. The saddle points can be any of three types mentioned above.

Since the Morse complex we construct requires only the maxima and index 2 saddles exterior to or on the surface $\Sigma$, we fix three classes of critical points:

$$\begin{aligned} C_{2,E} &= \{\text{index 2 saddles of } h_P \text{ exterior to } \Sigma\} \\ C_{2,S} &= \{\text{index 2 saddles of } h_P \text{ on the surface } \Sigma\} \\ C_{3,E} &= \{\text{maxima of } h_P \text{ exterior of to } \Sigma\} \end{aligned}$$

We include a point at infinity denoted $(m_\infty)$ in the set $C_{3,E}$ to compactify the copmlementary space structure.

As discussed in previous section, the points in the above classes come with a natural hierarchical structure based on stable manifolds. We construct a graph on the points based on this structure by the following rule: a maxima $m \in C_{3,E}$ is connected to a saddle $s \in C_{2,E} \cup C_{2,S}$ if and only if the stable manifold of $s$ lies on the boundary of the stable manifold of $m$. We use this graph to detect tunnels and pockets. The algorithm is depicted visually in Figure 1.2.

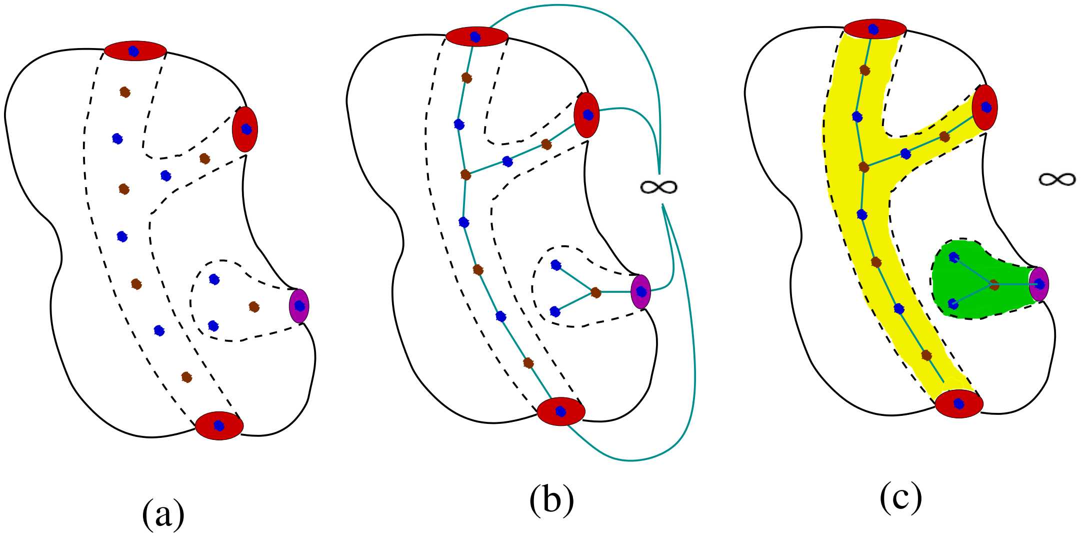


Figure 1.2: A visual depiction of our tunnel and pocket detection algorithm. An imaginary molecular surface is shown with a 3-mouth tunnel and a single pocket. (a) Critical points of $h_P$ are detected. Blue points are index 2 saddles and brown points are maxima. (b) A point at infinity is added and a graph constructed based on adjacency of stable manifolds. This graph approximates the Morse complex. (c) Breaking the edges to the point at infinity, we detect the tunnel (yellow with red mouths) and pocket (green with purple mouth).

### 1.4.1 Detection of Tunnels and Pockets

We first show that the graph constructed on the critical points of $h_P$ has $B_0 + 1$ components where $B_0$ is the 0-th Betti number. Any critical point inside a tunnel or pocket of the surface will have a path along the graph to $m_\infty$, the maximum at infinity. This reflects the fact that there is, by definition, a "way out" from a tunnel or pocket. A critical point in a void, on the other hand, will not have a path to $m_\infty$ and thus lies in a component separate from the tunnels and pockets. Since $B_0$ equals the number of voids captured by the surface, the graph has exactly $B_0 + 1$ components. Hence, the voids of $\Sigma$ are precisely the stable manifolds of the components not containing $m_\infty$.

With the component of the graph containing $m_\infty$, we cluster it into tunnels and pockets as follows. Observe that the point $m_\infty$ is connected only to index 2 saddles that lie on the mouth of a tunnel or pocket. Therefore, by "chopping around" the point $m_\infty$, we break apart the graph based on tunnels and pockets. More precisely, let $C_{2,\infty} \subset C_{2,E} \cup C_{2,S}$ denote the set of points connected to $m_\infty$. Removing all the edges from the point $m_\infty$, we are left with $n$ components of the graph, one corresponding to each tunnel or pocket of $\Sigma$. The stable manifolds of the points of $C_{2,\infty}$ form the *mouths* of the tunnels and pockets and we can now classify all components of our modified graph as follows.

- 0 **Mouth** indicates that the component belongs to a *void*.
- 1 **Mouth** indicates that the component belongs to a *pocket*.
- $k \geq 2$ **Mouths** indicate the component belongs to a *tunnel*. We call it a $k$-mouthed tunnel.

We use the algorithms described in [96] for computation of the stable manifolds of index 2 saddles. In order to have a computational description of the detected features, we also compute the stable manifolds of the maxima falling into every component using the algorithm described in [66]. This produces a tetrahedral decomposition of the features captured.

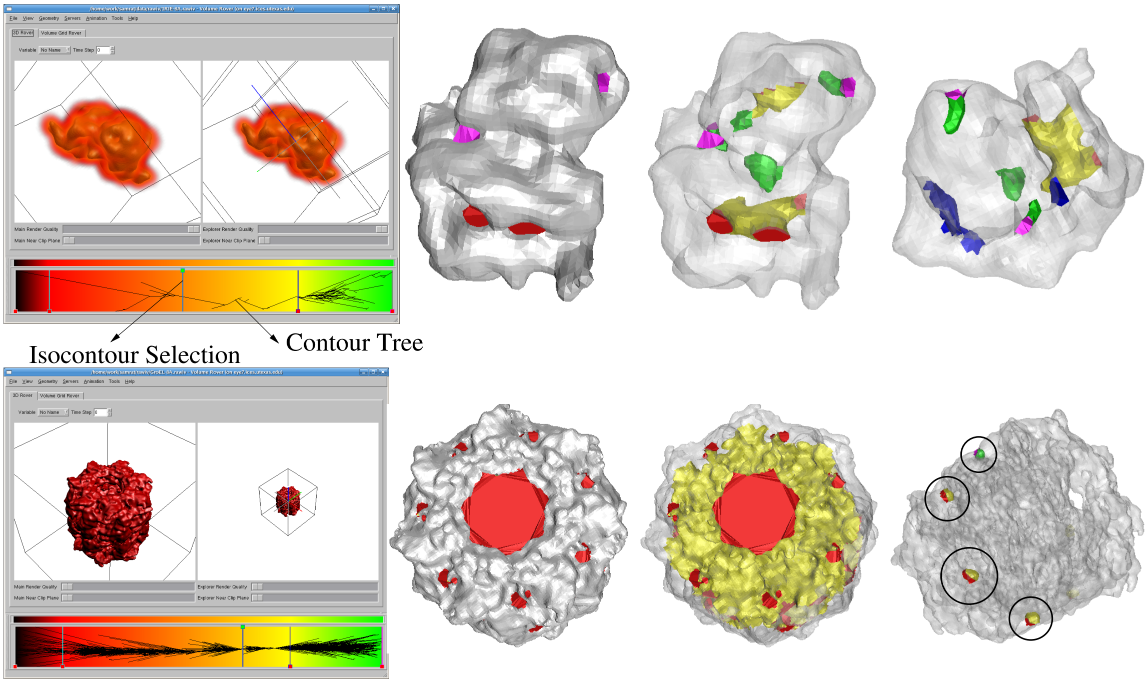


Figure 1.3: Results: Top row shows the interface selection for Rieske Iron-sulphur Protein molecule (PDB ID: 1RIE) from a blurred density map. Bottom row shows the isosurface selection for the virus particle (GroEL) from cryo-EM density map.

We compute the pockets, tunnels and voids of the molecular surface. The tetrahedral solids describing the pockets and tunnels provide a nice handle to those features and using these handles, the features can be ranked. We primarily use the geometric attributes of the features in order to rank them. Such attributes include, but are not limited to, the combined volume of the tetrahedra and the area of the mouths. The pockets and tunnels are then sorted in order of their increasing geometrically measured importance.

Removal of insignificant features are also made easy because of the volumetric description of the features. As dictated by the applications, a cut-off is set below which the features are considered *noise*. We remove the *topological noise* by marking the outside tetrahedra as inside and updating the surface triangles.

We show the results of our algorithm on two volumetric data. The top row in Figure 1.3 shows the electron density volume of Rieske Iron-Sulphur Protein (Protein Data Bank Id: 1RIE). The volume rendering using VolRover [61] is shown in the leftmost subfigure. The tool additionally supports the visualization and isosurface selection using CT. The other subfigures show the selected interface and the detected tunnels and pockets. Note, the mouth of the tunnel is drawn in red and the mouth of the pocket is drawn in purple. The rest of the tunnel surface is drawn in yellow while the pocket surface is drawn in green. The blue patches in the rightmost subfigure shows the filling of the smaller tunnels and pockets. The second row shows the results on the three dimensional scalar volume representing the electron density of the reconstructed image of a virus (GroEL) from a set of two dimensional electron micrographs. The resolution is $8\mathbf{r}A$.

Using VolRover, a suitable level set is chosen. Note the CT is very noisy and has many branches, because of which it is not possible to extract a single-component isosurface. Nevertheless only one component is vital and the rest of them are merely artifacts caused by noise. The main component along with the detected tunnel is shown next. The result is particularly useful in visualizing the symmetric structure of the virus particle as depicted in the symmetric set of mouths. In addition to detecting the principal topological feature, the algorithm detects few small tunnels and pockets which are shown separately for visual clarity (rightmost subfigure) and these are removed subsequently as part of the topological noise removal process. (Suggestion: Remove this paragraph)

However, so far we have discussed about curating the molecular surface by modifying only the complementary space topological features. This does not always serve the purpose. Consider a very thin interior surrounding a wide tunnel. The tunnel is significant but the thin portion surrounding it should disappear which we have not taken into account so far. Figure 1.4 shows a similar scenario where the inherent symmetry of the 3D map of nodavirus is shown in subfigure (a). Subfigure (b) shows that

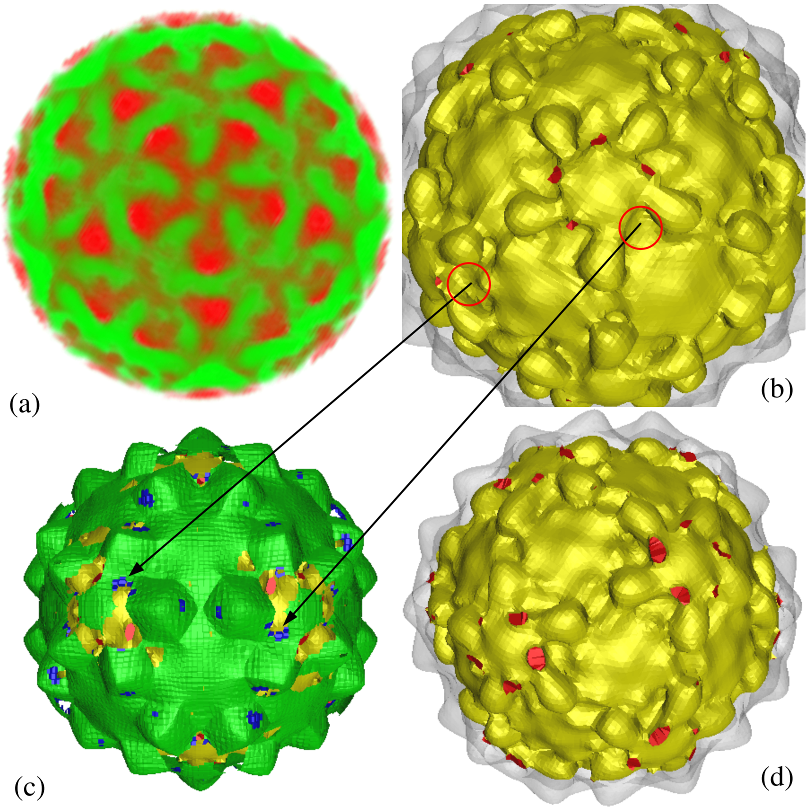


Figure 1.4: Identification of “thin” regions in the primal space. (a) The volume rendering of 3D image of nodavirus. (b) Tunnels are detected for the initial selection of the isosurface. Note, in some places of 5-fold symmetry, only 4 mouths of the tunnel are present. (c) The thin regions (blue) are identified as subsets of the unstable manifolds of the index 1 saddles identified on the interior medial axis. The arrow between (b) and (c) indicate that the places where the 5th mouth of the tunnel should open indeed have thin regions. (d) The final selected isosurface has complementary space topology consistent with the inherent symmetry of the 3D map.

due to wrong choice of isovalue only 4 mouths are open in the complementary space tunnel of a 5-fold symmetric region. To curate this surface, modification of the complementary space is not sufficient. To deal with such cases, we extend the curation process by detecting "thin" regions in the primal space.
Remarkably, distance function plays a very important role here also. In order to detect the thin regions, we first compute the interior medial axis by publicly available software [57]. Further we compute the index 1 ($U_1$)and index 2 ($U_2$) saddle points which lie on the interior medial axis and compute their unstable manifols using the algorithm described in [101]. $U_2$ produces linear subset of the medial axis and $U_1$ produces planar subset of the medial axis. We then sample the distance function on $U_1$ and $U_2$ and identify the subset corresponding to the "thin" regions measured by a suitable threshold parameter. Figure 1.4 (c) shows the thin regions (blue patches) on the $U_1$ (green) identified for noda virus. The reason for not using medial axis directly is that medial axis tends to be noisy and $U_1$ and $U_2$ usually describes the subset of the medial axis stable against the small undulation on the surface. Then we collect the interior maxima falling into the thin subset of $U_1$ and $U_2$ and compute their stable manifolds. Again, the stable manifolds are the collection of tetrahedra and therefore we cut the surface open by forming a channel interior to the volume bounded by the molecular surface by appropriately toggling the marking of the tetrahedra from inside to outside. Final selection of the isosurface for which the mouths of the tunnels respect the inherent symmetry of the density map of the virus is shown in Figure 1.4 (d).

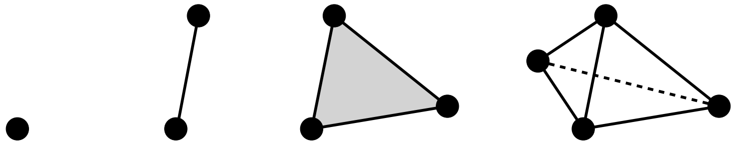
## 1.5 Primal and Dual Complexes

### 1.5.1 Primal Meshes

In algebraic topology, manifolds are discretized using simplicial complexes, a notion which guides the entire theory of discrete exterior calculus. We state the definition of simplicial complex here, along with supporting definitions to be used throughout.

**Definition 1.2.** A $k$**-simplex** $\sigma^k$ is the convex hull of $k+1$ geometrically independent points $v_0, \ldots, v_k \in \mathbb{R}^N$. Any simplex spanned by a (proper) subset of $\{v_0, \ldots, v_k\}$ is called a **(proper) face** of $\sigma^k$. The union of the proper faces of $\sigma^k$ is called its **boundary** and denoted Bd($\sigma^k$). The **interior** of $\sigma^k$ is Int($\sigma^k$) = $\sigma^k \backslash$Bd($\sigma^k$). Note that Int($\sigma^0$)=$\sigma^0$. The **volume** of $\sigma^k$ is denoted $|\sigma^k|$. Define $|\sigma^0 \models 1$. $\Diamond$

**Primal Simplicies**



Primal simplices of dimension 0, 1, 2, and 3 are shown. In general, a $k$-simplex is the convex hull of $k$ points in $\mathbb{R}^n$ in general position. We denote a $k$-simplex as $\sigma^k$.

We will indicate that a simplex has dimension $k$ with a superscript, e.g. $\sigma^k$, and will index simplices of any dimension with subscripts, e.g. $\sigma_i$.

**Definition 1.3.** A **simplicial complex** $K$ in $\mathbb{R}^N$ is a collection of simplices in $\mathbb{R}^N$ such that

1. Every face of a simplex of $K$ is in $K$.

2. The intersection of any two simplices of $K$ is either a face of each of them or it is empty.

The union of all simplices of $K$ treated as a subset of $\mathbb{R}^N$ is called the underlying space of $K$ and is denoted by $|K|$. $\Diamond$

**Definition 1.4.** A simplicial complex of dimension $n$ is called a **manifold-like simplicial complex** if and only if $|K|$ is a $C^0$-manifold, with or without boundary. More precisely,

1. All simplices of dimension $k$ with $0 \leq k \leq n-1$ must be a face of some simplex of dimension $n$ in $K$.

2. Each point on $|K|$ has a neighborhood homeomorphic to $\mathbb{R}^n$ or $n$-dimensional half-space. $\Diamond$

*Remark* 1.5. Since DEC is meant to treat discretizations of manifolds, we will assume all simplicial complexes are manifold-like from here forward. We note that $|K|$ is thought of as a piecewise linear approximation of a smooth manifold $\Omega$. Formally, this is taken to mean that there exists a homeomorphism $h$ between $|K|$ and $\Omega$ such that $h$ is isotopic to the identity. In applications, however, knowing $h$ or $\Omega$ explicitly may be irrelevant or impossible as $K$ often encodes everything known about $\Omega$. This emphasizes the usefulness of DEC as a theory built for discrete settings. $\Diamond$

**Orientation of Simplicial Complexes** We now review how to orient a simplicial complex $K$.

**Definition 1.6.** Define two orderings of the vertices of a simplex $\sigma^k$ ($k \geq 1$) to be equivalent if they differ by an even permutation. Thus, there are two equivalence classes of orderings, each of which is called an **orientation** of $\sigma^k$. If $\sigma^k$ is written as $[v_0, \ldots, v_k]$, the orientation of $\sigma^k$ is understood to be the equivalence class of this ordering. $\Diamond$

**Definition 1.7.** Let $\sigma^k = [v_0, \ldots, v_k]$ be an oriented simplex with $k \geq 2$. This gives an **induced orientation** on each of the $(k-1)$-dimensional faces of $\sigma^k$ as follows. Each face of $\sigma^k$ can be written uniquely as $[v_0, \ldots, \hat{v}_i, \ldots, v_k]$, where $\hat{v}_i$ means $v_i$ is omitted. If $i$ is even, the induced orientation on the face is the same as the oriented simplex $[v_0, \ldots, \hat{v}_i, \ldots, v_k]$. If $i$ is odd, it is the opposite. $\Diamond$

We note that this formal definition of induced orientation agrees with the notion of orientation induced by the boundary operator (Definition 4.12). In that setting, a 0-simplex can also receive an induced orientation.

*Remark* 1.8. We will need to be able to compare the orientation of two oriented $k$-simplices $\sigma^k$ and $\tau^k$. This is possible only if at least one of the following conditions holds:

1. There exists a $k$-dimensional affine subspace $P \subset \mathbb{R}^N$ containing both $\sigma^k$ and $\tau^k$.

2. $\sigma^k$ and $\tau^k$ share a face of dimension $k-1$.

In the first case, write $\sigma^k = [v_0, \ldots, v_k]$ and $\tau^k = [w_0, \ldots, w_k]$. Note that $\{v_1 - v_0, v_2 - v_0, \ldots, v_k - v_0\}$ and $\{w_1 - w_0, w_2 - w_0, \ldots, w_k - w_0\}$ are two ordered bases of $P$. We say $\sigma^k$ and $\tau^k$ have the same orientation if these bases orient $P$ the same way. Otherwise, we say they have opposite orientations. In the second case, $\sigma^k$ and $\tau^k$ are said to have the same orientation if the induced orientation on the shared $k-1$ face induced by $\sigma^k$ is *opposite* to that induced by $\tau^k$. $\Diamond$

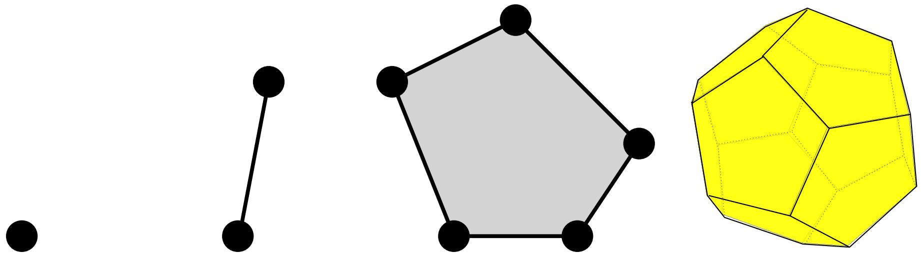
**Definition 1.9.** Let $\sigma^k$ and $\tau^k$ with $1 \leq k \leq n$ be two simplices whose orientations can be compared, as explained in Remark 1.8. If they have the same orientation, we say the simplices have a **relative orientation** of $+1$, otherwise $-1$. This is denoted as $\text{sgn}(\sigma^k, \tau^k) = +1$ or $-1$, respectively. $\Diamond$

**Definition 1.10.** A manifold-like simplicial complex $K$ of dimension $n$ is called an **oriented manifold-like simplicial complex** if adjacent $n$-simplices agree on the orientation of their shared face. Such a complex will be called a **primal mesh** from here forward. $\Diamond$

### 1.5.2 Dual Complexes

Dual complexes are defined relative to a primal mesh. While they represent the same subset of $\mathbb{R}^N$ as their associated primal mesh, they create a different data structure for the geometrical information and become essential in defining the various operators needed for DEC.

**Dual Cells**



Dual cells of dimension 0, 1, 2, and 3 are shown. In general, a $k$-cell is the convex hull of $k$ points in $\mathbb{R}^n$ homeomorphic to a filled $k$-ball such that the boundary of the $k$-cell is a collection of $k-1$ cells. If a $k$-cell is defined as the dual of an $n-k$ simplex, we denote it as $\star\sigma^{n-k}$.

**Definition 1.11.** The **circumcenter** of a $k$-simplex $\sigma^k$ is given by the center of the unique $k$-sphere that has all $k+1$ vertices of $\sigma^k$ on its surface. It is denoted $c(\sigma^k)$. A simplex $\sigma^k$ is said to be **well-centered** if $c(\sigma^k) \in \text{Int}(\sigma^k)$. A **well-centered simplicial complex** is one in which all simplices (of all dimensions) in the complex are well-centered. ♢

**Definition 1.12.** Let $K$ be a well-centered primal mesh of dimension $n$ and let $\sigma^k$ be a simplex in $K$. The **circumcentric dual cell** of $\sigma^k$, denoted $D(\sigma^k)$, is given by

$$D(\sigma^k) := \bigcup_{r=0}^{n-k} \bigcup_{\sigma^k \prec \sigma_1 \prec \cdots \prec \sigma_r} \text{Int}(c(\sigma^k)c(\sigma_1)\ldots c(\sigma_r)).$$

To clarify, the inner union is taken over all sequences of $r$ simplices such that $\sigma^k$ is the first element in the sequence and each sequence element is a proper face of its successor. Hence, $\sigma_1$ is a $(k+1)$ simplex and $\sigma_r$ is an $n$ simplex. For $r = 0$, this is to be interpreted as the sequence $\sigma^k$ only. The closure of the dual cell of $\sigma^k$ is denoted $\bar{D}(\sigma^k)$ and called the **closed dual cell**. We will use the notation $\star$ to indicate dual cells, i.e.

$$\star\sigma := \bar{D}(\sigma).$$

Each $(n-k)$-simplex on the points $c(\sigma^k), c(\sigma_1), \ldots, c(\sigma_r)$ is called an **elementary dual simplex** of $\sigma^k$. The collection of dual cells is called the **dual cell decomposition** of $K$ and denoted $D(K)$ or $\star K$. ♢

Note that the dual cell decomposition forms a CW complex.

**Orientation of Dual Complexes** Orientation of the dual complex must be done in a such a way that it "agrees" with the orientation of the primal mesh. This can be done canonically since a primal simplex and any of its elementary dual simplices have complementary dimension and live in orthogonal affine subspaces of $\mathbb{R}^N$. We make this more precise and fix the necessary conventions with the following definitions.

**Definition 1.13.** Let $K$ be a primal mesh containing a sequence of simplices $\sigma^0 \prec \sigma^1 \prec \cdots \prec \sigma^n$ and let $\sigma^k$ be one of these simplices with $1 \leq k \leq n-1$. The **orientation** of the elementary dual simplex with vertices $c(\sigma^k), \ldots, c(\sigma^n)$ is $s[c(\sigma^k), \ldots, c(\sigma^n)]$ where $s \in \{-1, +1\}$ is given by the formula

$$s := \text{sgn}\left([c(\sigma^0), \ldots, c(\sigma^k)], \sigma^k\right) \times \text{sgn}\left([c(\sigma^0), \ldots, c(\sigma^n)], \sigma^n\right).$$

The sgn function was defined in Definition 1.9.
For $k = n$, the dual element is a vertex which has no orientation. For $k = 0$, define $s := \text{sgn}\left([c(\sigma^0), \ldots, c(\sigma^n)], \sigma^n\right)$. ♢

The above definition serves to orient all the elementary dual simplices associated to $\sigma^k$ and hence all simplices in a dual cell decomposition. Further, the orientations on the elementary dual simplices induce orientations on the boundaries of dual cells in the same manner as given in Definition 1.7. The induced orientations on adjacent $(n-1)$ cells will agree since the dual cell decomposition comes from a primal mesh (see Definition 1.10).

**Definition 1.14.** The oriented dual cell decomposition of a primal mesh is called the **dual mesh**. ♢

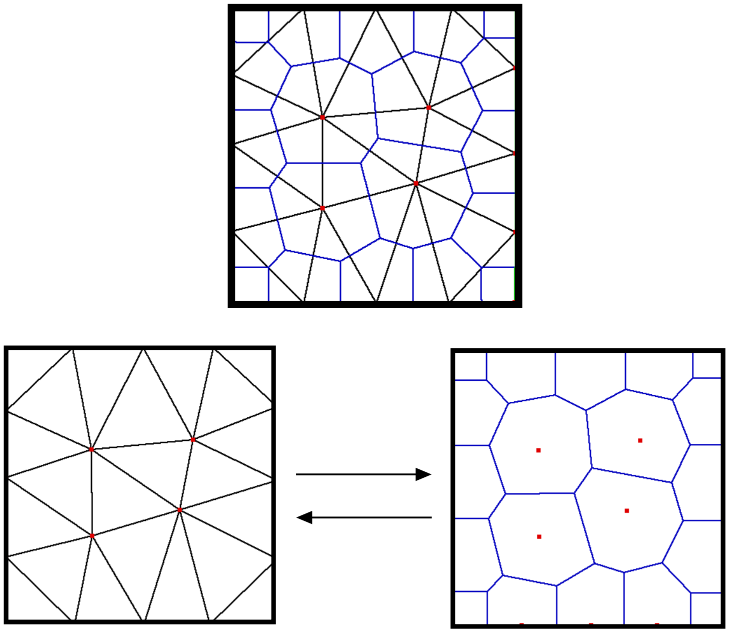
## 1.6 Voronoi and Delaunay Decompositions

For a finite set of points $P$ in $\mathbb{R}^3$, the Voronoi cell of $p \in P$ is

$$V_p = \{x \in \mathbb{R}^3 \,:\, \forall q \in P - \{p\},\ \|x-p\| \leq \|x-q\|)\}.$$

If the points are in general position, two Voronoi cells with non-empty intersection meet along a planar, convex Voronoi facet, three Voronoi cells with non-empty intersection meet along a common Voronoi edge and four Voronoi cells with non-empty intersection meet at a Voronoi vertex. A cell decomposition consisting of the *Voronoi objects*, that is, Voronoi cells, facets, edges and vertices is the **Voronoi diagram** Vor $P$ of the point set $P$.

**Voronoi and Delaunay Meshes**



Voronoi and Delaunay meshes are dual decompositions of the same domain. In the figure, the small red dots define the Voronoi cells and hence a dual mesh of the domain (shown at right) but also define the vertices of the Delaunay triangles and hence a primal mesh of the domain (shown at left).

The dual of Vor $P$ is the **Delaunay diagram** Del $P$ of $P$ which is a simplicial complex when the points are in general position. The tetrahedra are dual to the Voronoi vertices, the triangles are dual to the Voronoi edges, the edges are dual to the Voronoi facets and the vertices (sample points from $P$) are dual to the Voronoi cells. We also refer to the Delaunay simplices as *Delaunay objects*.

### 1.6.1 Euclidean vs. Power distance.

For MVC the choice of using the Power distance in place of the Euclidean distance is motivated by the the efficiency and simplicity of the construction of the power diagram together with the fact that the power distance can be proven to be an upper bound of the Euclidean distance.

Consider a point $p$ at distance $d$ from the center $c$ a ball $B$ of radius $r$ as in Figure 1.5. We define:

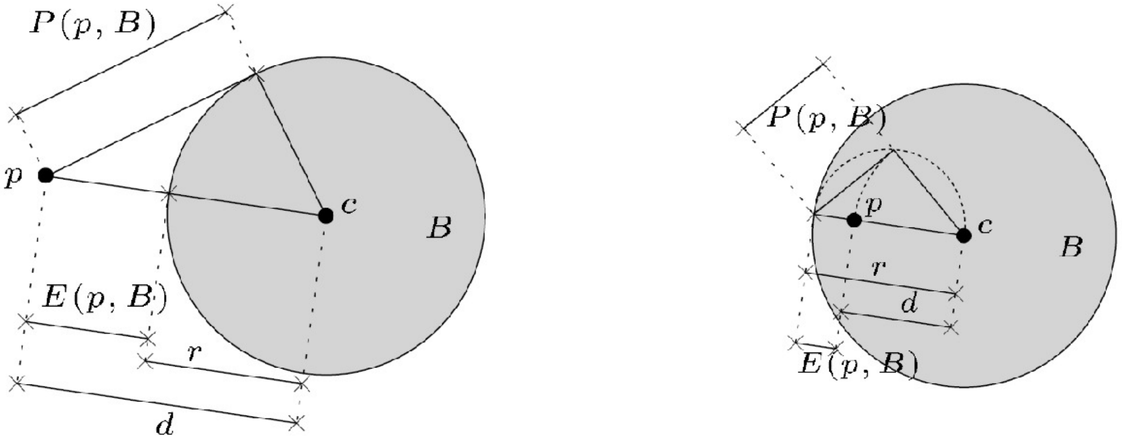
$$E(p,B) = |d-r| \ , \quad P(p,B) = \sqrt{|d^2 - r^2|} \ .$$

Then we have the following chain of inequalities (where $r$ and $d$ are positive numbers):

$$0 \leq 4dr(d-r)^2 = 4d^3r - 8d^2r^2 + 4dr^3$$

$$\begin{aligned} (d-r)^4 &= d^4 - 4d^3r + 6d^2r^2 - 4dr^3 + r^4 \\ &\leq d^4 - 2d^2r^2 + r^4 = (d^2-r^2)^2 \end{aligned}$$

$$E(p,B) = |d-r| \leq \sqrt{|d^2-r^2|} = P(p,B) \ .$$



(a)

(b)

Figure 1.5: Relationship between the Euclidean distance $E(p, B)$ between the point $p$ and the ball $B$ and their Power distance $P(p, B)$, (a) Configuration for $d > r$. (b) Configuration for $d < r$.

If the point $p$ is outside the ball $B$ the following inequality holds (both $r$ and $d$ are positive numbers):

$$\begin{aligned}
r &< d \\
r - d &\leq 0 \\
2r^2 - 2rd &\leq 0 \\
2r^2 - 2rd + d^2 &\leq d^2 \\
r^2 - 2rd + d^2 &\leq d^2 - r^2 \\
(r - d)^2 &\leq d^2 - r^2 \\
(r - d) &\leq \sqrt{d^2 - r^2} \\
E(p, B) &\leq P(p, B)
\end{aligned}$$

That is the $P(p, B)$ is larger than the Euclidean distance $E(p, B)$ . The same relation holds if $p$ is inside $B$:

$$\begin{aligned}
d &\leq r \\
d - r &\leq 0 \\
2d^2 - 2rd &\leq 0 \\
2d^2 - 2rd + r^2 &\leq r^2 \\
d^2 - 2rd + r^2 &\leq r^2 - d^2 \\
(d - r)^2 &\leq r^2 - d^2 \\
(d - r) &\leq \sqrt{r^2 - d^2} \\
E(p, B) &\leq P(p, B)
\end{aligned}$$

In conclusion we have that for any given ball $B$ and point $p$, the function $P(p, B)$ provides an upper bound on the distance $E(p, b)$:

$$E(p, B) \leq P(p, B) \ , \tag{1.1}$$

with equality holding only when $d = r$, i. e. the point is on the surface of the ball (and in trivial cases where $d$ or $r$ is zero). For a collection of $n$ balls $\mathcal{B} = \{B_1, \ldots, B_n\}$ the distance functions are extended as follows:

$$E(p, \mathcal{B}) = \min_{1 \leq i \leq n} |d_i - r_i| \tag{1.2}$$

$$P(p, \mathcal{B}) = \sqrt{\min_{1 \le i \le n} |d_i^2 - r_i^2|} \tag{1.3}$$

The problem in comparing $E(p, \mathcal{B})$ with $P(p, \mathcal{B})$ is that they may achieve their minimum for different values of $i$ because in general the Power diagram is not coincident with the Voronoi diagram. Figure 1.6.1 shows an example of comparison between the Voronoi diagram of two circles (in red) with the corresponding Power diagram (in blue). In this example the minimum distance of the point $p$ from the set $\mathcal{B} = \{B_1, B_2\}$ is achieved at $i = 1$ for $P(p, \mathcal{B})$ and at $i = 2$ for $E(p, \mathcal{B})$:

$$P(p, \mathcal{B}) = P(p, B_1)$$

$$E(p, \mathcal{B}) = E(p, B_2) \ .$$

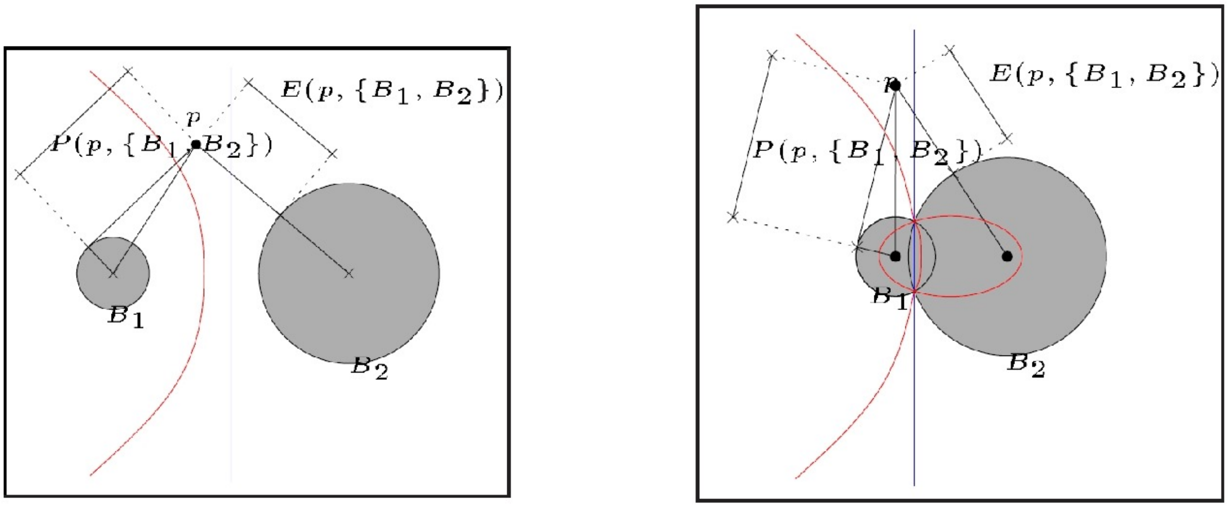
In general for a given point $p$ we call $i_P, i_E$ the two indices such that:

$$P(p, \mathcal{B}) = P(p, B_{i_P})$$

$$E(p, \mathcal{B}) = E(p, B_{i_E}) \ .$$

From equations (1.2) and (1.1) we have that:

$$\begin{aligned} E(p, \mathcal{B}) &= E(p, B_{i_E}) \le E(p, B_{i_P}) \\ &\le P(p, B_{i_P}) = P(p, \mathcal{B}) \ . \end{aligned}$$



(a)

(b)

Figure 1.6: Power diagram (in blue) and Voronoi diagram (in red) of two circles. (a) Case of nonintersecting circles. (b) Case of intersecting circles.

### 1.6.2 Weighted Alpha Shapes

A simplex $s$ in the regular triangulation of $\{P_i\}$ belongs to the $\alpha$-shape of $\{P_i\}$ only if the orthogonal center of (the weighted point orthogonal to the vertices of) $s$ is smaller than $\alpha$. The alpha shape where $\alpha = 0$, called the zero-shape, is the topological structure of molecules. For example, an edge $e = (u, v)$ is a part of the zero-shape only if $\|u - v\|^2 - w_u - w_v \ < 0$, which means that the two balls centered at $u$ and $v$ intersect (Figure 1.7(d)).

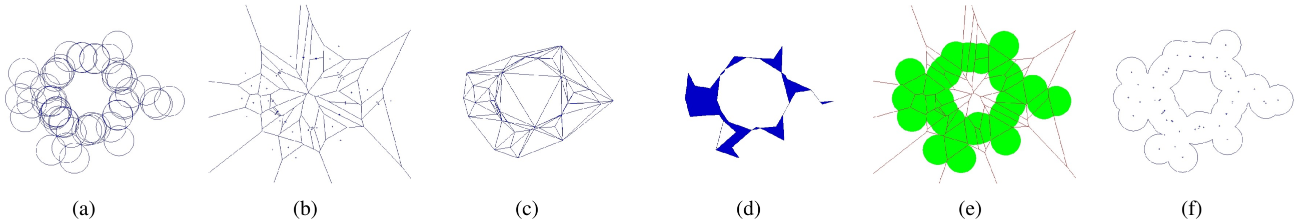
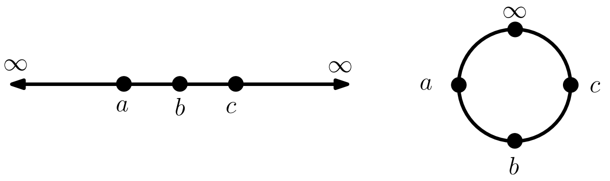


Figure 1.7: The combinatorial and geometric structures underlying a molecular shape: (a) The collection of balls (weighted points). (b) Power diagram of a set of the points. (c) Regular triangulation. (d) The $\alpha$-shape (with $\alpha = 0$) of the points. (e) Partitioning of the molecular body induced by the power diagram. (f) The boundary of the molecular body.

## 1.7 Biological Applications

### 1.7.1 Union of Balls Topology

**Stereographic Projection**



For any integer $n \geq 1$, the space $\mathbb{R}^n \cup \{\infty\}$ can be mapped to the $n$-dimensional sphere, commonly denoted $S^n$ by a mapping called **stereographic projection**. In the 1D case shown above, the real line is mapped to the circle $S^1$ by wrapping the points at infinity together to the top of the circle. The general form of the mapping is given by

$$s : S^n \to \mathbb{R}^n, \quad (x_1, \ldots, x_{n+1}) \to \frac{1}{1 - x_{n+1}}(x_1, \ldots, x_n)$$

with the convention $(0, \ldots, 1) \to \{\infty\}$. This process can be used to wrap 2D power diagrams into 3D polytopes.

**Power Diagram**

Given a weighted point $P = (p, w_p)$ where $p \in \mathrm{I\!R}^n$ and $w \in \mathrm{I\!R}$, the *power distance* from a point $x \in \mathrm{I\!R}^n$ to $P$ is defined as

$$\pi_P(x) = \sqrt{\|p - x\|^2 - w_p} \ ,$$

where $\|p - x\|$ is the ordinary Euclidean distance between $p$ and $x$.
In molecule context, we define the weight of an atom $B$ with center at $p$ and radius $r$ to be $w_B = r^2$. The *power distance* of $x$ to $B$ is

$$\pi_B(x) = \sqrt{\|p - x\|^2 - r^2} \ .$$

Given a set $\{P_i\}$ of weighted vertices (each vertex has a weight $w_i$ associated with it), the Power Diagram is a tiling of the space into convex regions where the $i$th tile is the set of points nearest to the vertex $P_i$, in the power distance metric The power diagram is similar to the Voronoi diagram using the power distance instead of Euclidean distance.
The weighted Voronoi cell of a ball $B$ in a molecule $\mathcal{B}$ is the set of points in space whose weighted distance to $B$ is less than or equal to their weighted distance to any other ball in $\mathcal{B}$:

$$V_B = \{x \in \mathrm{I\!R}^2 | \pi_B(x) \leq \pi_C(x) \ \forall C \in \mathcal{B}\} \ .$$

The *power diagram* of a molecule is the union of the weighted Voronoi cells for each of its atoms (Figure 1.7(b)).

### Regular Triangulation

The *regular triangulation*, or *weighted Delaunay triangulation*, is the dual (face adjacency graph) of the power diagram, just as the Delaunay triangulation is the dual shape of the Voronoi diagram. Vertices in the triangulation are connected if and only if their corresponding weighted Voronoi cells have a common face (Figure 1.7(c)). This implies that two vertices are connected if and only if they have a nearest neighbor relation measured in power distance metric
Given a set of $n$ 2D points with weights, it has been shown , that their regular triangulation can be computed in $O(n \log n)$ time, by incrementally inserting new points to the existing triangulation and correcting it using edge flips.

### Wrapping the Power Diagram

We can model a union of balls as an **embedded graph**. If two balls intersect, their circle of intersection can be defined by three points. The arcs connecting these points are directed edges and can be parametrized as portions of a circle. Each face is portion of a sphere and the circular arcs defining its boundary. This can be captured compactly as a graph structure (vertices, edges, and faces).
Moreover, the **weighted Delaunay triangulation** of the centers of the balls defines the topology of the volume. An edge in the complex corresponds to two balls intersecting and a face (triangle) corresponds to three balls intersecting. A cycle of three edges without a face corresponds to three balls which intersect pairwise but not mutually.
We can wrap the weighted Delaunay diagram onto the 4-sphere using sterographic projection (see box). This mapping results in a polytope that compactly represents the union of balls topology and surface patch embedded graph.

## 1.7.2 Meshing of Molecular Interfaces

In this subsection, we describe an approach to generate quality triangular/tetrahedral meshes for complicated biomolecular structures directly from the PDB format data, conforming to a good implicit solvation surface approximation. There are three main steps in our mesh generation process:

1. Implicit Solvation Surface Construction – A smooth implicit solvation model is constructed to approximate the Lee-Richards molecular surface by using weighted Gaussian isotropic atomic kernel functions and a two-level clustering techniques.

2. Mesh Generation – A modified dual contouring method is used to extract triangular and interior/exterior tetrahedral meshes, conforming to the implicit solvation surface. The dual contouring method is selected for mesh generation as it tends to yield meshes with better aspect ratio. In order to generate exterior meshes described by biophysical applications , we add a sphere or box outside the implicit solvation surface, and create an outer boundary. Our extracted tetrahedral mesh is spatially adaptive and attempts to preserve molecular surface features while minimizing the number of elements.

3. Quality Improvement – Geometric flows are used to improve the quality of extracted triangular and tetrahedral meshes.

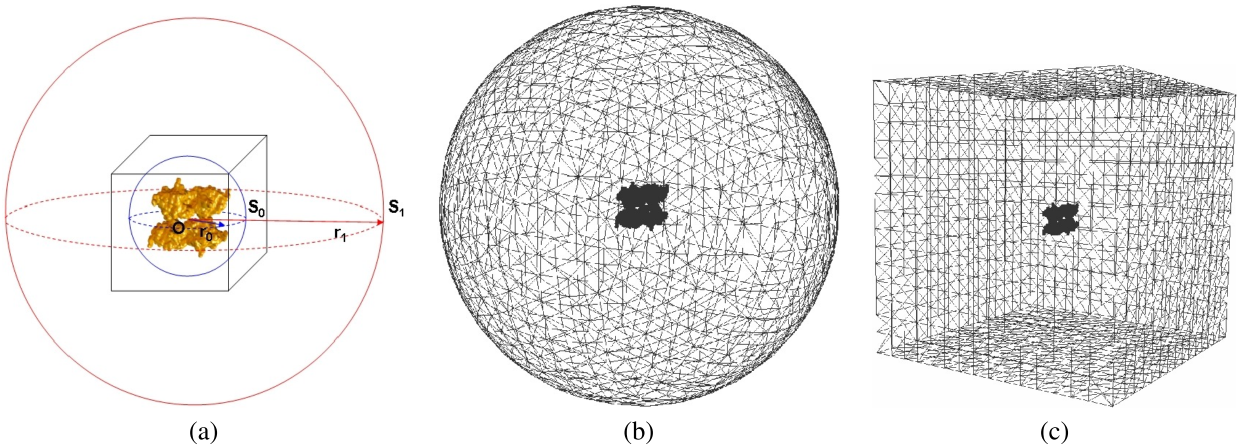
The generated tetrahedral meshes of the monomeric and tetrameric mouse acetylcholinesterase (mAChE) have been successfully used in solving the steady-state Smoluchowski equation using the finite element method .

### Mesh Generation

There are two main methods for contouring scalar fields, primal contouring and dual contouring . Both of them can be extended to tetrahedral mesh generation. The dual contouring method is often the method of choice as it tends to yield meshes with better aspect ratio.

**Triangular Meshing** Dual contouring uses an octree data structure, and analyzes those edges that have endpoints lying on different sides of the isosurface, called *sign change edges*. The mesh adaptivity is determined during a top-down octree construction. Each sign change edge is shared by either four (uniform case) or three (adaptive case) cells, and one minimizer point is calculated for each of them by minimizing a predefined Quadratic Error Function (QEF) :

$$QEF[x] = \sum_i \left[ n_i \cdot (x - p_i) \right]^2 , \tag{1.4}$$



(a) (b) (c)

Figure 1.8: The analysis domain of exterior meshes. (a) - '$O$' is the geometric center of the molecule, suppose the circum-sphere of the biomolecule has the radius of $r$. The box represents the volumetric data, and '$S_0$' is the maximum sphere inside the box, the radius is $r_0 (r_0 > r)$. '$S_1$' is an outer sphere with the radius of $r_1 (r_1 = (20 \sim 40)r)$. (b) - the diffusion domain is the interval volume between the molecular surface and the outer sphere '$S_1$', here we choose $r_1 = 5r$ for visualization. (c) - the outer boundary is a cubic box.

where $p_i$, $n_i$ represent the position and unit normal vectors of the intersubsection point respectively. For each sign change edge, a quad or triangle is constructed by connecting the minimizers. These quads and triangles provide a 'dual' approximation of the isosurface.

A recursive cell subdivision process was used to preserve the trilinear topology of the isosurface. During cell subdivision, the function value at each newly inserted grid point can be exactly calculated since we know the volumetric function. Additionally, we can generate a more accurate triangular mesh by projecting each generated minimizer point onto the isosurface.

**Tetrahedral Meshing** The dual contouring method has already been extended to extract tetrahedral meshes from volumetric scalar fields . The cells containing the isosurface are called boundary cells, and the interior cells are those cells whose eight vertices are inside the isosurface. In the tetrahedral mesh extraction process, all the boundary cells and the interior cells need to be analyzed in the octree data structure. There are two kinds of edges in boundary cells, one is a sign change edge, the other is an interior edge. Interior cells only have interior edges. In [237, 238], interior edges and interior faces in boundary cells are dealt with in a special way, and the volume inside boundary cells is tetrahedralized. For interior cells, we only need to split them into tetrahedra.

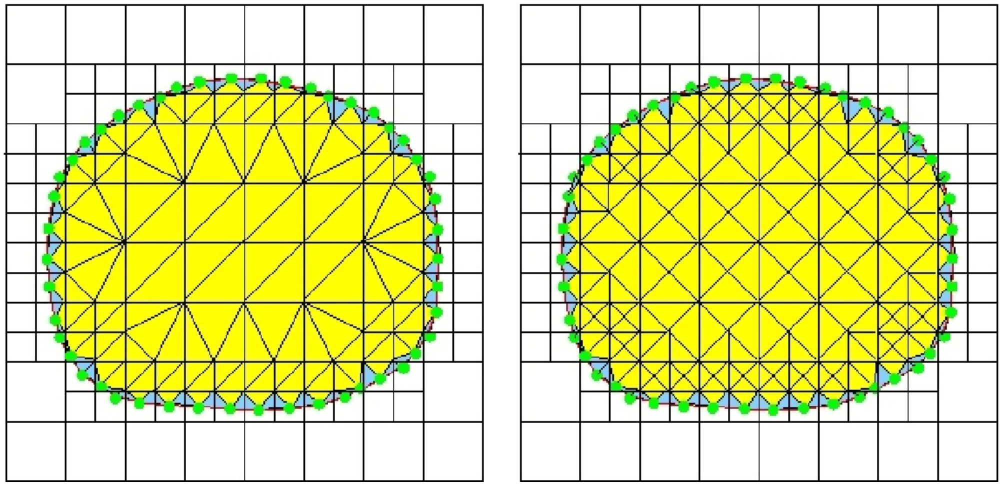
**Adding an Outer Boundary** In biological diffusion systems, we need to analyze the electrostatic potential field which is faraway from the molecular surface . Assume that the radius of the circum-sphere of a biomolecule is $r$. The computational model can be approximated by a field from an outer sphere $S_1$ with the radius of $(20 \sim 40)r$ to the molecular surface. Therefore the exterior mesh is defined as the tetrahedralization of the interval volume between the molecular surface and the outer sphere $S_1$ (Fig. 1.8(b)).

First we add a sphere $S_0$ with the radius of $r_0$ (where $r_0 > r$ and $r_0 = 2^n/2 = 2^{n-1}$) outside the molecular surface, and generate meshes between the molecular surface and the outer sphere $S_0$. Then we extend the tetrahedral meshes from the sphere $S_0$ to the outer bounding sphere $S_1$. For each data point inside the molecular surface, we keep the original function value. While for each data point outside the molecular surface, we reset the function value as the smaller one of $f(x) - \alpha$ and the shortest distance from the grid point to the sphere $S_0$. Eqn. (1.5) shows the newly constructed function $g(x)$ which provides a grid-based volumetric data containing the biomolecular surface and an outer sphere $S_0$.

$$g(x) = \begin{cases} \min(\|x - x_0\| - r_0, f(x) - \alpha), & \text{if} f(x) < \alpha, \|x - x_0\| < r_0, \\ \|x - x_0\| - r_0, & \text{if} f(x) < \alpha, \|x - x_0\| \geq r_0, \\ f(x) - \alpha, & \text{if} f(x) \geq \alpha, \end{cases} \tag{1.5}$$

where $x_0$ are coordinates of the molecular geometric center. The isovalue $\alpha$ = 0.5 for volumetric data generated from the characteristic function, and $\alpha$ = 1.0 for volumetric data generated from the summation of Gaussian kernels.

The molecular surface and the outer sphere $S_0$ can be extracted as an isosurface at the isovalue 0, $S_g(0) = \{x|g(x) = 0\}$. All the grid points inside the interval volume $I_g(0) = \{x|g(x) \leq 0\}$ have negative function values, and all the grid points outside it have positive values.



(a) (b)

Figure 1.9: 2D triangulation. (a) Old scheme, (b) New scheme. Blue and yellow triangles are generated for sign change edges and interior edges respectively. The red curve represents the molecular surface, and the green points represent minimizer points.

**Mesh Extraction**

Here we introduce a different scheme from the algorithm presented in [237, 238], in which we do not distinguish boundary cells and interior cells when we analyze edges. We only consider two kinds of edges - sign change edges and interior edges. For each boundary cell, we can obtain a minimizer point by minimizing its Quadratic Error Function. For each interior cell, we set the middle point of the cell as its minimizer point. Fig. 1.9(b) shows a simple 2D example. In 2D, there are two cells sharing each edge, and two minimizer points are obtained. For each sign change edge, the two minimizers and the interior vertex of this edge construct a triangle (blue triangles). For each interior edge, each minimizer point and this edge construct a triangle (yellow triangles). In 3D as shown in Fig. 1.10, there are three or four cells sharing each edge. Therefore, the three (or four) minimizers and the interior vertex of the sign change edge construct one (or two) tetrahedron, while the three (or four) minimizers and the interior edge construct two (or four) tetrahedra.

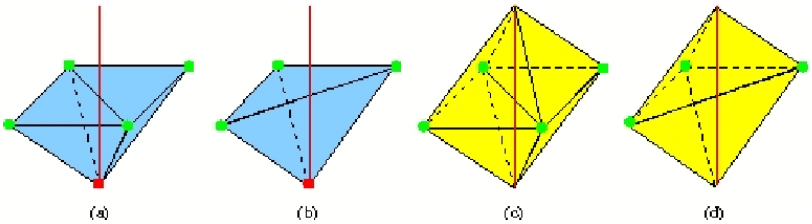


Figure 1.10: Sign change edges and interior edges are analyzed in 3D tetrahedralization. (a)(b) - sign change edge (the red edge); (c)(d) - interior edge (the red edge). The green solid points represent minimizer points, and the red solid points represent the interior vertex of the sign change edge.

Compared with the algorithm presented in [237, 238] as shown in Fig. 1.9(a), Fig. 1.9(b) generates the same surface meshes, and tends to generate more regular interior meshes with better aspect ratio, but a few more elements for interior cells. Fig. 1.9(b) can be easily extended to large volume decomposition. For arbitrary large volume data, it is difficult to import all the data into memory at the same time. So we first divide the large volume data into some small subvolumes, then mesh each subvolume separately. For those sign change edges and interior edges lying on the interfaces between subvolumes, we analyze them separately. Finally, the generated meshes are merged together to obtain the desired mesh. The mesh adaptivity is controlled by the structural properties of biomolecules. The extracted tetrahedral mesh is finer around the molecular surface, and gradually gets coarser from the molecular surface out towards the outer sphere, $S_0$. Furthermore, we generate the finest mesh around the active site, such as the cavity in the monomeric and tetrameric mAChE shown in Fig.**??** (a∼b), and a coarse mesh everywhere else.

**Mesh Extension**

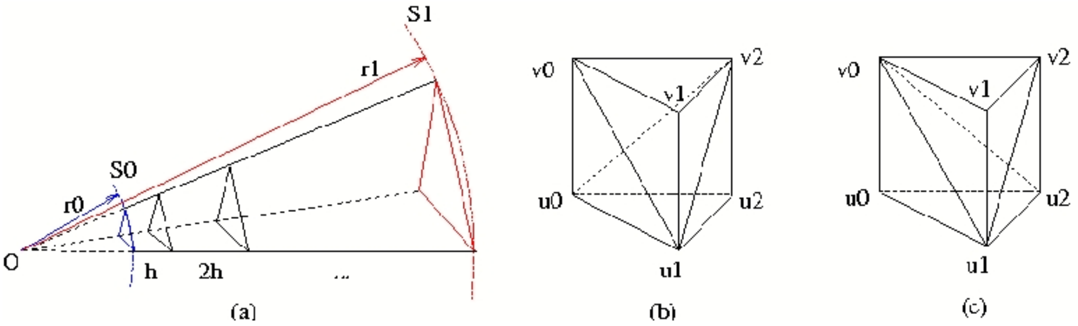


Figure 1.11: (a) - one triangle in the sphere $S_0$ (blue) is extended $n$ steps until arriving at the sphere $S_1$ (red); (b) and (c) - a prism is decomposed into three tetrahedra in two different ways.

We have generated meshes between the biomolecular surface and the outer sphere $S_0$, the next step is to construct tetrahedral meshes gradually from the sphere $S_0$ to the bounding sphere $S_1$ (Fig. 1.8). The sphere $S_0$ consists of triangles, so we extend each triangle radially as shown in Fig. 1.11 and a prism is obtained for each extending step. The prism can be divided into three tetrahedra. The extension step length $h$ can be calculated by Eqn. (1.6). It is better for the sphere $S_0$ to be triangulated uniformly since the step length is fixed for each extension step.

$$r_0 + h + 2h + \cdots + nh = r_1 \quad \Longrightarrow \quad h = \frac{2(r_1 - r_0)}{n(n+1)} \tag{1.6}$$

where $n$ is the step number. In Figure 1.11, suppose $u_0u_1u_2$ is a triangle on sphere $S_0$, and $u_0$, $u_1$, $u_2$ are the unique index numbers of the three vertices, where $u_1 < u_0$ and $u_1 < u_2$. For one extension step, $u_0u_1u_2$ is extended to $v_0v_1v_2$, and the two triangles construct a prism, which can be decomposed into three tetrahedra. In order to avoid the diagonal conflict problem, a different decomposition method (Fig. 1.11(b∼c)) is chosen based on the index number of the three vertices. If $u_0 < u_2$, then we choose Fig. 1.11(b) to split the prism into three tetrahedra. If $u_2 < u_0$, then Fig. 1.11(c) is selected

Assume there are $m$ triangles on the sphere $S_0$, which is extended $n$ steps to arrive at the sphere $S_1$. $m$ prisms or $3m$ tetrahedra are generated in each extending step, and a total of $3mn$ tetrahedra are constructed in the extension process. Therefore, it is better to keep a coarse and uniform triangular mesh on the sphere $S_0$.

**Union of Balls using Voronoi-Cell Complexes**

Several different approaches have been developed to achieve this efficiency for molecular surface computations [59, 183, 184, 185, 212, 214]. Other work on surface representations features the use of metaballs, molecular surfaces, and blobby models [4, 30, 223, 65, 89, 109, 122, 152, 160, 161, 228, 229, 231].

In previous work on dynamic triangulations the focus has been mostly on the simpler Delaunay/Voronoi structures (unweighted case) [17, 124, 50, 88, 111, 5, 178, 179]. Little has been done on the more general case of dynamic Regular Triangulation/Power Diagrams and for dimensions greater than two. Moreover, the kinds of dynamic operations developed are usually just the insertion/deletion of a single point. Such local operations become inefficient when we need to perform even a simple but global modification.

**Molecular Surface Computation using Adaptive Grids**

Since Richards introduced the SES definition, a number of techniques have been devised to compute the surface, both static and dynamic, implicit and explicit. Connolly introduced two algorithms to compute the surface. First, a dot based numerical surface construction and second, an enumeration of the patches that make up the analytical surface (See [59], [58] and his PhD thesis). In [214], the authors describe a distance function grid for computing surfaces of varying probe radii. Our data structure contains approaches similar to their idea. A number of algorithms were presented using the intersection information given by voronoi diagrams and the alpha shapes introduced by Edelsbrunner [73], including parallel algorithms in [212] and a triangulation scheme in [4]. Fast computations of SES is described in [184] and [183], using Reduced sets, which contains points where the probe is in contact with three atoms, and faces and edges connecting such points. Non Uniform Rational BSplines ( NURBs ) descriptions for the patches of the molecular surfaces are given in [22], [21] and [23]. You and Bashford in [232] defined a grid based algorithm to compute a set of volume elements which make up the Solvent Accessible Region.

### Maintaining Union of Balls Under Atom Movements

Though a number of techniques have been devised for the static construction of molecular surfaces (e.g., [59, 58, 214, 73, 212, 4, 184, 183, 232, 108, 22, 21, 239, 24]), not much work has been done on neighborhood data structures for the dynamic maintenance of molecular surfaces as needed in MD. In [23] Bajaj et al. considered limited dynamic maintenance of molecular surfaces based on Non Uniform Rational BSplines ( NURBS ) descriptions for the patches. Eyal and Halperin [79, 80] presented an algorithm based on dynamic graph connectivity that updates the union of balls molecular surface after a conformational change in $\mathcal{O}\left(\log^2 n\right)$ amortized time per affected (by this change) atom.

### Clustering and Decimation of Molecular Surfaces

Using multiresolution models for molecules can substantially improve rendering speed and interactive response rates in molecular interaction tools. Similar improvements in performance would be achieved when a set of balls is used as an approximate representation of a generic object either for modeling (meta-balls [109, 161], blobby models [231]) or for collision detection [122]. Direct application of previous approaches for the decimation and multiresolution representation of the surfaces themselves [184, 138] can have serious embedding and self-intersection problems and are specific to the surface definition. A possible solution if this problem has been addressed in [199] but limited to the case of the boundary surface of tetrahedral meshes. Our multiresolution scheme updates the underlying structure of the molecule, maintaining at any level of detail a regular triangulation of the current weighted point-set. In this way we explicitly track the topology of the molecular body at any adaptive level of resolution. Moreover this guarantees correct embedding in all resolutions and creates an approximation from which the surface boundary can be computed in any of the previous schemes.

There are many approaches for creating multiresolution representations of geometric data for graphics and visualization [181, 147, 133]. They vary in both the simplification scheme like vertex removal [63], edge contraction [120], triangle contraction [95], vertex clustering [188], wavelet analysis [67], and also in the structure used to organize the levels of detail (either a linear order or a using a DAG).

Maintaining the regular triangulation at all resolutions rules out the possibility of using decimation techniques like edge or triangle contraction, which do not guarantee the (weighted) Delaunay property. Other known decimation schemes that can guarantee this property such as vertex removal, do not seem appropriate in this case since they do not preserve the molecule features as a subset of the whole triangulation. Techniques which preserve features in the triangulation by tagging specific edges or vertices [49] are more suitable for preserving specific edges or regions. We are more interested in applying the decimation on a subset of the triangulation while this subset can change during the decimation.

Sphere trees have also been used in [121] for the purpose of fast collision detection. In this work, Sphere hierarchies are built around a given object either by replacing special octree regions or by placing balls on the medial-axis surfaces approximated using voronoi edges of some point sampling of the object. The basic approach of building the hierarchy by clustering pairs of balls for collision detection [122] is similar to ours. However in this scheme the simplification process does not update the underlying triangulation and hence does not track the topological changes induced by the decimation process. This make also the scheme unable to cluster balls that get in contact only after some simplification steps.

## Summary

## References and Further Reading

- Set theory background can be found in Rosen [180].

- Useful references for graph theory include Behzad and Chartrand [26], Chartrand and Lesniak [45], and Giblin [94]
- Minimal spanning trees are discussed in Graham and Hell [107] and algorithms for finding them in Kruskal[137]
- Useful algebraic topology texts include Armstrong [8] and Hatcher [115]. This includes more formal definitions of terms like homeomorphism, isomorphism, manifold, and homology.
- Some notation on primal and dual meshes has been adapted from Hirani [116].
- For practical computational topology, see Zomorodian [243]
- For more on CW complexes, see Munkres [157].
- The power distance metric is described in [15].
- More on regular triangulations in [70, 74] and on weighted alpha shapes in [68, 81].
- Some application references: [237, 238, 35, 36, 197, 198, 234, 126, 92, 119, 141].

## Exercises

# Chapter 2

# Sets, Functions and Mappings

## 2.1 Scalar, Vector and Tensor Functions

**Definition 2.1.** We use the following basic definitions

- A **scalar function** is a function whose values are in $\mathbb{R}$, i.e. $f : V \to \mathbb{R}$.
- A **vector function** is a function whose values are in $\mathbb{R}^k$ for some $k > 1$. i.e. $f : V \to \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{k \text{ copies}}$.

**Definition 2.2.** Let $V$ be a vector space and let $V^p$ denote the Cartesian product of $p$ copies of $V$. A (real) $p$**-tensor on** $V$ is a function $T : V^p \longrightarrow \mathbb{R}$ that it is linear in each variable.

The **tensor product** of a $p$-tensor $T$ and a $q$-tensor $S$ is defined by

$$T \otimes S(v_1, \ldots, v_p, v_{p+1}, \ldots, v_{p+q}) := T(v_1, \ldots, v_p) \cdot S(v_{p+1}, \ldots, v_{p+q})$$

Note that this operation is not symmetric. A tensor $T$ is called **alternating** or **anti-symmetric** if and only if the sign of $T$ is reversed whenever two variables are transposed. Let $S_p$ denote the symmetric group on $p$ elements. An arbitrary tensor $T$ is associated to the alternating tensor Alt $T$, defined by

$$\text{Alt}\, T := \frac{}{1}p! \sum_{\pi \in S_p} (-1)^\pi T^\pi,$$

where

$$T^\pi(v_1, \ldots, v_p) := T(v_{\pi(1)}, \ldots, v_{\pi(p)}).$$

Alternating $p$-tensors are closed under scalar multiplication and addition, thereby forming a vector space:

$$\Lambda^p(V^*) := \{\text{Alt}\, T : T \text{ is a } p\text{-tensor on } V\}$$

**Definition 2.3.** If $T \in \Lambda^p(V^*)$ and $S \in \Lambda^q(V^*)$, the **wedge product** of $T$ and $S$ is defined by

$$T \wedge S := \text{Alt}\,(T \otimes S) \in \Lambda^{p+q}(V^*)$$

◊

## 2.2 Inner Products and Norms

In this section we will introduce the norm, inner products and Hilbert Spaces.

### 2.2.1 Vector Space

**Definition 2.4.** Formally, a field is a set $\mathbb{F}$ together with two operations called addition and multiplication, $\mathbb{F}$ with addition forms an Abelian group with identity element "0" while $\mathbb{F}$ with multiplication forms an Abelian group with identity element "1"

**Definition 2.5.** A vector space is defined as $V = \{X, +, *, \mathbb{F}\}$, where $X$ is a set and $\mathbb{F}$ is a Field. $X$ and $+ : X \times X \to X$ forms an Abelian group and $* : \mathbb{F} \times X \to X$ satisfying the following:

- $\alpha * (a + b) = \alpha * a + \alpha * b$
- $(\alpha + \beta) * a = \alpha * a + \beta * a$
- $\alpha * (\beta * a) = (\alpha\beta)a$
- $1 * a = a$
- $0 * a = 0$

where $\alpha, \beta \in \mathbb{F}$, $a, b \in X$

Vector spaces show us to speak linear transformations, summation, subspace and duality.

### 2.2.2 Topological Space

**Definition 2.6.** $X$ is an nonempty set, $\mathcal{X}$ is the class of subsets of $X$ such that:

- $X \in \mathcal{X}$
- $\emptyset \in \mathcal{X}$
- $X_1, X_2, \ldots, X_n \in \mathcal{X} \implies \bigcap_{i=1}^{n} X_i \in \mathcal{X}$ (finite intersection)
- $\bigcup_{i \in \mathcal{I}} X_i \in \mathcal{X}$ (any union)

Then $\mathcal{X}$ defines the topology on X, $\forall x \in \mathcal{X}$ is called an open set in $X$, and $\mathcal{V} = (X, \mathcal{X})$ forms a **topological space**.

**Definition 2.7.** $x_1 \in X$, $B_{x_1}$ is defined as neighborhood of $x_1$ if $B_{x_1}$ is a subset at $X$ and there exists an open set $U \in \mathcal{X}$ containing $x_1$ s.t. $U \subset B_{x_1}$

**Definition 2.8.** $\mathcal{V} = (X, \mathcal{X})$ is **Hausdorff** if and only if, $\forall$ pair of points $x_1, x_2 \in X$, $\exists$ neighborhood $B_{x_1}$,$B_{x_2}$ such that:

$$B_{x_1} \cap B_{x_2} = \emptyset$$

(Point) Topological spaces allow us to speak of open sets, closed sets, compactness, convergence of sequences, continuity of functions, etc.

**Example 2.9.** *Let $\{X, \mathcal{X}\}$,$\{Y, \mathcal{Y}\}$ be two topological spaces. $F : X \to Y$ is a continuous mapping at $x_0 \in X$ if and only if : $\forall$ open set $Y_0 \in \mathcal{Y}$ containing $F(x_0)$ contains an open set $B$ that is the image of an open set containing $x_0$. (An open set's original image is an open set)*

### 2.2.3 Metric Space

**Definition 2.10.** A metric space is an ordered pair $(X, d)$ where $X$ is the set and $d$ is a function defined on $X \times X$:

$$d : X \times X \to R$$

such that for $\forall x, y, z \in X$, the following holds:

- $d(x, y) \geqslant 0$

- $d(x, y) = 0 \implies x = y$
- $d(x, y) = d(y, x)$
- $d(x, z) \leqslant d(x, y) + d(y, z)$

**Example 2.11.** *Every* ***metric space*** *(denoted as $(X, d)$) is a topological space. Since we can define open sets*

$$B_r(x_0) = \{y \in X : d(x_0, y) = r\}$$

*like the balls on metric space. In this case:*

- $x_n \to x_0 \iff \forall \epsilon > 0, \exists n \in \mathbb{N}$ *such that* $d(x_0, x_n) < \epsilon$ *for all* $m > n$
- $F$ *is continuous* $\iff \forall \epsilon > 0, \exists \delta > 0$ *such that* $d(F(x), F(x_0)) < \epsilon$ *whenever* $d(x, x_0) < \delta$

### 2.2.4 Topological Vector Space

**Definition 2.12.** $\mathcal{V}$ is called a topological vector space if and only if:

- $\mathcal{V}$ is a vector space
- The underlying set $V$ of vectors in $\mathcal{V}$ is endowed with a topology $\mathcal{U}$ such that:
  - $(V,\mathcal{U})$ is a Hausdorff topological space
  - vector addition is continuous: $u + v \in V$ if $u, v \in V$
  - scalar multiplication is continuous: $\alpha u \in V$ if $\alpha \in F, u \in V$

### 2.2.5 Normed Space

**Definition 2.13.** $V$ is a vector space, $N : V \to \mathbb{R}$ is a **norm** of $V$ if :

- $N(v) \geqslant 0$, and $N(v) = 0$ if and only if $v = 0$
- $N(\alpha v) = |\alpha| N(v)$
- $N(u + v) \leqslant N(u) + N(v)$

We always denote $\|u\| := N(u)$.

*Remark* 2.14. You can verify that $\|u - v\|$ is a metric on $V$, thus every normed space is a metric space.

*Remark* 2.15. You can also verify that every normed space is a Topological Vector Space:

- let we assume there are two convergent sequence $\{u_n\}, \{v_n\} \subset V$:

  $$u_n \to u, v_n \to v$$

  where $u, v \in V$, then we can verify that:

  $$\|(u_n + v_n) - (u + v)\| \leqslant \|u - u_n\| + \|v - v_n\| \to 0 \text{ as } n \to 0$$

- Suppose $\alpha_n \to \alpha$ in $\mathbb{F}$, then:

  $$\|\alpha_n u_n - \alpha u\| \leqslant |\alpha - \alpha_n| \|u_n\| + |\alpha| \|u - u_n\| \to 0 \text{ as } n \to \infty$$

Therefore, in normed spaces, we have the concept that adapted both from linear spaces and topological spaces. Next is the definition for a Banach Space.

**Definition 2.16.** A complete normed space is a **Banach** space, or a $B$ space. Here complete means : every Cauchy sequence in a metric space converges in that metric space.

Here are some properties pertinent to normed spaces:

- $A : U \to V$, $U, V$ are underlying sets of normed spaces with norms $\|\cdot\|_U, \|\cdot\|_V$, respectively.
- $A$ is linear if and only if
$$A(\alpha u_1 + \beta u_2) = \alpha A(u_1) + \beta A(u_2) \forall u_1, u_2 \in U$$
- $A$ is bounded if and only if $A$ maps a bounded sets in $U$ into bounded sets in $V$:
$$\|u\|_U \leqslant C_1 \implies \exists C_2 \text{ such that } \|Au\|_V \leqslant C_2$$
- $A$ is continuous if and only if $\forall \epsilon > 0, \exists \delta > 0$ such that:
$$\|u - v\|_U < \delta \implies \|Au - Av\|_V < \epsilon$$
or if and only if , whenever $u_n \to u$ ($\|u - u_n\|_V \to 0$ as $n \to \infty$), we have:
$$\|Au - Av\|_V \to 0 \text{ as } n \to \infty$$

**Theorem 2.17.** *Let $(U, \|\cdot\|_U), (V, \|\cdot\|)_V$ be normed spaces over the same field. Let $A : U \to V$ be a linear function. Then the following are equivalent:*

*1) A is continuous*

*2) A is continuous at $u = 0$*

*3) A is bounded*

*4) $\exists C > 0$ such that:*
$$\|Au\|_V \leqslant C\|u\|_U \quad \forall u \in U$$

*Proof.*
1) $\Rightarrow$ 2) is obvious.
2) $\Rightarrow$ 3):
Let $\|u\|_U < r$. Since $A$ is continuous at 0, $\forall \epsilon > 0, \exists \delta > 0$ such that
$$\|Au\|_V < \epsilon \implies \|u\|_U < \delta$$
Pick $\epsilon = 1$, then $\exists\ \delta$ such that $\|u\|_U < \delta \Rightarrow \|Au\|_V < 1$.
If $\|u\|_U < r$,
$$\|\frac{\delta}{r}u\|_U = \frac{\delta}{r}\|u\|_U \leqslant \delta$$
Thus
$$\|A(\frac{\delta}{r}u)\|_V \leqslant 1 \implies \|Au\|_V \leqslant \frac{r}{\delta} = \text{constant}$$
Hence, $A$ is bounded.
3) $\Rightarrow$ 4):
Since $A$ is bounde, $\exists C > 0$ such that $\|Au\|_V \leqslant C$ whenever $\|u\|_U \leqslant 1$.
Thus, $\forall u \neq 0$,
$$\|A(\frac{u}{\|u\|_U})\|_V \leqslant C$$
and therefore:
$$\|Au\|_V \leqslant C\|u\|_U$$

4)⇒ 1):
If $u_n \to u$, then:

$$\|Au - Au_n\|_V \leqslant C\|u - u_n\|_V \to 0 \text{ as } n \to \infty$$

□

### 2.2.6 Inner Product Space

**Definition 2.18.** Let $V$ be a vector space, and define $p : V \times V \to \mathbb{F}(\mathbb{C} \text{ or } \mathbb{R})$. Then $p$ is an **inner product** on $V$ if it satisfies the following:

- $\forall u \in V, p(u,u) \geqslant 0; p(u,u) = 0 \iff u = 0$
- $\forall u, v \in V, p(u,v) = \overline{p(v,u)}$ (Conjugate Symmetry)
- $\forall u_1, u_2, v \in V, \forall \alpha_1, \alpha_2 \in \mathbb{F}, p(\alpha_1 u_1 + \alpha_2 u_2, v) = \alpha_1 p(u_1, v) + \alpha_2 p(u_2, v)$

Denote as $(u,v) = p(u,v)$. A vector space on which an inner product has been defined is called an **inner product space**. Denote the inner product space as $(V, (\cdot,\cdot))$

*Remark* 2.19. You can verify that an inner product also satisfies the following:

$$p(u, \beta_1 v_1 + \beta_2 v_2) = \bar{\beta}_1 p(u, v_1) + \bar{\beta}_2 p(u, v_2)$$

where $u, v_1, v_2 \in V, \beta_1, \beta_2 \in \mathbb{F}$

**Definition 2.20.** Let $(V, (\cdot,\cdot))$ be an inner product space. Pick $u, v \in V$, we claim that $u$ and $v$ are **orthogonal** if

$$(u,v) = 0$$

One important property for the inner product is that it satisfies the Cauthy-Schwarz Inequality.

**Theorem 2.21** (Cauthy-Schwarz Inequality)**.** *Let $(V, (\cdot,\cdot))$ be an inner product space. If $u, v \in V$, then:*

$$|(u,v)| \leqslant \sqrt{(u,u)(v,v)}$$

*Proof.* Suppose $\mathbb{F} = \mathbb{C}$, pick $\alpha = \frac{\overline{(v,u)}}{(v,v)} \in \mathbb{C}$, then:

$$\begin{aligned}
0 &\leqslant (u - \alpha v, u - \alpha v) \\
&= (u,u) - \alpha(v,u) - \bar{\alpha}(u,v) + \alpha\bar{\alpha}(v,v) \\
&= (u,u) - \frac{\overline{(v,u)}}{(v,v)}(v,u) - \frac{\overline{(u,v)}}{(v,v)}(u,v) + \frac{\overline{(v,u)(u,v)}}{(v,v)^2}(v,v) \\
&= \frac{1}{(v,v)}\left[(u,u)(v,v) - 2|(u,v)|^2 + |(u,v)|^2\right]
\end{aligned}$$

Therefore $|(u,v)|^2 \leqslant (u,u)(v,v)$. □

Next, we want to connect the inner product space with normed space.

**Theorem 2.22.** *Every inner product space is a normed space with norm :*

$$\sqrt{(u,u)} = \|u\|$$

*Proof.* Recall the definition of the norm, all you need is to verify that

- $\|u\| \geqslant 0$ and $\|u\| = 0 \iff u = 0$

- $\|u+v\| \leqslant \|u\| + \|v\|$
- $\forall \alpha \in \mathbb{F}, \|\alpha u\| = |\alpha| \|u\|$

□

*Remark* 2.23. It is understood that the inner product space $V$ is induced with the topology induced by the norm $(u,u)^{\frac{1}{2}}$

Now, we introduce an important type of space:

**Definition 2.24.** An inner product space is a **Hilbert Space** if and only if it is complete (with respect to the norm induced by the inner product)

A typical example of a Hilbert Space will be the Euclidean Space $\mathbb{R}^d$ with an inner product defined as:

$$(x,y) = \sum_{i=1}^{d} x_i y_i$$

**Theorem 2.25.** *Suppose an inner product space $(V, (\cdot,\cdot))$ has two convergence sequence in norm:*

$$v_m \to v \text{ and } u_m \to u$$

*Then*

$$(v_m, u_m) \to (v,u)$$

*Proof.* In fact, we have:

$$\begin{aligned} |(v_m,u_m) - (v,u)| &= |(v_m,u_m) + (v_m,u) - (v_m,u) - (v,u)| \\ &= |(v_m, u_m - u) + (v_m - v, u)| \\ &\leqslant \|v_m\|\|u_m - u\| + \|v_m - v\|\|u\| \\ &\to 0 \text{ as } m \to \infty \end{aligned} \tag{2.1}$$

□

*Remark* 2.26. Similar as Euclidean Space, inner product shares some geometric properties in general vector space:

- $\cos\theta \overset{def}{=} \frac{(u,v)}{\|u\|\|v\|} \quad (\mathbb{F} = \mathbb{R})$
- Pythagoras: $(u,v) = 0 \Rightarrow \|u+v\|^2 = \|u\|^2 + \|v\|^2$
- Sphere: $(u - u_0, u - u_0) = a^2$
- Hyperplane: $(u - a, n) = 0$
- Parallelogram Law: $\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2$

## 2.3 Piecewise-defined Functions

## 2.4 Homogeneous and Barycentric coordinates

### 2.4.1 Homogeneous coordinates

A point in complex projective space $\mathbf{CP}^n$ is given by a nonzero *homogeneous coordinate vector* $(X_0, X_1, \ldots, X_n)$ of $n+1$ complex numbers. A point in complex affine space $\mathbf{CA}^n$ is given by the *non-homogeneous coordinate vector* $(x_1, x_2, \ldots, x_n)$ = $(\frac{X_1}{X_0}, \frac{X_2}{X_0}, \ldots, \frac{X_n}{X_0})$ of $n$ complex numbers. The set of points $Z_d^n(f)$ of $\mathbf{CA}^n$ whose coordinates satisfy a single non-homogeneous polynomial equation $f(x_1, x_2, \ldots, x_n) = 0$ of degree $d$, is called an $n-1$ dimension, affine hypersurface

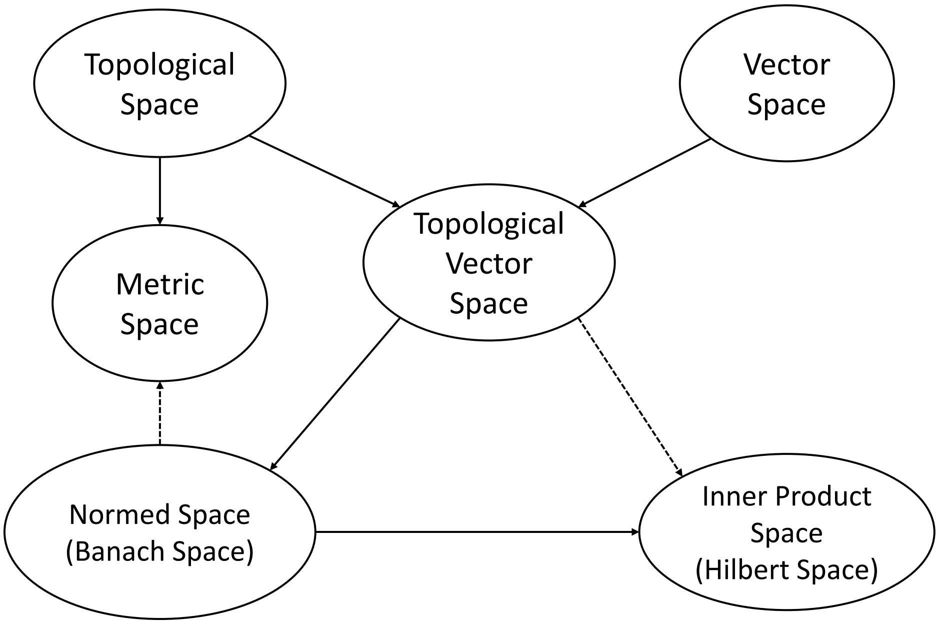


Figure 2.1: Relationship among spaces

of degree $d$. The hypersurface $Z_1^n(f)$ is also known as a *flat* or a *hyperplane*, a $Z_2^n(f)$ is known as a *quadric* hypersurface, and a $Z_3^n(f)$ is known as a *cubic* hypersurface. The hypersurface $Z_d^2$ is a plane *curve* of degree $d$, a $Z_d^3$ is known as a *surface* of degree $d$, and $Z_d^4$ is known as a *threefold* of degree $d$. A hypersurface $Z_d^n$ is *reducible* or *irreducible* based upon whether $f(x_1, x_2, ..., x_n) = 0$ factors or not, over the field of complex numbers. An algebraic variety $Z^n\{f_1, ..., f_n\}$ is then an irreducible common intersection of a collection of hypersurfaces $Z_{d_i}^n(f_i)$.

An irreducible *rational* hypersurface $Z_d^n(f)$, can additionally be defined by rational parametric equations which are given as $(x_1 = G_1(u_1, u_2, \ldots, u_{n-1})$, $x_2 = G_2(u_1, u_2, \ldots, u_{n-1})$, …, $x_n = G_n(u_1, u_2, \ldots, u_{n-1}))$, where $G_1$, $G_2$, …, $G_n$ are rational functions of degree $d$ in $\mathbf{u} = (u_1, u_2, \ldots, u_{n-1})$, i.e., each is a quotient of polynomials in $\mathbf{u}$ of maximum degree $d$.

*Multi-polynomial Resultant*: Consider $F_1 = 0, ..., F_m = 0$ polynomial equations in $n + 1$ variables ($X_0$, ..., $X_n$) and homogeneous in $m$ variables ($X_0$, ..., $X_{m-1}$). These equations could be the homogenization of the earlier system (**??**) with $X_0$ acting as the homogenizing variable. The multi-polynomial *resultant* $R(F_1, ..., F_m)$ is a polynomial in the *coefficients* of the $F_i$ that vanishes if and only if the $F_i$ have a common zero in projective space. For this reason, the resultant is also often called the eliminant. Geometrically, the resultant vanishes if and only if the $n$ hypersurfaces $Z_d^n(F_i)$ have a common intersection in projective space.

The resultant of several equations has several different characterizations. Probably the most elegant was discovered by Macaulay [145]. He shows that the multi-polynomial resultant can be expressed as the quotient of the determinant of two matrices whose entries are coefficients of the polynomials. In the case of two equations, the matrix for the denominator always has determinant 1 and the matrix for the numerator is the traditional Sylvester matrix[182]. In computing the multi-polynomial resultant, the $F_i$ are multiplied by suitable monomials to transform the problem of determining whether the polynomials have a common zero into a problem in linear algebra. We construct a matrix whose entries are the coefficients of the $F_1, ..., F_m$. The determinant of this matrix will be the product of the resultant and the determinant of a specific minor of the matrix.

The general construction due to [145] is as follows: In the system $F_1 = 0$, ..., $F_m = 0$ of polynomial equations, homogenous in variables $X_0$, ..., $X_{m-1}$, let $F_i$ be of degree $d_i$. The coefficients of the $F_i$'s are treated as indeterminates. Let

$$d = 1 + \sum(d_i - 1).$$

and let the $m$-vector $\alpha$ denote the exponents of a monomial in $X_0$, ..., $X_{m-1}$. For example, if $\alpha = (\alpha_0, ..., \alpha_{m-1})$, then

$$X^\alpha = X_0^{\alpha_0} ... X_{m-1}^{\alpha_{m-1}}.$$

Thus, the set of all monomials of degree $d$ in $m$ variables is

$$\mathcal{X}^d = \{X^\alpha | \alpha_0 + ... + \alpha_{m-1} = d.\}$$

If $N$ denotes the number of monomials in this set, then the monomials will index the columns of an $N$ by $N$ matrix.

$$N = \begin{pmatrix} d+m-1 \\ d \end{pmatrix}$$

Partition $\mathcal{X}^d$ into $n$ disjoint sets. These sets are

$$\mathcal{X}_i^d = \{X^\alpha | \alpha_i \geq d_i \text{ and } \alpha_j < d_j \ , \forall j < i\}.$$

Next, for each set $\mathcal{X}_i^d$, construct a set $F_i^d$ of polynomials from $F_i$ using monomials in $\mathcal{X}_i^d$. Specifically, let

$$F_i^d = \frac{\mathcal{X}_i^d}{x_i^{d_i}} f_i.$$

The $F_i^d$ are sets of homogeneous polynomials in $m$ variables of degree $d$. Moreover, each of the polynomials in the union of the $F_i^d$, equated to zero, collectively yields a set of $N$ homogeneous polynomial equations. Construct an $N$ by $N$ matrix (call it $A$) whose columns are indexed by monomials in $\mathcal{X}^d$ and whose rows correspond to the polynomials in the $F_i^d$'s. For a given polynomial $P$ in $F_i^d$, its row consists of the symbolic coefficients $a_{ik}$, $b_{jk}$ etc., of each monomial in $P$.

$$A \begin{pmatrix} X_0^d \\ . \\ . \\ . \\ X_{m-1}^d \end{pmatrix} = \begin{pmatrix} . & a_{i1} & a_{i2} & a_{i3} & . & . & . & . \\ . & . & b_{j1} & b_{j2} & b_{j3} & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \end{pmatrix} \begin{pmatrix} X_0^d \\ . \\ . \\ . \\ X_{m-1}^d \end{pmatrix} = \begin{pmatrix} 0 \\ . \\ . \\ . \\ 0 \end{pmatrix} \tag{2.2}$$

Now, if the $F_i$ have a common root $(\hat{X}_0, ..., \hat{X}_{m-1})$, then this root must satisfy all of the polynomial equations in the $F_i^d$'s. This fact implies that the nontrivial vector $(\hat{X}_0, ..., \hat{X}_{m-1})$ must be in the null space of $A$. Thus, $A$ must be singular or equivalently, the determinant of $A$ (call it $D$) must be zero. This argument establishes that the resultant $R$ is a factor of $D$. The remaining factors of $D$ are extraneous and have no bearing on whether the original equations have a common root. The beauty of Macaulay's result is that he established that the extraneous factors are the determinant of a *minor* of $A$. This minor (call it $B$) can be constructed from $A$ in the following manner. Delete all columns of $A$ that correspond to monomials $X^\alpha$ where $\alpha_i < d_i$ for all but one value of $i$. (Note there must at least one such $i$ due to the manner in which $d$ was chosen.) Delete all rows of $A$ that correspond to polynomials in $F_i$ whose multipliers $X^\alpha$ have $\alpha_j < d_j$ for $i < j \leq n$.
Macaulay shows that the resultant $R$ satisfies

$$R = \frac{det(A)}{det(B)}$$

where this division is carried out before the indeterminates forming the entries of $A$ and $B$ are specialized. The reason for specializing after division is that $det(A)$ and $det(B)$ may evaluate to zero even though $R$ is not identically zero. Techniques for computing $R$ by specializing before division has recently been considered in [42, 174].
*Multi-polynomial Remainder Sequence*: Consider first two polynomial equations $f_1(x_1, \ldots, x_n) = 0$ and $f_2(x_1, \ldots, x_n) = 0$. Treating them as polynomials in $x_1$, the psuedo-remainder $(f_1/f_2)$ = $g(x_1, \ldots, x_n)$ for degree$_{x_1}(f_2) \leq$ degree$_{x_1}(f_1)$, is the result of one step of psuedo-division in the ring $C$ of coefficient polynomials in $n-1$ variables $(x_2, \ldots, x_n)$, i.e. $\alpha f_1 = \beta f_2 - g$ with $\alpha, \beta \epsilon C$ and degree$_{x_1}(g) <$ degree$_{x_1}(f_2)$. Repeating the psuedo-division with $f_2$ and $g$ and ensuring that the factors $\alpha$ and $\beta$ are 'primitve', one can compute a subresultant polynomial remainder sequence (p.r.s):

$$f_1, f_2, g = S_{k-1}, \ldots, S_1, S_0 \tag{2.3}$$

where $S_i$ is the psuedo-remainder of the two polynomials preceding it in the sequence and is known as the $i^{th}$ subresultant of $f_1$ and $f_2$, with respect to $x_1$, see for e.g [117, 144]. Here $S_0$ is a polynomial independent of $x_1$ and is the resultant of $f_1$ and $f_2$, with respect to $x_1$. (Note in the homogeneous case $S_0$ is the polynomial resultant of $F_1$ and $F_2$, with respect to $X_0$ and $X_1$.

For the set of polynomial equations ( **??**), treating them as polynomials in $x_1$, we select the polynomial, say $f_k$, of minimum degree in $x_1$. We then compute the subresultant psuedo-remainder for each pair $(f_i/f_k) = g_i$, $1 \leq i \leq m$ and $i \neq k$, yielding a new system of equations $g_i$ and $f_k$. We repeat the above, first selecting from the new system, a polynomial of minimum degree in $x_1$, and then computing pairwise subresultant psuedo-remainders. Eventually, we obtain a system of $m-1$ polynomial equations, say $S^{m-1}$

$$\begin{aligned} \tilde{f}_1(x_2, ..., x_n) &= 0 \\ &... \\ \tilde{f}_{m-1}(x_2, ..., x_n) &= 0 \end{aligned} \tag{2.4}$$

independent of $x_1$.
The above is then one (macro) step of the multi-equational polynomial remainder sequence (m.p.r.s). For the new set of polynomial equations (2.4), treating them as polynomials in $x_2$, we repeat the entire process above and obtain yet another reduced system $S^{m-2}$ of $m-2$ polynomial equations, all independent of $x_2$, and so on. This sequence of systems of multi-equational polynomial equations

$$S = S^m, S^{m-1}, S^{m-2}, \ldots, S^1, S^0 \tag{2.5}$$

is what we term the multi-equational polynomial remainder sequence.

### 2.4.2 Barycentric coordinates

Barycentric coordinates are a natural method for describing a function defined on a triangular domain and we can extend the idea of barycentric coordinates over polygon or in higher dimensional simplex. Barycentric coordinates on general polygons (denotes as **generalized barycentric coordinates**) are any set of functions satisfying certain key properties.

**Definition 2.27.** Functions $\lambda_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, n$ are **barycentric coordinates** on $\Omega$ if they satisfy two properties.

B1. **Non-negative**: $\lambda_i \geq 0$ on $\Omega$.

B2. **Linear Completeness**: For any linear function $L : \Omega \to \mathbb{R}$, $L = \displaystyle\sum_{i=1}^{n} L(\mathbf{v}_i)\lambda_i$.

Most commonly used barycentric coordinates, including the mean value coordinates, are invariant under rigid transformation and simple scaling which we will state precisely. Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be a composition of rotation, translation, and uniform scaling transformations and let $\{\lambda_i^T\}$ denote a set of barycentric coordinates on $T\Omega$.

B3. **Invariance:** $\lambda_i(\mathbf{x}) = \lambda_i^T(T(\mathbf{x}))$.

The invariance property can be easily passed through Sobolev norms and semi-norms, allowing attention to be restricted to domains $\Omega$ with diameter one without loss of generality. The essential case in our analysis is the $H^1$-norm, $|u|_{H^1(\Omega)} = \sqrt{\int |\nabla u(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x}}$ where $\nabla u = (\partial u/\partial x, \partial u/\partial y)^T$ is the vector of first partial derivatives of $u$, and for simplicity $T$ is a uniform transformation, $T(\mathbf{x}) := h\mathbf{x}$. For simplicity , the Euclidean norm of vectors will be denoted with single bars $|\cdot|$ without any subscript. Applying the chain rule and change of variables in the integral gives the equality:

$$\begin{aligned} \left|\lambda_i^T\right|^2_{H^1(T\Omega)} &= \int_{T\Omega} \left|\nabla\lambda_i^T(\mathbf{x})\right|^2 \mathrm{d}\mathbf{x} = \int_{T\Omega} \left|\frac{1}{h}\nabla\left(\lambda_i^T(h\mathbf{x})\right)\right|^2 \mathrm{d}\mathbf{x} \\ &= h^{d-2}\int_{\Omega} |\nabla\lambda_i(\mathbf{y})|^2 \, \mathrm{d}\mathbf{y} = h^{d-2}\, |\lambda_i|^2_{H^1(\Omega)} \,. \end{aligned}$$

The scaling factor $h^d$ resulting from the Jacobian when changing variables in the integral is the same for any Sobolev seminorm, while the factor of $h^{-2}$ from the chain rule depends on the order of differentiation in the norm (1, in this case) and the $L^p$ semi-norm used ($p = 2$, in this case). When developing interpolation error estimates, which are ratios of Sobolev norms, the former term (i.e., the chain of variables portion) cancels out and latter term (i.e., the chain rule portion) determines the convergence rate.
Several other familiar properties immediately result from the definition of generalized barycentric coordinates (B1 and B2).

B4. **Partition of unity:** $\sum_{i=1}^{n} \lambda_i \equiv 1$.

B5. **Linear precision:** $\sum_{i=1}^{n} \mathbf{v}_i \lambda_i(\mathbf{x}) = \mathbf{x}$.

B6. **Interpolation:** $\lambda_i(\mathbf{v}_j) = \delta_{ij}$.

**Proposition 2.28.** *Suppose B1 and B2 hold. Then B4, B5, and B6 hold as well.*

### Example: Triangulation Barycentric Coordinates

We begin with the simplest possible case of describing a line passing through a triangle. Let $T$ be a triangle with vertices $(x_1, y_1)$, $(x_2, y_2)$, $(x_3, y_3)$. To represent a line $C$ implicitly, we could find real coefficients $c_{ij} \in \mathbb{R}$ of a function

$$g(x, y) = \sum_{i+j \leq 1} c_{ij} x^i y^j = c_{00} + c_{10}x + c_{01}y$$

such that $C = \{g = 0\}$. Since the function $g$ is defined in terms of the global coordinates $x$ and $y$, it is not immediately obvious given the coefficients $c_{ij}$ whether $C$ passes through $T$ at all. Thus, we transform the $(x, y)$ coordinates to **real barycentric coordinates** $(\lambda_1, \lambda_2, \lambda_3)$ via

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix}$$

Under the mapping, $(x_1, y_1)$ becomes $(1, 0, 0)$, $(x_2, y_2)$ becomes $(0, 1, 0)$, and $(x_3, y_3)$ becomes $(0, 0, 1)$. We can now seek real coefficients $\gamma_i \in \mathbb{R}$ of a function

$$g(\lambda_1, \lambda_2, \lambda_3) = \sum_{i=1}^{3} \gamma_i \lambda_i$$

such that $C = \{g = 0\}$. The coefficients are very easily described: $\gamma_i$ is the value of $g$ at $(x_i, y_i)$. Accordingly, if at least one $\gamma_i$ is positive and at least one is negative, $C$ will pass through $T$, intersecting at the edges between vertices with opposite signs. Thus, barycentric coordinates give us an easy way to define and control the shape of lines through a triangle.

To describe more complicated curves through $T$, we use a generalization of barycentric coordinates. Fix a degree $n \geq 1$ and compute the **trinomial expansion** of

$$(\lambda_1 + \lambda_2 + \lambda_3)^d = 1.$$

This will yield $\binom{d+2}{2}$ terms of the form $\lambda_1^i \lambda_2^j \lambda_3^k$ with $i + j + k = n$. These functions, called the **Bernstein polynomials**, form a basis for degree $n$ polynomials in $\mathbb{R}^2$ and can be used analogously to the linear case.(Which will be covered later) In standard coordinates, we could find real coefficients $c_{ij} \in \mathbb{R}$

$$g(x, y) = \sum_{i+j \leq n} c_{ij} x^i y^j$$

such that $C = \{g = 0\}$. This problem is much more difficult than the linear case, making the barycentric coordinate change essential. We seek instead the real **Bernstein-Bézier coefficients** $\gamma_{ijk} \in \mathbb{R}$ of the function

$$g(\lambda_1, \lambda_2, \lambda_3) = \sum_{i+j+k=n} \gamma_{ijk} \frac{d!}{i!j!k!} \lambda_1^i \lambda_2^j \lambda_3^k \tag{2.6}$$

such that $C = \{g = 0\}$. As with barycentric coordinates, the coefficient at a vertex of the triangle is exactly the value of $g$ at the vertex, for example

$$\gamma_{300} = g(1, 0, 0) = g(x_1, y_1).$$

The remaining coefficients control the properties of $g$ (and hence its level sets) within $T$. The coefficients are associated to the **domain points** on a regular subdivision of the triangle. We show examples of such subdivisions for $n = 2$ and $n = 3$ in Figure 2.2.

Besides triangulation coordinates, there are several other type of coordinates that we would like to introduce:

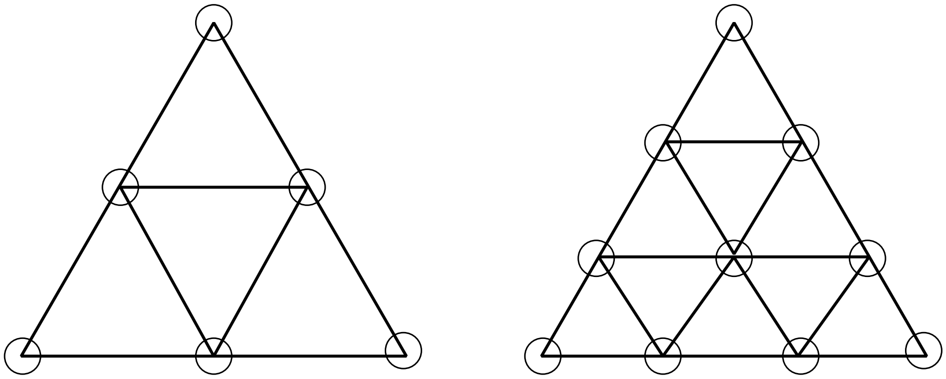


Figure 2.2: Domain points associated to Bernstein-Bézier coefficients for $n = 2$ (left) and $n = 3$ (right).

**Harmonic Coordinates**

**Mean Value Coordinates**

**Wachspress Coordinate**

## 2.5 Polynomials, Piecewise Polynomials, Splines

In this section, we will bring examples of basis from functional space that can be used to interpret geometric objects (in practice, i.e. are lines and surfaces). The bases are first defined for restricted subdomains of the defining space as opposed to the power basis which is defined for all points of the space. They are mostly compacted supported, not infinitively supported, but they have better properties other than simplest power basis. The example formulations given below are defined for values of each of the variables $x$, $y$ and $z$ in the unit interval [0,1]. We will introduce each the following.

### 2.5.1 Univariate case

**Bernstein-Bezier**

$$P(x) = \sum_{j=0}^{m} w_j B_j^m(x)$$

where

$$B_i^m(x) = \binom{m}{i} x^i (1-x)^{m-i}$$

**B-Spline**

The B-spline basis over the unit interval [0,1] is easily generated by a fractional linear recurrence as given below for the univariate case. The bivariate and trivariate forms can also be similarly generated from this in either tensor product or barycentric form. (See examples in BB form)

The univariate B-spline form is defined by linear combination of **control points** $\{\mathbf{p}_l\}_{l=0}^{n}$ :

$$\mathbf{P}_n = \sum_{l=0}^{m} \mathbf{p}_l N_l^n(x)$$

where $N_l^n(x)$ is defined via **knot sequence** $0 = u_0 \leq u_1 < \ldots < u_{m+1} = 1$ :

$$N_l^1(x) = \begin{cases} 1 \text{ for } \ u_l \leq u_{l+1} \\ 0 \text{ otherwise.} \end{cases}$$

$$N_l^n(x) = \frac{x - u_{l-1}}{u_{l+n-1} - u_{l-1}} N_l^{n-1}(x) + \frac{u_{l+n} - x}{u_{l+n} - u_l} N_{l+1}^{n-1}(x)$$

### 2.5.2 Bivariate case

**Tensor Product**

$$P(x, y) = \sum_{i=0}^{m} \sum_{j=0}^{n} w_{ij} B_i^n(x) B_j^n(y)$$

**Generalized Barycentric Coordinate on convex Polygon**

$$P(x, y) = \sum_{i=0}^{m} \sum_{j=0}^{m-i} w_{ij} B_{ij}^n(x, y)$$

where

$$B_{ij}^m(x, y) = \binom{m}{ij} x^i y^j (1 - x - y)^{m-i-j}$$

Here $(x, y) \to (x, y, 1 - x - y)$ is a naive mapping from world coordinate to the barycentric coordinate.

### 2.5.3 Multivariate case

**Tensor Product**

$$p(x_1, \ldots, x_d) = \sum_{i_1=0}^{n_1} \cdots \sum_{i_d=0}^{n_d} b_{i_1 i_2 \ldots i_d}\, B_{i_1}^{n_1}(t_1) B_{i_2}^{n_2}(t_2) \ldots B_{i_d}^{n_d}(t_d) \tag{2.7}$$

where

$$(x_1, \ldots, x_d)^T \in [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_d, b_d]$$

$$t_i = \tfrac{x_i - a_i}{b_i - a_i}, \quad i = 1, 2, \ldots, d$$

$$B_i^n(t) = \tfrac{n!}{i!(n-i)!}\, t^i (1 - t)^{n-i}$$

**Generalized Barycentric Coordinate on Simplex**

$$P(x, y, z) = \sum_{i=0}^{m} \sum_{j=0}^{m-i} \sum_{k=0}^{m-i-j} w_{ijk} B_{ijk}^m(x, y, z)$$

where

$$B_{ijk}^m(x, y, z) = \binom{m}{ijk} x^i y^j z^k (1 - x - y - z)^{m-i-j-k}$$

**Mixed Bernstein Form**

Take the simplest tetrahedron as the example:

$$P(x, y, z) = \sum_{i=0}^{m} \sum_{j=0}^{m-i} \sum_{k=0}^{p} \mathbf{b}_{ijk} B_{ij}^n(x, y) B_k^n(z)$$

Let $d = d_1 + d_2$, $\mathbf{p}_0, \ldots, \mathbf{p}_{d_1} \in \mathbb{R}^{d_1}$ be affine independent. Then the mixed Bernstein form is

$$p(x_1, \ldots, x_d) = \sum_{i_1+\ldots+i_{d_1} \le m} \sum_{j_1=0}^{n_1} \cdots \sum_{j_{d_2}=0}^{n_{d_2}} b_{i_1 \ldots i_{d_1} j_1 \ldots j_{d_2}} \, \tilde{B}^m_{i_1 \ldots i_{d_1}}(\alpha_1, \ldots, \alpha_{d_1}) \, B^{n_2}_{j_1}(t_1) \ldots B^{n_{d_2}}_{j_{d_2}}(t_{d_2}) \tag{2.8}$$

where

$$(x_1, \ldots x_d)^T \in [\mathbf{p}_0, \ldots, \mathbf{p}_{d_1}] \times [a_1, b_1] \times \ldots \times [a_{d_2}, b_{d_2}]$$

$$\begin{bmatrix} x_1 \\ \vdots \\ x_{d_1} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{p}_0 & \mathbf{p}_1 & \cdots & \mathbf{p}_{d_1} \\ 1 & 1 & \ldots & 1 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{d_1} \end{bmatrix}$$

and

$$t_i = \frac{x_{d_1+i} - a_i}{} b_i - a_i \qquad i = 1, 2, \ldots, d_2$$

If $d_1 = 0$, then $p$ is the Bernstein form on hypercube. If $d_2 = 0$, then $p$ is the Bernstein form on simplex.

## 2.6 Parametric and Implicit Representation

We will continue discuss two different representation based on definition and notation in 2.4 and 2.5.3 : what is parametric and what is implicit? How will they represent 2d and 3d objects (or in higher degree)? And how they are different from.

For notation simplicity, all function $f_i$ , no matter univariate,bivariate or multivariate, will be generalized as the linear combination we proposed in 2.5.3.

### 2.6.1 Curves

A real implicit algebraic plane curve $f(x, y) = 0$ is a hypersurface of dimension 1 in $\mathbb{R}^2$, while a parametric plane curve $[f_3(s)x - f_1(s) = 0, f_3(s)y - f_2(s) = 0]$ is an algebraic variety of dimension 1 in $\mathbb{R}^3$, defined by the two independent algebraic equations in the three variables $x, y, s$.

A plane parametric curve is a very special algebraic variety of dimension 1 in $x, y, s$ space, since the curve lies in the 2-dimensional subspace defined by $x, y$ and furthermore points on the curve can be put in $(1, 1)$ rational correspondence with points on the 1-dimensional sub-space defined by $s$. Parametric curves are thus a special subset of algebraic curves, and are often also called rational algebraic curves. Figure 2.3 depicts the relationship between the set of parametric curves and non-parametric curves at various degrees.

Example parametric (rational algebraic) curves are degree two algebraic curves (conics) and degree three algebraic curves (cubics) with a singular point. The non-singular cubics are not rational and are also known as elliptic cubics. In general, a necessary and sufficient condition for the rationality of an algebraic curve of arbitrary degree is given by the Cayley-Riemann criterion: a curve is rational if and only if $g = 0$, where $g$, the genus of the curve is a measure of the deficiency of the curve's singularities from its maximum allowable limit [217]. Algorithms for computing the genus of an algebraic curve and for symbolically deriving the parametric equations of genus 0 curves, are given for example in [1].

For implicit algebraic plane curves and surfaces defined by polynomials of degree $d$, the maximum number of intersections between the curve and a line in the plane or the surface and a line in space, is equal to the maximum number of roots of a polynomial of degree $d$. Hence, here the geometric degree is the same as the algebraic degree which is equal to $d$. For parametric curves defined by polynomials of degree $d$, the maximum number of intersections between the curve and a line in the plane is also equal to the maximum number of roots of a polynomial of degree $d$. Hence here again the geometric degree is the same as the algebraic degree.

For parametric surfaces defined by polynomials of degree $d$ the geometric degree can be as large as $d^2$, the square of the algebraic degree $d$. This can be seen as follows. Consider the intersection of a generic line in space $[a_1x + b_1y + c_1z - d_1 = 0, a_2x + b_2y + c_2z - d_2 = 0]$ with the parametric surface. The intersection yields two implicit algebraic curves of degree $d$ which intersect in $O(d^2)$ points (via Bezout's theorem), corresponding to the intersection points of the line and the parametric surface.

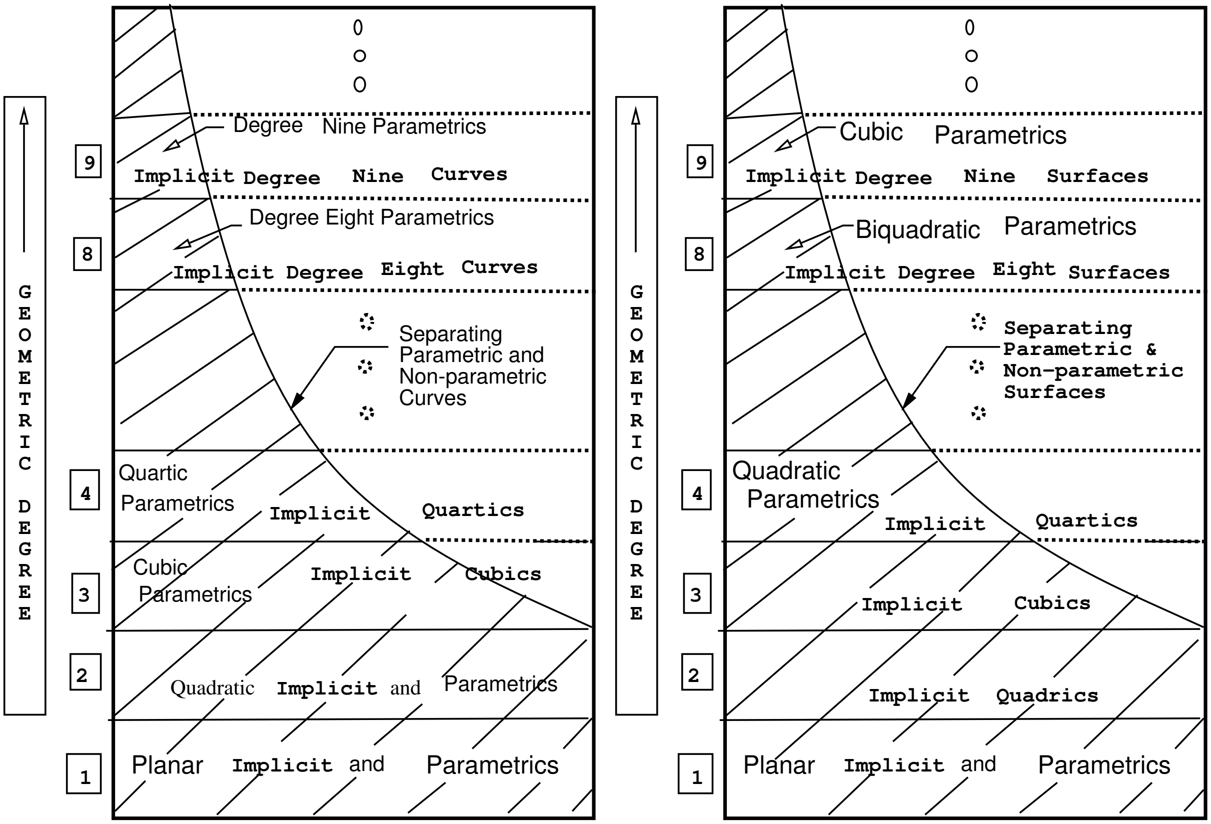

GEOMETRIC DEGREE
0
o
O
9
8
4
3
2
1
Degree Nine Parametrics
Implicit Degree Nine Curves
Degree Eight Parametrics
Implicit Degree Eight Curves
Separating Parametric and Non-parametric Curves
Quartic Parametrics
Implicit Quartics
Cubic Parametrics
Implicit Cubics
Quadratic Implicit and Parametrics
Planar Implicit and Parametrics
GEOMETRIC DEGREE
0
o
O
9
8
4
3
2
1
Cubic Parametrics
Implicit Degree Nine Surfaces
Biquadratic Parametrics
Implicit Degree Eight Surfaces
Separating Parametric & Non-parametric Surfaces
Quadratic Parametrics
Implicit Quartics
Implicit Cubics
Implicit Quadrics
Planar Implicit and Parametrics

Figure 2.3: A classification of low degree algebraic curves (left) and surfaces (right)

A parametric curve of algebraic degree $d$ is an algebraic curve of genus 0 and so have $\frac{(d-1)(d-2)}{2} = O(d^2)$ singular (double) points. This number is the maximum number of singular points an algebraic curve of degree $d$ may have. From Bezout's theorem, we realize that the intersection of two implicit surfaces of algebraic degree $d$ can be a curve of geometric degree $O(d^2)$. Furthermore the same theorem implies that the intersection of two parametric surfaces of algebraic degree $d$ (and geometric degree $O(d^2)$) can be a curve of geometric degree $O(d^4)$. Hence, while the potential singularities of the space curve defined by the intersection of two implicit surfaces defined by polynomials of degree $d$ can be as many as $O(d^4)$, the potential singularities of the space curve defined by the intersection of two parametric surfaces defined by polynomials of degree $d$ can be as many as $O(d^8)$.

Let $\mathbf{C} : (f_1(x,y,z) = 0, f_2(x,y,z) = 0)$ implicitly define an irreducible algebraic space curve of degree $d$. The irreducibility of the curve is not really a restriction, since reducible curves can be handled similarly by treating each irreducible component in turn. The situation is slightly more complicated if in the real setting, we may wish to achieve separate containment of each real component of an irreducible curve. We defer a solution to this problem, and for the time being consider it reduced to the problem of choosing appropriate clipping surfaces to isolate that real component, after the interpolated surface is computed. Note for parametrically defined curves, this problem does not arise.

### 2.6.2 Surface

Similarly, a real implicit algebraic surface $f(x,y,z) = 0$ is a hypersurface of dimension two in $\mathbb{R}^3$, while a parametric surface $[f_4(s,t)x - f_1(s,t) = 0, f_4(s,t)y - f_2(s,t) = 0, f_4(s,t)z - f_3(s,t) = 0]$ is an algebraic variety of dimension 2 in $\mathbb{R}^5$, defined by three independent algebraic equations in the five variables $x, y, z, s, t$.

When a curve is given in rational parametric form, its equations can be used directly to produce a linear system for interpolation, instead of first computing $nd + 1$ points on the curve. Let $\mathbf{C} : (x = G_1(t), y = G_2(t), z = G_3(t))$ be a rational curve of degree $d$. An interpolating surface $S : f(x,y,z) = 0$ of degree $n$ which contains $C$ is computed as follows:

B1. Substitute $(x = G_1(t), y = G_2(t), z = G_3(t))$ into the equation $f(x,y,z) = 0$.

B2. Simplify and rationalize to obtain $Q(t) = 0$, where $Q$ is a polynomial in $t$, of degree at most $nd$, and with coefficients which are linear expressions in the coefficients of $f$. For $Q$ to be identically zero, each of its coefficents must be zero, and hence we obtain a system of at most $nd + 1$ linear equations, where the unknowns are the coefficients of $f$. Any non-trivial solution of this linear system will represent a surface $S$ which interpolates $\mathbf{C}$.

The proof of correctness of the algorithm follows from the lemma below.

**Lemma 2.29.** *The containment condition is satisfied by step 2. of the above algorithm*

**Proof**: We omit this here and refer the reader to the full paper. □

**Parametric Curves**

**Parametric Surface**

### 2.6.3 Examples

**Conics**

The general conic implicit equation is given by

$$C(x,y) = ax^2 + by^2 + cxy + dx + ey + f = 0.$$

The non-trivial case in converting this to a rational parameterization arises when $a$ and $b$ are both non-zero. Otherwise one already has one variable ( $x$, or $y$) in linear form and expressible as a rational polynomial expression of the other, and hence a rational parameterization. This then suggests that to obtain a rational parameterization all we need to do is to make $C(x,y)$ non-regular in $x$ or $y$. That is, eliminate the $x^2$ or the $y^2$ term through a coordinate transformation. For then one of the variables is again in linear form and is expressible as a rational polynomial expression of the other. We choose to eliminate the $y^2$ term, by an appropriate coordinate transformation applied to $C(x,y)$. This is always possible and the algorithm is now described below. (The entire algorithm which also handles all trivial and degenerate cases of the conic is implemented on a VAX-780 using VAXIMA.)

Geometrically speaking, a conic being irregular in $x$ or $y$ means that most lines parallel to the $x$ or $y$ axis respectively, intersect the curve in one point. Also, most lines through a point $(b_1, b_2)$ on the conic meet the conic in one additional point. By sending this point ( $b_1, b_2$) to infinity we make all these lines parallel to some axis and the curve irregular in one of the variables ( $x$, or $y$) and hence amenable to parameterization. The coordinate transformation we select is thus one which sends the point ( $b_1, b_2$) on the conic to infinity. The rational parameterization we obtain is global, of degree at most 2 and with parameter $t$ corresponding to the slopes of the lines through the point $(b_1, b_2)$ on the conic. Further $t$ ranges from $(-\infty, \infty)$ and covers the entire curve. The selection of the point ( $b_1, b_2$) on the conic becomes important and may be made appropriately, when the parameterization is desired only for a specific piece of the conic.

**Step (1)** If $C(x, y)$ has a real root at infinity, a *linear* transformation of the type $x' = a_1x + b_1y + c_1$ and $y' = a_2x + b_2y + c_2$ will suffice. If $C(x, y)$ has no real root at infinity, we must use a *linear transformation* of the type $x'$ = ( $a_1x + b_1y + c_1$)/( $a_3x + b_3y + c_3$) and $y' = a_2x + b_2y + c_2$)/( $a_3x + b_3y + c_3$). This is equivalent to a *homogeneous linear transformation* of the type $X' = a_1X + b_1Y + c_1H$, $Y' = a_2X + b_2Y + c_2H$ and $H' = a_3X + b_3Y + c_3H$ applied to the homogeneous conic $C(X, Y, H) = aX^2 + bY^2 + cXY + dXH + eYH + fH^2 = 0$.

**Step (2)** Points at infinity for $C(x, y)$ are given by the linear factors of the *degree form* (highest degree terms) of $I$. For the conic this corresponds to a real root at infinity if $c^2 \;>=\; 4ab$, (e.g. parabolas and hyperbolas). For otherwise both roots at infinity are complex , (complex roots arise in conjugate pairs). Further for $c^2 \;=\; 4ab$, (e.g. parabolas), the degree form is a perfect square and this gives a polynomial parameterization for the curve.

**Step (3)** Applying a linear transformation for $c^2 \;>=\; 4ab$, gives rise to $C(x', y') = I($ $a_1x + b_1y + c_1, a_2x + b_2y + c_2$). To eliminate the $y^2$ term we need to choose $b_1$ and $b_2$ such that $ab_1^2 \;+\; cb_1b_2 \;+\; bb_2^2 \;=\; 0$. Here both the values of $b_1$ and $b_2$ can always be chosen to be real.

**Step (4)** Applying a homogeneous linear transformation for $c^2 \;<\; 4ab$, gives rise to $C(X', Y', H') = C(a_1X + b_1Y + c_1H$, $a_2X + b_2Y + c_2H, a_3X + b_3Y + c_3H)$. To eliminate the $Y^2$ term we need to choose $b_1$, $b_2$ and $b_3$ such that $ab_1^2 + bb_2^2 + cb_1b_2 + db_1b_3 + eb_2b_3 + fb_3^2 = 0$. This is equivalent to finding a point ( $b_1$, $b_2$, $b_3$) on the homogeneous conic. The values of $b_1$ and $b_2$ are both real if $(cd - 2ae)$ is not less than the $geometric\ mean$ of $4af \;-\; d^2$ and $4ab \;-\; c^2$, or alternatively $(ce - 2bd)$ is not less than the $geometric\ mean$ of $4bf \;-\; e^2$ and $4ab \;-\; c^2$.

**Step (5)** Finally choose the remaining coefficients $a_i$' $s$, $c_i$' $s$, ensuring that the appropriate transformation is well defined. In the case of a linear transformation, this corresponds to ensuring that the matrix

$$\begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}$$

is non-singular. Hence $c_i$' $s$ can be chosen to be 0 and $a_1 = 1$, $a_2 = 0$. In the case of a homogeneous linear transformation, one needs to ensure that the matrix

$$\begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix}$$

is non-singular. Here $a_1 = 1$, $c_2 = 1$ and the rest set to 0 suffices. These remaining coefficients provide a measure of local control for the curve and may also be chosen in a way that gives specific local parameterizations for pieces of the curve, appropriate for particular applications.

### Conicoids

The case of the conicoid is a generalization of the method of the conic. The general conicoid implicit equation is given by $C(x, y, z) = ax^2 + by^2 + cz^2 + dxy + exz + fyz + gx + hy + iz + j = 0$. Again the main case of concern is when $a$, $b$ and $c$ are all non-zero. Otherwise one already has one of the variables ( $x$, $y$, or $z$) in linear form and expressible as a rational polynomial expression of the other two. This then suggests that to obtain the rational parameterization all we need to do again is to make $C(x, y, z)$ *non-regular* in say, $y$. That is, eliminate the $y^2$ term through a coordinate transformation. For then $y$ is in linear form

and is expressible as a rational polynomial expression of the other two. We eliminate the $y^2$ term by an appropriate coordinate transformation applied to $C(x, y, z)$. This is always possible and the algorithm is now described below. (The entire algorithm which also handles all trivial and degenerate cases of the conicoid is implemented on a VAX-780 using VAXIMA. )
Geometrically speaking, a conicoid being irregular in $x$, $y$ or $z$ means that most lines parallel to the $x$, $y$ or $z$ axis respectively, intersect the surface in one point. Also, most lines through a point $(b_1, b_2, b_3)$ on the conicoid meet the conicoid in one additional point. By sending this point ( $b_1, b_2, b_3$) to infinity we make all these lines parallel to some axis and the surface irregular in one of the variables ( $x$, or $y$) and hence amenable to parameterization. The coordinate transformation we select is thus one which sends the point ( $b_1, b_2, b_3$) on the conicoid to infinity. The rational parameterization we obtain is global, of degree at most 2 and with parameters $s$ and $t$ corresponding to the ratio of the direction cosines of the lines through the point $(b_1, b_2, b_3)$ on the conicoid. Further $s$ and $t$ both range from $(-\infty, \infty)$ and cover the entire surface. The selection of the point ( $b_1, b_2, b_3$) on the conicoid becomes important and may be made appropriately, when the parameterization is desired only for a specific patch of the conicoid.

**Step (1)** If $C(x, y, z)$ has a real root at infinity, a*linear transformation* of the type $x' = a_1x + b_1y + c_1z + d_1$, $y' = a_2x + b_2y + c_2z + d_2$ and $z' = a_3x + b_3y + c_3z + d_3$ will suffice. If $C(x, y, z)$ has no real root at infinity, we must use a *linear transformation* of the type $x'$ = ( $a_1x + b_1y + c_1z + d_1$)/( $a_4x + b_4y + c_4z + d_4$), $y' = a_2x + b_2y + c_2z + d_2$)/( $a_4x + b_4y + c_4z + d_4$). and $z'$ = ( $a_3x + b_3y + c_3z + d_3$)/( $a_4x + b_4y + c_4z + d_4$). This is equivalent to a *homogeneous linear transformation* of the type $X' = a_1X + b_1Y + c_1Z + d_1H$, $Y' = a_2X + b_2Y + c_2Z + d_2H$, $Z' = a_3X + b_3Y + c_3Z + d_3H$ and $H' = a_4X + b_4Y + c_4Z + d_4H$ applied to the homogeneous conicoid $C(X, Y, Z, H) = aX^2 + bY^2 + cZ^2 + dXY + eXZ + fYZ + gXH + hYH + iZH + jH^2 = 0$,

**Step (2)** Points at infinity for $C(x, y)$ are given by the linear factors of the *degree form* (highest degree terms) of $I$. For the conicoid this corresponds to the roots of the homogeneous conic equation $C(x, y, z) = ax^2 + by^2 + dxy + exz + fyz + cz^2 = 0$. Also, here the simultaneous truth of $d^2 = 4ab$, $e^2 = 4ac$ and $f^2 = 4bc$ corresponds to the existence of a polynomial parameterization for the conicoid, as then the degree form is a perfect square.

**Step (3)** Apply a linear transformation if a real root $(r_x, r_y, r_z)$ exists for the homogeneous conic $C(x, y, z)$ of (2). This gives rise to $C(x', y', z') = I($ $a_1x + b_1y + c_1z + d_1, a_2x + b_2y + c_2z + d_2, a_3x + b_3y + c_3z + d_3$). To eliminate the $y^2$ term we can take ( $b_1$, $b_2$, $b_3$) = ( $r_x, r_y, r_z$), the real point on $C(x, y, z)$.

**Step (4)** Apply a homogeneous linear transformation if only complex roots exist for the homogeneous conic $C(x, y, z)$ of (2). This gives rise to $C(X', Y', Z', H') = I($ $a_1X + b_1Y + c_1Z + d_1H, a_2X + b_2Y + c_2Z + d_2H, a_3X + b_3Y + c_3Z + d_3H, a_4X + b_4Y + c_4Z + d_4H$). To eliminate the $Y^2$ term we choose $b_4 = 1$ and ( $b_1$, $b_2$) to be a point on either the conic $ax^2 + by^2 + dxy + gxz + hyz + jz^2 = 0$ with $b_3 = 0$ or a point on the conic $ax^2 + by^2 + dxy + (e + g)xz + (f + h)yz + (c + i + j)z^2 = 0$ with $b_3 = 1$. Real values exist for $b_1$ and $b_2$ if there exists a real point on either of the above conics.

**Step (5)** Finally choose the remaining coefficients $a_i$' $s$, $c_i$' $s$, and $d_i$' $s$, ensuring that the appropriate transformation is well defined. In the case of a linear transformation, this corresponds to ensuring that the matrix

$$\begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix}$$

is non-singular. Here the $d_i$' $s$ can be chosen to be $0$. Further $a_2 = 1$, $c_3 = 1$ if $b_1$ is non-zero or else $a_1 = 1$, $c_3 = 1$ if $b_2$ is non-zero or else $a_1 = 1$, $c_2 = 1$, with the rest set to $0$. In the case of a homogeneous linear transformation one needs to ensure that the matrix

$$\begin{pmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ a_4 & b_4 & c_4 & d_4 \end{pmatrix}$$

is non-singular. Here $a_1 = 1$, $c_3 = 1$, $d_2 = 1$ with the rest set to 0 suffices. These remaining coefficients provide a measure of local control for the surface and may also be chosen in a way that gives specific local parameterizations for pieces of the surface, appropriate for particular applications.

## 2.7 Finite Elements and Error Estimation

### 2.7.1 Tensor Product Over The Domain: Irregular Triangular Prism

**Definition**

Given the triangulation mesh $\mathcal{T}$, let $[\mathbf{v}_i\mathbf{v}_j\mathbf{v}_k]$ be one of the triangles where $\mathbf{v}_i,\ \mathbf{v}_j,\ \mathbf{v}_k$ are the vertices of the triangle. Suppose the unit normals of the surface at the vertices are also known, denoted as $\mathbf{n}_l,\ (l = i,j,k)$. Let $\mathbf{v}_l(\lambda) = \mathbf{v}_l + \lambda\mathbf{n}_l$. First we define a prism (Figure 2.4) $D_{ijk} := \{\mathbf{p}:\ \mathbf{p} = b_1\mathbf{v}_i(\lambda) + b_2\mathbf{v}_j(\lambda) + b_3\mathbf{v}_k(\lambda),\ \lambda \in I_{ijk}\}$, where $(b_1, b_2, b_3)$ are the barycentric coordinates of points in $[\mathbf{v}_i\mathbf{v}_j\mathbf{v}_k]$, and $I_{ijk}$ is a maximal open interval containing 0 and for any $\lambda \in I_{ijk}$, $\mathbf{v}_i(\lambda)$, $\mathbf{v}_j(\lambda)$, $\mathbf{v}_k(\lambda)$ are not collinear and $\mathbf{n}_i$, $\mathbf{n}_j$, $\mathbf{n}_k$ point to the same side of the plane $P_{ijk}(\lambda) := \{\mathbf{p}:\ \mathbf{p} = b_1\mathbf{v}_i(\lambda) + b_2\mathbf{v}_j(\lambda) + b_3\mathbf{v}_k(\lambda)\}$. Next we define a function in the Bernstein-Bezier (BB) basis over the prism $D_{ijk}$:

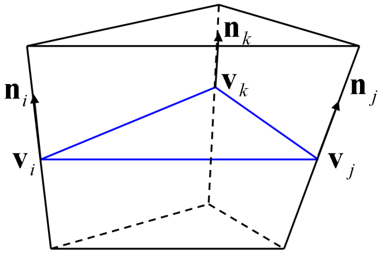


Figure 2.4: A prism $D_{ijk}$ constructed based on the triangle $[\mathbf{v}_i\mathbf{v}_j\mathbf{v}_k]$.

$$F(b_1, b_2, b_3, \lambda) = \sum_{i+j+k=n} b_{ijk}(\lambda) B^n_{ijk}(b_1, b_2, b_3), \tag{2.9}$$

where $B^n_{ijk}(b_1, b_2, b_3)$ is the Bezier basis

$$B^n_{ijk}(b_1, b_2, b_3) = \frac{n!}{i!j!k!} b_1^i b_2^j b_3^k.$$

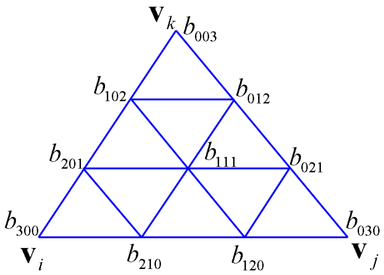


Figure 2.5: The control coefficients of the cubic Bezier basis of function $F$.

### Non-degenerancy

(yiwang) : I am not sure if the subsection title is correct Let $p_{ijk}^{(l)}(\lambda) = \det[n_l, v_j(\lambda) - v_i(\lambda), v_k(\lambda) - v_i(\lambda)], \quad l = i, j, k$. Assume

$$p_{ijk}^{(l)}(\lambda) > 0, \qquad \forall \lambda \in [0,1], \qquad l = i, j, k \; . \tag{2.10}$$

Consider the real numbers $\lambda_1, \cdots, \lambda_s$ $(s \leq 6)$ that solve one of these three equations of degree 2: $p_{ijk}^{(l)}(\lambda) = 0, \ \ l = i, j, k$, and define $a = \max(-\infty, \{\lambda_l : \lambda_l < 0\})$, $b = \min(+\infty, \{\lambda_l : \lambda_l > 1\})$, and $I_{ijk} = (a, b)$. Then $I_{ijk}$ is the largest interval containing $[0,1]$ such that $P_{ijk}(I_{ijk})$ is non-degenerate. To show this fact, note that a triangle $T_{ijk}(\lambda)$ is non-degenerate if and only if

$$n_l^T [v_j(\lambda) - v_i(\lambda)] \times [v_k(\lambda) - v_i(\lambda)] = p_{ijk}^{(l)}(\lambda) > 0, \tag{2.11}$$

$l = i, j, k$, where × denotes the cross product of two vectors. The assumption ((2.10)) implies that $[0,1] \subset I$. Since $p_{ijk}^{(l)}(0) > 0$ and $p_{ijk}^{(l)}(1) > 0$, for $l = i, j, k$, then $p_{ijk}^{(l)}(\lambda) > 0$ for $\lambda \in (a, b)$ and $l = i, j, k$. Since $p_{ijk}^{(l)}(a) = 0$ for $l = i$ or $l = j$ or $l = k$ if $a > -\infty$, $a$ is the infimum of the interval of $\lambda$ that contains $[0,1]$ and makes (2.11) hold. Similarly, $b$ is the supremum of such an interval. Therefore $I_{ijk}$ is the largest interval such that $P_{ijk}(I_{ijk})$ is non-degenerate.

### Smoothness

We can then approximate the given surface by the zero contour of $F$, denoted as $S$. In order to make $S$ smooth, the degree of the Bezier basis $n$ should be no less than 3. For simplicity, here we consider the case of $n = 3$. The control coefficients $b_{ijk}(\lambda)$ should be properly defined such that $S$ is continuous. In Figure 2.5 we show the relationship of the control coefficients and the points of the triangle when $n = 3$. Next we are going to discuss these coefficients are defined.
Since $S$ passes through the vertices $\mathbf{v}_i, \ \mathbf{v}_j, \ \mathbf{v}_k$, we define

$$b_{300} = b_{030} = b_{003} = \lambda. \tag{2.12}$$

Next we are going to define the coefficients on the edges of the triangle in Figure 2.5. To obtain $C^1$ continuity at $\mathbf{v}_i$, we require that the directional derivatives of $F$ at $\mathbf{v}_i$ in the direction of $b_2$ and $b_3$ are equal to $\nabla F \cdot (\mathbf{v}_j - \mathbf{v}_i)$ and $\nabla F \cdot (\mathbf{v}_k - \mathbf{v}_i)$, respectively. Noticing that F has the form of (2.9) and $(b_1, b_2, b_3) = (1, 0, 0)$ at $\mathbf{v}_i$, one can derive that $b_{210} - b_{300} = \frac{1}{3}\nabla F(v_i) \cdot (\mathbf{v}_j(\lambda) - \mathbf{v}_i(\lambda))$, where $\nabla F(v_i) = \mathbf{n}_i$. Therefore

$$b_{210} = \lambda + \frac{1}{3}\mathbf{n}_i \cdot (\mathbf{v}_j(\lambda) - \mathbf{v}_i(\lambda)). \tag{2.13}$$

$b_{120}, b_{201}, b_{102}, b_{021}, b_{012}$ are defined similarly.
To obtain the $C^1$ continuity at the midpoints of the edges of $\mathcal{T}$, we define $b_{111}$ by using the side-vertex scheme [**?**]:

$$b_{111} = w_1 b_{111}^{(1)} + w_2 b_{111}^{(2)} + w_3 b_{111}^{(3)}, \tag{2.14}$$

where

$$w_i = \frac{b_j^2 b_k^2}{b_2^2 b_3^2 + b_1^2 b_3^2 + b_1^2 b_2^2}, \qquad i = 1, 2, 3, \ i \neq j \neq k.$$

Next we are going to define $b_{111}^{(1)}$, $b_{111}^{(2)}$ and $b_{111}^{(3)}$. In Appendix **??** we prove that our scheme of defining this three coefficients can guarantee the $C^1$ continuity at the midpoints of the edges $\mathbf{v}_j\mathbf{v}_k$, $\mathbf{v}_i\mathbf{v}_k$ and $\mathbf{v}_i\mathbf{v}_j$. Consider the edge $\mathbf{v}_i\mathbf{v}_j$. Recall that any point $\mathbf{p} = (x, y, z)$ in $D_{ijk}$ can be represented by

$$(x, y, z)^{\mathrm{T}} = b_1\mathbf{v}_i(\lambda) + b_2\mathbf{v}_j(\lambda) + b_3\mathbf{v}_k(\lambda). \tag{2.15}$$

Therefore differentiating both sides of (2.15) with respect to $x$, $y$ and $z$, respectively, yields

$$I_3 = \begin{pmatrix} \frac{\partial b_1}{\partial x} & \frac{\partial b_2}{\partial x} & \frac{\partial \lambda}{\partial x} \\ \frac{\partial b_1}{\partial y} & \frac{\partial b_2}{\partial y} & \frac{\partial \lambda}{\partial y} \\ \frac{\partial b_1}{\partial z} & \frac{\partial b_2}{\partial z} & \frac{\partial \lambda}{\partial z} \end{pmatrix} \begin{pmatrix} (\mathbf{v}_i(\lambda) - \mathbf{v}_k(\lambda))^{\mathrm{T}} \\ (\mathbf{v}_j(\lambda) - \mathbf{v}_k(\lambda))^{\mathrm{T}} \\ (b_1\mathbf{n}_i + b_2\mathbf{n}_j + b_3\mathbf{n}_k)^{\mathrm{T}} \end{pmatrix}, \tag{2.16}$$

where $I_3$ is a $3 \times 3$ unit matrix. Denote

$$M := \begin{pmatrix} (\mathbf{v}_i(\lambda) - \mathbf{v}_k(\lambda))^{\mathrm{T}} \\ (\mathbf{v}_j(\lambda) - \mathbf{v}_k(\lambda))^{\mathrm{T}} \\ (b_1\mathbf{n}_i + b_2\mathbf{n}_j + b_3\mathbf{n}_k)^{\mathrm{T}} \end{pmatrix}, \tag{2.17}$$

and let $A = \mathbf{v}_i(\lambda) - \mathbf{v}_k(\lambda)$, $B = \mathbf{v}_j(\lambda) - \mathbf{v}_k(\lambda)$ and $C = b_1\mathbf{n}_i + b_2\mathbf{n}_j + b_3\mathbf{n}_k$, then $M = (A\ B\ C)^{\mathrm{T}}$. From (2.16) we have

$$\begin{pmatrix} \frac{\partial b_1}{\partial x} & \frac{\partial b_2}{\partial x} & \frac{\partial \lambda}{\partial x} \\ \frac{\partial b_1}{\partial y} & \frac{\partial b_2}{\partial y} & \frac{\partial \lambda}{\partial y} \\ \frac{\partial b_1}{\partial z} & \frac{\partial b_2}{\partial z} & \frac{\partial \lambda}{\partial z} \end{pmatrix} = M^{-1} = \frac{1}{\det(M)}\left(B \times C,\ C \times A,\ A \times B\right). \tag{2.18}$$

According to (2.9), at the midpoint of $\mathbf{v}_i\mathbf{v}_j$, $(b_1, b_2, b_3) = (\frac{1}{2}, \frac{1}{2}, 0)$, we have

$$\begin{pmatrix} \frac{\partial F}{\partial b_1} \\ \frac{\partial F}{\partial b_2} \\ \frac{\partial F}{\partial \lambda} \end{pmatrix} = \begin{pmatrix} (\mathbf{v}_i(\lambda) - \mathbf{v}_k(\lambda))^{\mathrm{T}} \\ (\mathbf{v}_j(\lambda) - \mathbf{v}_k(\lambda))^{\mathrm{T}} \\ (\mathbf{n}_i + \mathbf{n}_j)^{\mathrm{T}}/2 \end{pmatrix} \left(\frac{\mathbf{n}_i + \mathbf{n}_j}{4}\right) + \begin{pmatrix} \frac{3}{2}(b_{210} - b_{111}) \\ \frac{3}{2}(b_{120} - b_{111}) \\ \frac{1}{2} \end{pmatrix}.$$

By (2.14), at $(b_1, b_2, b_3) = (\frac{1}{2}, \frac{1}{2}, 0)$ we have $b_{111} = b_{111}^{(3)}$. Therefore the gradient at $(\frac{1}{2}, \frac{1}{2}, 0)$ is

$$\begin{aligned} \nabla F &= M^{-1}(\frac{\partial F}{\partial b_1}, \frac{\partial F}{\partial b_2}, \frac{\partial F}{\partial \lambda})^{\mathrm{T}} \\ &= \frac{\mathbf{n}_i + \mathbf{n}_j}{4} + \frac{1}{2\det(M)}[3(b_{210} - b_{111}^{(3)})B \times C +\ 3(b_{120} - b_{111}^{(3)})C \times A + A \times B] \end{aligned} \tag{2.19}$$

Define vectors

$$\begin{aligned} &\mathbf{d}_1(\lambda) = \mathbf{v}_j(\lambda) - \mathbf{v}_i(\lambda) = B - A, \\ &\mathbf{d}_2(b_1, b_2, b_3) = b_1\mathbf{n}_i + b_2\mathbf{n}_j + b_3\mathbf{n}_k = C, \\ &\mathbf{d}_3(b_1, b_2, b_3, \lambda) = \mathbf{d}_1 \times \mathbf{d}_2 = B \times C + C \times A. \end{aligned} \tag{2.20}$$

Let

$$\mathbf{c} = C(\frac{1}{2}, \frac{1}{2}, 0), \tag{2.21}$$

$$\mathbf{d}_3(\lambda) = \mathbf{d}_3(\frac{1}{2}, \frac{1}{2}, 0, \lambda) = B \times \mathbf{c} + \mathbf{c} \times A. \tag{2.22}$$

Let $\nabla F = \nabla F(\frac{1}{2}, \frac{1}{2}, 0)$. In order to have $C^1$ continuity at $(\frac{1}{2}, \frac{1}{2}, 0)$, we should have $\nabla F \cdot \mathbf{d}_3(\lambda) = 0$. Therefore, by (2.19) and (2.22), we have

$$b_{111}^{(3)} = \frac{\mathbf{d}_3(\lambda)^{\mathrm{T}}(3b_{210}B \times \mathbf{c} + 3b_{120}\mathbf{c} \times A + A \times B)}{3\|\mathbf{d}_3(\lambda)\|^2}. \tag{2.23}$$

Similarly, we may define $b_{111}^{(1)}$ and $b_{111}^{(2)}$.

Now the function $F(b_1, b_2, b_3, \lambda)$ is well defined. The next step is to extract the zero level set $S$. Given the barycentric coordinates $(b_1, b_2, b_3)$ of a point in the triangle $[\mathbf{v}_i\mathbf{v}_j\mathbf{v}_k]$, we find the corresponding $\lambda$ by solving the equation $F(b_1, b_2, b_3, \lambda) = 0$ for $\lambda$ and this could be done by the Newton's method. Then we may get the corresponding point on $S$ as

$$(x,\ y,\ z)^{\mathrm{T}} = b_1\mathbf{v}_i(\lambda) + b_2\mathbf{v}_j(\lambda) + b_3\mathbf{v}_k(\lambda). \tag{2.24}$$

We have the following property for surface $S$:

*Theorem* 2.7.1. $S$ is $C^1$ at the vertices of $\mathcal{T}$ and the midpoints of the edges of $\mathcal{T}$.

*Theorem* 2.7.2. $S$ is $C^1$ everywhere if every edge $\mathbf{v}_i\mathbf{v}_j$ of $\mathcal{T}$ satisfies $\mathbf{n}_i \cdot (\mathbf{v}_i - \mathbf{v}_j) = \mathbf{n}_j \cdot (\mathbf{v}_j - \mathbf{v}_i)$.

*Theorem* 2.7.3. $S$ is $C^1$ everywhere if the unit normals at the vertices of $\mathcal{T}$ are the same.

Proofs of the theorems are shown in the [240].

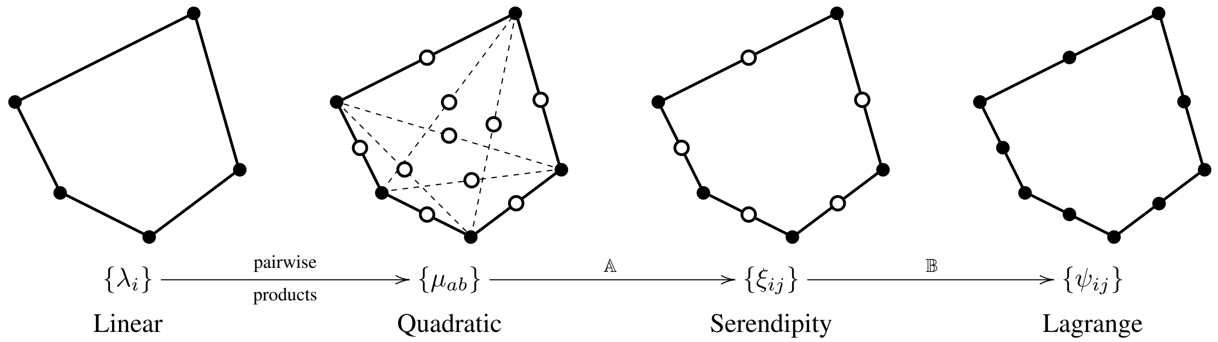


Figure 2.6: Overview of the construction process. In each figure, the dots are in one-to-one correspondence with the set of functions listed below it. At filled dots, all functions in the set evaluate to zero except for the function corresponding to the dot which evaluates to one. The rightmost element has quadratic precision with only these types of 'Lagrange-like' basis functions.

### 2.7.2 Generalized Barycentric Coordinate and Serendipity Elements

Barycentric coordinates provide a basis for linear finite elements on simplices, and generalized barycentric coordinates naturally produce a suitable basis for linear finite elements on general polygons. Various applications make use of this technique [98, 99, 151, 153, 172, 193, 203, 204, 207, 221], but in each case, only linear error estimates can be asserted. A quadratic finite element can easily be constructed by taking pairwise products of the basis functions from the linear element, yet this approach has not been pursued, primarily since the requisite number of basis functions grows quadratically in the number of vertices of the polygon. Still, many of the pairwise products are zero along the entire polygonal boundary and thus are unimportant for inter-element continuity, a key ingredient in finite element theory. For quadrilateral elements, these 'extra' basis functions are well understood and, for quadrilaterals that can be affinely mapped to a square, the so-called 'serendipity element' yields an acceptable basis consisting of only those basis functions needed to guarantee inter-element continuity [242, 13, 12]. We generalize this construction to produce a quadratic serendipity element for a class of shape-regular convex polygons derived from generalized barycentric coordinates.

Our construction yields a set of Lagrange-like basis functions $\{\psi_{ij}\}$ – one per vertex and one per edge midpoint – using a linear combination of pairwise products of generalized barycentric functions $\{\lambda_i\}$. We show that this set spans all constant, linear, and quadratic polynomials, making it suitable for finite element analysis via the Bramble-Hilbert lemma. Further, given uniform bounds on the aspect ratio, edge lengths, and interior angles of the polygon, we bound $||\psi_{ij}||_{H^1(\Omega)}$ uniformly with respect to $||\lambda_i||_{H^1(\Omega)}$. Since our previous work shows that $||\lambda_i||_{H^1(\Omega)}$ is bounded uniformly under these geometric hypotheses for typical definitions of $\lambda_i$ [100, 170], this proves that the $\psi_{ij}$ functions are well-behaved.

Figure 2.6 gives a visual depiction of the construction process. Starting with one generalized barycentric function $\lambda_i$ per vertex of an $n$-gon, take all pairwise products yielding a total of $n(n+1)/2$ functions $\mu_{ab} := \lambda_a\lambda_b$. The linear transformation $\mathbb{A}$ reduces the set $\{\mu_{ab}\}$ to the $2n$ element set $\{\xi_{ij}\}$, indexed over vertices and edge midpoints of the polygon. A simple bounded linear transformation $\mathbb{B}$ converts $\{\xi_{ij}\}$ into a basis $\{\psi_{ij}\}$ which satisfies the "Lagrange property" meaning each function takes the value 1 at its associated node and 0 at all other nodes.

The paper is organized as follows. In Section 2.7.2 we review relevant background on finite element theory, serendipity elements, and generalized barycentric functions. In Section 2.7.2, we show that if the entries of matrix $\mathbb{A}$ satisfy certain linear constraints Q$_c$1-Q$_c$3, the resulting set of functions $\{\xi_{ij}\}$ span all constant, linear and quadratic monomials in two variables, a requirement for quadratic finite elements. In Section 2.7.2, we show how the constraints Q$_c$1-Q$_c$3 can be satisfied in the special cases of the unit square, regular polygons, and convex quadrilaterals. In Section 2.7.2, we show how Q$_c$1-Q$_c$3 can be satisfied on a simple convex polygon. We also prove that the resulting value of $||\mathbb{A}||$ is bounded uniformly, provided the convex polygon satisfies certain geometric quality conditions. In Section 2.7.2 we define $\mathbb{B}$ and show that the final $\{\psi_{ij}\}$ basis is Lagrange-like. Finally, in Section 2.7.2, we describe practical applications, give numerical evidence, and consider future directions.

#### Background and Notation

Let $\Omega$ be a convex polygon with $n$ vertices $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ ordered counter-clockwise. Denote the interior angle at $\mathbf{v}_i$ by $\beta_i$. The largest distance between two points in $\Omega$ (the diameter of $\Omega$) is denoted diam$(\Omega)$ and the radius of the largest inscribed circle

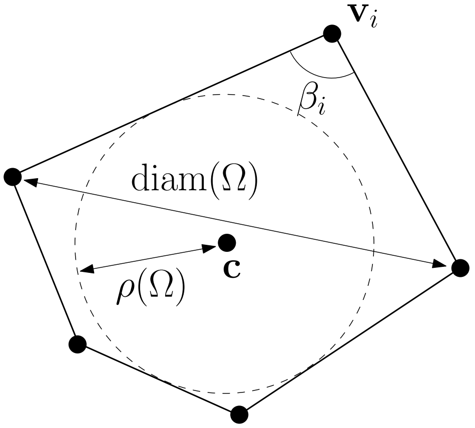


Figure 2.7: Notation used to describe polygonal geometry.

is denoted $\rho(\Omega)$. The center of this circle is denoted **c** and is selected arbitrarily when no unique circle exists. The **aspect ratio** (or chunkiness parameter) $\gamma$ is the ratio of the diameter to the radius of the largest inscribed circle, i.e.

$$\gamma := \frac{\text{diam}(\Omega)}{\rho(\Omega)}.$$

The notation is shown in Figure 2.7.

For a multi-index $\alpha = (\alpha_1, \alpha_2)$ and point $\mathbf{x} = (x, y)$, define $\mathbf{x}^\alpha := x^{\alpha_1} y^{\alpha_2}$, $\alpha! := \alpha_1 \alpha_2$, $|\alpha| := \alpha_1 + \alpha_2$, and $D^\alpha u := \partial^{|\alpha|} u / \partial x^{\alpha_1} \partial y^{\alpha_2}$. The Sobolev semi-norms and norms over an open set $\Omega$ for a non-negative integer $m$ are defined by

$$|u|^2_{H^m(\Omega)} := \int_\Omega \sum_{|\alpha|=m} |D^\alpha u(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x} \qquad \text{and} \qquad ||u||^2_{H^m(\Omega)} := \sum_{0 \le k \le m} |u|^2_{H^k(\Omega)} .$$

The $H^0$-norm is the $L^2$-norm and will be denoted $||\cdot||_{L^2(\Omega)}$. The space of polynomials of degree $\le k$ on a domain is denoted $\mathcal{P}_k$.

**The Bramble-Hilbert Lemma** A finite element method approximates a function $u$ from an infinite-dimensional functional space $V$ by a function $u_h$ from a finite-dimensional subspace $V_h \subset V$. One goal of such approaches is to prove that the error of the numerical solution $u_h$ is bounded *a priori* by the error of the best approximation available in $V_h$, i.e. $||u - u_h||_V \le C \inf_{w \in V_h} ||u - w||_V$. In this paper, $V = H^1$ and $V_h$ is the span of a set of functions defined piecewise over a 2D mesh of convex polygons. The parameter $h$ indicates the maximum diameter of an element in the mesh. Further details on the finite element method can be found in a number of textbooks [55, 37, 76, 242].

A quadratic finite element method in this context means that when $h \to 0$, the best approximation error ($\inf_{w \in V_h} ||u - w||_V$) converges to zero with order $h^2$. This means the space $V_h$ is 'dense enough' in $V$ to allow for quadratic convergence. Such arguments are usually proved via the Bramble-Hilbert lemma which guarantees that if $V_h$ contains polynomials up to a certain degree, a bound on the approximation error can be found. The variant of the Bramble-Hilbert lemma stated below includes a uniform constant over all convex domains which is a necessary detail in the context of general polygonal elements and generalized barycentric functions.

**Lemma 2.30** (Bramble-Hilbert [213, 64])**.** *There exists a uniform constant $C_{BH}$ such that for all convex polygons $\Omega$ and for all $u \in H^{k+1}(\Omega)$, there exists a degree $k$ polynomial $p_u$ with $||u - p_u||_{H^{k'}(\Omega)} \le C_{BH}\, \text{diam}(\Omega)^{k+1-k'} |u|_{H^{k+1}(\Omega)}$ for any $k' \le k$.*

Our focus is on quadratic elements (i.e., $k = 2$) and error estimates in the $H^1$-norm (i.e., $k' = 1$) which yields an estimate that scales with diam$(\Omega)^2$. Our methods extend to more general Sobolev spaces (i.e., $W^{k,p}$, the space of functions with all derivatives of order $\le k$ in $L^p$) whenever the Bramble-Hilbert lemma holds. Extensions to higher order elements ($k > 2$) will be briefly discussed in Section 2.7.2.

Observe that if $\Omega$ is transformed by any invertible affine map $T$, the polynomial $p \circ T^{-1}$ on $T\Omega$ has the same degree as the polynomial $p$ on $\Omega$. This fact is often exploited in the simpler and well-studied case of triangular meshes; an estimate on a

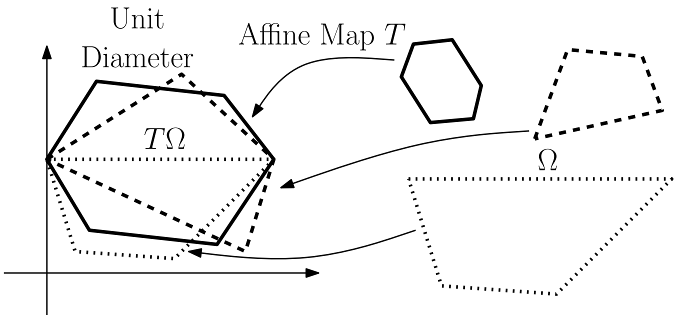


Figure 2.8: Using affine transformation, analysis can be restricted to a class of unit diameter polygons.

reference triangle $\hat{K}$ becomes an estimate on any physical triangle $K$ by passing through an affine transformation taking $\hat{K}$ to $K$. For $n > 3$, however, two generic $n$-gons may differ by a non-affine transformation and thus, as we will see in the next section, the use of a single reference element can become overly restrictive on element geometry. In our arguments, we instead analyze classes of "reference" elements, namely, diameter one convex quadrilaterals or convex polygons of diameter one satisfying the geometric criteria given in Section 2.7.2; see Figure 2.8. Using a class of reference elements allows us to establish uniform error estimates over all affine transformations of this class.

**Serendipity Quadratic Elements** The term 'serendipity element' refers to a long-standing observation in the finite element community that tensor product bases of polynomials on rectangular meshes of quadrilaterals in 2D or cubes in 3D can obtain higher order convergence rates with fewer than the 'expected' number of basis functions resulting from tensor products. This phenomenon is discussed in many finite element textbooks, e.g. [201, 123, 55], and was recently characterized precisely by Arnold and Awanou [12]. For instance, the degree $r$ tensor product basis on a square reference element has $(r+1)^2$ basis functions and can have guaranteed convergence rates of order $r+1$ when transformed to a rectangular mesh via bilinear isomorphisms [13]. By the Bramble-Hilbert lemma, however, the function space spanned by this basis may be unnecessarily large as the dimension of $\mathcal{P}_r$ is only $(r+1)(r+2)/2$ and only $4r$ degrees of freedom associated to the boundary are needed to ensure sufficient inter-element continuity in $H^1$.

This motivates the construction of the serendipity element for quadrilaterals. By a judicious choice of basis functions, an order $r$ convergence rate can be obtained with one basis function associated to each vertex, $(r-1)$ basis functions associated to each edge, and $q$ additional functions associated to interior points of the quadrilateral, where $q = 0$ for $r < 4$ and $q = (r-2)(r-1)/2$ for $r \geq 4$ [12]. Such an approach only works if the reference element is mapped via an affine transformation; it has been demonstrated that the serendipity element fails on trapezoidal elements, such as those shown in Figure 2.15 [146, 131, 242, 235]. Some very specific serendipity elements have been constructed for quadrilaterals and regular hexagons based on the Wachspress coordinates (discussed in the next sections) [216, 7, 102, 6, 103]. Our work generalizes this construction to arbitrary polygons without dependence on the type of generalized barycentric coordinate selected and with uniform bounds under certain geometric criteria.

**Generalized Barycentric Elements** To avoid non-affine transformations associated with tensor products constructions on a single reference element, we use generalized barycentric coordinates to define our basis functions. These coordinates are any functions satisfying the following agreed-upon definition in the literature.

**Definition 2.31.** Functions $\lambda_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, n$ are **barycentric coordinates** on $\Omega$ if they satisfy two properties.

B1. **Non-negative**: $\lambda_i \geq 0$ on $\Omega$.

B2. **Linear Completeness**: For any linear function $L : \Omega \to \mathbb{R}$, $L = \displaystyle\sum_{i=1}^{n} L(\mathbf{v}_i)\lambda_i$.

We will further restrict our attention to barycentric coordinates satisfying the following invariance property. Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be a composition of translation, rotation, and uniform scaling transformations and let $\{\lambda_i^T\}$ denote a set of barycentric coordinates on $T\Omega$.

B3. **Invariance:** $\lambda_i(\mathbf{x}) = \lambda_i^T(T(\mathbf{x}))$.

This assumption will allow estimates over the class of convex sets with diameter one to be immediately extended to generic sizes since translation, rotation and uniform scaling operations can be easily passed through Sobolev norms. At the expense of requiring uniform bounds over a class of diameter-one domains rather than a single reference element, we avoid having to handle non-affine mappings between reference and physical elements.

A set of barycentric coordinates $\{\lambda_i\}$ also satisfies three additional familiar properties. A proof that B1 and B2 imply the additional properties B4-B6 can be found in [100]. Note that B4 and B5 follow immediately by setting $L = 1$ or $L = \mathbf{x}$ in B2.

B4. **Partition of unity:** $\displaystyle\sum_{i=1}^{n} \lambda_i \equiv 1$.

B5. **Linear precision:** $\displaystyle\sum_{i=1}^{n} \mathbf{v}_i \lambda_i(\mathbf{x}) = \mathbf{x}$.

B6. **Interpolation:** $\lambda_i(\mathbf{v}_j) = \delta_{ij}$.

Various particular barycentric coordinates have been constructed in the literature. We briefly mention a few of the more prominent kinds and associated references here; readers are referred to our prior work [100, Section 2] as well as the survey papers of Cueto et al. [60] and Sukumar and Tabarraei [204] for further details. The triangulation coordinates $\lambda^{\text{Tri}}$ are defined by triangulating the polygon and using the standard barycentric coordinates over each triangle [85]. Harmonic coordinates $\lambda^{\text{Har}}$ are defined as the solution to Laplace's equation on the polygon with piecewise linear boundary data satisfying B6 [125, 151, 53]. Explicitly constructed functions include the rational Wachspress coordinates $\lambda^{\text{Wach}}$ [216], the Sibson coordinates $\lambda^{\text{Sibs}}$ defined in terms of the Voronoi diagram of the vertices of the polygon [192, 82], and the mean value coordinates $\lambda^{\text{MVal}}$ defined by Floater [83, 85].

To obtain convergence estimates with any of these functions, certain geometric conditions must be satisfied by a generic mesh element. We will consider domains satisfying the following three geometric conditions.

G1. **Bounded aspect ratio:** There exists $\gamma^* \in \mathbb{R}$ such that $\gamma < \gamma^*$.

G2. **Minimum edge length:** There exists $d_* \in \mathbb{R}$ such that $|\mathbf{v}_i - \mathbf{v}_j| > d_* > 0$ for all $i \neq j$.

G3. **Maximum interior angle:** There exists $\beta^* \in \mathbb{R}$ such that $\beta_i < \beta^* < \pi$ for all $i$.

Under some set of these conditions, the $H^1$-norm of many generalized barycentric coordinates are bounded in $H^1$ norm. This is a key estimate in asserting the expected (linear) convergence rate in the typical finite element setting.

**Theorem 2.32** ([170] for $\lambda^{\text{MVal}}$ and [100] for others)**.** *For any convex polygon $\Omega$ satisfying G1, G2, and G3, $\lambda^{\text{Tri}}$, $\lambda^{\text{Har}}$, $\lambda^{\text{Wach}}$, $\lambda^{\text{Sibs}}$, and $\lambda^{\text{MVal}}$ are all bounded in $H^1$, i.e. there exists a constant $C > 0$ such that*

$$||\lambda_i||_{H^1(\Omega)} \leq C. \tag{2.25}$$

The results in [100] and [170] are somewhat stronger than the statement of Theorem 2.32, namely, not all of the geometric hypotheses are necessary for every coordinate type. Our results, however, rely generically on any set of barycentric coordinates satisfying (2.25). For instance, the degenerate pentagon formed by adding an additional vertex in the center of the side of a square does not satisfy G3, but some choices of barycentric coordinates, such as $\lambda^{\text{MVal}}$ and $\lambda^{\text{Har}}$, will still admit an estimate like (2.25) on this geometry. We analyze the potential weakening of the geometric hypotheses in Section 2.7.2.

**Quadratic Precision Barycentric Functions** Since generalized barycentric coordinates are only guaranteed to have linear precision (property B5), they cannot provide greater than linear order error estimates. Pairwise products of barycentric coordinates, however, provide quadratic precision as the following simple proposition explains.

**Proposition 2.33.** *Given a set of barycentric coordinates $\{\lambda_i\}_{i=1}^n$, the set of functions $\{\mu_{ab}\} := \{\lambda_a\lambda_b\}_{a,b=1}^n$ has constant, linear, and quadratic precision*[1]*, i.e.*

$$\sum_{a=1}^{n}\sum_{b=1}^{n} \mu_{ab} = 1, \qquad \sum_{a=1}^{n}\sum_{b=1}^{n} \mathbf{v}_a \mu_{ab} = \boldsymbol{x} \qquad \text{and} \qquad \sum_{a=1}^{n}\sum_{b=1}^{n} \mathbf{v}_a \mathbf{v}_b^T \mu_{ab} = \boldsymbol{x}\boldsymbol{x}^T. \tag{2.26}$$

[1] Note that $\mathbf{x}\mathbf{x}^T$ is a symmetric matrix of quadratic monomials.

*Proof.* The result is immediate from properties B4 and B5 of the $\lambda_i$ functions. □

The product rule ensures that Theorem 2.32 extends immediately to the pairwise product functions.

**Corollary 2.34.** *Let $\Omega$ be a convex polygon satisfying G1, G2, and G3, and let $\lambda_i$ denote a set of barycentric coordinates satisfying the result of Theorem 2.32 (e.g. $\lambda^{\mathrm{Tri}}$, $\lambda^{\mathrm{Har}}$, $\lambda^{\mathrm{Wach}}$, $\lambda^{\mathrm{Sibs}}$, or $\lambda^{\mathrm{MVal}}$). Then pairwise products of the $\lambda_i$ functions are all bounded in $H^1$, i.e. there exists a constant $C > 0$ such that*

$$||\mu_{ab}||_{H^1(\Omega)} \leq C. \tag{2.27}$$

While the $\{\mu_{ab}\}$ functions are commonly used on triangles to provide a quadratic Lagrange element, they have not been considered in the context of generalized barycentric coordinates on convex polygons as considered here. Langer and Seidel have considered higher order barycentric interpolation in the computer graphics literature [140]; their approach, however, is for problems requiring $C^1$-continuous interpolation rather than the weaker $H^1$-continuity required for finite element theory.

In the remainder of this section, we describe notation that will be used to index functions throughout the rest of the paper. Since $\mu_{ab} = \mu_{ba}$, the summations from (2.26) can be written in a symmetric expansion. Define the paired index set

$$I := \{\{a, b\} \,|\, a, b \in \{1, \ldots, n\}\}.$$

Note that sets with cardinality 1 occur when $a = b$ and *are* included in $I$. We partition $I$ into three subsets corresponding to geometrical features of the polygon: vertices, edges of the boundary, and interior diagonals. More precisely, $I = V \cup E \cup D$, a disjoint union, where

$$\begin{aligned} V &:= \{\{a, a\} \,|\, a \in \{1, \ldots, n\}\}; \\ E &:= \{\{a, a+1\} \,|\, a \in \{1, \ldots, n\}\}; \\ D &:= I \setminus (V \cup E). \end{aligned}$$

In the definition of $E$ above (and in general for indices throughout the paper), values are interpreted modulo $n$, i.e. $\{n, n+1\}$, $\{n, 1\}$, and $\{0, 1\}$ all correspond to the edge between vertex 1 and vertex $n$. To simplify notation, we will omit the braces and commas when referring to elements of the index set $I$. For instance, instead of $\mu_{\{a,b\}}$, we write just $\mu_{ab}$. We emphasize that $ab \in I$ refers to an unordered and possibly non-distinct pair of vertices. Occasionally we will also use the abbreviated notation

$$\mathbf{v}_{ab} := \frac{\mathbf{v}_a + \mathbf{v}_b}{2},$$

so that $\mathbf{v}_{aa}$ is just a different expression for $\mathbf{v}_a$. Under these conventions, the precision properties from (2.26) can be rewritten as follows.

Q1. **Constant Precision**: $\sum_{aa\in V} \mu_{aa} + \sum_{ab\in E\cup D} 2\mu_{ab} = 1$

Q2. **Linear Precision**: $\sum_{aa\in V} \mathbf{v}_{aa}\mu_{aa} + \sum_{ab\in E\cup D} 2\mathbf{v}_{ab}\mu_{ab} = \mathbf{x}$

Q3. **Quadratic Precision**: $\sum_{aa\in V} \mathbf{v}_a\mathbf{v}_a^T\mu_{aa} + \sum_{ab\in E\cup D} (\mathbf{v}_a\mathbf{v}_b^T + \mathbf{v}_b\mathbf{v}_a^T)\mu_{ab} = \mathbf{x}\mathbf{x}^T$

**Reducing Quadratic Elements to Serendipity Elements**

We now seek to reduce the set of pairwise product functions $\{\mu_{ab}\}$ to a basis $\{\xi_{ij}\}$ for a serendipity quadratic finite element space. Our desired basis must

(i) span all quadratic polynomials of two variables on $\Omega$,

(ii) be exactly the space of quadratic polynomials (of one variable) when restricted to edges of $\Omega$, and

(iii) contain only $2n$ basis functions.

The intuition for how to achieve this is seen from the number of distinct pairwise products:

$$|\{\mu_{ab}\}| = |I| = |V| + |E| + |D| = n + n + \frac{}{n(n-3)}2 = n + \binom{n}{2}$$

On $\partial\Omega$, functions with indices in $V$ vanish on all but two adjacent edges, functions with indices in $E$ vanish on all but one edge, and functions with indices in $D$ vanish on all edges. Since Q1-Q3 hold on all of $\Omega$, including $\partial\Omega$, the set $\{\mu_{ab} : ab \in V \cup E\}$ satisfies (ii) and (iii), but not necessarily (i). Thus, our goal is to add linear combinations of functions with indices in $D$ to those with indices in $V$ or $E$ such that (i) is ensured.

We formalize this goal as a linear algebra problem: find a matrix $\mathbb{A}$ for the equation

$$[\xi_{ij}] := \mathbb{A}[\mu_{ab}] \tag{2.28}$$

such that $[\xi_{ij}]$ satisfies the following conditions analogous to Q1-Q3:

Q$_\xi$1. **Constant Precision**: $\sum_{ii\in V} \xi_{ii} + \sum_{i(i+1)\in E} 2\xi_{i(i+1)} = 1.$

Q$_\xi$2. **Linear Precision**: $\sum_{ii\in V} \mathbf{v}_{ii}\xi_{ii} + \sum_{i(i+1)\in E} 2\mathbf{v}_{i(i+1)}\xi_{i(i+1)} = \mathbf{x}.$

Q$_\xi$3. **Quadratic Precision**:
$$\sum_{ii\in V} \mathbf{v}_i\mathbf{v}_i^T\xi_{ii} + \sum_{i(i+1)\in E} (\mathbf{v}_i\mathbf{v}_{i+1}^T + \mathbf{v}_{i+1}\mathbf{v}_i^T)\xi_{i(i+1)} = \mathbf{x}\mathbf{x}^T.$$

Since (2.28) is a linear relationship, we are still able to restrict our analysis to a reference set of unit diameter polygons (recall Figure 2.8). Specifically if matrix $\mathbb{A}$ yields a "reference" basis $T[\xi_{ij}] = \mathbb{A}T[\mu_{ab}]$ satisfying Q$_\xi$1-Q$_\xi$3, then the "physical" basis $[\xi_{ij}] = \mathbb{A}[\mu_{ab}]$ also satisfies Q$_\xi$1-Q$_\xi$3.

To specify $\mathbb{A}$ in (2.28), we will use the specific basis orderings

$$[\xi_{ij}] := [\,\underbrace{\xi_{11}, \xi_{22}, \ldots, \xi_{nn}}_{\text{indices in } V},\ \underbrace{\xi_{12}, \xi_{23}, \ldots, \xi_{(n-1)n}, \xi_{n(n+1)}}_{\text{indices in } E}\,], \tag{2.29}$$

$$[\mu_{ab}] := [\,\underbrace{\mu_{11}, \mu_{22}, \ldots, \mu_{nn}}_{\text{indices in } V},\ \underbrace{\mu_{12}, \mu_{23}, \ldots, \mu_{(n-1)n}, \mu_{n(n+1)}}_{\text{indices in } E}, \tag{2.30}$$

$$\underbrace{\mu_{13}, \ldots, \text{(lexicographical)}, \ldots, \mu_{(n-2)n}}_{\text{indices in } D}\,].$$

The entries of $\mathbb{A}$ are denoted $c_{ab}^{ij}$ following the orderings given in (2.29)-(2.30) so that

$$\mathbb{A} := \begin{bmatrix} c_{11}^{11} & \cdots & c_{ab}^{11} & \cdots & c_{(n-2)n}^{11} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{11}^{ij} & \cdots & c_{ab}^{ij} & \cdots & c_{(n-2)n}^{ij} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{11}^{n(n+1)} & \cdots & c_{ab}^{n(n+1)} & \cdots & c_{(n-2)n}^{n(n+1)} \end{bmatrix}. \tag{2.31}$$

A sufficient set of constraints on the coefficients of $\mathbb{A}$ to ensure Q$_\xi$1-Q$_\xi$3 is given by the following lemma.

**Lemma 2.35.** *The constraints $Q_c1$-$Q_c3$ listed below imply $Q_\xi1$-$Q_\xi3$, respectively. That is, $Q_c1 \Rightarrow Q_\xi1$, $Q_c2 \Rightarrow Q_\xi2$, and $Q_c3 \Rightarrow Q_\xi3$.*

*$Q_c1$.* $\displaystyle\sum_{ii\in V} c_{aa}^{ii} + \sum_{i(i+1)\in E} 2c_{aa}^{i(i+1)} = 1\ \forall aa \in V$, *and*

$$\sum_{ii\in V} c_{ab}^{ii} + \sum_{i(i+1)\in E} 2c_{ab}^{i(i+1)} = 2,\ \forall ab \in E \cup D.$$

*$Q_c2$.* $\displaystyle\sum_{ii\in V} c_{aa}^{ii}\mathbf{v}_{ii} + \sum_{i(i+1)\in E} 2c_{aa}^{i(i+1)}\mathbf{v}_{i(i+1)} = \mathbf{v}_{aa}\ \forall aa \in V$, *and*

$$\sum_{ii\in V} c_{ab}^{ii}\mathbf{v}_{ii} + \sum_{i(i+1)\in E} 2c_{ab}^{i(i+1)}\mathbf{v}_{i(i+1)} = 2\mathbf{v}_{ab},\ \forall ab \in E \cup D.$$

*$Q_c3$.* $\displaystyle\sum_{ii\in V} c_{aa}^{ii}\mathbf{v}_i\mathbf{v}_i^T + \sum_{i(i+1)\in E} c_{aa}^{i(i+1)}(\mathbf{v}_i\mathbf{v}_{i+1}^T + \mathbf{v}_{i+1}\mathbf{v}_i^T) = \mathbf{v}_a\mathbf{v}_a^T\ \forall a \in V$, *and*

$$\sum_{ii\in V} c_{ab}^{ii}\mathbf{v}_i\mathbf{v}_i^T + \sum_{i(i+1)\in E} c_{ab}^{i(i+1)}(\mathbf{v}_i\mathbf{v}_{i+1}^T + \mathbf{v}_{i+1}\mathbf{v}_i^T) = \mathbf{v}_a\mathbf{v}_b^T + \mathbf{v}_b\mathbf{v}_a^T,\ \forall ab \in E \cup D.$$

*Proof.* Suppose Q$_c$1 holds. Substituting the expressions from Q$_c$1 into the coefficients of Q1 (from the end of Section 2.7.2), we get

$$\sum_{aa\in V}\left(\sum_{ii\in V} c_{aa}^{ii} + \sum_{i(i+1)\in E} 2c_{aa}^{i(i+1)}\right)\mu_{aa} +$$
$$\sum_{ab\in E\cup D}\left(\sum_{ii\in V} c_{ab}^{ii} + \sum_{i(i+1)\in E} 2c_{ab}^{i(i+1)}\right)\mu_{ab} = 1.$$

Regrouping this summation over $ij$ indices instead of $ab$ indices, we have

$$\sum_{ii\in V}\left(\sum_{ab\in I} c_{ab}^{ii}\mu_{ab}\right) + \sum_{i(i+1)\in E} 2\left(\sum_{ab\in I} c_{ab}^{i(i+1)}\mu_{ab}\right) = 1. \tag{2.32}$$

Since (2.28) defines $\xi_{ij} = \sum_{ab\in I} c_{ab}^{ij}\mu_{ab}$, (2.32) is exactly the statement of Q$_\xi$1. The other two cases follow by the same technique of regrouping summations. □

We now give some remarks about our approach to finding coefficients satisfying Q$_c$1-Q$_c$3. Observe that the first equation in each of Q$_c$1-Q$_c$3 is satisfied by

$$c_{aa}^{ii} := \delta_{ia} \quad \text{and} \quad c_{aa}^{i(i+1)} := 0 \tag{2.33}$$

Further, if $ab = a(a+1) \in E$, the second equation in each of Q$_c$1-Q$_c$3 is satisfied by

$$c_{a(a+1)}^{ii} := 0 \quad \text{and} \quad c_{a(a+1)}^{i(i+1)} := \delta_{ia} \tag{2.34}$$

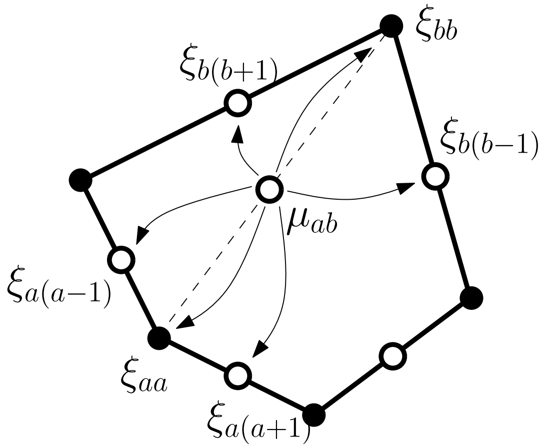


Figure 2.9: When constructing the matrix $\mathbb{A}$, only six non-zero elements are used in each column corresponding to an interior diagonal of the pairwise product basis. In the serendipity basis, the interior diagonal function $\mu_{ab}$ only contributes to six basis functions as shown, corresponding to the vertices of the diagonal's endpoints and the midpoints of adjacent boundary edges.

The choices in (2.33) and (2.34) give $\mathbb{A}$ the simple structure

$$\mathbb{A} := \left[\ \mathbb{I} \mid \mathbb{A}'\ \right], \tag{2.35}$$

where $\mathbb{I}$ is the $2n \times 2n$ identity matrix. Note that this corresponds exactly to our intuitive approach of setting each $\xi_{ij}$ function to be the corresponding $\mu_{ij}$ function plus a linear combination of $\mu_{ab}$ functions with $ab \in D$. Also, with this selection, we can verify that many of the conditions which are part of $\mathrm{Q}_c1$, $\mathrm{Q}_c2$ and $\mathrm{Q}_c3$ hold. Specifically, whenever $ab \in V \cup E$, the corresponding conditions hold, as we prove in the following lemma.

**Lemma 2.36.** *The first $2n$ columns of the matrix $\mathbb{A}$ given by (2.35), i.e., the identity portion, ensure $Q_c1$, $Q_c2$ and $Q_c3$ hold for $ab \in V \cup E$.*

*Proof.* This lemma follows from direct substitution. In each case, there is only one nonzero element $c_{aa}^{aa}$ or $c_{a(a+1)}^{a(a+1)}$ on the hand side of the equation from $\mathrm{Q}_c1$, $\mathrm{Q}_c2$ or $\mathrm{Q}_c3$ and substituting 1 for that coefficient gives the desired equality. □

It remains to define $\mathbb{A}'$, i.e. those coefficients $c_{ab}^{ij}$ with $ab \in D$ and verify the corresponding equations in $\mathrm{Q}_c1$, $\mathrm{Q}_c2$, and $\mathrm{Q}_c3$. For each column of $\mathbb{A}'$, $\mathrm{Q}_c1$, $\mathrm{Q}_c2$, and $\mathrm{Q}_c3$ yield a system of six scalar equations for the $2n$ variables $\{c_{ab}^{ij}\}_{ij \in V \cup E}$. Since we have many more variables than equations, there remains significant flexibility in the construction of a solution. In the upcoming sections, we will present such a solution where all but six of the coefficients in each column of $\mathbb{A}'$ are set to zero. The non-zero coefficients are chosen to be $c_{ab}^{a(a-1)}$, $c_{ab}^{aa}$, $c_{ab}^{a(a+1)}$, $c_{ab}^{b(b-1)}$, $c_{ab}^{bb}$, and $c_{ab}^{b(b+1)}$ as these have a natural correspondence to the geometry of the polygon and the edge $ab$; see Figure 2.9.
We will show that the system of equations $\mathrm{Q}_c1$-$\mathrm{Q}_c3$ with this selection of non-zero coefficients for $\mathbb{A}'$ has an explicitly constructible solution. The solution is presented for special classes of polygons in Section 2.7.2 and for generic convex polygons in Section 2.7.2. In each case, we prove a uniform bound on the size of the coefficients of $\mathbb{A}$, a sufficient result to control $||\xi_{ij}||_{H^1(\Omega)}$, as the following lemma shows.

**Lemma 2.37.** *Let $\Omega$ be a convex polygon satisfying G1, G2, and G3, and let $\lambda_i$ denote a set of barycentric coordinates satisfying the result of Theorem 2.32 (e.g. $\lambda^{\mathrm{Tri}}$, $\lambda^{\mathrm{Har}}$, $\lambda^{\mathrm{Wach}}$, $\lambda^{\mathrm{Sibs}}$, or $\lambda^{\mathrm{MVal}}$). Suppose there exists $M > 1$ such that for all entries of $\mathbb{A}'$, $|c_{ab}^{ij}| < M$. Then the functions $\xi_{ij}$ are all bounded in $H^1$, i.e. there exists a constant $B > 0$ such that*

$$||\xi_{ij}||_{H^1(\Omega)} \leq B. \tag{2.36}$$

*Proof.* Since $\xi_{ij}$ is defined by (2.28), Corollary 2.34 implies that there exists $C > 0$ such that

$$||\xi_{ij}||_{H^1(\Omega)} \leq ||\mathbb{A}|| \max_{ab} ||\mu_{ab}||_{H^1(\Omega)} < C||\mathbb{A}||.$$

Since the space of linear transformations from $\mathbb{R}^{n(n+1)/2}$ to $\mathbb{R}^{2n}$ is finite-dimensional, all norms on $\mathbb{A}$ are equivalent. Thus, without loss of generality, we interpret $||\mathbb{A}||$ as the maximum absolute row sum norm, i.e.

$$||\mathbb{A}|| := \max_{ij} \sum_{ab} |c_{ab}^{ij}|. \tag{2.37}$$

By the structure of $\mathbb{A}$ from (2.35) and the hypothesis, we have

$$||\mathbb{A}|| \leq \frac{n(n+1)}{2} M$$

□

**Special Cases of the Serendipity Reduction**

Before showing that $\text{Q}_c1$-$\text{Q}_c3$ can be satisfied in a general setting, we study some simpler special cases in which symmetry reduces the number of equations that must be satisfied simultaneously.

**Unit Square** We begin with the case where serendipity elements were first examined, namely over meshes of squares. Strang and Fix [201] gave one of the first discussions of the serendipity element; in this paper we will use the modern notation introduced by Arnold and Awanou [12]. Here, the quadratic serendipity space on the unit square, denoted $\mathcal{S}_2(I^2)$, is defined as the span of eight monomials:

$$\mathcal{S}_2(I^2) := \text{span}\left\{1, x, y, x^2, xy, y^2, x^2y, xy^2\right\} \tag{2.38}$$

We will now show how our construction process recovers the same space of monomials. Denote vertices and midpoints on $[0,1]^2$ by

$$\begin{gathered} \mathbf{v}_1 = (0,0) \quad \mathbf{v}_2 = (1,0) \quad \mathbf{v}_3 = (1,1) \quad \mathbf{v}_4 = (0,1) \\ \mathbf{v}_{12} = (1/2,0) \quad \mathbf{v}_{23} = (1,1/2) \quad \mathbf{v}_{34} = (1/2,1) \quad \mathbf{v}_{14} = (0,1/2) \\ \mathbf{v}_{13} = \mathbf{v}_{24} = (1/2,1/2) \end{gathered} \tag{2.39}$$

The standard bilinear basis for the square is

$$\begin{aligned} \lambda_1 &= (1-x)(1-y) & \lambda_2 &= x(1-y) \\ \lambda_4 &= (1-x)y & \lambda_3 &= xy \end{aligned}$$

Since the $\lambda_i$ have vanishing second derivatives and satisfy the definition of barycentric coordinates, they are in fact the harmonic coordinates $\lambda^{\text{Har}}$ in this special case. Pairwise products give us the following 10 (not linearly independent) functions

$$\begin{aligned} \mu_{11} &= (1-x)^2(1-y)^2 & \mu_{12} &= (1-x)x(1-y)^2 \\ \mu_{22} &= x^2(1-y)^2 & \mu_{23} &= x^2(1-y)y \\ \mu_{33} &= x^2y^2 & \mu_{34} &= (1-x)xy^2 \\ \mu_{44} &= (1-x)^2y^2 & \mu_{14} &= (1-x)^2(1-y)y \\ \mu_{13} &= (1-x)x(1-y)y & \mu_{24} &= (1-x)x(1-y)y \end{aligned}$$

For the *special* geometry of the square, $\mu_{13} = \mu_{24}$, but this is not true for general quadrilaterals as we see in Section 2.7.2. The serendipity construction eliminates the functions $\mu_{13}$ and $\mu_{24}$ to give an 8-dimensional space. The basis reduction via the $\mathbb{A}$ matrix is given by

$$\begin{bmatrix} \xi_{11} \\ \xi_{22} \\ \xi_{33} \\ \xi_{44} \\ \xi_{12} \\ \xi_{23} \\ \xi_{34} \\ \xi_{14} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{22} \\ \mu_{33} \\ \mu_{44} \\ \mu_{12} \\ \mu_{23} \\ \mu_{34} \\ \mu_{14} \\ \mu_{13} \\ \mu_{24} \end{bmatrix} \tag{2.40}$$

It can be confirmed directly that (2.40) follows from the definitions of $\mathbb{A}$ given in the increasingly generic settings examined in Section 2.7.2, Section 2.7.2 and Section 2.7.2. The resulting functions are

$$\begin{aligned} \xi_{11} &= (1-x)(1-y)(1-x-y) & \xi_{12} &= (1-x)x(1-y) \\ \xi_{22} &= x(1-y)(x-y) & \xi_{23} &= x(1-y)y \\ \xi_{33} &= xy(-1+x+y) & \xi_{34} &= (1-x)xy \\ \xi_{44} &= (1-x)y(y-x) & \xi_{14} &= (1-x)(1-y)y \end{aligned} \tag{2.41}$$

**Theorem 2.38.** *For the unit square, the basis functions $\{\xi_{ij}\}$ defined in (2.41) satisfy $Q_\xi 1$-$Q_\xi 3$.*

*Proof.* A simple proof is to observe that the coefficients of the matrix in (2.40) satisfy $\mathrm{Q}_c 1$-$\mathrm{Q}_c 3$ and then apply Lemma 2.35. To illuminate the construction in this special case of common interest, we state some explicit calculations. The constant precision condition $\mathrm{Q}_\xi 1$ is verified by the calculation

$$\xi_{11} + \xi_{22} + \xi_{33} + \xi_{44} + 2\xi_{12} + 2\xi_{23} + 2\xi_{34} + 2\xi_{14} = 1.$$

The $x$ component of the linear precision condition $\mathrm{Q}_\xi 2$ is verified by the calculation

$$\begin{aligned} &(\mathbf{v}_1)_x\xi_{11} + (\mathbf{v}_2)_x\xi_{22} + (\mathbf{v}_3)_x\xi_{33} + (\mathbf{v}_4)_x\xi_{44} + \\ &\qquad 2(\mathbf{v}_{12})_x\xi_{12} + 2(\mathbf{v}_{23})_x\xi_{23} + 2(\mathbf{v}_{34})_x\xi_{34} + 2(\mathbf{v}_{14})_x\xi_{14} \\ &\qquad\qquad = \xi_{22} + \xi_{33} + 2\cdot\frac{1}{2}\xi_{12} + 2\cdot 1\xi_{23} + 2\cdot\frac{1}{2}\xi_{34} \\ &\qquad\qquad = x. \end{aligned}$$

The verification for the $y$ component is similar. The $xy$ component of the quadratic precision condition $\mathrm{Q}_\xi 3$ is verified by

$$\begin{aligned} &(\mathbf{v}_1)_x(\mathbf{v}_1)_y\xi_{11} + (\mathbf{v}_2)_x(\mathbf{v}_2)_y\xi_{22} + (\mathbf{v}_3)_x(\mathbf{v}_3)_y\xi_{33} + (\mathbf{v}_4)_x(\mathbf{v}_4)_y\xi_{44} \\ &+ \left[(\mathbf{v}_1)_x(\mathbf{v}_2)_y + (\mathbf{v}_2)_x(\mathbf{v}_1)_y\right]\xi_{12} + \left[(\mathbf{v}_2)_x(\mathbf{v}_3)_y + (\mathbf{v}_3)_x(\mathbf{v}_2)_y\right]\xi_{23} \\ &+ \left[(\mathbf{v}_3)_x(\mathbf{v}_4)_y + (\mathbf{v}_4)_x(\mathbf{v}_3)_y\right]\xi_{34} + \left[(\mathbf{v}_4)_x(\mathbf{v}_1)_y + (\mathbf{v}_1)_x(\mathbf{v}_4)_y\right]\xi_{14} \\ &= \xi_{33} + \xi_{23} + \xi_{34} = xy. \end{aligned}$$

The monomials $x^2$ and $y^2$ can be expressed as a linear combination of the $\xi_{ij}$ similarly, via the formula given in $\mathrm{Q}_\xi 3$. □

**Corollary 2.39.** *The span of the $\xi_{ij}$ functions defined by (2.41) is the standard serendipity space, i.e.*

$$\operatorname{span}\left\{\xi_{ii}, \xi_{i(i+1)}\right\} = \mathcal{S}_2(I^2)$$

*Proof.* Observe that $x^2y = \xi_{23} + \xi_{33}$ and $xy^2 = \xi_{33} + \xi_{34}$. By the definition of $\mathcal{S}_2(I^2)$ in (2.38) and the theorem, $\operatorname{span}\left\{\xi_{ii}, \xi_{i(i+1)}\right\} \supset \mathcal{S}_2(I^2)$. Since both spaces are dimension eight, they are identical. □

**Regular Polygons** We now generalize our construction to any regular polygon with $n$ vertices. Without loss of generality, this configuration can be described by two parameters $0 < \sigma \leq \theta \leq \pi/2$ as shown in Figure 2.10. Note that the $n$ vertices of the polygon are located at angles of the form $k\sigma$ where $k = 0, 1, \ldots, n-1$.

For two generic non-adjacent vertices $\mathbf{v}_a$ and $\mathbf{v}_b$, the coordinates of the six relevant vertices (recalling Figure 2.9) are:

$$\begin{aligned} \mathbf{v}_a &= \begin{bmatrix}\cos\theta \\ \sin\theta\end{bmatrix}; & \mathbf{v}_{a-1} &= \begin{bmatrix}\cos(\theta-\sigma) \\ \sin(\theta-\sigma)\end{bmatrix}; & \mathbf{v}_{a+1} &= \begin{bmatrix}\cos(\theta+\sigma) \\ \sin(\theta+\sigma)\end{bmatrix}; \\ \mathbf{v}_b &= \begin{bmatrix}\cos\theta \\ -\sin\theta\end{bmatrix}; & \mathbf{v}_{b-1} &= \begin{bmatrix}\cos(\theta+\sigma) \\ -\sin(\theta+\sigma)\end{bmatrix}; & \mathbf{v}_{b+1} &= \begin{bmatrix}\cos(\theta-\sigma) \\ -\sin(\theta-\sigma)\end{bmatrix}. \end{aligned}$$

We seek to establish the existence of suitable constants $c_{ab}^{aa}$, $c_{ab}^{a,a+1}$, $c_{ab}^{a-1,a}$, $c_{ab}^{bb}$, $c_{ab}^{b-1,b}$, $c_{ab}^{b,b+1}$ which preserve quadratic precision and to investigate the geometric conditions under which these constants become large. The symmetry of this configuration suggests that $c_{ab}^{aa} = c_{ab}^{bb}$, $c_{ab}^{a-1,a} = c_{ab}^{b,b+1}$, and $c_{ab}^{a,a+1} = c_{ab}^{b-1,b}$ are reasonable requirements. For simplicity we will denote these constants by $c_0 := c_{ab}^{aa}$, $c_- := c_{ab}^{a-1,a}$, and $c_+ := c_{ab}^{a,a+1}$.

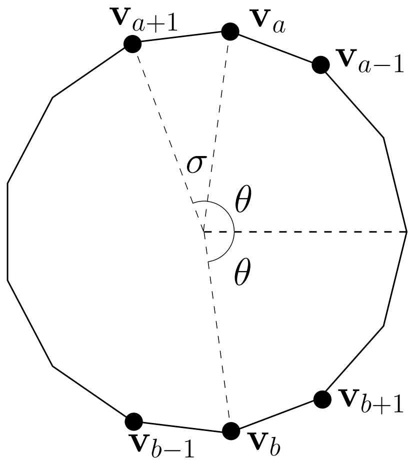


Figure 2.10: Notation for the construction for a regular polygon.

Thus equation Q$_c$1 (which contains only six non-zero elements) reduces to:

$$2c_0 + 4c_- + 4c_+ = 2. \tag{2.42}$$

Q$_c$2 involves two equations, one of which is trivially satisfied in our symmetric configuration. Thus, the only restriction to maintain is

$$2\cos\theta c_0 + 2\left[\cos\theta + \cos(\theta-\sigma)\right]c_- + 2\left[\cos\theta + \cos(\theta+\sigma)\right]c_+ = 2\cos\theta. \tag{2.43}$$

Q$_c$3 gives three more requirements, one of which is again trivially satisfied. This gives two remaining restrictions:

$$2\cos^2\theta c_0 + 4\cos\theta\cos(\theta-\sigma)c_- + 4\cos\theta\cos(\theta+\sigma)c_+ = 2\cos^2\theta; \tag{2.44}$$

$$2\sin^2\theta c_0 + 4\sin\theta\sin(\theta-\sigma)c_- + 4\sin\theta\sin(\theta+\sigma)c_+ = -2\sin^2\theta. \tag{2.45}$$

Now we have four equations (2.42)-(2.45) and three unknowns $c_0$, $c_-$ and $c_+$. Fortunately, equation (2.43) is a simple linear combination of (2.42) and(2.44); specifically (2.43) is $\frac{\cos\theta}{2}$ times (2.42) plus $\frac{1}{2\cos\theta}$ times (2.44). With a little algebra, we can produce the system:

$$\begin{bmatrix} 1 & 2 & 2 \\ 1 & 2(\cos\sigma + \sin\sigma\tan\theta) & 2(\cos\sigma - \sin\sigma\tan\theta) \\ 1 & 2(\cos\sigma - \sin\sigma\cot\theta) & 2(\cos\sigma + \sin\sigma\cot\theta) \end{bmatrix} \begin{bmatrix} c_0 \\ c_- \\ c_+ \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}. \tag{2.46}$$

The solution of this system can be computed:

$$c_0 = \frac{(-1+\cos\sigma)\cot\theta + (1+\cos\sigma)\tan\theta}{(-1+\cos\sigma)(\cot\theta+\tan\theta)};$$

$$c_- = \frac{\cos\sigma - \sin\sigma\tan\theta - 1}{2\,(\tan\theta+\cot\theta)\sin\sigma\,(\cos\sigma-1)}; \qquad c_+ = \frac{1-\cos\sigma-\sin\sigma\tan\theta}{2\,(\tan\theta+\cot\theta)\sin\sigma\,(\cos\sigma-1)}.$$

Although $\tan\theta$ (and thus the solution above) is not defined for $\theta = \pi/2$, the solution in this boundary case can be defined by the limiting value which always exists. We can now prove the following.

**Theorem 2.40.** *For any regular polygon, the basis functions $\{\xi_{ij}\}$ constructed using the coefficients $c_{ab}^{aa} = c_{ab}^{bb} = c_0$, $c_{ab}^{a-1,a} = c_{ab}^{b,b+1} = c_-$, $c_{ab}^{a,a+1} = c_{ab}^{b-1,b} = c_+$ satisfy Q$_\xi$1-Q$_\xi$3.*

*Proof.* The construction above ensures that the solution satisfies Q$_c$1, Q$_c$2, and Q$_c$3. □

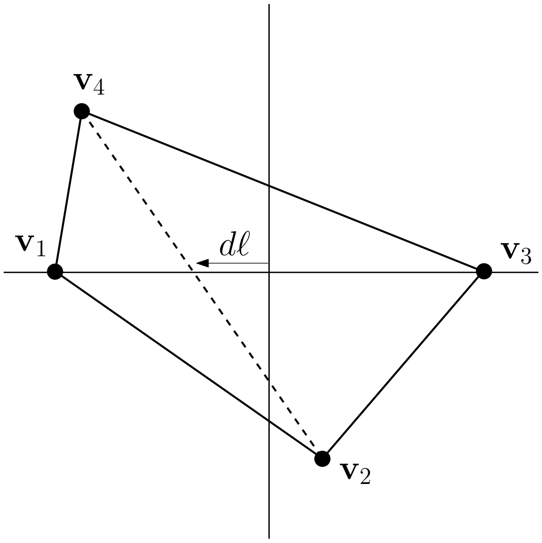


Figure 2.11: A generic convex quadrilateral, rotated so that one of its diagonals lies on the $x$-axis. Geometrically, $c_{13}^{12}$ and $c_{13}^{34}$ are selected to be coefficients of the convex combination of $\mathbf{v}_2$ and $\mathbf{v}_4$ that lies on the $x$-axis.

The serendipity element for regular polygons can be used for meshes consisting of only one regular polygon or a finite number of regular polygons. The former occurs only in meshes of triangles, squares and hexagons as these are the only regular polygons that can tile the plane. On the other hand, many tilings consisting of several regular polygons can be constructed using multiple regular polygons. Examples include the snub square tiling (octagons and squares), the truncated hexagonal tiling (dodecahedra and triangles), the rhombitrihexagonal tiling (hexagons, squares, and triangles), and the truncated trihexagonal tiling (dodecagons, hexagons, and squares); see e.g. [46]. The construction process outlined above opens up the possibility of finite element methods applied over these types of mixed-geometry meshes, a mostly unexplored field.

**Generic Quadrilaterals** Fix a convex quadrilateral $\Omega$ with vertices $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, and $\mathbf{v}_4$, ordered counterclockwise. We will describe how to set the coefficients of the submatrix $\mathbb{A}'$ in (2.35). It suffices to describe how to set the coefficients in the '13'-column of the matrix, i.e., those of the form $c_{13}^{ij}$. The '$24'$-column can be filled using the same construction after permuting the indices. Thus, without loss of generality, suppose that $\mathbf{v}_1 := (-\ell, 0)$ and $\mathbf{v}_3 := (\ell, 0)$ so that $\mathbf{v}_2$ is below the $x$-axis and $\mathbf{v}_4$ is above the $x$-axis, as shown in Figure 2.11. We have eight coefficients to set:

$$c_{13}^{11},\ c_{13}^{22},\ c_{13}^{33},\ c_{13}^{44},\ c_{13}^{12},\ c_{13}^{23},\ c_{13}^{34},\ \text{and}\ c_{13}^{14}.$$

Using a subscript $x$ or $y$ to denote the corresponding component of a vertex, define the coefficients as follows.

$$c_{13}^{22} := 0 \qquad c_{13}^{44} := 0 \tag{2.47}$$

$$c_{13}^{12} := \frac{(\mathbf{v}_4)_y}{(\mathbf{v}_4)_y - (\mathbf{v}_2)_y} \qquad c_{13}^{34} := \frac{(\mathbf{v}_2)_y}{(\mathbf{v}_2)_y - (\mathbf{v}_4)_y} \tag{2.48}$$

$$c_{13}^{23} := c_{13}^{12} \qquad c_{13}^{14} := c_{13}^{34} \tag{2.49}$$

$$c_{13}^{11} := \frac{c_{13}^{12}(\mathbf{v}_2)_x + c_{13}^{34}(\mathbf{v}_4)_x}{\ell} - 1 \quad c_{13}^{33} := -\frac{c_{13}^{12}(\mathbf{v}_2)_x + c_{13}^{34}(\mathbf{v}_4)_x}{\ell} - 1 \tag{2.50}$$

Note that there following the strategy shown in Figure 2.9, there are only six non-zero entries. For ease of notation in the rest of this section, we define the quantity

$$d := \frac{c_{13}^{12}(\mathbf{v}_2)_x + c_{13}^{34}(\mathbf{v}_4)_x}{\ell}.$$

First we assert that the resulting basis does span all quadratic polynomials.

**Theorem 2.41.** *For any quadrilateral, the basis functions $\{\xi_{ij}\}$ constructed using the coefficients given in (2.47)-(2.50) satisfy $Q_\xi 1$-$Q_\xi 3$.*

*Proof.* Considering Lemmas 2.35 and 2.36, we only must verify Q$_c$1-Q$_c$3 in the cases when $ab \in D = \{13, 24\}$. This will be verified directly by substituting (2.47)-(2.50) into the constraints Q$_c$1-Q$_c$3 in the case $ab = 13$. As noted before, the $ab = 24$ case is identical, requiring only a permutation of indices. First note that

$$c^{11}_{13} + c^{33}_{13} = -2 \quad \text{and} \quad c^{11}_{13} - c^{33}_{13} = 2d. \tag{2.51}$$

For Q$_c$1, the sum reduces to

$$c^{11}_{13} + c^{33}_{13} + 4(c^{12}_{13} + c^{34}_{13}) = -2 + 4(1) = 2,$$

as required. For Q$_c$2, the $x$-coordinate equation reduces to

$$\ell(c^{33}_{13} - c^{11}_{13}) + 2d\ell = 0$$

by (2.51) which is the desired inequality since we fixed (without loss of generality) $\mathbf{v}_a b = (0, 0)$. The $y$-coordinate equation reduces to $2(c^{12}_{13}(\mathbf{v}_2)_y + c^{34}_{13}(\mathbf{v}_4)_y) = 0$ which holds by (2.48). Finally, a bit of algebra reduces the matrix equality of Q$_c$3 to only the equality $\ell^2(c^{11}_{13} + c^{33}_{13}) = -2\ell^2$ of its first entry (all other entries are zero), which holds by (2.51). □

**Theorem 2.42.** *Over all convex quadrilaterals,* $||\mathbb{A}||$ *is uniformly bounded.*

*Proof.* By Lemma 2.37, it suffices to bound $|c^{ij}_{13}|$ uniformly. First observe that the convex combination of the vertices $\mathbf{v}_2$ and $\mathbf{v}_4$ using coefficients $c^{12}_{13}$ and $c^{34}_{13}$ produces a point lying on the $x$-axis, i.e.,

$$1 = c^{12}_{13} + c^{34}_{13}, \text{ and} \tag{2.52}$$

$$0 = c^{12}_{13}(\mathbf{v}_2)_y + c^{34}_{13}(\mathbf{v}_4)_y. \tag{2.53}$$

Since $(\mathbf{v}_2)_y > 0$ and $(\mathbf{v}_4)_y < 0$, (2.49) implies that $c^{12}_{13}, c^{34}_{13} \in (0, 1)$. By (2.49), it also follows that $c^{23}_{13}, c^{14}_{13} \in (0, 1)$.
For $c^{11}_{13}$ and $c^{33}_{13}$, note that the quantity $d\ell$ is the $x$-intercept of the line segment connecting $\mathbf{v}_2$ and $\mathbf{v}_4$. Thus $d\ell \in [-\ell, \ell]$ by convexity. So $d \in [-1, 1]$ and thus (2.50) implies $|c^{11}_{13} \models |d - 1| \le 2$ and $|c^{33}_{13} \models\vdash d - 1| \le 2$. □

### Proof of the Serendipity Reduction on Generic Convex Polygons

We now define the sub-matrix $\mathbb{A}'$ from (2.35) in the case of a generic polygon. Pick a column of $\mathbb{A}'$, i.e., fix $ab \in D$. The coefficients $c^{ij}_{ab}$ are constrained by a total of six equations Q$_c$1, Q$_c$2, and Q$_c$3. As before (recall Figure 2.9), six non-zero coefficients will be selected in each column to satisfy these constraints. Specifically,

$$c^{ii}_{ab} := 0, \text{ for } i \notin \{a, b\} \quad \text{and} \quad c^{i(i+1)}_{ab} = 0, \text{ for } i \notin \{a-1, a, b-1, b\}, \tag{2.54}$$

leaving only the following six coefficients to be determined:

$$c^{aa}_{ab},\ c^{bb}_{ab},\ c^{(a-1)a}_{ab},\ c^{a(a+1)}_{ab},\ c^{(b-1)b}_{ab}, \text{ and } c^{b(b+1)}_{ab}.$$

For the remainder of this section, we will omit the subscript $ab$ to ease the notation. Writing out Q$_c$1-Q$_c$3 for this fixed $ab$ pair, we have six equations with six unknowns:

$$c^{aa} + c^{bb} + 2c^{(a-1)a} + 2c^{a(a+1)} + 2c^{(b-1)b} + 2c^{b(b+1)} = 2;$$

$$c^{aa}\mathbf{v}_{aa} + 2c^{(a-1)a}\mathbf{v}_{(a-1)a} + 2c^{a(a+1)}\mathbf{v}_{a(a+1)} + \\ c^{bb}\mathbf{v}_{bb} + 2c^{(b-1)b}\mathbf{v}_{(b-1)b} + 2c^{b(b+1)}\mathbf{v}_{b(b+1)} = 2\mathbf{v}_{ab};$$

$$c^{aa}\mathbf{v}_a\mathbf{v}_a^T + c^{(a-1)a}(\mathbf{v}_{a-1}\mathbf{v}_a^T + \mathbf{v}_a\mathbf{v}_{a-1}^T) + c^{a(a+1)}(\mathbf{v}_a\mathbf{v}_{a+1}^T + \mathbf{v}_{a+1}\mathbf{v}_a^T) + \\ c^{bb}\mathbf{v}_b\mathbf{v}_b^T + c^{(b-1)b}(\mathbf{v}_{b-1}\mathbf{v}_b^T + \mathbf{v}_b\mathbf{v}_{b-1}^T) + c^{b(b+1)}(\mathbf{v}_b\mathbf{v}_{b+1}^T + \mathbf{v}_{b+1}\mathbf{v}_b^T) = \mathbf{v}_a\mathbf{v}_b^T + \mathbf{v}_b\mathbf{v}_a^T.$$

Assume without loss of generality that $\mathbf{v}_a = (-\ell, 0)$ and $\mathbf{v}_b = (\ell, 0)$ with $\ell < 1/2$ (since $\Omega$ has diameter 1). We introduce the terms $d_a$ and $d_b$ defined by

$$d_a := \frac{(\mathbf{v}_{a-1})_x(\mathbf{v}_{a+1})_y - (\mathbf{v}_{a+1})_x(\mathbf{v}_{a-1})_y}{(\mathbf{v}_{a-1})_y - (\mathbf{v}_{a+1})_y} \cdot \frac{1}{\ell}, \text{ and} \tag{2.55}$$

$$d_b := \frac{(\mathbf{v}_{b+1})_x(\mathbf{v}_{b-1})_y - (\mathbf{v}_{b-1})_x(\mathbf{v}_{b+1})_y}{(\mathbf{v}_{b-1})_y - (\mathbf{v}_{b+1})_y} \cdot \frac{1}{\ell}. \tag{2.56}$$

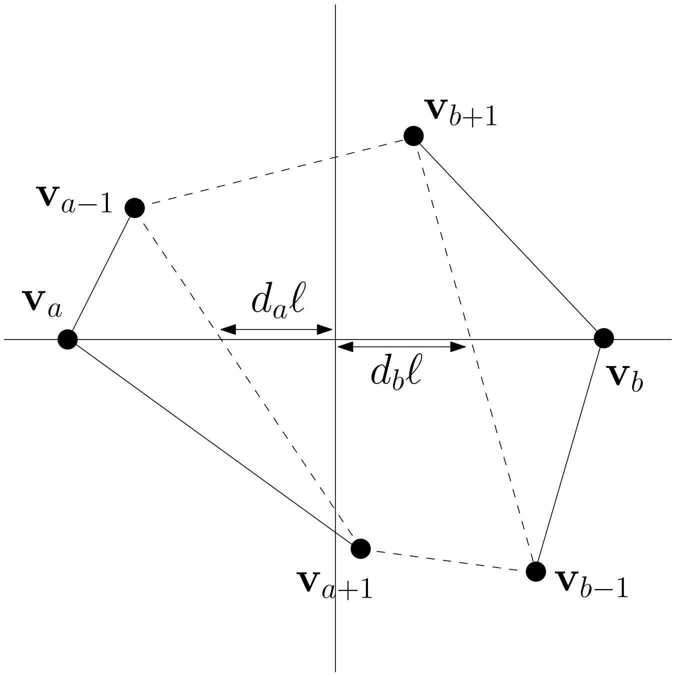


Figure 2.12: Generic convex polygon, rotated so that $\mathbf{v}_a = (-\ell, 0)$ and $\mathbf{v}_b = (\ell, 0)$. The $x$-intercept of the line between $\mathbf{v}_{a-1}$ and $\mathbf{v}_{a+1}$ is defined to be $-d_a\ell$ and the $x$-intercept of the line between $\mathbf{v}_{b-1}$ and $\mathbf{v}_{b+1}$ is defined to be $d_b\ell$.

These terms have a concrete geometrical interpretation as shown in Figure 2.12: $-d_a\ell$ is the $x$-intercept of the line between $\mathbf{v}_{a-1}$ and $\mathbf{v}_{a+1}$, while $d_b\ell$ is the $x$-intercept of the line between $\mathbf{v}_{b-1}$ and $\mathbf{v}_{b+1}$. Thus, by the convexity assumption, $d_a, d_b \in [-1, 1]$. Additionally, $-d_a \leq d_b$ with equality only in the case of a quadrilateral which was dealt with previously. For ease of notation and subsequent explanation, we also define

$$s := \frac{}{2} 2 - (d_a + d_b). \tag{2.57}$$

First we choose $c^{(a-1)a}$ and $c^{a(a+1)}$ as the solution to the following system of equations:

$$c^{(a-1)a} + c^{a(a+1)} = s; \tag{2.58}$$

$$c^{(a-1)a}\mathbf{v}_{a-1} + c^{a(a+1)}\mathbf{v}_{a+1} = s d_a \mathbf{v}_a. \tag{2.59}$$

There are a total of three equations since (2.59) equates vectors, but it can be verified directly that this system of equations is only rank two. Moreover, any two of the equations from (2.58) and (2.59) suffice to give the same unique solution for $c^{(a-1)a}$ and $c^{a(a+1)}$.

Similarly, we select $c^{(b-1)b}$ and $c^{b(b-1)}$ as the solution to the system:

$$c^{(b-1)b} + c^{b(b+1)} = s; \tag{2.60}$$

$$c^{(b-1)b}\mathbf{v}_{b-1} + c^{b(b+1)}\mathbf{v}_{b+1} = s d_b \mathbf{v}_b. \tag{2.61}$$

Finally, we assign $c^{aa}$ and $c^{bb}$ by

$$c^{aa} = \frac{-2 - 2d_a}{2 - (d_a + d_b)} \quad \text{and} \tag{2.62}$$

$$c^{bb} = \frac{-2 - 2d_b}{2 - (d_a + d_b)}, \tag{2.63}$$

and claim that this set of coefficients leads to a basis with quadratic precision.

**Theorem 2.43.** *For any convex polygon, the basis functions $\{\xi_{ij}\}$ constructed using the coefficients defined by (2.58)-(2.63) satisfy Q$_\xi$1-Q$_\xi$3.*

*Proof.* Based on Lemmas 2.35 and 2.36, it only remains to verify that Q$_c$1, Q$_c$2, and Q$_c$3 hold when $ab \in D$. Observe that $c^{aa}$ and $c^{bb}$ satisfy the following equations:

$$c^{aa} + c^{bb} + 4s = 2; \tag{2.64}$$

$$c^{aa} - c^{bb} + s(d_a - d_b) = 0; \tag{2.65}$$

$$c^{aa} + c^{bb} + 2s(d_a + d_b) = -2. \tag{2.66}$$

First, note that Q$_c$1 follows immediately from (2.58), (2.60) and (2.64).
The linear precision conditions (Q$_c$2) are just a matter of algebra. Equations (2.58)-(2.61) yield

$$\begin{aligned}
&c^{aa}\mathbf{v}_{aa} + c^{bb}\mathbf{v}_{bb} + 2c^{(a-1)a}\mathbf{v}_{(a-1)a} + 2c^{a(a+1)}\mathbf{v}_{a(a+1)} + 2c^{(b-1)b}\mathbf{v}_{(b-1)b} + 2c^{b(b+1)}\mathbf{v}_{b(b+1)} \\
&\qquad = (c^{aa} + c^{(a-1)a} + c^{a(a+1)})\mathbf{v}_a + (c^{bb} + c^{(b-1)b} + c^{b(b+1)})\mathbf{v}_b \\
&\qquad\quad + c^{(a-1)a}\mathbf{v}_{a-1} + c^{a(a+1)}\mathbf{v}_{a+1} + c^{(b-1)b}\mathbf{v}_{b-1} + c^{b(b+1)}\mathbf{v}_{b+1} \\
&\qquad = (c^{aa} + c^{(a-1)a} + c^{a(a+1)})\mathbf{v}_a + (c^{bb} + c^{(b-1)b} + c^{b(b+1)})\mathbf{v}_b + sd_a\mathbf{v}_a + sd_b\mathbf{v}_b \\
&\qquad = (c^{aa} + s + sd_a)\mathbf{v}_a + (c^{bb} + s + sd_b)\mathbf{v}_b.
\end{aligned}$$

Substituting the fixed coordinates of $\mathbf{v}_a = (-\ell, 0)$ and $\mathbf{v}_b = (\ell, 0)$ reduces this expression to the vector

$$\begin{bmatrix} (-c^{aa} - s - sd_a + c^{bb} + s + sd_b)\ell \\ 0 \end{bmatrix}.$$

Finally, we address Q$_c$3. Factoring the left side gives,

$$\begin{aligned}
&c^{aa}\mathbf{v}_a\mathbf{v}_a^T + c^{bb}\mathbf{v}_b\mathbf{v}_b^T + c^{(a-1)a}(\mathbf{v}_{a-1}\mathbf{v}_a^T + \mathbf{v}_a\mathbf{v}_{a-1}^T) + \cdots + c^{b(b+1)}(\mathbf{v}_b\mathbf{v}_{b+1}^T + \mathbf{v}_{b+1}\mathbf{v}_b^T) \\
&\qquad = c^{aa}\mathbf{v}_a\mathbf{v}_a^T + c^{bb}\mathbf{v}_b\mathbf{v}_b^T \\
&\qquad\qquad + (c^{(a-1)a}\mathbf{v}_{a-1} + c^{a(a+1)}\mathbf{v}_{a+1})\mathbf{v}_a^T + \mathbf{v}_a(c^{(a-1)a}\mathbf{v}_{a-1}^T + c^{a(a+1)}\mathbf{v}_{a+1}^T) \\
&\qquad\qquad + (c^{(b-1)b}\mathbf{v}_{b-1} + c^{b(b+1)}\mathbf{v}_{b+1})\mathbf{v}_b^T + \mathbf{v}_b(c^{(b-1)b}\mathbf{v}_{b-1}^T + c^{b(b+1)}\mathbf{v}_{b+1}^T) \\
&\qquad = c^{aa}\mathbf{v}_a\mathbf{v}_a^T + c^{bb}\mathbf{v}_b\mathbf{v}_b^T + sd_a\mathbf{v}_a\mathbf{v}_a^T + sd_b\mathbf{v}_b\mathbf{v}_b^T + \mathbf{v}_a(sd_a\mathbf{v}_a^T) + \mathbf{v}_b(sd_b\mathbf{v}_b^T) \\
&\qquad = (c^{aa} + 2sd_a)\mathbf{v}_a\mathbf{v}_a^T + (c^{bb} + 2sd_b)\mathbf{v}_b\mathbf{v}_b^T.
\end{aligned}$$

Again substituting the coordinates of $\mathbf{v}_a$ and $\mathbf{v}_b$, we obtain the matrix

$$\begin{bmatrix} \left(c^{aa} + 2sd_a + c^{bb} + 2sd_b\right)\ell^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

The right side of Q$_c$3 is

$$\mathbf{v}_a\mathbf{v}_b^T + \mathbf{v}_b\mathbf{v}_a^T = \begin{bmatrix} -2\ell^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence the only equation that must be satisfied is exactly (2.66). □

*Remark* 2.44. We note that $s$ was specifically chosen so that (2.64)-(2.66) would hold. The case $s = 1$ happens when $d_a = -d_b$, i.e. only for the quadrilateral.

**Theorem 2.45.** *Given a convex polygon satisfying G1, G2 and G3, $||\mathbb{A}||$ is uniformly bounded.*

*Proof.* By Lemma 2.37, it suffices to show a uniform bound on the six coefficients defined by equations (2.58)-(2.63). First we prove a uniform bound on $d_a$ and $d_b$ given G1-G3.

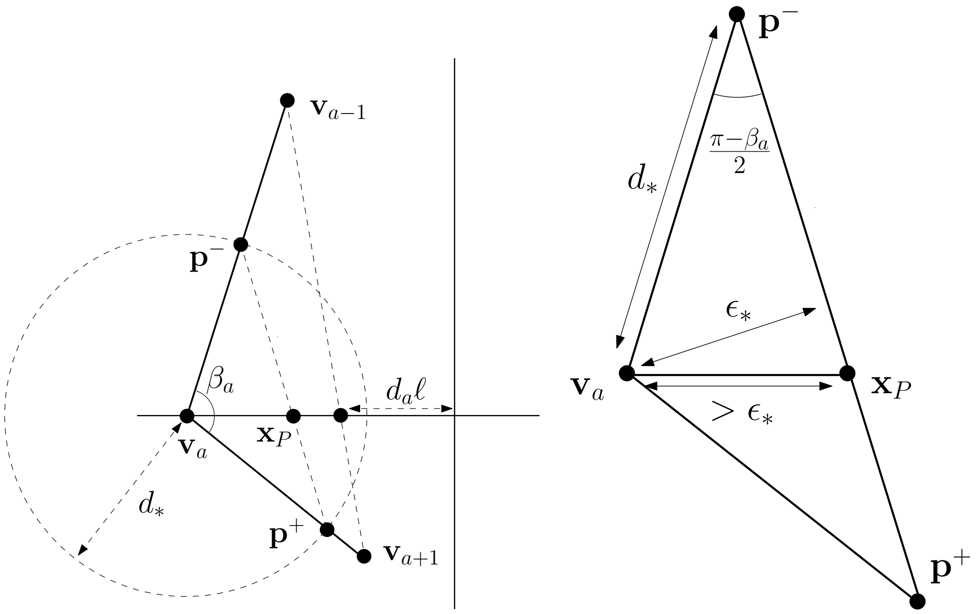


Figure 2.13: Notation used in proof of Theorem 2.45.

We fix some notation as shown in Figure 2.13. Let $C(\mathbf{v}_a, d_*)$ be the circle of radius $d_*$ (from G2) around $\mathbf{v}_a$. Let $\mathbf{p}^- := (p_x^-, p_y^-)$ and $\mathbf{p}^+ := (p_x^+, p_y^+)$ be the points on $C(\mathbf{v}_a, d_*)$ where the line segments to $\mathbf{v}_a$ from $\mathbf{v}_{a-1}$ and $\mathbf{v}_{a+1}$, respectively, intersect. The chord on $C(\mathbf{v}_a, d_*)$ between $\mathbf{p}^-$ and $\mathbf{p}^+$ intersects the $x$-axis at $\mathbf{x}_p := (x_p, 0)$. By convexity, $(\mathbf{v}_a)_x < x_p$.

To bound $x_p - (\mathbf{v}_a)_x$ below, note that the triangle $\mathbf{v}_a\mathbf{p}^-\mathbf{p}^+$ with angle $\beta_a$ at $\mathbf{v}_a$ is isosceles. Thus, the triangle $\mathbf{v}_a\mathbf{p}^-\mathbf{x}_p$ has angle $\measuredangle\mathbf{v}_a\mathbf{p}^-\mathbf{x}_p = (\pi - \beta_a)/2$, as shown at the right of Figure 2.13. The distance to the nearest point on the line segment between $\mathbf{p}^-$ and $\mathbf{p}^+$ is $d_* \sin\left(\frac{\pi-\beta_a}{2}\right)$. Based on G3, $\epsilon_* > 0$ is defined to be

$$x_p - (\mathbf{v}_a)_x \geq d_* \sin\left(\frac{\pi - \beta_a}{2}\right) > d_* \sin\left(\frac{\pi - \beta^*}{2}\right) =: \epsilon_* > 0. \tag{2.67}$$

Since $-d_a\ell < 1$ is the $x$-intercept of the line between $\mathbf{v}_{a-1}$ and $\mathbf{v}_{a+1}$, we have $x_p \leq -d_a\ell$. Then we rewrite $(\mathbf{v}_a)_x = -\ell$ in the geometrically suggestive form

$$(x_p - (\mathbf{v}_a)_x) + (-d_a\ell - x_p) + (0 + d_a\ell) = \ell.$$

Since $-d_a\ell - x_p \geq 0$, we have $\mathbf{x}_p - (\mathbf{v}_a)_x + d_a\ell \leq \ell$. Using (2.67), this becomes $d_a\ell < \ell - \epsilon_*$. Recall from Figure 2.12 and previous discussion that $d_a, d_b \in [-1, 1]$ and $-d_a \leq d_b$. By symmetry, $d_b\ell < \ell - \epsilon_*$ and hence $d_a + d_b < 2\ell - 2\epsilon_* < 1 - 2\epsilon_*$. We use the definition of $c^{aa}$ from (2.62), the derived bounds on $d_a$ and $d_b$, and the fact that $\ell \leq 1/2$ to conclude that

$$|c^{aa}| < \frac{|2 + 2d_a|}{1 + 2\epsilon_*} < \frac{2 + 2(1 - (\epsilon_*/\ell))}{1 + 2\epsilon_*} \leq \frac{4 - 4\epsilon_*}{1 + 2\epsilon_*} < 4.$$

Similarly, $|c^{bb}| < \frac{4-4\epsilon_*}{1+2\epsilon_*} < 4$. For the remaining coefficients, observe that the definition of $s$ in (2.57) implies that $0 < s < 2/(1 + 2\epsilon_*)$. Equation (2.58) and the $y$-component of equation (2.59) ensure that $c^{(a-1)a}/s$ and $c^{a(a+1)}/s$ are the coefficients of a convex combination of $\mathbf{v}_{a-1}$ and $\mathbf{v}_{a+1}$. Thus $c^{(a-1)a}, c^{(a+1)a} \in (0, s)$ and $s$ serves as an upper bound on the norms of each coefficient. Likewise, $|c^{(b-1)b}|, |c^{b(b+1)}| < s$. Therefore,

$$\max\left(\frac{4 - 4\epsilon_*}{1 + 2\epsilon_*}, \frac{2}{1 + 2\epsilon_*}, 1\right)$$

is a uniform bound on all the coefficients of $\mathbb{A}$. □

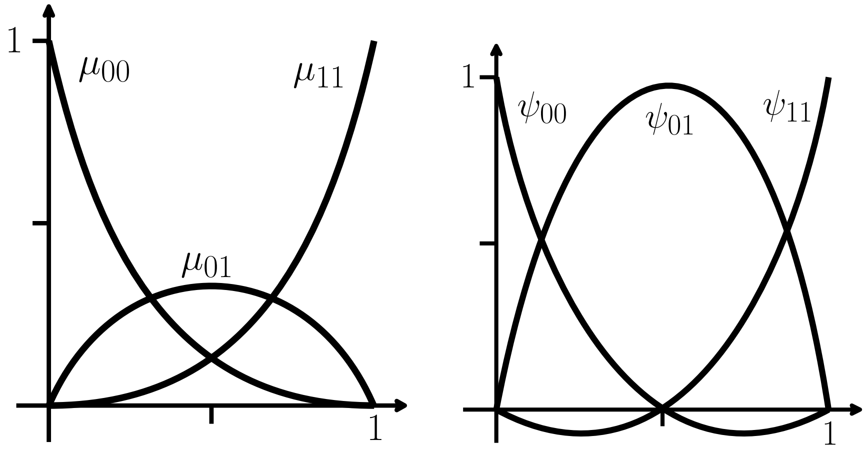


Figure 2.14: A comparison of the product barycentric basis (left) with the standard Lagrange basis (right) for quadratic polynomials in one dimension.

### Converting Serendipity Elements to Lagrange-like Elements

The $2n$ basis functions constructed thus far naturally correspond to vertices and edges of the polygon, but the functions associated to midpoints are not Lagrange-like. This is due to the fact that functions of the form $\xi_{i(i+1)}$ may not evaluate to 1 at $\mathbf{v}_{i(i+1)}$ or $\xi_{ii}$ may not evaluate to 0 at $\mathbf{v}_{i(i+1)}$, even though the set of $\{\xi_{ij}\}$ satisfies the partition of unity property $\mathrm{Q}_\xi 1$. To fix this, we apply a simple bounded linear transformation given by the matrix $\mathbb{B}$ defined below.

To motivate our approach, we first consider a simpler setting: polynomial bases over the unit segment $[0,1] \subset \mathbb{R}$. The barycentric functions on this domain are $\lambda_0(x) = 1 - x$, and $\lambda_1(x) = x$. Taking pairwise products, we get the quadratic basis $\mu_{00}(x) := (\lambda_0(x))^2 = (1-x)^2$, $\mu_{01}(x) := \lambda_0(x)\lambda_1(x) = (1-x)x$, and $\mu_{11}(x) := (\lambda_1(x))^2 = x^2$, shown on the left of Figure 2.14. This basis is not Lagrange-like since $\mu_{01}(1/2) \neq 1$ and $\mu_{00}(1/2), \mu_{11}(1/2) \neq 0$. The quadratic Lagrange basis is given by $\psi_{00}(x) := 2(1-x)\left(\frac{1}{2} - x\right)$, $\psi_{01}(x) := 4(1-x)x$, and $\psi_{11}(x) := 2\left(x - \frac{1}{2}\right)x$, shown on the right of Figure 2.14. These two bases are related by the linear transformation $\mathbb{B}_{\mathrm{1D}}$:

$$[\psi_{ij}] = \begin{bmatrix} \psi_{00} \\ \psi_{11} \\ \psi_{01} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} \mu_{00} \\ \mu_{11} \\ \mu_{01} \end{bmatrix} = \mathbb{B}_{\mathrm{1D}}[\mu_{ij}]. \tag{2.68}$$

This procedure generalizes to the case of converting the 2D serendipity basis $\{\xi_{ij}\}$ to a Lagrange like basis $\{\psi_{ij}\}$. Define

$$\psi_{ii} := \xi_{ii} - \xi_{i,i+1} - \xi_{i-1,i} \quad \text{and} \quad \psi_{i,i+1} = 4\,\xi_{i,i+1}.$$

Using our conventions for basis ordering and index notation, the transformation matrix $\mathbb{B}$ taking $[\xi_{ij}]$ to $[\psi_{ij}]$ has the structure

$$[\psi_{ij}] = \begin{bmatrix} \psi_{11} \\ \psi_{22} \\ \vdots \\ \psi_{nn} \\ \psi_{12} \\ \psi_{23} \\ \vdots \\ \psi_{n1} \end{bmatrix} = \left[\begin{array}{ccccc|ccccc}
1 & & & & & -1 & & \cdots & & -1 \\
 & 1 & & & & -1 & -1 & \cdots & & \\
 & & \ddots & & & & \ddots & \ddots & & \\
 & & & \ddots & & & & \ddots & \ddots & \\
 & & & & 1 & & & & -1 & -1 \\
\hline
 & & & & & 4 & & & & \\
 & & & & & & 4 & & & \\
 & & 0 & & & & & \ddots & & \\
 & & & & & & & & \ddots & \\
 & & & & & & & & & 4
\end{array}\right] \begin{bmatrix} \xi_{11} \\ \xi_{22} \\ \vdots \\ \xi_{nn} \\ \xi_{12} \\ \xi_{23} \\ \vdots \\ \xi_{n1} \end{bmatrix} = \mathbb{B}[\xi_{ij}].$$

The following proposition says that the functions $\{\psi_{ij}\}$ defined by the above transformation are Lagrange-like.

**Proposition 2.46.** *For all $i, j \in \{1, \ldots, n\}$, $\psi_{ii}(\mathbf{v}_j) = \delta_i^j$, $\psi_{ii}(\mathbf{v}_{j,j+1}) = 0$, $\psi_{i(i+1)}(\mathbf{v}_j) = 0$, and $\psi_{i(i+1)}(\mathbf{v}_{j,j+1}) = \delta_i^j$.*

*Proof.* We show the last claim first. By the definitions of $\mathbb{B}$ and $\mathbb{A}$, we have

$$\psi_{i(i+1)}(\mathbf{v}_{j,j+1}) = 4\,\xi_{i(i+1)}(\mathbf{v}_{j,j+1}) = 4\left(\sum_{a=1}^{n} c_{aa}^{i(i+1)}\mu_{aa}(\mathbf{v}_{j,j+1}) + \sum_{a<b} c_{ab}^{i(i+1)}\mu_{ab}(\mathbf{v}_{j,j+1})\right)$$

Since $\lambda_a$ is piecewise linear on the boundary of the polygon and $\lambda_a(\mathbf{v}_j) = \delta_a^j$ (by B6), we have that $\lambda_a(\mathbf{v}_{j,j+1}) = 1/2$ if $a \in \{j, j+1\}$ and zero otherwise. Accordingly, $\mu_{aa}(\mathbf{v}_{j,j+1}) = 1/4$ if $a \in \{j, j+1\}$ and zero otherwise, while $\mu_{ab}(\mathbf{v}_{j,j+1}) = 1/4$ if $\{a, b\} = \{j, j+1\}$ and zero otherwise.

$$\psi_{i(i+1)}(\mathbf{v}_{j,j+1}) = 4\left(\left(c_{jj}^{i(i+1)} + c_{(j+1)(j+1)}^{i(i+1)}\right)\cdot\frac{1}{4} + c_{j(j+1)}^{i(i+1)}\cdot\frac{1}{4}\right) = c_{j(j+1)}^{i(i+1)} = \delta_{ij},$$

since the identity structure of $\mathbb{A}$ as given in (2.35) implies that $c_{jj}^{i(i+1)} = c_{(j+1)(j+1)}^{i(i+1)} = 0$ and that $c_{j(j+1)}^{i(i+1)} = \delta_{ij}$.
Next, observe that $\mu_{ab}(\mathbf{v}_j) = \lambda_a(\mathbf{v}_j)\lambda_b(\mathbf{v}_j) = 1$ if $a = b = j$ and 0 otherwise. Hence, any term of the form $c_{ab}^{**}\mu_{ab}(\mathbf{v}_j)$ for $a \neq b$ is necessarily zero. Therefore, by a similar expansion, $\psi_{i(i+1)}(\mathbf{v}_j) = c_{jj}^{i(i+1)} = 0$, proving the penultimate claim.
For the first two claims, similar analysis yields

$$\begin{aligned}\psi_{ii}(\mathbf{v}_j) &= \xi_{ii}(\mathbf{v}_j) - \xi_{i(i+1)}(\mathbf{v}_j) - \xi_{(i-1)i}(\mathbf{v}_j)\\ &= c_{jj}^{ii}\cdot 1 - c_{jj}^{i(i+1)}\cdot 1 - c_{jj}^{(i-1)i}\cdot 1\\ &= c_{jj}^{ii} = \delta_{ij},\end{aligned}$$

again by the identity structure of $\mathbb{A}$. Finally, by similar analysis, we have that

$$\begin{aligned}\psi_{ii}(\mathbf{v}_{j,j+1}) &= \xi_{ii}(\mathbf{v}_{j,j+1}) - \xi_{i(i+1)}(\mathbf{v}_{j,j+1}) - \xi_{(i-1)i}(\mathbf{v}_{j,j+1})\\ &= (c_{jj}^{ii} + c_{(j+1)(j+1)}^{ii} + c_{j(j+1)}^{ii})\frac{1}{4} - \xi_{i(i+1)}(\mathbf{v}_{j,j+1}) - \xi_{(i-1)i}(\mathbf{v}_{j,j+1})\\ &= (\delta_{ij} + \delta_{i(j+1)})\frac{1}{4} - \frac{1}{4}\delta_{ij} - \frac{1}{4}\delta_{i(j+1)} = 0,\end{aligned}$$

completing the proof. □

In closing, note that $||\mathbb{B}||$ is uniformly bounded since its entries all lie in $\{-1, 0, 1, 4\}$.

**Applications and Extensions** Our quadratic serendipity element construction has a number of uses in modern finite element application contexts. First, the construction for quadrilaterals given in Section 2.7.2 allows for quadratic order methods on *arbitrary* quadrilateral meshes with only eight basis functions per element instead of the nine used in a bilinear map of the biquadratic tensor product basis on a square. In particular, we show that our approach maintains quadratic convergence on a mesh of convex quadrilaterals known to result in only linear convergence when traditional serendipity elements are mapped non-affinely [13].

We solve Poisson's equation on a square domain composed of $n^2$ trapezoidal elements as shown in Figure 2.15 (left). Boundary conditions are prescribed according to the solution $u(x, y) = \sin(x)e^y$; we use our construction from Section 2.7.2 starting with mean value coordinates $\{\lambda_i^{\text{MVal}}\}$. Mean value coordinates were selected based on a few advantages they have over other types: they are easy to compute based on an explicit formula and the coordinate gradients do not degrade based on large interior angles [170]. For this particular example, where no interior angles asymptotically approach 180°, Wachspress coordinates give very similar results. As shown in Figure 2.15 (right), the expected convergence rates from our theoretical analysis are observed, namely, cubic in the $L^2$-norm and quadratic in the $H^1$-norm.

An additional application of our method is to adaptive finite elements, such as the one shown in Figure 2.16. This is possible since the result of Theorem 2.45 still holds if G3 fails to hold only on a set of consecutive vertices of the polygon. This weakened condition suffices since consecutive large angles in the polygon do not cause the coefficients $c_{ab}^{ij}$ to blow up. For instance, consider the degenerate pentagon shown in Figure 2.16 which satisfies this weaker condition but not G3. Examining the potentially problematic coefficients $c_{25}^{ij}$, observe that the lines through $\mathbf{v}_1$, $\mathbf{v}_4$ and $\mathbf{v}_1$, $\mathbf{v}_3$ both intersect the midpoint of the line through $\mathbf{v}_2$, $\mathbf{v}_5$ (which happens to be $\mathbf{v}_1$). In the computation of the $c^{ij}_{25}$ coefficients, the associated values $d_2$ and $d_5$ are both zero and hence $s = 1$ (recall Figure 2.12 and formula (2.57)). Since $s$ is bounded away from $\infty$, the analysis from the proof of Theorem 2.45 holds as stated for these coefficients and hence for the entire element. A more detailed analysis of such large-angle elements is an open question for future study.

Nevertheless, the geometric hypotheses of Theorem 2.45 cannot be relaxed entirely. Arbitrarily large non-consecutive large angles as well as very short edges, can cause a blowup in the coefficients used in the construction of $\mathbb{A}$, as shown in Figure 2.17. In the left figure, as edges $\mathbf{v}_{a-1}\mathbf{v}_a$ and $\mathbf{v}_{b-1}\mathbf{v}_b$ approach length zero, $d_a$ and $d_b$ both approach one, meaning $s$ (in the construction of Section 2.7.2) approaches $\infty$. In this case, the coefficients $c_{ij}^{(a-1)a}$ and $c_{ij}^{(b-1)b}$ grow larger without bound, thereby violating the result of Theorem 2.45. In the right figure, as the overall shape approaches a square, $d_a$ and $d_b$ again approach one so that $s$ again approaches $\infty$. In this case, all the coefficients $c_{ij}^{(a-1)a}$, $c_{ij}^{a(a+1)}$, $c_{ij}^{(b-1)b}$ and $c_{ij}^{b(b+1)}$ all grow without bound. Nevertheless, if these types of extreme geometries are required, it may be possible to devise alternative definitions of the $c_{ij}^{ab}$ coefficients satisfying Q$_c$1-Q$_c$3 with controlled norm estimates since the set of restrictions Q$_c$1-Q$_c$3 does not have full rank. Note that this flexibility has lead to multiple constructions of the traditional serendipity square [146, 131]. Cursory numerical experimentation suggests that some bounded construction exists even in the degenerate situation.

The computational cost of our method is an important consideration to application contexts. A typical finite element method using our approach would involve the following steps: (1) selecting $\lambda_i$ coordinates and implementing the corresponding $\psi_{ij}$ basis functions, (2) defining a quadrature rule for each affine-equivalent class of shapes appearing in the domain mesh, (3) assembling a matrix $\mathbb{L}$ representing the discrete version of linear operator, and (4) solving a linear system of the form $\mathbb{L}u = f$. The quadrature step may incur some computational effort, however, if only a few shape templates are needed, this is a one-time fixed pre-preprocessing cost. In the trapezoidal mesh example from Figure 2.15, for instance, we only needed one quadrature rule as all domain shapes were affinity equivalent. Assembling the matrix $\mathbb{L}$ may also be expensive as the entries involve integrals of products of gradients of $\psi_{ij}$ functions. Again, however, this cost is incurred only once per affine-equivalent domain shape and thus can be reasonable to allow, depending on the application context.

The computational advantage to our approach comes in the final linear solve. The size of the matrix $\mathbb{L}$ is proportional to the number of edges in the mesh, matching the size of the corresponding matrix for quadratic Lagrange elements on triangles or quadratic serendipity elements on squares. If the pairwise products $\mu_{ab}$ were used instead of the $\psi_{ij}$ functions, the size of $\mathbb{L}$ would be proportional to the *square* of the number of edges in the mesh, a substantial difference.

Finally, we note that although this construction is specific to quadratic elements, the approach seems adaptable, with some effort, to the construction of cubic and higher order serendipity elements on generic convex polygons. As a larger linear system must be satisfied, stating an explicit solution becomes complex. Further research along these lines should probably assert the existence of a uniformly bounded solution without specifying the construction. In practice, a least squares solver could be used to construct such a basis numerically.

### 2.7.3 Construction of Scalar and Vector Finite Element Families on Polygonal and Polyhedral Meshes

In this work, we join and expand three threads of research in the analysis of modern finite element methods: polytope domain meshing, generalized barycentric coordinates, and families of finite-dimensional solution spaces characterized by finite element exterior calculus. It is well-known that on simplicial meshes, standard barycentric coordinates provide a local basis for the lowest-order $H^1$-conforming scalar-valued finite element spaces, commonly called the Lagrange elements. Further, local bases for the lowest-order vector-valued Brezzi-Douglas-Marini[40], Raviart-Thomas[173], and Nédélec[39, 158, 159] finite element spaces on simplices can also be defined in a canonical fashion from an associated set of standard barycentric functions. Here, we use generalized barycentric coordinates in an analogous fashion on meshes of convex polytopes, in dimensions 2 and 3, to construct local bases with the same global continuity and polynomial reproduction properties as their simplicial counterparts.

We have previously analyzed linear order, scalar-valued methods on polygonal meshes[100, 170] using four different types of generalized barycentric coordinates: Wachspress[216, 215], Sibson[82, 192], harmonic[53, 125, 151], and mean value[83, 85, 86]. The analysis was extended by Gillette, Floater and Sukumar in the case of Wachspress coordinates to convex polytopes in any dimension[84], based on work by Warren and colleagues[127, 218, 219]. We have also shown how taking pairwise products of generalized barycentric coordinates can be used to construct quadratic order methods on polygons[171].

Our expansion in this paper to vector-valued methods is inspired by Whitney differential forms, first defined in[220]. Bossavit first recognized that Whitney forms could be used to construct basis functions for computational electromagnetics[34], sparking a long chain of research, recently unified by the theory of finite element exterior calculus[9, 10, 11]. Some prior work has

| n | k | functions |
|---|---|---|
| 2 | 0 | $\lambda_i$ |
|  | 1 | $\lambda_i \nabla \lambda_j$ |
|  |  | $\mathcal{W}_{ij}$ |
|  |  | rot $\lambda_i \nabla \lambda_j$ |
|  |  | rot $\mathcal{W}_{ij}$ |
|  | 2 | $\lambda_i \nabla \lambda_j \cdot \text{rot}\nabla \lambda_k$ |
|  |  | $\mathcal{W}_{ijk}$ |
| 3 | 0 | $\lambda_i$ |
|  | 1 | $\lambda_i \nabla \lambda_j$ |
|  |  | $\mathcal{W}_{ij}$ |
|  | 2 | $\lambda_i \nabla \lambda_j \times \nabla \lambda_k$ |
|  |  | $\mathcal{W}_{ijk}$ |
|  | 3 | $\lambda_i \nabla \lambda_j \cdot (\nabla \lambda_k \times \nabla \lambda_\ell)$ |
|  |  | $\mathcal{W}_{ijk\ell}$ |

Table 2.1: For meshes of convex $n$-dimensional polytopes in $\mathbb{R}^n$, $n = 2$ or $3$, computational basis functions for each differential form order $0 \leq k \leq n$. The generalized barycentric coordinate functions $\lambda$, Whitney-like functions $\mathcal{W}$, and polynomial spaces $\mathcal{P}_r \Lambda^k$ are defined in Section 2.7.3.

explored the possibility of Whitney functions over non-simplicial elements in specific cases of rectangular grids[104], square-base pyramids[105], and prisms[32]. Other authors have examined the ability of generalized Whitney functions to recover constant-valued forms in certain cases[78, 132], whereas here we show their ability to reproduce ***all*** the elements of the spaces denoted $\mathcal{P}_1^- \Lambda^k$ in finite element exterior calculus. Gillette and Bajaj considered the use of generalized Whitney forms on polytope meshes defined by duality from a simplicial mesh[98, 99], which illustrated potential benefits to discrete exterior calculus[116], computational magnetostatics, and Darcy flow modeling.

| n | k | global continuity | polynomial reproduction |
|---|---|---|---|
| 2 | 0 | $H^1(\mathcal{M})$ | $\mathcal{P}_1 \Lambda^0(\mathcal{M})$ |
|  | 1 | $H(\text{curl}, \mathcal{M})$, by Theorem 2.49 | $\mathcal{P}_1 \Lambda^1(\mathcal{M})$, by Theorem 2.52 |
|  |  | $H(\text{curl}, \mathcal{M})$, by Theorem 2.49 | $\mathcal{P}_1^- \Lambda^1(\mathcal{M})$, by Theorem 2.57 |
|  |  | $H(\text{div}, \mathcal{M})$, see Remark 2.50 | $\mathcal{P}_1 \Lambda^1(\mathcal{M})$, by Corollary 2.53 |
|  |  | $H(\text{div}, \mathcal{M})$, see Remark 2.50 | $\mathcal{P}_1^- \Lambda^1(\mathcal{M})$, by Corollary 2.58 |
|  | 2 | none (piecewise linear) | $\mathcal{P}_1 \Lambda^2(\mathcal{M})$, by Theorem 2.55 |
|  |  | none (piecewise constant) | $\mathcal{P}_1^- \Lambda^2(\mathcal{M})$, see Remark 2.60 |
| 3 | 0 | $H^1(\mathcal{M})$ | $\mathcal{P}_1 \Lambda^0(\mathcal{M})$ |
|  | 1 | $H(\text{curl}, \mathcal{M})$, by Theorem 2.49 | $\mathcal{P}_1 \Lambda^1(\mathcal{M})$, by Theorem 2.52 |
|  |  | $H(\text{curl}, \mathcal{M})$, by Theorem 2.49 | $\mathcal{P}_1^- \Lambda^1(\mathcal{M})$, by Theorem 2.57 |
|  | 2 | $H(\text{div}, \mathcal{M})$, by Theorem 2.51 | $\mathcal{P}_1 \Lambda^2(\mathcal{M})$, by Theorem 2.54 |
|  |  | $H(\text{div}, \mathcal{M})$, by Theorem 2.51 | $\mathcal{P}_1^- \Lambda^2(\mathcal{M})$, by Theorem 2.59 |
|  | 3 | none (piecewise linear) | $\mathcal{P}_1 \Lambda^3(\mathcal{M})$, see Remark 2.60 |
|  |  | none (piecewise constant) | $\mathcal{P}_1^- \Lambda^3(\mathcal{M})$, see Remark 2.60 |

Table 2.2: Summary of the global continuity and polynomial reproduction properties of the spaces considered.

Using the bases defined in Table 2.1, our main results are summarized in Table 2.2. On a mesh of convex $n$-dimensional polytopes in $\mathbb{R}^n$ with $n = 2$ or $3$, we construct computational basis functions associated to the polytope elements for each differential form order $k$ as indicated. Each function is built from generalized barycentric coordinates, denoted $\lambda_i$, and their gradients; formulae for the Whitney-like functions, denoted $\mathcal{W}$, are given in Section 2.7.3. In the vector-valued cases ($0 < k < n$), we prove that the functions agree on tangential or normal components at inter-element boundaries, providing global continuity in $H(\text{curl})$ or $H(\text{div})$. The two families of polynomial differential forms that are reproduced, $\mathcal{P}_r \Lambda^k$ and $\mathcal{P}_r^- \Lambda^k$, were shown to recover and generalize the classical simplicial finite element spaces mentioned previously, via the theory of finite element exterior calculus[9, 11].

The outline of the paper is as follows. In Section 2.7.3, we describe relevant theory and prior work in the areas of finite element exterior calculus, generalized barycentric coordinates, and Whitney forms. In Section 2.7.3, we show how the functions listed in Table 2.1 can be used to build piecewise-defined functions with global continuity in $H^1$, $H(\text{curl})$ or $H(\text{div})$, as indicated. In Section 2.7.3, we show how these same functions can reproduce the requisite polynomial differential forms from $\mathcal{P}_1\Lambda^k$ or $\mathcal{P}_1^-\Lambda^k$, as indicated in Table 2.1, by exhibiting explicit linear combinations whose coefficients depend only on the location of the vertices of the mesh. In Section 9, we count the basis functions constructed by our approach on generic polygons and polyhedra and explain how the size of the basis could be reduced in certain cases. We close with a discussion relating this work to serendipity and virtual element methods.

**Background and prior work**

**Spaces from Finite Element Exterior Calculus** Finite element spaces can be broadly classified according to three parameters: $n$, the spatial dimension of the domain, $r$, the order of error decay, and $k$, the differential form order of the solution space. The $k$ parameter can be understood in terms of the classical finite element sequence for a domain $\Omega \subset \mathbb{R}^n$ with $n = 2$ or 3, commonly written as

$$\begin{tikzcd}
n = 2 : & H^1 \arrow[r, "\text{grad}"] & H(\text{curl}) \arrow[r, leftrightarrow, "\text{rot}"] & H(\text{div}) \arrow[r, "\text{div}"] & L^2 \\
n = 3 : & H^1 \arrow[r, "\text{grad}"] & H(\text{curl}) \arrow[r, "\text{curl}"] & H(\text{div}) \arrow[r, "\text{div}"] & L^2
\end{tikzcd}$$

Note that for $n = 2$, we use the definitions

$$\text{curl}\,\vec{F} := \frac{\partial F_1}{\partial y} - \frac{\partial F_2}{\partial x} \quad \text{and} \quad \text{rot}\vec{F} := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\vec{F} \quad \text{where} \quad \vec{F}(x,y) := \begin{bmatrix} F_1(x,y) \\ F_2(x,y) \end{bmatrix}.$$

Thus, in $\mathbb{R}^2$, we have both $\text{curl}\nabla\phi = 0$ and $\text{divrot}\,\nabla\phi = 0$ for any $\phi \in H^2$. Put differently, rot gives an isomorphism from $H(\text{curl})$ to $H(\text{div})$ in $\mathbb{R}^2$. In some cases we will write $H(\text{curl}, \Omega)$ and $H(\text{div}, \Omega)$ if we wish to emphasize the domain in consideration.

In the terminology of differential topology, the applicable sequence is described more simply as the $L^2$ deRham complex of $\Omega$. The spaces are re-cast as differential form spaces $H\Lambda^k$ and the operators as instances of the exterior derivative $d_k$, yielding

$$\begin{tikzcd}
n = 2 : & H\Lambda^0 \arrow[r, "d_0"] & H\Lambda^1 \arrow[r, leftrightarrow, "\cong"] & H\Lambda^1 \arrow[r, "d_1"] & H\Lambda^2 \\
n = 3 : & H\Lambda^0 \arrow[r, "d_0"] & H\Lambda^1 \arrow[r, "d_1"] & H\Lambda^2 \arrow[r, "d_2"] & H\Lambda^3
\end{tikzcd}$$

Finite element methods seek approximate solutions to a PDE in finite dimensional subspaces $\Lambda_h^k$ of the $H\Lambda^k$ spaces, where $h$ denotes the maximum width of a domain element associated to the subspace. The theory of finite element exterior calculus classifies two families of suitable choices of $\Lambda_h^k$ spaces on meshes of simplices, denoted $\mathcal{P}_r\Lambda^k$ and $\mathcal{P}_r^-\Lambda^k$[9, 11]. The space $\mathcal{P}_r\Lambda^k$ is defined as "those differential forms which, when applied to a constant vector field, have the indicated polynomial dependence"[11, p. 328], which can be interpreted informally as the set of differential $k$ forms with polynomial coefficients of total degree at most $r$. The space $\mathcal{P}_r^-\Lambda^k$ is then defined as the direct sum

$$\mathcal{P}_r^-\Lambda^k := \mathcal{P}_{r-1}\Lambda^k \oplus \kappa\mathcal{H}_{r-1}\Lambda^{k+1}, \tag{2.69}$$

where $\kappa$ is the Koszul operator and $\mathcal{H}_r$ denotes homogeneous polynomials of degree $r$[11, p. 331]. We will use the coordinate formulation of $\kappa$, given in[11, p. 329] as follows. Let $\omega \in \Lambda^k$ and suppose that it can be written in local coordinates as $\omega_x = a(x)dx_{\sigma_1} \wedge \cdots \wedge dx_{\sigma_k}$. Then $\kappa\omega$ is written as

$$(\kappa\omega)_x := \sum_{i=1}^{k}(-1)^{i+1}a(x)x_{\sigma(i)}dx_{\sigma_1} \wedge \cdots \wedge \widehat{dx_{\sigma_i}} \wedge \cdots \wedge dx_{\sigma_k}, \tag{2.70}$$

where $\wedge$ denotes the wedge product and $\widehat{dx_{\sigma_i}}$ means that the term is omitted. We summarize the relationship between the spaces $\mathcal{P}_1\Lambda^k$, $\mathcal{P}_1^-\Lambda^k$ and certain well-known finite element families in dimension $n = 2$ or 3 in Table 2.3.

A crucial property of $\mathcal{P}_r\Lambda^k$ and $\mathcal{P}_r^-\Lambda^k$ is that each includes in its span a sufficient number of polynomial differential $k$-forms to ensure an *a priori* error estimate of order $r$ in $H\Lambda^k$ norm. In the classical description of finite element spaces, this approximation

| n | k | dim | space | classical description | reference |
|---|---|---|---|---|---|
| 2 | 0 | 3 | $\mathcal{P}_1\Lambda^0(\mathcal{T})$ | Lagrange, degree $\leq 1$ | |
| | | 3 | $\mathcal{P}_1^-\Lambda^0(\mathcal{T})$ | Lagrange, degree $\leq 1$ | |
| | 1 | 6 | $\mathcal{P}_1\Lambda^1(\mathcal{T})$ | Brezzi-Douglas-Marini, degree $\leq 1$ | [40] |
| | | 3 | $\mathcal{P}_1^-\Lambda^1(\mathcal{T})$ | Raviart-Thomas, order 0 | [173] |
| | 2 | 3 | $\mathcal{P}_1\Lambda^2(\mathcal{T})$ | discontinuous linear | |
| | | 1 | $\mathcal{P}_1^-\Lambda^2(\mathcal{T})$ | discontinuous piecewise constant | |
| 3 | 0 | 4 | $\mathcal{P}_1\Lambda^0(\mathcal{T})$ | Lagrange, degree $\leq 1$ | |
| | | 4 | $\mathcal{P}_1^-\Lambda^0(\mathcal{T})$ | Lagrange, degree $\leq 1$ | |
| | 1 | 12 | $\mathcal{P}_1\Lambda^1(\mathcal{T})$ | Nédélec second kind $H$(curl), degree $\leq 1$ | [159, 39] |
| | | 6 | $\mathcal{P}_1^-\Lambda^1(\mathcal{T})$ | Nédélec first kind $H$(curl), order 0 | [158] |
| | 2 | 12 | $\mathcal{P}_1\Lambda^2(\mathcal{T})$ | Nédélec second kind $H$(div), degree $\leq 1$ | [159, 39] |
| | | 4 | $\mathcal{P}_1^-\Lambda^2(\mathcal{T})$ | Nédélec first kind $H$(div), order 0 | [158] |
| | 3 | 4 | $\mathcal{P}_1\Lambda^3(\mathcal{T})$ | discontinuous linear | |
| | | 1 | $\mathcal{P}_1^-\Lambda^3(\mathcal{T})$ | discontinuous piecewise constant | |

Table 2.3: Correspondence between $\mathcal{P}_1\Lambda^k(\mathcal{T})$, $\mathcal{P}_1^-\Lambda^k(\mathcal{T})$ and common finite element spaces associated to a simplex $\mathcal{T}$ of dimension $n$. Further explanation of these relationships can be found in [*Arnold, Falk, Winther 2006 and 2010*]. On simplices, our constructions recover known local bases for each of these spaces.

power is immediate; any computational or 'local' basis used for implementation of these spaces must, by definition, span the requisite polynomial differential forms. The main results of this paper are proofs that generalized barycentric coordinates can be used as local bases on polygonal and polyhedral element geometries to create analogues to the lowest order $\mathcal{P}_r\Lambda^k$ and $\mathcal{P}_r^-\Lambda^k$ spaces with the same polynomial approximation power and global continuity properties.

In the remainder of the paper, we will frequently use standard vector proxies[2] in place of differential form notation, as indicated here:

$$\begin{aligned}
\begin{bmatrix} u_1 & u_2 \end{bmatrix}^T &\longleftrightarrow u_1 dx_1 + u_2 dx_2 \in \Lambda^1(\mathbb{R}^2), \\
\begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}^T &\longleftrightarrow v_1 dx_1 + v_2 dx_2 + v_3 dx_3 \in \Lambda^1(\mathbb{R}^3), \\
\begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix}^T &\longleftrightarrow w_1 dx_2 dx_3 + w_2 dx_3 dx_1 + w_3 dx_1 dx_2 \in \Lambda^2(\mathbb{R}^3).
\end{aligned}$$

**Generalized Barycentric Coordinates** Let $\mathbf{m}$ be a convex $n$-dimensional polytope in $\mathbb{R}^n$ with vertex set $\{\mathbf{v}_i\}$, written as column vectors. A set of non-negative functions $\{\lambda_i\} : \mathbf{m} \to \mathbb{R}$ are called **generalized barycentric coordinates** on $\mathbf{m}$ if for any linear function $L : \mathbf{m} \to \mathbb{R}$, we can write

$$L = \sum_i L(\mathbf{v}_i)\lambda_i. \tag{2.71}$$

We will use the notation $\mathbb{I}$ to denote the $n \times n$ identity matrix and **x** to denote the vector $(x_1, x_2, \cdots, x_n)^T$ where $x_i$ is the $i$th coordinate in $\mathbb{R}^n$. We have the following useful identities:

$$\sum_i \lambda_i = 1 \tag{2.72}$$

$$\sum_i \mathbf{v}_i \lambda_i(\mathbf{x}) = \mathbf{x} \tag{2.73}$$

$$\sum_i \nabla\lambda_i(\mathbf{x}) = 0 \tag{2.74}$$

$$\sum_i \mathbf{v}_i \nabla\lambda_i^T(\mathbf{x}) = \mathbb{I} \tag{2.75}$$

Equations (2.72) and (2.73) follow immediately from (2.71) while (2.74) and (2.75) follow by taking the gradient of equations (2.72) and (2.73), respectively.

Applications of generalized barycentric coordinates to finite element methods have primarily focused on scalar-valued PDE problems[153, 172, 203, 204, 221]. By incorporating gradients of the $\lambda_i$ functions, we can exploit the above identities to build functions for vector-valued problems.

**Whitney forms** Let $\mathbf{m}$ be a convex $n$-dimensional polytope in $\mathbb{R}^n$ with vertex set $\{\mathbf{v}_i\}$ and an associated set of generalized barycentric coordinates $\{\lambda_i\}$. Define associated sets of index pairs and triples by

$$E := \{(i,j) \; : \; \mathbf{v}_i, \mathbf{v}_j \in \mathbf{m}\}, \tag{2.76}$$

$$T := \{(i,j,k) \; : \; \mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k \in \mathbf{m}\}. \tag{2.77}$$

If $\mathbf{m}$ is a *simplex*, the elements of the set

$$\{\lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i \; : \; (i,j) \in E\}$$

are called **Whitney 1-forms** and are part of a more general construction[220], which we now present. Again, if $\mathbf{m}$ is a *simplex*, the **Whitney $k$-forms** are elements of the set

$$\left\{ k! \sum_{i=0}^{k} (-1)^i \; \lambda_{j_i} \; d\lambda_{j_0} \wedge \ldots \wedge \widehat{d\lambda_{j_i}} \wedge \ldots \wedge d\lambda_{j_k} \right\}, \tag{2.78}$$

where $j_0, \ldots, j_k$ are indices of vertices of $\mathbf{m}$. Up to sign, this yields a set of $\binom{n+1}{k+1}$ distinct functions and provides a local basis for $\mathcal{P}_1^- \Lambda^k$[10].
We now generalize these definitions to the case where $\mathbf{m}$ is non necessarily a simplex. For any $(i,j) \in E$, define a generalized Whitney 1-form on $\mathbf{m}$ by

$$\mathcal{W}_{ij} := \lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i. \tag{2.79}$$

If $n = 3$, then for any $(i,j,k) \in T$, define a generalized Whitney 2-form on $\mathbf{m}$ by

$$\mathcal{W}_{ijk} := (w_i \nabla w_j \times \nabla w_k) + (w_j \nabla w_k \times \nabla w_i) + (w_k \nabla w_i \times \nabla w_j). \tag{2.80}$$

Note that $\mathcal{W}_{ii} = 0$ and if $i$, $j$, and $k$ are not distinct then $\mathcal{W}_{ijk} = 0$.
Whitney forms have natural interpretations as vector fields when $k = 1$ or $n - 1$. Interpolation of vector fields requires less data regularity than the canonical scalar interpolation theory using nodal values. Averaged interpolation developed for scalar spaces[56, 189] has been extended to families of spaces from finite element exterior calculus[54]. Recent results on polygons and polyhedra can be extended to less regular data with average interpolation following the framework in[169], based on affine invariance of the coordinates.

### Global Continuity Results

We first present results about the global continuity properties of vector-valued functions defined in terms of generalized barycentric coordinates and their gradients over meshes of $n$-dimensional polytopes in $\mathbb{R}^n$ with $n = 2$ or 3. When we say that a function is defined 'piecewise with respect to a mesh,' we mean that the definition of the function on the interior of a mesh element depends only on geometrical properties of the element (as opposed to depending on adjacent elements, for instance). We begin with a general result about global continuity in such a setting.

**Proposition 2.47.** *Fix a mesh $\mathcal{M}$ of $n$-dimensional polytopes in $\mathbb{R}^n$ with $n = 2$ or 3. Let $\mathbf{u}$ be a vector field defined piecewise with respect to $\mathcal{M}$. Let $\mathfrak{f}$ be a face of codimension 1 with $\mathbf{u}_1$, $\mathbf{u}_2$ denoting the values of $\mathbf{u}$ on $\mathfrak{f}$ as defined by the two $n$-dimensional mesh elements sharing $\mathfrak{f}$. Write $\mathbf{u}_i = T_{\mathfrak{f}}(\mathbf{u}_i) + N_{\mathfrak{f}}(\mathbf{u}_i)$ where $T_{\mathfrak{f}}(\mathbf{u}_i)$ and $N_{\mathfrak{f}}(\mathbf{u}_i)$ are the vector projections of $\mathbf{u}_i$ onto $\mathfrak{f}$ and its outward normal, respectively.*

*(i.) If $T_{\mathfrak{f}}(\mathbf{u}_1) = T_{\mathfrak{f}}(\mathbf{u}_2)$ for all $\mathfrak{f} \in \mathcal{M}$ then $\mathbf{u} \in H(curl, \mathcal{M})$.*

*(ii.) If $N_{\mathfrak{f}}(\mathbf{u}_1) = N_{\mathfrak{f}}(\mathbf{u}_2)$ for all $\mathfrak{f} \in \mathcal{M}$ then $\mathbf{u} \in H(div, \mathcal{M})$.*

The results of Proposition 2.47 are well-known in the finite element community. A proof employing the notation used here can be found in[97, Section 2.4], based on the presentation in the textbook by Ern and Guermond[76, Section 1.4].

**Proposition 2.48.** *Let $\mathbf{m}$ be a convex $n$-dimensional polytope in $\mathbb{R}^n$ with vertex set $\{\mathbf{v}_i\}_{i\in I}$ and an associated set of generalized barycentric coordinates $\{\lambda_i\}$. Let $\mathfrak{f}$ be a face of $\mathbf{m}$ of codimension 1 whose vertices are indexed by $J \subsetneq I$. If $k \notin J$ then $\lambda_k \equiv 0$ on $\mathfrak{f}$ and $\nabla\lambda_k$ is normal to $\mathfrak{f}$ on $\mathfrak{f}$, pointing inward.*

*Proof.* Fix a point $\mathbf{x}_0 \in \mathbf{m}$. Observe that $\sum_{i\in I}\mathbf{v}_i\lambda_i(\mathbf{x}_0)$ is a point in $\mathbf{m}$ lying in the interior of the convex hull of those $\mathbf{v}_i$ for which $\lambda_i(\mathbf{x}_0) > 0$, since the $\lambda_i$ are non-negative by definition. By (2.73), this summation is equal to $\mathbf{x}_0$. Hence, if $\mathbf{x}_0 \in \mathfrak{f}$, then $\lambda_k \equiv 0$ on $\mathfrak{f}$ unless $k \in J$, proving the first claim. The same argument implies that for any $k \notin$J, $\mathfrak{f}$ is part of the zero level set of $\lambda_k$, meaning $\nabla\lambda_k$ is orthogonal to $\mathfrak{f}$ on $\mathfrak{f}$. In that case, $\nabla\lambda_k$ points inward since $\lambda_k$ has support inside $\mathbf{m}$ but not on the other side of $\mathfrak{f}$. □

We now show that generalized barycentric coordinates and their gradients defined over individual elements in a mesh of polytopes naturally stitch together to build conforming finite elements with global continuity of the expected kind. To be clear about the context, we introduce notation for generalized barycentric hat functions, defined piecewise over a mesh of polytopes $\{\mathbf{m}\}$ by

$$\hat{\lambda}_i(\mathbf{x}) = \begin{cases} \lambda_i(\mathbf{x}) \text{ as defined on } \mathbf{m} & \text{if } \mathbf{x}\in\mathbf{m} \text{ and } \mathbf{v}_i \in \mathbf{m};\\ 0 & \text{if } \mathbf{x}\in\mathbf{m} \text{ but } \mathbf{v}_i \notin \mathbf{m}.\end{cases}$$

We note a slight abuse of notation above: generalized barycentric coordinates $\{\lambda_i\}$ are usually indexed locally on a particular polygon while the above functions require a global indexing of the verticies to consistently identify matching functions across element boundaries. Further, $\hat{\lambda}_i$ is well-defined at vertices and edges of the mesh as any choice of generalized barycentric coordinates on a particular element will give the same value at such points. If $\mathbf{x}$ belongs to the interior of shared faces between polyhedra in $\mathbb{R}^3$ (or higher order analogues), $\hat{\lambda}_i(\mathbf{x})$ is well-defined so long as the same *kind* of coordinates are chosen on each of the incident polyhedra (e.g. Wachspress or mean value).
Our first result about global continuity concerns functions of the form $\hat{\lambda}_i\nabla\hat{\lambda}_j$, where $i$ and $j$ are indices of vertices belonging to at least one fixed mesh element $\mathbf{m}$. Note that the vertices $\mathbf{v}_i$ and $\mathbf{v}_j$ need not define an edge of $\mathbf{m}$.

**Theorem 2.49.** *Fix a mesh $\mathcal{M}$ of $n$-dimensional polytopes $\{\mathbf{m}\}$ in $\mathbb{R}^n$ with $n = 2$ or $3$ and assign some ordering $\mathbf{v}_1,\dots,\mathbf{v}_p$ to all the vertices in the mesh. Fix an associated set of generalized barycentric coordinate hat functions $\hat{\lambda}_1,\dots,\hat{\lambda}_p$. Let*

$$\mathbf{u} \in \text{span}\left\{\hat{\lambda}_i\nabla\hat{\lambda}_j \;:\; \exists\, \mathbf{m}\in\mathcal{M} \text{ such that } \mathbf{v}_i,\mathbf{v}_j \in \mathbf{m}\right\}.$$

*Then $\mathbf{u} \in H(curl,\mathcal{M})$.*

*Proof.* Following the notation of Proposition 2.47, it suffices to show that $T_{\mathfrak{f}}(\mathbf{u}_1) = T_{\mathfrak{f}}(\mathbf{u}_2)$ for an arbitrary face $\mathfrak{f}\in\mathcal{M}$ of codimension 1. Consider an arbitrary term $c_{ij}\hat{\lambda}_i\nabla\hat{\lambda}_j$ in the linear combination defining $\mathbf{u}$. Observe that if $\mathbf{v}_i \notin \mathfrak{f}$, then by Proposition 2.48, $\hat{\lambda}_i \equiv 0$ on $\mathfrak{f}$ and hence $\mathbf{u}\equiv 0$ on $\mathfrak{f}$. Further, if $\mathbf{v}_j \notin \mathfrak{f}$, then $\nabla\hat{\lambda}_j$ is orthogonal to $\mathfrak{f}$. Therefore, without loss of generality, we can reduce to the case where $\mathbf{v}_i,\mathbf{v}_j \in \mathfrak{f}$. Since $\hat{\lambda}_i$ and $\hat{\lambda}_j$ are both $C^0$ on $\mathcal{M}$, their well-defined values on $\mathfrak{f}$ suffice to determine the projection of $\hat{\lambda}_i\nabla\hat{\lambda}_j$ to $\mathfrak{f}$. Since the choice of pair $ij$ was arbitrary, we have $T_{\mathfrak{f}}(\mathbf{u}_1) = T_{\mathfrak{f}}(\mathbf{u}_2)$, completing the proof. □

*Remark* 2.50. *When $n = 2$, we may replace $\hat{\lambda}_i\nabla\hat{\lambda}_j$ in the statement Theorem 2.49 by* rot $\hat{\lambda}_i\nabla\hat{\lambda}_j$ *and conclude that $\mathbf{u} \in H(div,\mathcal{M})$. This is immediate since* rot *gives an isomorphism between $H(\mathrm{curl})$ and $H(\mathrm{div})$ in $\mathbb{R}^2$, as discussed in Section 2.7.3. When $n = 3$, we construct functions in $H(div,\mathcal{M})$ using triples of indices associated to vertices of mesh elements, according to the next result.*

**Theorem 2.51.** *Fix a mesh $\mathcal{M}$ of polyhedra $\{\mathbf{m}\}$ in $\mathbb{R}^3$ and assign some ordering $\mathbf{v}_1,\dots,\mathbf{v}_p$ to all the vertices in the mesh. Fix an associated set of generalized barycentric coordinate hat functions $\hat{\lambda}_1,\dots,\hat{\lambda}_p$. Let*

$$\mathbf{u} \in \text{span}\left\{\hat{\lambda}_i\nabla\hat{\lambda}_j\times\nabla\hat{\lambda}_k \;:\; \exists\, \mathbf{m}\in\mathcal{M} \text{ such that } \mathbf{v}_i,\mathbf{v}_j,\mathbf{v}_k \in \mathbf{m}\right\}.$$

*Then $\mathbf{u} \in H(div,\mathcal{M})$.*

*Proof.* Again following the notation of Proposition 2.47, it suffices to show that $N_{\mathfrak{f}}(\mathbf{u}_1) = N_{\mathfrak{f}}(\mathbf{u}_2)$ for an arbitrary face $\mathfrak{f}\in\mathcal{M}$ of codimension one whose vertices are indexed by $J$. We will use the shorthand notation

$$\xi_{ijk} := \hat{\lambda}\nabla\hat{\lambda}_i \times \nabla\hat{\lambda}_j k.$$

Consider an arbitrary term $c_{ijk}\xi_{ijk}$ in the linear combination defining $\mathbf{u}$. We will first show that $\xi_{ijk}$ has a non-zero normal component on $\mathfrak{f}$ only if $i, j, k \in J$. If $i \notin J$ then $\hat{\lambda}_i \equiv 0$ on $\mathfrak{f}$ by Proposition 2.48, making $\xi_{ijk} \equiv 0$ on $\mathfrak{f}$, as well. If $i \in J$ but $j, k \notin J$, then $\nabla\hat{\lambda}_j$ and $\nabla\hat{\lambda}_k$ are both normal to $\mathfrak{f}$ on $\mathfrak{f}$ by Proposition 2.48. Hence, their cross product is zero and again $\xi_{ijk} \equiv 0$ on $F$. If $i, j \in J$ but $k \notin J$ then again $\nabla\hat{\lambda}_k \perp \mathfrak{f}$ on $\mathfrak{f}$. Since $\nabla\hat{\lambda}_j \times \nabla\hat{\lambda}_k \perp \nabla\hat{\lambda}_k$, we conclude that $\xi_{ijk}$ has no normal component on $\mathfrak{f}$. The same argument holds for the case $i, k \in J$, $j \notin J$. The only remaining case is $i, j, k \in J$, proving the claim.

Thus, without loss of generality, we assume that $i, j, k \in J$. Since $\hat{\lambda}_j$ and $\hat{\lambda}_k$ are both $C^0$ on $\mathcal{M}$, their well-defined values on $\mathfrak{f}$ suffice to determine the projection of $\nabla\hat{\lambda}_j$ and $\nabla\hat{\lambda}_k$ to $\mathfrak{f}$, which then uniquely defines the normal component of $\nabla\hat{\lambda}_j \times \nabla\hat{\lambda}_k$ on $\mathfrak{f}$. Since $\hat{\lambda}_i$ is also $C^0$ on $\mathcal{M}$, and the choice of $i, j, k$ was arbitrary, we have $N_{\mathfrak{f}}(\mathbf{u}_1) = N_{\mathfrak{f}}(\mathbf{u}_2)$, completing the proof. □

### Polynomial Reproduction Results

We now show how generalized barycentric coordinate functions $\lambda_i$ and their gradients can reproduce all the polynomial differential forms in $\mathcal{P}_1\Lambda^k$ and $\mathcal{P}_1^-\Lambda^k$ for $0 \leq k \leq n$ with $n = 2$ or 3. The results for the functions $\lambda_i\nabla\lambda_j$ and $\mathcal{W}_{ij}$ extend immediately to any value of $n \geq 2$ since those functions do not use any dimension-specific operators like × or rot.

**Theorem 2.52.** *Fix $n \geq 2$. Let $\mathbf{m}$ be a convex $n$-dimensional polytope in $\mathbb{R}^n$ with vertex set $\{\mathbf{v}_i\}$. Given any set of generalized barycentric coordinates $\{\lambda_i\}$ associated to $\mathbf{m}$,*

$$\sum_{i,j} \lambda_i \nabla\lambda_j (\mathbf{v}_j - \mathbf{v}_i)^T = \mathbb{I}, \tag{2.81}$$

*where $\mathbb{I}$ is the $n \times n$ identity matrix. Further, for any $n \times n$ matrix $\mathbb{A}$,*

$$\sum_{i,j} (\mathbb{A}\mathbf{v}_i \cdot \mathbf{v}_j)(\lambda_i \nabla\lambda_j) = \mathbb{A}\boldsymbol{x}. \tag{2.82}$$

*Thus,* span $\{\lambda_i\nabla\lambda_j \ : \ \mathbf{v}_i, \mathbf{v}_j \in \mathbf{m}\} \supseteq \mathcal{P}_1\Lambda^1(\mathbf{m})$.

*Proof.* From (2.72) - (2.75), we see that

$$\begin{aligned}\sum_{i,j} \lambda_i \nabla\lambda_j (\mathbf{v}_j - \mathbf{v}_i)^T &= \left(\sum_i \lambda_i\right)\left(\sum_j \nabla\lambda_j \mathbf{v}_j^T\right) - \left(\sum_j \nabla\lambda_j\right)\left(\sum_i \lambda_i \mathbf{v}_i^T\right) \\ &= 1(\mathbb{I}^T) - 0(\mathbf{x}^T) = \mathbb{I},\end{aligned}$$

establishing (2.81). Similarly for (2.82), a bit of algebra yields

$$\begin{aligned}\sum_{i,j} (\mathbb{A}\mathbf{v}_i \cdot \mathbf{v}_j)(\lambda_i \nabla\lambda_j) &= \sum_{i,j} (\lambda_i \nabla\lambda_j)\mathbf{v}_j^T \mathbb{A}\mathbf{v}_i = \sum_{i,j} \nabla\lambda_j \mathbf{v}_j^T \mathbb{A}\mathbf{v}_i \lambda_i \\ &= \left(\sum_j \nabla\lambda_j \mathbf{v}_j^T\right) \mathbb{A} \left(\sum_i \mathbf{v}_i \lambda_i\right) = \mathbb{I}^T \mathbb{A}\mathbf{x} = \mathbb{A}\mathbf{x}\end{aligned}$$

We have shown that any vector of linear polynomials can be written as a linear combination of $\lambda_i\nabla\lambda_j$ functions, hence the span of these functions contains the vector proxies for all elements of $\mathcal{P}_1\Lambda^1(\mathbf{m})$. □

**Corollary 2.53.** *Let $\mathbf{m}$ be a convex polygon in $\mathbb{R}^2$ with vertex set $\{\mathbf{v}_i\}$. Given any set of generalized barycentric coordinates $\{\lambda_i\}$ associated to $\mathbf{m}$,*

$$\sum_{i,j} \operatorname{rot} \lambda_i \nabla\lambda_j (\operatorname{rot}(\mathbf{v}_j - \mathbf{v}_i))^T = \mathbb{I}, \tag{2.83}$$

*where $\mathbb{I}$ is the $2 \times 2$ identity matrix. Further, for any $2 \times 2$ matrix $\mathbb{A}$,*

$$\sum_{i,j} (-\operatorname{rot} \mathbb{A}\, \mathbf{v}_i \cdot \mathbf{v}_j)(\operatorname{rot} \lambda_i \nabla\lambda_j) = \mathbb{A}\mathbf{x}. \tag{2.84}$$

*Thus,* span $\{\operatorname{rot}\lambda_i\nabla\lambda_j \ : \ \mathbf{v}_i, \mathbf{v}_j \in \mathbf{m}\} \supseteq \mathcal{P}_1\Lambda^1(\mathbf{m})$.

*Proof.* For (2.83), observe that for any $\mathbf{w}, \mathbf{y} \in \mathbb{R}^2$, $\mathbf{w}\mathbf{y}^T = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ implies $(\text{rot}\,\mathbf{w})(\text{rot}\,\mathbf{y})^T = \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}$. Hence, the result follows immediately from (2.81). For (2.84), note $\text{rot}^{-1} = -\text{rot}$ and define $\mathbb{B} := -\text{rot}\,\mathbb{A}$. Using $\mathbb{B}$ as the matrix in (2.82), we have

$$\sum_{i,j}(\mathbb{B}\mathbf{v}_i \cdot \mathbf{v}_j)(\lambda_i \nabla \lambda_j) = \mathbb{B}\mathbf{x}$$

Applying rot to both sides of the above yields the result. □

**Theorem 2.54.** *Let* $\mathbf{m}$ *be a convex polyhedron in* $\mathbb{R}^3$ *with vertex set* $\{\mathbf{v}_i\}$. *Given any set of generalized barycentric coordinates* $\{\lambda_i\}$ *associated to* $\mathbf{m}$,

$$\frac{1}{2}\sum_{i,j,k}\lambda_i \nabla\lambda_j \times \nabla\lambda_k \left((\mathbf{v}_j - \mathbf{v}_i) \times (\mathbf{v}_k - \mathbf{v}_i)\right)^T = \mathbb{I}, \tag{2.85}$$

*where* $\mathbb{I}$ *is the* $n \times n$ *identity matrix. Further, for any* $n \times n$ *matrix* $\mathbb{A}$,

$$\frac{1}{2}\sum_{i,j,k}(\mathbb{A}\mathbf{v}_i \cdot (\mathbf{v}_j \times \mathbf{v}_k))(\lambda_i \nabla\lambda_j \times \nabla\lambda_k) = \mathbb{A}\mathbf{x}. \tag{2.86}$$

*Thus,* span $\{\lambda_i \nabla\lambda_j \times \nabla\lambda_k \ : \ \mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k \in \mathbf{m}\} \supseteq \mathcal{P}_1\Lambda^2(\mathbf{m})$.

*Proof.* We start with (2.85). First, observe that

$$(\mathbf{v}_j - \mathbf{v}_i) \times (\mathbf{v}_k - \mathbf{v}_i) = \mathbf{v}_i \times \mathbf{v}_j + \mathbf{v}_j \times \mathbf{v}_k + \mathbf{v}_k \times \mathbf{v}_i.$$

By (2.74), we have that

$$\sum_{i,j,k}\lambda_i \nabla\lambda_j \times \nabla\lambda_k \left(\mathbf{v}_i \times \mathbf{v}_j\right)^T = \sum_{i,j}\lambda_i \left(\nabla\lambda_j \times \left(\sum_k \nabla\lambda_k\right)\right)\left(\mathbf{v}_i \times \mathbf{v}_j\right)^T = 0.$$

A similar argument shows that replacing $\mathbf{v}_i \times \mathbf{v}_j$ with $\mathbf{v}_k \times \mathbf{v}_i$ also yields the zero matrix. Hence,

$$\begin{aligned}
\sum_{i,j,k}\lambda_i \nabla\lambda_j \times \nabla\lambda_k \left((\mathbf{v}_j - \mathbf{v}_i) \times (\mathbf{v}_k - \mathbf{v}_i)\right)^T &= \sum_{i,j,k}\lambda_i \nabla\lambda_j \times \nabla\lambda_k \left(\mathbf{v}_j \times \mathbf{v}_k\right)^T \\
&= \sum_i \lambda_i \sum_{j,k}\left(\nabla\lambda_j \times \nabla\lambda_k\right)\left(\mathbf{v}_j \times \mathbf{v}_k\right)^T \\
&= \sum_{j,k}\left(\nabla\lambda_j \times \nabla\lambda_k\right)\left(\mathbf{v}_j \times \mathbf{v}_k\right)^T.
\end{aligned}$$

To simplify this further, we use the Kronecker delta symbol $\delta_{i_1 i_2}$ and the 3D Levi-Civita symbol $\varepsilon_{i_1 i_2 i_3}$. It suffices to show that the entry in row $r$, column $c$ of the matrix $\sum_{j,k}(\nabla\lambda_j \times \nabla\lambda_k)\left(\mathbf{v}_j \times \mathbf{v}_k\right)^T$ is $2\delta_{rc}$. We see that

$$\begin{aligned}
\left[\sum_{j,k}(\nabla\lambda_j \times \nabla\lambda_k)\left(\mathbf{v}_j \times \mathbf{v}_k\right)^T\right]_{rc} &= \sum_{j,k}\varepsilon_{r\ell m}\mathbf{1}(j)_{\ell}\mathbf{2}(k)_m \varepsilon_{cpq}(\mathbf{v}_j)_p(\mathbf{v}_k)_q \\
&= \varepsilon_{r\ell m}\varepsilon_{cpq}\sum_j (\mathbf{v}_j)_p\mathbf{3}(j)_\ell \sum_k (\mathbf{v}_k)_q\mathbf{4}(k)_m \\
&= \varepsilon_{r\ell m}\varepsilon_{cpq}\delta_{\ell p}\delta_{mq}.
\end{aligned}$$

The last step in the above chain of equalities follows from (2.75). Observe that $\varepsilon_{r\ell m}\varepsilon_{cpq}\delta_{\ell p}\delta_{mq} = \varepsilon_{r\ell m}\varepsilon_{c\ell m} = 2\delta_{rc}$, as desired. For (2.86), observe that

$$\begin{aligned}
\sum_{i,j,k}(\mathbb{A}\mathbf{v}_i \cdot (\mathbf{v}_j \times \mathbf{v}_k))(\lambda_i \nabla\lambda_j \times \nabla\lambda_k) &= \left(\sum_i \mathbb{A}\mathbf{v}_i\lambda_i\right) \cdot \sum_{j,k}(\mathbf{v}_j \times \mathbf{v}_k)(\nabla\lambda_j \times \nabla\lambda_k) \\
&= \sum_{j,k}(\nabla\lambda_j \times \nabla\lambda_k)(\mathbf{v}_j \times \mathbf{v}_k)^T\left(\mathbb{A}\sum_i \mathbf{v}_i\lambda_i\right) \\
&= 2\ \mathbb{I}\ \mathbb{A}\mathbf{x} = 2\mathbb{A}\mathbf{x}.
\end{aligned}$$

Note that we used the proof of (2.85) to rewrite the sum over $j, k$ as $2\mathbb{I}$. We have shown that any vector of linear polynomials can be written as a linear combination of $\lambda_i \nabla\lambda_j \times \nabla\lambda_k$ functions, hence the span of these functions contains the vector proxies for all elements of $\mathcal{P}_1\Lambda^2(\mathbf{m})$. □

**Theorem 2.55.** *Let* $\mathbf{m}$ *be a convex polygon in* $\mathbb{R}^2$ *with vertex set* $\{\mathbf{v}_i\}$. *Given any set of generalized barycentric coordinates* $\{\lambda_i\}$ *associated to* $\mathbf{m}$*,*

$$\frac{1}{2}\sum_{i,j,k} \lambda_i \nabla\lambda_j \cdot \mathrm{rot}\nabla\lambda_k \left((\mathbf{v}_j - \mathbf{v}_i)\cdot \mathrm{rot}(\mathbf{v}_k - \mathbf{v}_i)\right) = 1. \tag{2.87}$$

*Further, for any vector* $\boldsymbol{\alpha} \in \mathbb{R}^2$*,*

$$\frac{1}{2}\sum_{i,j,k} (\boldsymbol{\alpha}^T \mathbf{v}_i (\mathbf{v}_j \cdot \mathrm{rot}\mathbf{v}_k))(\lambda_i \nabla\lambda_j \cdot \mathrm{rot}\nabla\lambda_k) = \boldsymbol{\alpha}^T \mathbf{x}. \tag{2.88}$$

*Thus,* span $\{\lambda_i \nabla\lambda_j \cdot \mathrm{rot}\nabla\lambda_k \ : \ \mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k \in \mathbf{m}\} \supseteq \mathcal{P}_1\Lambda^2(\mathbf{m})$.

*Proof.* The proof is essentially identical to that of Theorem 2.54. First,

$$(\mathbf{v}_j - \mathbf{v}_i)\cdot \mathrm{rot}(\mathbf{v}_k - \mathbf{v}_i) = \mathbf{v}_i \cdot \mathrm{rot}\mathbf{v}_j + \mathbf{v}_j \cdot \mathrm{rot}\mathbf{v}_k + \mathbf{v}_k \cdot \mathrm{rot}\mathbf{v}_i,$$

and by (2.74),

$$\sum_{i,j,k} \lambda_i \nabla\lambda_j \cdot \mathrm{rot}\nabla\lambda_k \left(\mathbf{v}_i \cdot \mathrm{rot}\mathbf{v}_j\right) = \sum_{i,j} \lambda_i \left(\nabla\lambda_j \cdot \mathrm{rot}\left(\sum_k \nabla\lambda_k\right)\right)\left(\mathbf{v}_i \cdot \mathrm{rot}\mathbf{v}_j\right) = 0.$$

A similar argument shows that replacing $\mathbf{v}_i \cdot \mathrm{rot}\mathbf{v}_j$ with $\mathbf{v}_k \cdot \mathrm{rot}\mathbf{v}_i$ also yields zero. Hence as before,

$$\sum_{i,j,k} \lambda_i \nabla\lambda_j \cdot \mathrm{rot}\nabla\lambda_k \left((\mathbf{v}_j - \mathbf{v}_i)\cdot \mathrm{rot}(\mathbf{v}_k - \mathbf{v}_i)\right)^T = \sum_{j,k} \left(\nabla\lambda_j \cdot \mathrm{rot}\nabla\lambda_k\right)\left(\mathbf{v}_j \cdot \mathrm{rot}\mathbf{v}_k\right)^T.$$

Finally, the same argument holds using the 2D Levi-Civita symbol:

$$\begin{aligned}\sum_{j,k} \left(\nabla\lambda_j \cdot \mathrm{rot}\nabla\lambda_k\right)\left(\mathbf{v}_j \cdot \mathrm{rot}\mathbf{v}_k\right) &= \sum_{j,k} \varepsilon_{\ell m}\mathbf{5}(j)_\ell \mathbf{6}(k)_m \varepsilon_{pq} (\mathbf{v}_j)_p (\mathbf{v}_k)_q \\ &= \varepsilon_{\ell m}\varepsilon_{pq} \sum_j (\mathbf{v}_j)_p \mathbf{7}(j)_\ell \sum_k (\mathbf{v}_k)_q \mathbf{8}(k)_m \\ &= \varepsilon_{\ell m}\varepsilon_{pq}\delta_{\ell p}\delta_{mq} = \varepsilon_{\ell m}\varepsilon_{\ell m} = 2,\end{aligned}$$

establishing (2.87). For (2.88), observe that

$$\begin{aligned}\sum_{i,j,k} (\boldsymbol{\alpha}^T \mathbf{v}_i (\mathbf{v}_j \cdot \mathrm{rot}\mathbf{v}_k))(\lambda_i \nabla\lambda_j \cdot \mathrm{rot}\nabla\lambda_k) &= \left(\sum_i \boldsymbol{\alpha}^T \mathbf{v}_i \lambda_i\right) \sum_{j,k} (\mathbf{v}_j \cdot \mathrm{rot}\mathbf{v}_k)(\nabla\lambda_j \cdot \mathrm{rot}\nabla\lambda_k) \\ &= \sum_{j,k} (\nabla\lambda_j \cdot \mathrm{rot}\nabla\lambda_k)(\mathbf{v}_j \cdot \mathrm{rot}\mathbf{v}_k)^T \left(\boldsymbol{\alpha}^T \sum_i \mathbf{v}_i \lambda_i\right) = 2\boldsymbol{\alpha}^T \mathbf{x}.\end{aligned}$$

□

*Remark* 2.56. *The proof of Theorem 2.55 can also be obtained by augmenting the 2D vectors and matrices with zeros to make 3D vectors and matrices and recognizing (2.87) as the element equality in the third row and third column of (2.85).*

We also have polynomial reproduction results using the Whitney-like basis functions (2.79) and (2.80). Recall that $\mathcal{H}_r$ denotes homogeneous polynomials of degree $r$ and let $\mathbb{M}_{n\times n}$ denote $n \times n$ matrices. We have the following theorems.

**Theorem 2.57.** *Fix $n \geq 2$. Let $\mathbf{m}$ be a convex $n$-dimensional polytope in $\mathbb{R}^n$ with vertex set $\{\mathbf{v}_i\}$ and an associated set of generalized barycentric coordinates $\{\lambda_i\}$. Then*

$$\sum_{i<j} \mathcal{W}_{ij}(\mathbf{v}_j - \mathbf{v}_i)^T = \mathbb{I}. \tag{2.89}$$

*Further, define a map $\Phi : \mathcal{H}_1\Lambda^1(\mathbb{R}^n) \to \mathbb{M}_{n\times n}$ by*

$$\sum_{j=1}^{n} \left( \sum_{i=1}^{n} a_{ij}x_j \right) dx_i \longmapsto \left[\text{sign } (a_{ij})\right].$$

*Then for all $\omega \in \mathcal{H}_0\Lambda^2(\mathbb{R}^n)$,*

$$\sum_{i<j} \left(\Phi(\kappa\omega)\mathbf{v}_i) \cdot \mathbf{v}_j\right) \mathcal{W}_{ij} = (\Phi(\kappa\omega))\mathbf{x}. \tag{2.90}$$

*Thus,* span $\{\mathcal{W}_{ij} \ : \ \mathbf{v}_i, \mathbf{v}_j \in \mathbf{m}\} \supseteq \mathcal{P}_1^-\Lambda^1(\mathbf{m})$.

*Proof.* For (2.89), we reorganize the summation and apply (2.81) to see that

$$\begin{aligned}
\sum_{i<j} \mathcal{W}_{ij}(\mathbf{v}_j - \mathbf{v}_i)^T &= \sum_{i<j} \lambda_i \nabla\lambda_j (\mathbf{v}_j - \mathbf{v}_i)^T - \sum_{i<j} \mathbf{9}\, i(\mathbf{v}_j - \mathbf{v}_i)^T \\
&= \sum_{i<j} \lambda_i \nabla\lambda_j (\mathbf{v}_j - \mathbf{v}_i)^T + \sum_{j<i} \lambda_i \nabla\lambda_j (\mathbf{v}_j - \mathbf{v}_i)^T \\
&= \sum_{i,j} \lambda_i \nabla\lambda_j (\mathbf{v}_j - \mathbf{v}_i)^T = \mathbb{I}.
\end{aligned}$$

For (2.90), fix $\omega \in \mathcal{H}_0\Lambda^2(\mathbb{R}^n)$ and express it as

$$\omega = \sum_{i<j} a_{ij} dx_i dx_j,$$

for some coefficients $a_{ij} \in \mathbb{R}$. Then

$$\kappa\omega = \sum_{i<j} a_{ij}(x_i dx_j - x_j dx_i).$$

The entries of the matrix $\Phi(\kappa\omega)$ are thus given by

$$[\Phi(\kappa\omega)]_{ij} = \begin{cases} \text{sign } (a_{ij}) & \text{if } i < j, \\ -\text{sign } (a_{ij}) & \text{if } i > j, \\ 0 & \text{if } i = j. \end{cases} \tag{2.91}$$

From (2.82), we have that

$$\sum_{i,j} \left(\Phi(\kappa\omega)\mathbf{v}_i) \cdot \mathbf{v}_j\right) \lambda_i \nabla\lambda_j = (\Phi(\kappa\omega))\mathbf{x}, \qquad \forall \omega \in \mathcal{H}_0\Lambda^2(\mathbb{R}^n)$$

Since $\Phi(\kappa\omega)$ is anti-symmetric by (2.91), we have that

$$\begin{aligned}
\sum_{i,j} \left(\Phi(\kappa\omega)\mathbf{v}_i) \cdot \mathbf{v}_j\right) \lambda_i \nabla\lambda_j &= \sum_{i<j} \left(\Phi(\kappa\omega)\mathbf{v}_i) \cdot \mathbf{v}_j\right) \lambda_i \nabla\lambda_j + \sum_{j<i} \left(\Phi(\kappa\omega)\mathbf{v}_i) \cdot \mathbf{v}_j\right) \lambda_i \nabla\lambda_j \\
&= \sum_{i<j} \left(\Phi(\kappa\omega)\mathbf{v}_i) \cdot \mathbf{v}_j\right) \mathcal{W}_{ij}.
\end{aligned}$$

We have shown that any vector proxy of an element of $\mathcal{P}_0\Lambda^1(\mathbf{m})$ or $\kappa\mathcal{H}_0\Lambda^2(\mathbf{m})$ can be written as a linear combination of $\mathcal{W}_{ij}$ functions. By (2.69), we conclude that the span of the $\mathcal{W}_{ij}$ functions contains the vector proxies for all elements of $\mathcal{P}_1^-\Lambda^1(\mathbf{m})$. □

**Corollary 2.58.** *Let* $\mathbf{m}$ *be a convex polygon in* $\mathbb{R}^2$ *with vertex set* $\{\mathbf{v}_i\}$. *Given any set of generalized barycentric coordinates* $\{\lambda_i\}$ *associated to* $\mathbf{m}$*,*

$$\sum_{i<j} \text{rot}\,\mathcal{W}_{ij}\ \text{rot}(\mathbf{v}_j - \mathbf{v}_i)^T = \mathbb{I}, \tag{2.92}$$

*where* $\mathbb{I}$ *is the* $2 \times 2$ *identity matrix. Further,*

$$\sum_{i<j} ((\text{rot}\,\mathbf{v}_i)\cdot\mathbf{v}_j)\,\text{rot}\,\mathcal{W}_{ij} = \mathbf{x}. \tag{2.93}$$

*Thus,* $\text{span}\,\{\text{rot}\,\mathcal{W}_{ij}\ :\ \mathbf{v}_i, \mathbf{v}_j \in \mathbf{m}\} \supseteq \mathcal{P}_1^-\Lambda^1(\mathbf{m})$.

*Proof.* By the same argument as the proof of (2.83) in Corollary 2.53, the identity (2.92) follows immediately from (2.89). For (2.93), observe that setting $\omega := 1 \in \mathcal{H}_0\Lambda^2(\mathbb{R}^2)$, we have that $\Phi(\kappa\omega) = \text{rot}$. Therefore, (2.90) implies that

$$\sum_{i<j} (\text{rot}\,\mathbf{v}_i)\cdot\mathbf{v}_j)\,\mathcal{W}_{ij} = \text{rot}\,\mathbf{x}.$$

Applying rot to both sides of the above equation completes the proof. □

**Theorem 2.59.** *Let* $\mathbf{m}$ *be a convex polyhedron in* $\mathbb{R}^3$ *with vertex set* $\{\mathbf{v}_i\}$ *and an associated set of generalized barycentric coordinates* $\{\lambda_i\}$. *Then*

$$\sum_{i<j<k} \mathcal{W}_{ijk}\left((\mathbf{v}_j - \mathbf{v}_i)\times(\mathbf{v}_k - \mathbf{v}_i)\right)^T = \mathbb{I}, \tag{2.94}$$

*and*

$$\sum_{i<j<k} (\mathbf{v}_i\cdot(\mathbf{v}_j\times\mathbf{v}_k))\mathcal{W}_{ijk} = \mathbf{x}. \tag{2.95}$$

*Thus,* $\text{span}\,\{\mathcal{W}_{ijk}\ :\ \mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k \in \mathbf{m}\} \supseteq \mathcal{P}_1^-\Lambda^2(\mathbf{m})$.

*Proof.* We adopt the shorthand notations

$$\xi_{ijk} := \lambda_i\nabla\lambda_j\times\nabla\lambda_k, \quad \mathbf{z}_{ijk} := (\mathbf{v}_j - \mathbf{v}_i)\times(\mathbf{v}_k - \mathbf{v}_i), \quad \mathbf{v}_{ijk} := \mathbf{v}_i\cdot(\mathbf{v}_j\times\mathbf{v}_k).$$

For (2.94), we re-write (2.85) as

$$\sum_{i,j,k} \xi_{ijk}\mathbf{z}_{ijk}{}^T = 2\mathbb{I}.$$

Observe that $\xi_{ijk}\mathbf{z}_{ijk}{}^T = (-\xi_{ikj})(-\mathbf{z}_{ikj})^T = \xi_{ikj}\mathbf{z}_{ikj}{}^T$ and $\mathbf{z}_{ijk} = 0$ if $i$, $j$, $k$ are not distinct. Thus,

$$2\mathbb{I} = \sum_{\substack{i<j<k\\ k<i<j\\ j<k<i}} \xi_{ijk}\mathbf{z}_{ijk}{}^T + \sum_{\substack{i<k<j\\ k<j<i\\ j<i<k}} \xi_{ikj}\mathbf{z}_{ikj}{}^T.$$

The two summations have different labels for the indices but are otherwise identical. Therefore,

$$\begin{aligned}
\mathbb{I} &= \sum_{i<j<k} \xi_{ijk}\mathbf{z}_{ijk}{}^T + \sum_{k<i<j} \xi_{ijk}\mathbf{z}_{ijk}{}^T + \sum_{j<k<i} \xi_{ijk}\mathbf{z}_{ijk}{}^T\\
&= \sum_{i<j<k} \xi_{ijk}\mathbf{z}_{ijk}{}^T + \xi_{jki}\mathbf{z}_{jki}{}^T + \xi_{kij}\mathbf{z}_{kij}{}^T\\
&= \sum_{i<j<k} (\xi_{ijk} + \xi_{jki} + \xi_{kij})\mathbf{z}_{ijk}{}^T\\
&= \sum_{i<j<k} \mathcal{W}_{ijk}\left((\mathbf{v}_j - \mathbf{v}_i)\times(\mathbf{v}_k - \mathbf{v}_i)\right)^T.
\end{aligned}$$

For (2.95), we take $\mathbb{A}$ as the identity, and re-write (2.86) as

$$\sum_{i,j,k} \mathbf{v}_{ijk}\xi_{ijk} = 2\mathbf{x}.$$

Observe that $\mathbf{v}_{ijk}\xi_{ijk} = (-\mathbf{v}_{ikj})(-\xi_{ikj}) = \mathbf{v}_{ikj}\xi_{ikj}$ and $\mathbf{v}_{ijk} = 0$ if $i$, $j$, $k$ are not distinct. Thus,

$$2\mathbf{x} = \sum_{\substack{i<j<k\\ k<i<j\\ j<k<i}} \mathbf{v}_{ijk}\xi_{ijk} + \sum_{\substack{i<k<j\\ k<j<i\\ j<i<k}} \mathbf{v}_{ikj}\xi_{ikj}.$$

The rest of the argument follows similarly, yielding

$$\mathbf{x} = \sum_{i<j<k} \mathbf{v}_{ijk}\xi_{ijk} + \sum_{k<i<j} \mathbf{v}_{ijk}\xi_{ijk} + \sum_{j<k<i} \mathbf{v}_{ijk}\xi_{ijk} = \sum_{i<j<k} (\mathbf{v}_i \cdot (\mathbf{v}_j \times \mathbf{v}_k))\mathcal{W}_{ijk}.$$

Note that $\mathcal{H}_0\Lambda^3(\mathbf{m})$ is generated by the volume form $\eta = dxdydz$ and that $\kappa\eta$ has vector proxy $\mathbf{x}$. Thus, by (2.69), we have shown that the span of the $\mathcal{W}_{ijk}$ functions contains the vector proxy of any element of $\mathcal{P}_1^-\Lambda^2(\mathbf{m})$. □

*Remark* 2.60. *There are some additional constructions in this same vein that could be considered. On a polygon in $\mathbb{R}^2$, we can define $\mathcal{W}_{ijk}$ in the same way as (2.80), interpreting $\times$ as the two dimensional cross product. Likewise, on a polyhedron in $\mathbb{R}^3$, we can define $\mathcal{W}_{ijk\ell}$ according to formula (2.78), yielding functions that are summations of terms like $\lambda_i(\nabla\lambda_j \cdot (\nabla\lambda_k \times \nabla\lambda_\ell)$. These constructions will yield the expected polynomial reproduction results, yet they are not of practical interest in finite element contexts, as we will see in the next section.*

### Polygonal and Polyhedral Finite Element Families

Let $\mathcal{M}$ be a mesh of convex $n$-dimensional polytopes $\{\mathbf{m}\}$ in $\mathbb{R}^n$ with $n = 2$ or 3 and assign some ordering $\mathbf{v}_1, \ldots, \mathbf{v}_p$ to all the vertices in the mesh. Fix an associated set of generalized barycentric coordinates $\lambda_1, \ldots, \lambda_p$ where $\lambda_i$ is defined piecewise over the set of polytopes incident to $\mathbf{v}_i$. In Table 2.1, we list all the types of scalar-valued and vector-valued functions that we have defined this setting. When used over all elements in a mesh of polygons or polyhedra, these functions have global continuity and polynomial reproduction properties as indicated in the table.

These two properties – global continuity and polynomial reproduction – are essential and intertwined necessities in the construction of $H\Lambda^k$-conforming finite element methods on *any* type of domain mesh. Global continuity of type $H\Lambda^k$ ensures that the piecewise-defined approximate solution is an element of the function space $H\Lambda^k$ in which a solution is sought. Polynomial reproduction of type $\mathcal{P}_1\Lambda^k$ or $\mathcal{P}_1^-\Lambda^k$ ensures that the error between the true solution and the approximate solution decays linearly with respect to the maximum diameter of a mesh element, as measured in $H\Lambda^k$ norm. On meshes of simplicial elements, the basis functions listed in Table 2.1 are known and often used as local bases for the corresponding classical finite element spaces listed in Table 2.3, meaning our approach recapitulates known methods on simplicial meshes.

**Relation to polygonal serendipity elements.** As a brief aside, consider the scalar bi-quadratic element on *rectangles*, which has nine degrees of freedom: one associated to each vertex, one to each edge midpoint, and one to the center of the square. It has long been known that the 'serendipity' element, which has only the eight degrees of freedom associated to the vertices and edge midpoints of the rectangle, is also an $H^1$-conforming, quadratic order method. In this case, polynomial reproduction requires the containment of $\mathcal{P}_2\Lambda^0(\mathbf{m})$ in the span of the basis functions, meaning at least six functions are required per element $\mathbf{m} \in \mathcal{M}$. To ensure global continuity of $H^1$, however, the method must agree 'up to quadratics' on each edge, which necessitates the eight degrees of freedom associated to the boundary. Therefore, the serendipity space associated to the scalar bi-quadratic element on a rectangle has dimension eight.

In a previous work[171], we generalized this 'serendipity' reduction to $\mathcal{P}_2\Lambda^0(\mathcal{M})$ where $\mathcal{M}$ is a mesh of strictly convex polygons in $\mathbb{R}^2$. For a simple polygon with $n$ vertices (and thus $n$ edges), polynomial reproduction still only requires 6 basis functions, while global continuity of $H^1$ still requires reproduction of quadratics on edges, leading to a total of $2n$ basis functions required per element $\mathbf{m} \in \mathcal{M}$. Given a convex polygon, our approach takes the $n + \binom{n}{2}$ pairwise products of all the $\lambda_i$ functions and forms explicit linear combinations to yield a set of $2n$ basis functions with the required global $H^1$ continuity and polynomial reproduction properties.

| n | k | space | # construction | # boundary | # polynomial |
|---|---|---|---|---|---|
| 2 | 0 | $\mathcal{P}_1\Lambda^0(\mathbf{m})$ | $v$ | $v$ | 3 |
| | | $\mathcal{P}_1^-\Lambda^0(\mathbf{m})$ | $v$ | $v$ | 3 |
| | 1 | $\mathcal{P}_1\Lambda^1(\mathbf{m})$ | $2\binom{v}{2}$ | $2e$ | 6 |
| | | $\mathcal{P}_1^-\Lambda^1(\mathbf{m})$ | $\binom{v}{2}$ | $e$ | 3 |
| | 2 | $\mathcal{P}_1\Lambda^2(\mathbf{m})$ | $3\binom{v}{3}$ | 0 | 3 |
| | | $\mathcal{P}_1^-\Lambda^2(\mathbf{m})$ | $\binom{v}{3}$ | 0 | 1 |

Table 2.4: Dimension counts relevant to serendipity-style reductions in basis size. Here, $v$ and $e$ denote the number of vertices and edges in the polygonal element $\mathbf{m}$. The column '# construction' gives the number of basis functions we define (cf. Table 2.1), '# boundary' gives the number of basis functions related to inter-element continuity, and '# polynomial' gives the dimension of the contained space of polynomial differential forms.

**Reduction of basis size.** A similar reduction procedure can be applied to the polygonal and polyhedral spaces described in Table 2.1. A key observation is that the continuity results of Theorems 2.49 and 2.51 only rely on the agreement of basis functions whose indices are of vertices on a shared boundary edge (in 2D) or face (in 3D). For example, if vertices $\mathbf{v}_i$ and $\mathbf{v}_j$ form the edge of a polygon in a 2D mesh, $H(\text{curl}, \mathcal{M})$ continuity across the edge comes from identical tangential contributions in the $\lambda_i\nabla\lambda_j$ and $i$ functions from either element containing this edge and zero tangential contributions from all other basis functions. Thus, basis functions whose indices do not belong to a single polyon edge (in 2D) or polyhedral face (in 3D) do not contribute to inter-element continuity, allowing the basis size to be reduced.

To quantify the extent to which the bases we have defined could be reduced without affecting the global continuity properties, we count the number of functions associated with codimension 1 faces for each space considered. For a polygon in 2D, the results are summarized in Table 2.4. The $k = 0$ case is optimal in the sense that every basis function $\lambda_i$ contributes to the $H^1$-continuity in some way, meaning no basis reduction is available. In the $k = 1$ cases, the number of basis functions we construct is quadratic in the number of vertices, $v$, of the polygon, but the number associated with the boundary is only linear in the number of edges, $e$. Since $e = v$ for a simple polygon, this suggests a basis reduction procedure would be both relevant and useful; the description of such a reduction will be the focus of a future work. In the $k = 2$ cases, our procedure constructs $O(v^3)$ basis functions but no inter-element continuity is required; in these cases, a discontinuous Galerkin or other type of finite element method would be more practical.

For a polyhedron $\mathbf{m}$ in 3D, the results are summarized in Table 2.5. As in 2D, the basis for the $k = 0$ case cannot be reduced while the bases for the $k = n$ cases would not be practical for implementation since no inter-element continuity is required. In the $k = 1$ cases, the number of basis functions we construct is again quadratic in $v$, while the number of basis functions required for continuity can be reduced for non-simplicial polyhedra. For instance, if $\mathbf{m}$ is a hexahedron, our construction for $\mathcal{P}_1\Lambda^1$ gives 56 functions but only 48 are relevant to continuity; in the $\mathcal{P}_1^-\Lambda^1$ case, we construct 28 functions but only 20 are relevant to continuity. In the $k = 2$ cases, a similar reduction is possible for non-simplicial polyhedra. Again in the case of a hexahedron, we construct 168 functions for $\mathcal{P}_1\Lambda^1$ and 56 functions for $\mathcal{P}_1^-\Lambda^1$, but the elements require only 72 and 24 functions, respectively, for inter-element continuity. Chen and Wang[48] have proposed an approach for such constructions by making use of many of the results from this manuscript.

An additional line of inquiry along these lines is the derivation of geometry-independent bounds on reduced bases. While Sobolev-norm estimates on the original basis can be directly inferred from bounds on the underlying generalized barycentric coordinates, a serendipity-style reduction requires further analysis to ensure that the coefficients of the linear combinations used can be bounded uniformly over the geometric class of elements considered. In practice, a least-squares solution for constructing the optimal matrix is straightforward to implement, e.g.[202], but rigorous analysis of robustness will require an explicit construction of the reduction processes.

| n | k | space | # construction | # boundary | # polynomial |
|---|---|---|---|---|---|
| 3 | 0 | $\mathcal{P}_1\Lambda^0(\mathbf{m})$ | $v$ | $v$ | 4 |
| | | $\mathcal{P}_1^-\Lambda^0(\mathbf{m})$ | $v$ | $v$ | 4 |
| | 1 | $\mathcal{P}_1\Lambda^1(\mathbf{m})$ | $2\binom{v}{2}$ | $\left(\sum_{a=1}^{f} v_a(v_a-1)\right)-2e$ | 12 |
| | | $\mathcal{P}_1^-\Lambda^1(\mathbf{m})$ | $\binom{v}{2}$ | $\left(\sum_{a=1}^{f}\binom{v_a}{2}\right)-e$ | 6 |
| | 2 | $\mathcal{P}_1\Lambda^2(\mathbf{m})$ | $3\binom{v}{3}$ | $\sum_{a=1}^{f}\frac{v_a(v_a-1)(v_a-2)}{2}$ | 12 |
| | | $\mathcal{P}_1^-\Lambda^2(\mathbf{m})$ | $\binom{v}{3}$ | $\sum_{a=1}^{f}\binom{v}{3}$ | 4 |
| | 3 | $\mathcal{P}_1\Lambda^3(\mathbf{m})$ | $4\binom{v}{4}$ | 0 | 4 |
| | | $\mathcal{P}_1^-\Lambda^3(\mathbf{m})$ | $\binom{v}{4}$ | 0 | 1 |

Table 2.5: The $n = 3$ version of Table 2.4. Here, $f$ denotes the number of faces on a polyhedral element $\mathbf{m}$ and $v_a$ denotes the number of vertices on a particular face $\mathfrak{f}_a$. The entries of the '# boundary' column are determined by counting functions associated to each face of the polyhedron and, in the $k = 1$ cases, accounting for double-counting by subtraction.

**Relation to mimetic and virtual element methods** Similar to the mimetic finite difference method[33, 38, 41], the recently developed virtual element method[27] considers the linear system $\mathbb{K}\mathbf{u}_h = \mathbf{f}$ that is used to compute the finite element solution $\mathbf{u}_h$ on a mesh of polygons or polyhedra. In a first order method for the Poisson equation on a mesh of polygons, the entries of matrix $\mathbb{K}$ have the form

$$\mathbb{K}_{ij} = \int_{\mathcal{M}} \nabla\lambda_i \cdot \nabla\lambda_j,$$

where $\lambda_i$ are the harmonic generalized barycentric coordinates associated to the mesh. This approach is shown to be consistent, meaning it recovers linear polynomial data exactly, and stable, meaning the discrete norm associated with the method is equivalent to the continuous norm for the problem. A description of how to apply the method to $H(\text{curl})$ and $H(\text{div})$ problems has been provided[28] and some comparisons to generalized barycentric approaches have been provided[150]. We expect that the explicit basis functions constructed for polygonal and polyhedral elements here will inform the implementation of new types of vector-valued virtual element methods.

**Error Estimation** We consider first-order interpolation operators from some generalization of barycentric coordinates to arbitrary convex polygons. A set of barycentric coordinates $\{\lambda_i\}$ for $\Omega$ associated with the interpolation operator $I : H^2(\Omega) \to \text{span}\{\lambda_i\} \subset H^1(\Omega)$ is given by

$$Iu := \sum_i u(\mathbf{v}_i)\lambda_i. \tag{2.96}$$

Since barycentric coordinates are unique on triangles, this is merely the standard linear Lagrange interpolation operator when $\Omega$ is a triangle.

Before stating any error estimates, we fix some notation. For multi-index $\alpha = (\alpha_1, \alpha_2)$ and point $\mathbf{x} = (x, y)$, define $\mathbf{x}^\alpha := x^{\alpha_1}y^{\alpha_2}$, $\alpha! := \alpha_1\alpha_2$, $|\alpha| := \alpha_1 + \alpha_2$, and $D^\alpha u := \partial^{|\alpha|}u/\partial x^{\alpha_1}\partial y^{\alpha_2}$. The Sobolev semi-norms and norms over an open set $\Omega$ are defined by

$$|u|^2_{H^m(\Omega)} := \int_\Omega \sum_{|\alpha|=m} |D^\alpha u(\mathbf{x})|^2 \partial\mathbf{x} \qquad \text{and} \qquad ||u||^2_{H^m(\Omega)} := \sum_{0\le k\le m} |u|^2_{H^m(\Omega)}.$$

The $H^0$-norm is the $L^2$-norm and will be denoted $||\cdot||_{L^2(\Omega)}$.

Analysis of the finite element method often yields bounds on the solution error in terms of the best possible approximation in the finite-dimensional solution space. Thus the challenge of bounding the solution error is reduced to a problem of finding a good interpolant. In many cases Lagrange interpolation can provide a suitable estimate which is asymptotically optimal. For first-order interpolants that we consider, this **optimal convergence estimate** has the form

$$||u - Iu||_{H^1(\Omega)} \leq C \, \mathrm{diam}(\Omega) \, |u|_{H^2(\Omega)}, \quad \forall u \in H^2(\Omega). \tag{2.97}$$

To prove estimate (2.97) in our setting, it is sufficient to restrict the analysis to a class of domains with diameter one and show that $I$ is a bounded operator from $H^2(\Omega)$ into $H^1(\Omega)$, that is

$$||Iu||_{H^1(\Omega)} \leq C_I \, ||u||_{H^2(\Omega)}, \quad \forall u \in H^1(\Omega). \tag{2.98}$$

We call equation (2.98) the $H^1$ **interpolant estimate** associated to the barycentric coordinates $\lambda_i$ used to define $I$.

The optimal convergence estimate (2.97) does not hold uniformly over all possible domains; a suitable geometric restriction must be selected to produce a uniform bound. Even in the simplest case (Lagrange interpolation on triangles), there is a gap between geometric criteria which are simple to analyze (e.g. the minimum angle condition) and those that encompass the largest possible set of domains (e.g. the maximum angle condition).

## 2.8 Biological Applications

Complementary space visualization can be used for model checking, error analysis, detection of topologically uncertain regions, topological preservation in model reduction, and dynamic deformation visualization, as we outline in the following subsections.

### 2.8.1 Tertiary Motif Detection

Using the Morse-Smale complex and stable and unstable manifolds, we can detect helices and sheets in molecular structures as well as large scale ‘tertiary motifs’.

### 2.8.2 Ion channel models

Ion channels are a cell’s mechanism for regulating the flow of ions into and out of the cell. They usually have two main structural confirmations: the “open” configuration, in which the tunnel through its center is wide enough to allow passage of the ions, and a “closed” configuration in which it is not. We look at the acetylcholine receptor (PDB ID 2BG9) as a particular example of an ion channel. This molecule is embedded in a cell membrane, as shown in Figure 2.18a, and is a control mechanism for the flow of sodium and potassium ions into the cell. It is made up of five homologous (in the biological sense) subunits. A conformational change from closed to open occurs when acetylcholine, a small neurotransmitter ligand, docks into the five small pockets on the exterior of the molecule near the tunnel opening in the extracellular region. In particular, when acetylcholine fills one of these pockets, it causes the attached chain subunit of 2BG9 to twist slightly. The combined effect from rotations in all five chains is a widening of the mouth of the tunnel, akin to the opening of a shutter on a traditional camera.

From this description of the action of the acetylcholine receptor, the importance of accurate complementary space topology becomes evident. First, an accurate model of the channel must feature a tunnel passing completely through the length of the surface. Put differently, the complementary space should include a connected component running the length of the molecule with mouths at opposite ends. Such a requirement can be quickly verified by a complementary space visualization as shown in Figure 2.18. Furthermore, the diameter of this tunnel at its narrowest point should be within the range of biological feasibility, i.e. it should be wide enough to accommodate sodium and potassium ions in the open confirmation and narrow enough to block them in the closed confirmation. The margin of error here is quite small as the channel, when open, selectively allows the desired ions and not ions of similar size, e.g. magnesium or calcium. Measuring this width is straightforward with a geometrical representation of the tunnel.

Additionally, the model must have correct geometry at the ligand binding site. In terms of complementary space, this implies the existence of a small component with one mouth on each of the subunits such that its volume and mouth diameter are of plausible size compared to the acetylcholine molecule. We show a visualization of the pocket in one subunit in Figure 2.18 d. While such features are difficult to visualize and measure with a model based primal space, they much easier to detect and manipulate with a model based on complementary space.

### 2.8.3 Ribosome models

The ribosome molecule provides another example of natural structural questions best answered with a complementary space model. Ribosomes live inside cells and are the construction equipment for proteins made within the cell. When a ribosome receives the end of a mRNA chain, it passes the mRNA through a small tunnel in its surface. (More precisely, it opens a flap that pulls the end of the mRNA in, then closes around it). The mRNA is a copy of the DNA data from the nucleus of the cell and codes for the construction of a specific protein. The presence of the mRNA in the ribosome allows specific amino acids to enter a larger tunnel through the ribosome; the type of amino acid permitted to enter depends on the portion of mRNA code in the ribosome at the moment. As the mRNA is fed through, amino acids are linked into a chain, forming the desired protein. Figure 2.19 shows three visualizations of the ribosome molecule: a primal space view, a cut-away view, and a complementary space view.

Since the ribosome has two tunnels of biological significance, an obvious question is to determine the widths of each tunnel. In particular, it would be interesting to compare the width of the mRNA tunnel to the diameter of the mRNA molecule to get a sense of the variation in width that the ribosome will tolerate. Similar questions could be asked of the larger tunnel accommodating the amino acids. The complementary space of the ribosome's surface includes a component for each tunnel. Visualizing these components gives clues to the tunnels' structure and applicable measurements of mouth sizes and tunnel lengths can then be made.

### 2.8.4 Topological Agreement of Reduced Models

Model reduction or decimation is the process of removing geometrical information from a model while attempting to keep sufficient data for maintenance of important features. This is used, for example, in coarse-grained models of proteins, used prominently in electrostatic simulations [25]. Protein surfaces are often defined based on atomic positions and radii, obtained from the PDB. For large proteins, a significant speed-up in computational time can be achieved by grouping atoms into clusters and treating the clusters as single atoms with an averaged radius. Model reduction is also common for point-sampled surfaces such as CAD models and geometries acquired from three-dimensional scanners. If points on the surface can be culled without dramatic effect on the shape of the surface, subsequent visualization and simulation pipelines will experience a reduction in computational cost.

In all model reduction contexts, a primary concern is whether the reduction has changed the topology. Put more precisely, we would like to know when, if ever, the original topology is lost in the progressive decimation of a surface. Complementary space visualization is a natural tool in this context. We consider, for example, the industrial part model shown in Figure 2.20. We use the software QSlim [91] to decimate the model from 106,708 triangles to only 1000 and then only 500. At 1000 triangles, the model has lost some geometrical precision, but still has the same number of tunnels. At 500 triangles, however, some of the smaller tunnels have collapsed, representing a fundamental change in the model. Such changes would be evident from a complementary space visualization.

### 2.8.5 Dynamic Deformation Visualization

Complementary space aids in visualizing and quantifying dynamic deformations of models in addition to its aid for static models previously discussed. The omnipresent consideration in a computer generated simulation of real movement is always whether the dynamics are realistically plausible. In the context of molecular modeling, such considerations are especially difficult to formalize as current video technology cannot capture a molecule in vivo for comparison. As a result, various techniques have been developed for automated animation of molecular models, including the popular of which is Normal Mode Analysis (NMA) [143, 208]. To determine whether the fluctuations simulated by these means have any functional significance to the molecule, we must be able to measure the extent of changes in particular features of the surface. This is especially important in molecules which perform specific actions by modifying their complementary space features, such as the ribosome. With a model of complementary space, we can measure the area of the mouth of a tunnel or pocket used in the various processes and compare the sizes before and after a conformational change. This gives insight into the relative magnitude of different aspects of the shape reconfiguration; a seemingly significant deformation may only involve a small change in the size of a pocket mouth or vice versa.

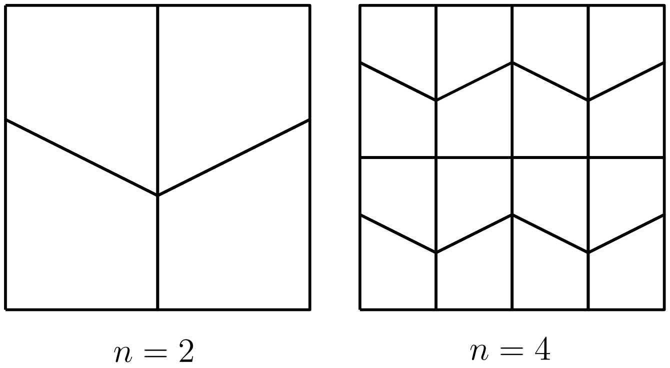
## Summary

## References and Further Reading

Algebraic curves are handled here in real projective space but the interested reader should consider how they can be understood in the additional structure provided by complex projective space.

The exposition of exterior calculus in the continuous setting is based on the presentations in [2, 112].

## Exercises



| | $\|\|u - u_h\|\|_{L^2}$ | | $\|\|\nabla(u - u_h)\|\|_{L^2}$ | |
|---|---|---|---|---|
| n | error | rate | error | rate |
| 2 | 2.34e-3 | | 2.22e-2 | |
| 4 | 3.03e-4 | 2.95 | 6.10e-3 | 1.87 |
| 8 | 3.87e-5 | 2.97 | 1.59e-3 | 1.94 |
| 16 | 4.88e-6 | 2.99 | 4.04e-4 | 1.97 |
| 32 | 6.13e-7 | 3.00 | 1.02e-4 | 1.99 |
| 64 | 7.67e-8 | 3.00 | 2.56e-5 | 1.99 |
| 128 | 9.59e-9 | 3.00 | 6.40e-6 | 2.00 |
| 256 | 1.20e-9 | 3.00 | 1.64e-6 | 1.96 |

Figure 2.15: Trapezoidal meshes (left) fail to produce quadratic convergence with traditional serendipity elements; see [13]. Since our construction begins with affinely-invariant generalized barycentric functions, the expected quadratic convergence rate can be recovered (right). The results shown were generated using the basis $\{\psi_{ij}\}$ resulting from the selection of the mean value coordinates as the initial barycentric functions.

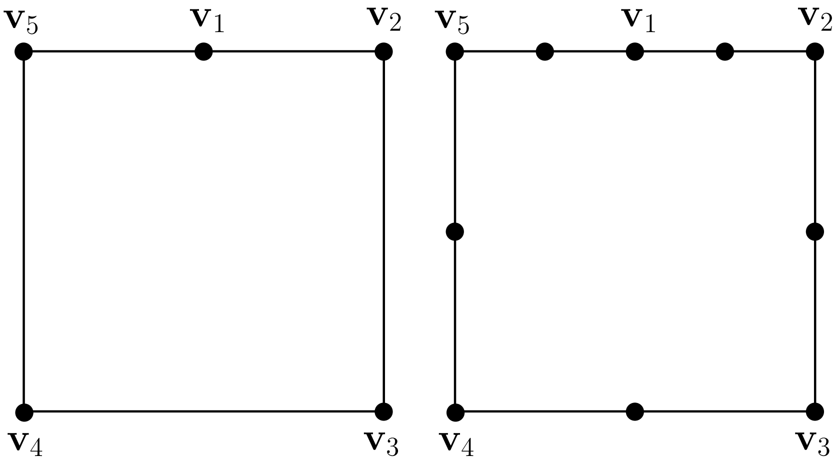


Figure 2.16: Theorem 2.45 can be generalized to allow certain types of geometries that do not satisfy G3. The degenerate pentagon (left), widely used in adaptive finite element methods for quadrilateral meshes, satisfies G1 and G2, but only satisfies G3 for four of its vertices. The bounds on the coefficients $c_{ab}^{ij}$ from Section 2.7.2 still hold on this geometry, resulting in the Lagrange-like quadratic element (right).

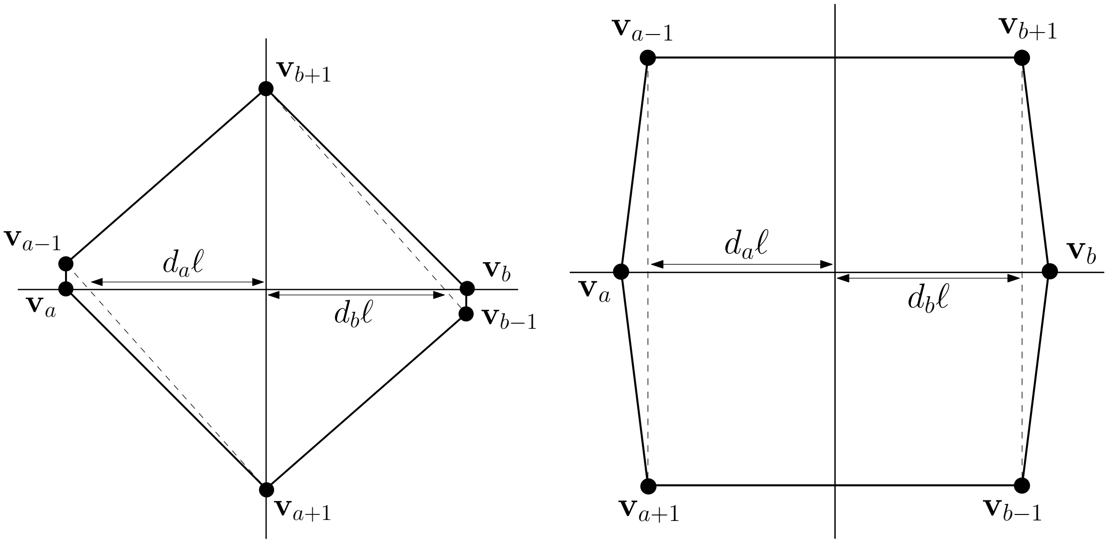


Figure 2.17: The hypotheses of Theorem 2.45 cannot be relaxed entirely as demonstrated by these shapes. If G2 does not hold, arbitrarily small edges can cause a blowup in the coefficients $c_{ab}^{ij}$ (left). If G3 does not hold, non-consecutive angles approaching $\pi$ can cause a similar blowup.

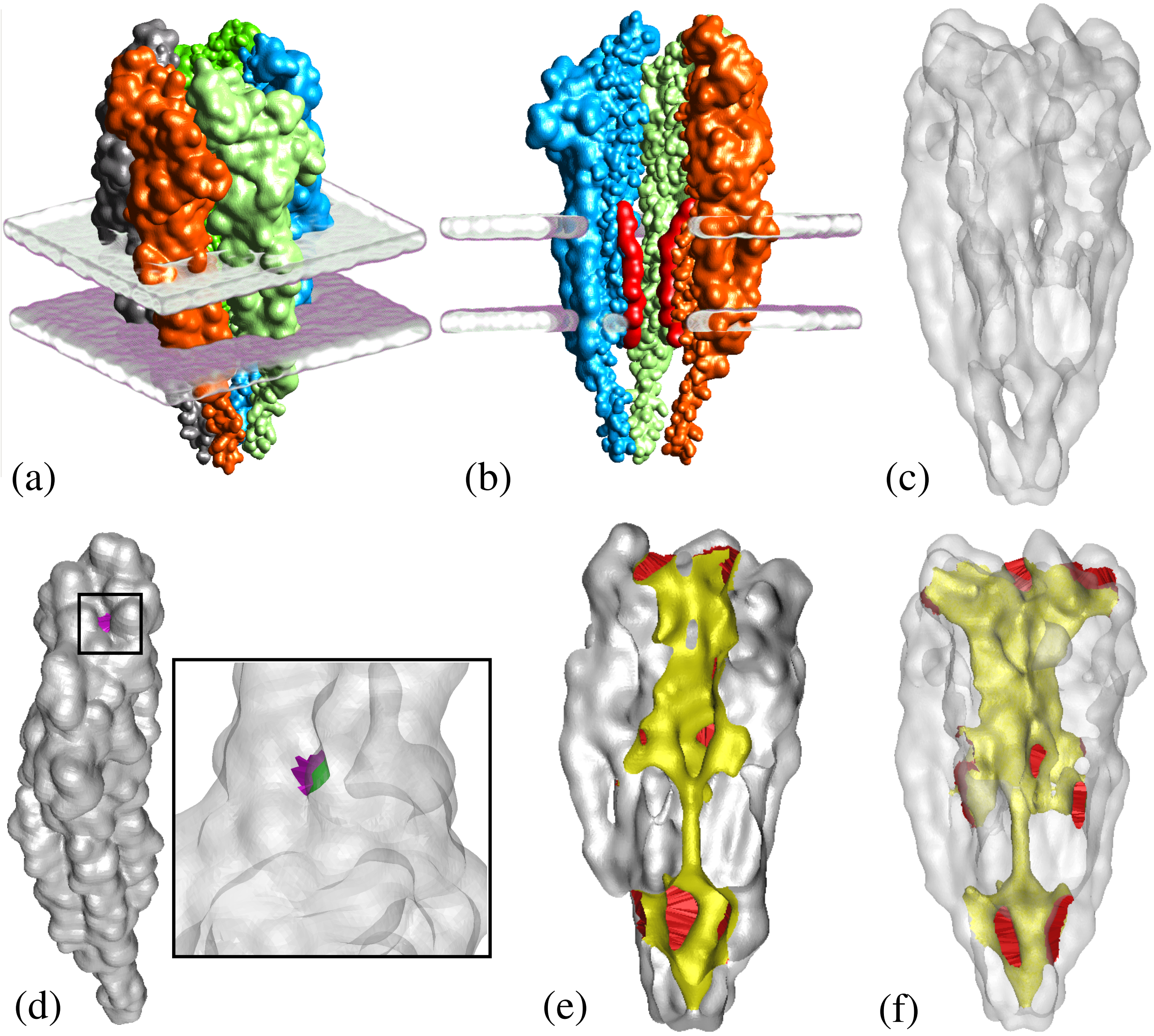


Figure 2.18: Various visualizations of the acetylcholine receptor molecule. The top and bottom rows shows primal and complementary space visualizations, respectively. **(a)** The molecule is shown as it would sit embedded in a bilipid cell membrane (grey) with the five identical subunit colored for identification. **(b)** A cut-away view of the same model showing where ions may pass through the center. **(c)** A transparent view of the molecular surface. **(d)** Each subunit contains a pocket where acetylcholine binds. The pocket interior (green) and its mouth (purple) are shown in a zoomed in view after the surface has been made transparent. **(e)** A cut-away view of the surface with the interior of the tunnel (yellow) and its mouths (red) identified. **(f)** The same view as (c) with the tunnel geometry opaque, showing how it lies inside the surface.

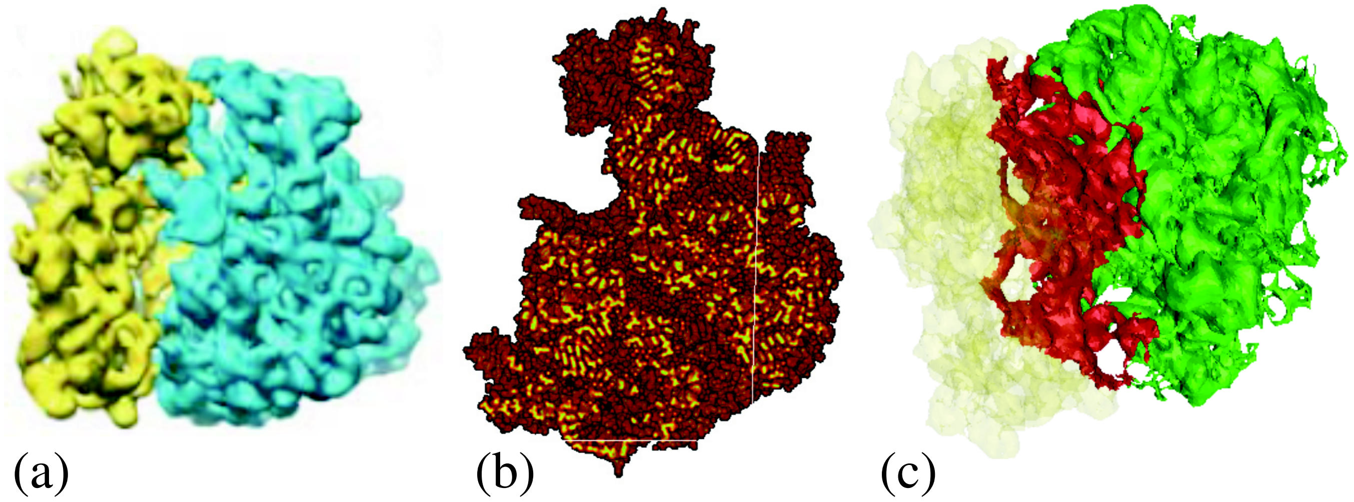


Figure 2.19: Three visualizations of the ribosome molecule. **(a)** A primal space visualization showing the two subunits in their joined state. **(b)** A cut-away view of the molecule with the protein exit tunnel visible diagonally from upper left to bottom right. **(c)** A complementary space view of the exit tunnel indicates its intricate three-dimensional geometry in a way that the primal and cut-away views do not.

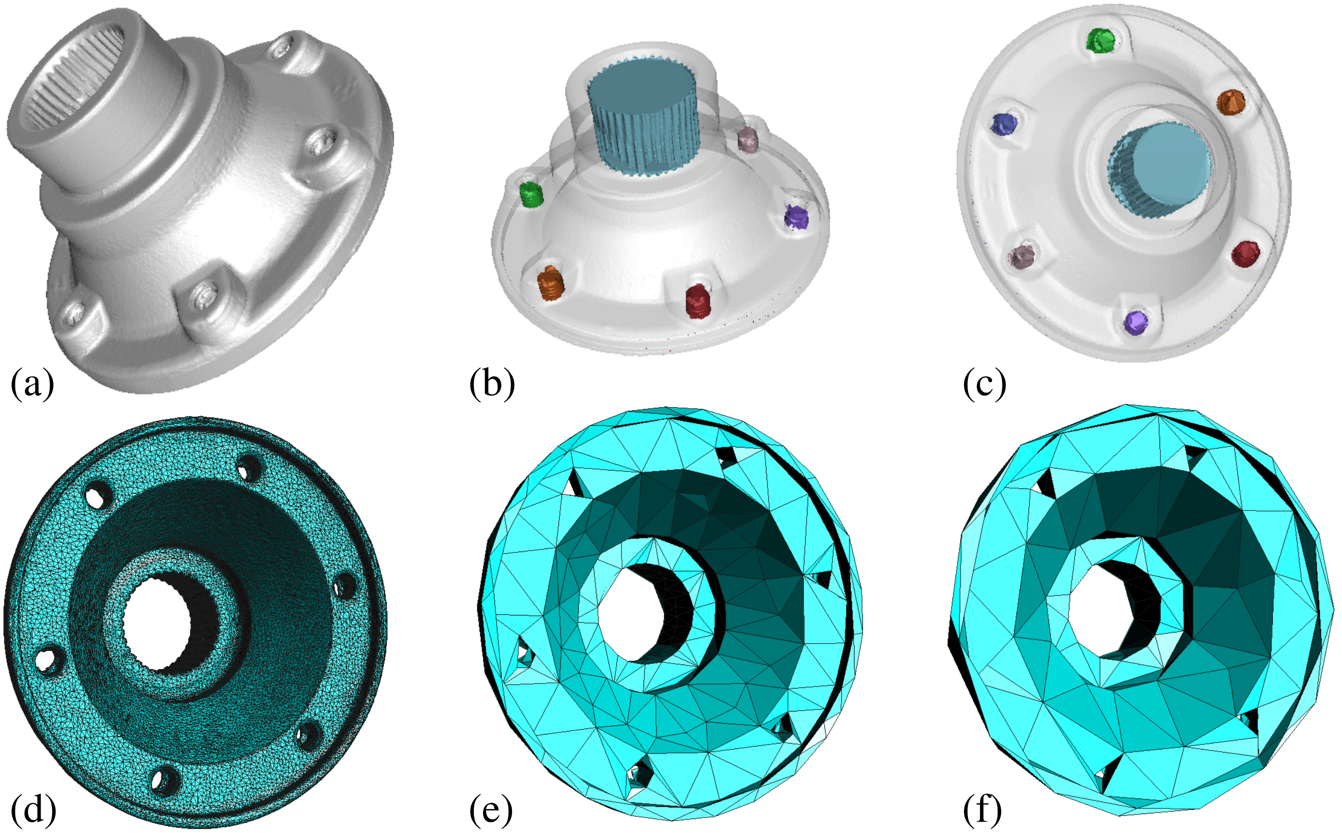


Figure 2.20: Visualizations of the Carter dataset. **(a)** A basic primal space visualization of the mechanical part . **(b-c)** Complementary space features identified and visualized. **(d)** A visualization of the dense mesh representing the surface reveals that at 106,708 triangles, it is probably amenable to decimation. **(e)** Using QSlim [91], the mesh is decimated to 1000 triangles. Prominent topological and geometrical features are still present, though the geometry of the smaller tunnels has changed. **(f)** Decimated to 500 triangles, some of the smaller tunnels have collapsed, causing a topological change in the model. Visualizing complementary space could aid in detecting such changes.

# Chapter 3

# Differential Geometry, Operators

## 3.1 Shape Operators, First and Second Fundamental Forms

### 3.1.1 Curvature: Gaussian, Mean

### 3.1.2 The Shape of Space: convex, planar, hyperbolic

### 3.1.3 Laplacian Eigenfunctions

## 3.2 Finite Element Basis, Functional Spaces, Inner Products

### 3.2.1 Hilbert Complexes

**Definition 3.1.** A real **Hilbert space** $W$ is a vector space with a real-valued inner product $(\cdot, \cdot)$ such that $W$ is complete with respect to the norm given by

$$||w||_W := (w, w)^{1/2}.$$

A **Hilbert complex** $(W, d)$ is a sequence of Hilbert spaces $W^k$ and a sequence of closed, densely defined linear operators $d_k : W^k \to W^{k+1}$ such that the range of $d_k$ is contained in the kernel of $d_{k+1}$, i.e.

$$d_{k+1} \circ d_k = 0.$$

The **domain complex** $(V, d)$ associated to $(W, d)$ is the sequence of spaces $V^k := \text{domain}(d_k) \subset W^k$ along with the **graph norm** defined via the inner product

$$(u, v)_{V^k} := (u, v)_{W^k} + (d_k u, d_k v)_{W^{k+1}}.$$

$\Diamond$

**Definition 3.2.** The space of $L^2$-bounded differential forms along with the exterior derivative map define a Hilbert complex $(L^2\Lambda, d)$. The associated domain complex, denoted $(H\Lambda, d)$ is called the $L^2$ **deRham complex**:

$$\begin{tikzcd} 0 \arrow[r] & H\Lambda^0 \arrow[r, "d_0"] & H\Lambda^1 \arrow[r, "d_1"] & \cdots \arrow[r, "d_{n-1}"] & H\Lambda^n \arrow[r] & 0 \end{tikzcd}$$

$\Diamond$

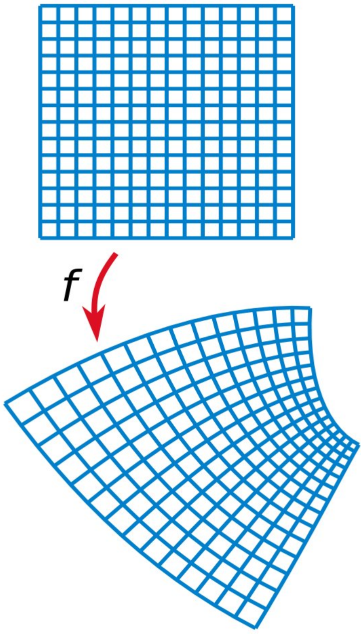
## 3.3 Topology of Function Spaces

## 3.4 Differential Operators and their Discretization formulas

discrete and continuous formulas

## 3.5 Conformal Mappings from Intrinsic Curvature

**Conformal Maps**



A conformal map $f : X \to Y$ is function which preserves angles. In the mapping shown in the figure, the rectangular grid is distorted by $f$ but the 90 angles at each grid point (and indeed everywhere) are preserved in an infintesimal sense. Conformal maps are useful when paramaterizing molecular surfaces as a triangulation with good angle bounds will preserve the angle bounds when passed to through the parametrization to a surface triangulation.

To create a global conformal parameterization of a surface $\Omega$ with arbitrary topology, one must sovle the following problem. Let $\Omega$ be a surface of genus $g$ and let $\{L_1, \ldots, L_{2g}\}$ be a set of loops providing a basis for the 1-homology group of $\Omega$. Let $c_1, \ldots, c_{2g} \in \mathbb{R}$ where $c_i$ represents the desired value of the integral of the gradient field around $e_i$.

The goal is to find a conformal gradient field $\omega + \sqrt{-1} \star \omega$ where $\omega$ and $\star\omega$ are real gradient fields on $\Omega$. Further, $\omega$ and $\star\omega$ should be closed, harmonic, determined by their values of integration over the homology basis, and orthogonal to each other. This means $\omega$ should solve the following PDE:

$$\left\{ \begin{array}{rcl} d\omega & = & 0 \\ \Delta\omega & = & 0 \\ \int_{e_i} \omega & = & c_i \end{array} \right.$$

Given a solution to the above, one can construct an appropriate $\star\omega$ as well. The approach below is based on that of Gu and Yau [110] Section 3.2.

**Approach:** Let $M$ be a mesh of $\Omega$ with the correct topology. The computation of the homology basis is as follows .

$Q_\xi$1. Compute the dual mesh $\bar{M}$ of $M$. The dual mesh has a face for every vertex of $M$, a vertex for every face of $M$, and an edge for every edge of $M$ with connectivity provided in the standard manner.

$Q_\xi$2. Find a minimal spanning tree $\bar{T}$ of the vertices of $\bar{M}$.

$Q_\xi$3. Define the graph $G$ to be those edges of $M$ whose dual edge is not in $\bar{T}$. Then $G$ is a cut graph of $M$, meaning $M/G$ is topologically a disk and $G$ has $2g$ loops, corresponding to homology basis elements.

$Q_\xi$4. Construct a maximal spanning tree $T$ of $G$. Since $G$ has $2g$ loops, $G - T$ is exactly $2g$ edges $\{e_1, \ldots, e_{2g}\}$, one per loop of $G$.

Q$_\xi$5. Each $e_i$ connects two leaves of the tree $T$. Let $L_i$ denote the loop in $G$ consisting of the path from one of these leaves to the root of $T$, down the other leaf, and across $e_i$. Then $\{L_1, \dots, L_{2g}\}$ are non-trivial, independent (i.e. not homotopic nor homologous) loops in $G$, and hence form a basis for the first homology group $H_1(M, \mathbb{Z})$

The PDE is discretized as follows. Let $[u, v]$ denote an oriented edge and $[u, v, w]$ denote an oriented face in the mesh $M$. For an edge $[u, v]$, let $\alpha, \beta$ be the angles against the edge at $u$ and define

$$k_{u,v} := -\frac{\;}{1}2(\cot \alpha + \cot \beta).$$

Also, for each homology basis element $L_i$, write

$$L_i = \sum_{j=1}^{n_i} [u^i_{j-1}, u^i_j], \quad u_0 = u_{n_i}.$$

Then the PDE is discretized as

$$\begin{cases} \sum_{j=1}^{3} \omega([u_{j-1}, u_j]) &= 0 \quad \forall [u_0, u_1, u_2] \in M, u_0 = u_3 \\ \sum_{[u,v] \in M} k_{u,v} \omega([u, v]) &= 0 \quad \forall u \in M \\ \sum_{j=1}^{n_i} \omega([u^i_{j-1}, u^i_j]) &= c_i \quad \forall L_i \end{cases}$$

The authors prove this linear system is of full rank, hence it has a solution. In words, the above equations seek a vector field $\omega$ such that:

- $\omega$ integrated around any face is zero (so $d\omega = 0$)
- At each vertex $u$ of the mesh, summing the values of $\omega$ on the edges around $u$ with appropriate cotangent weights is zero (so $\omega$ is harmonic in the discrete sense)
- The integral of $\omega$ around each homological basis element has a prescribed value.

A basis for the solution set would be $\{\omega_i\}$ where for $\omega_i$, set $c_i = 1$ and set $c_j = 0$ for $j \neq i$ and solve the above. Then $\omega_i$ has value 1 when integrated around $L_i$ and value 0 when integrated around any combination of basis elements besides $L_i$.

# 3.6 Biological Applications

## 3.6.1 Molecular Surface Analysis

## 3.6.2 Solving PDEs in Biology

# Summary

# References and Further Reading

For more on conformal mapping, see [187, 110, 130, 191]

# Exercises

# Chapter 4

# Differential Forms and Homology of Discrete Functions

## 4.1 Exterior Calculus

Recall the definitions of tensors and exterior algebra given in Section 2.1.
Let $\Omega$ be an $n$-manifold embedded in some $\mathbb{R}^N$ with $n \leq N$. Minimally, we will assume $\Omega$ is a bounded subset, but we will usually consider the case $n = N = 3$ and assume $\Omega$ has a piecewise smooth, Lipschitz boundary as this allows us to identify $\Omega$ with its primal mesh (Definition 1.10) or dual mesh (Definition 1.14).

**Definition 4.1.** Let $\Omega$ be a manifold of dimension $n$. Given a point $x \in \Omega$, we denote the **tangent space of** $\Omega$ **at** $x$ by $T_x(\Omega)$. Let $0 \leq k \leq n$. A **k-form** $\omega$ is a mapping from $\Omega$ to the space of alternating $k$-tensors on the tangent space of $\Omega$ at the input point. We use the notation

$$\omega : \Omega \to \Lambda^k[T_x(\Omega)^*], \qquad \omega(x) : \bigoplus_{i=1}^{k} T_x(\Omega) \to \mathbb{R},$$

where $\omega(x)$ is an alternating $k$-tensor. A 0-form is taken to mean a real-valued function on $\Omega$. We denote **the space of continuous differential k-forms** on $\Omega$ by $\Lambda^k(\Omega)$. $\Diamond$

**Definition 4.2.** A **differential** $dx_i$ is a 1-form whose action at $x \in M$ is to assign the $i^{\text{th}}$ value of the input vector from $T_x(M)$. Let $I = \{i_1, \ldots, i_k\}$ be a list of indices. Define

$$dx_I := dx_{i_1} \wedge \cdots \wedge dx_{i_k}.$$

We use the notation $a_I$ to a real-valued function in the variables of $I$.

**Theorem 4.3.** *If $\{dx_1, \ldots, dx_n\}$ is an orthonormal basis for $T_x(\Omega)$ then*

$$\{dx_I : \quad I = \{i_1, \ldots, i_k\}, \quad 1 \leq i_1 < \cdots < i_k \leq n\}$$

*is a basis for $\Lambda^k(\Omega)$. Put differently, any k-form $\omega \in \Lambda^k(\Omega)$ can be written in the form*

$$\omega = \sum_I a_I dx_I$$

*where $I$ ranges over all strictly increasing sequences of $k$ indices.*

The theorem is a standard result from differential topology.

**Definition 4.4.** The space of $L^2$**-bounded continuous differential k-forms** on $\Omega$ is given by

$$L^2\Lambda^k(\Omega) := \left\{ \sum_I a_I dx_I \in \Lambda^k(\Omega) \ : \ a_I \in L^2(\Omega) \quad \forall I \right\}$$

$\Diamond$

**Definition 4.5.** The **exterior derivative** operator denoted by $d$ is a map

$$d : \Lambda^k(\Omega) \to \Lambda^{k+1}(\Omega),$$

defined as follows. Let $I := \{i_1, \dots, i_k\}$ denote an increasing sequence of $k$ indices ($i_j < i_{j+1}$) and let $dx_I = dx_{i_1} \wedge \cdots \wedge dx_{i_k}$. Given $\omega = \sum_I a_I dx_I$ define

$$d\omega := \sum_I da_I \wedge dx_I \quad \text{where} \quad da_I := \sum_{i \in I} \frac{\partial a_I}{\partial x_i} dx_i. \tag{4.1}$$

◊

We note that $d$ commutes with pullbacks (that is, $df^*\omega = f^*d\omega$) and that if $\omega$ is a $k$-form and $\theta$ is any form,

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

The exterior derivative plays a prominent role in Stokes' Theorem, which we now state.

**Theorem 4.6.** ***(Stokes)*** *Given a compact, oriented $n$-dimensional manifold $\Omega$ with boundary $\partial\Omega$ and a smooth $(n-1)$ form $\omega$ on $\Omega$, the following equality holds:*

$$\int_{\partial\Omega} \omega = \int_\Omega d\omega.$$

Stokes' Theorem provides an alternative definition for the exterior derivative.

**Definition 4.7.** **(Alternative Definition)** Let $\omega$ be a $k$-form on a compact oriented $n$-manifold $\Omega$ $(0 \le k < n)$. The **exterior derivative** of $\omega$ is the unique $(k+1)$-form $d\omega$ such that on any $(k+1)$-dimensional submanifold $\Pi \subset \Omega$ the following equality holds:

$$\int_\Pi d\omega = \int_{\partial\Pi} \omega.$$

◊

It can be shown that $d\omega$ is well-defined in this way by proving the existence and uniqueness of the $d$ map via the definition (4.1). We note that this definition will motivate the discrete exterior derivative in Definition 4.13.

**Definition 4.8.** The continuous Hodge star $*$ maps between forms of complementary and orthogonal dimensions, i.e. $* : \Lambda^k \to \Lambda^{n-k}$. For domains in $\mathbb{R}^3$ as considered here, $*$ is defined by the equations

$$*dx_1 = dx_2 dx_3, \quad *dx_2 = -dx_1 dx_3, \quad *dx_3 = dx_1 dx_2,$$

$$*1 = dx_1 dx_2 dx_3, \quad ** = 1,$$

where $\{dx_1, dx_2, dx_3\}$ is an orthonormal basis for $\Lambda^1(\Omega)$. ◊

## 4.2 deRham Cohomology

## 4.3 $k$-forms and $k$-cochains

### 4.3.1 Discrete Differential Forms

**Definition 4.9.** Let $K$ be a primal mesh of a compact $n$-manifold $\Omega$. Let $K_k$ denote the $k$-simplices of $K$. A **primal $k$-chain** $c$ is a linear combination of the elements of $K_k$:

$$c = \sum_{\sigma \in K_k} c_\sigma \sigma,$$

where $c_\sigma \in \mathbb{R}$. The set of all such chains form the **vector space of primal $k$-chains**, denoted $\mathcal{C}_k$. It has dimension $|\mathcal{C}_k|$, equal to the number of elements of $K_k$. A $k$-chain $c$ is represented as a column vector of length $|\mathcal{C}_k|$.
Similarly, a **dual $k$-chain** is a linear combination of $k$-cells of the dual complex $\star K$. The vector space of dual $k$-chains is denoted $\overline{\mathcal{C}}_k$. ◊

**Definition 4.10.** A **primal $k$-cochain** $\mathrm{w}$ is a linear functional on primal $k$-chains, i.e.

$$\mathrm{w} : \mathcal{C}_k \to \mathbb{R} \quad \text{via} \quad c \mapsto \mathrm{w}(c),$$

where $\mathrm{w}$ is a linear mapping. It is represented as a column vector of length $|\mathcal{C}_k|$ so that the action of $\mathrm{w}$ on a $k$-chain $c$ is the matrix multiplication $\mathrm{w}^T c$, yielding the scalar $\mathrm{w}(c)$. The space of primal cochains is denoted $\mathcal{C}^k$.
A **dual $k$-cochain** $\overline{\mathrm{w}}$ is a linear functional on dual $k$-chains, i.e.

$$\overline{\mathrm{w}} : \overline{\mathcal{C}}_k \to \mathbb{R} \quad \text{via} \quad c \mapsto \overline{\mathrm{w}}(c),$$

where $\overline{\mathrm{w}}$ is a linear mapping. The space of dual cochains is denoted $\overline{\mathcal{C}}^k$. ◊

Cochains are the discrete analogues of differential forms as they can be evaluated over $k$-dimensional subspaces. To make this precise, we define the integration of a cochain over a chain to be the evaluation of the cochain as a function.

**Definition 4.11.** The **integral** of a primal $k$-cochain $\mathrm{w}$ over a primal $k$-chain $c$ is defined to be

$$\int_c \mathrm{w} := \; w^T c = \; w(c).$$

Hence, the integration of $\mathrm{w}$ over $c$ is exactly the same as the evaluation of $\mathrm{w}$ on $c$. ◊

### 4.3.2 Discrete Exterior Derivative

The definition of a discrete exterior derivative is motivated by the alternative definition of the continuous operator (Definition 4.7). First we define the boundary operator in the discrete case.

**Definition 4.12.** The **$k$th boundary operator** denoted by $\partial_k$ takes a primal $k$-chain to its primal $(k-1)$-chain boundary. It is defined by its action on an oriented $k$-simplex:

$$\partial_k[v_0, v_1, \cdots, v_k] := \sum_{i=0}^{k}(-1)^i[v_0, \cdots, \widehat{v_i}, \cdots, v_k]$$

where $\widehat{v_i}$ indicates that $v_i$ is omitted. The primal boundary operator is represented as a matrix of size $|\mathcal{C}_{k-1}| \times |\mathcal{C}_k|$ so that the action of $\partial_k$ on a $k$-chain $c$ is the usual matrix multiplication $\partial_k c$. ◊

**Definition 4.13.** The **$k$th discrete exterior derivative** of a primal $k$-cochain $\mathrm{w}$ is the transpose of the $(k+1)$st boundary operator:

$$\mathbb{D}_k = \partial_{k+1}^T.$$

This is also referred to in the literature as the **coboundary operator**. It is represented as a matrix of size $|\mathcal{C}_{k+1}| \times |\mathcal{C}_k|$ so that the action of $\mathbb{D}_k$ on a primal $k$-cochain $\mathrm{w}$ is the usual matrix multiplication $\mathbb{D}_k \mathrm{w} := \partial_{k+1}^T \, w$. ◊

The discrete exterior derivative satisfies the discrete version of Stokes' theorem.

**Lemma 4.14.** *Let $\mathrm{w}$ be a primal $k$-cochain and $c \in \mathcal{C}_{k+1}$ any primal $(k+1)$-chain. Then*

$$\int_c \mathbb{D}_k \mathrm{w} = \int_{\partial_{k+1} c} w.$$

*Proof.* By Definition 4.11 we see that

$$\int_{\partial_{k+1} c} \mathrm{w} = \; w^T \partial_{k+1} c = (\partial_{k+1}^T \, w)^T c = (D_k \, w)^T c = \int_c D_k \, w.$$

□

We now consider the analogous constructions for dual cochains. Observe that mesh duality allows us to view a dual $k$-chain $\overline{c}$ as a primal $(n-k)$-chain $c$. Hence $\partial_{n-k+1}^T$ serves as a boundary operator on dual $k$-cochains, giving us the following definition.

**Definition 4.15.** The **$k$th discrete exterior derivative** of a dual $k$-cochain $\overline{\mathrm{w}}$ is $\mathbb{D}_{n-k-1}^T$, which is equal to $\partial_{n-k}$. It is represented as a matrix of size $|\overline{\mathcal{C}}_{k+1}| \times |\overline{\mathcal{C}}_k|$. ◊

## 4.4 Types of $k$-form Finite Elements

### 4.4.1 Nédélec Elements

The original Nédélec paper [158] introduced what is now called the $H(\text{curl})$ and $H(\text{div})$ Nédélec elements of the first kind. Phrased in the notation of this document, Nédélec defines

$$Y^k_{p-1} := S_p \Lambda^k_{p-1}$$

where $S_p \subset \widetilde{\mathbb{P}}_p$ is defined for domains embedded in $\mathbb{R}^3$ by

$$S_p := \{f \in \widetilde{\mathbb{P}}_p \ : \ (x_1, x_2, x_3) \cdot f = 0\}.$$

He uses the formal definition of finite elements given below.

**Definition 4.16.** A **finite element** is a triple $(K,P,A)$ where

- $K$ is a domain
- $P$ is a space of polynomials on $K$ of dimension $N$
- $A$ is a set of $N$ linear functionals acting on $P$ called **degrees of freedom**

His $H(\text{curl})$ finite element of degree $p$ from [158] is defined as follows. Set

- $K$ to be a tetrahedron
- $P$ to be $\mathbb{P}_{p-1} \oplus S_p$
- $A$ to be the following linear functionals acting on an element $u \in \mathbb{P}p$

  1 $\int_e (u \cdot \vec{e}) \ \ q \ \ ds, \forall q \in \mathbb{P}_{k-1}$
  where $\vec{e}$ is the unit vector directed along the edge $e$ of $K$;

  2 $\int_f (u \times \hat{n} \cdot q) \ \ d\gamma, \forall q \in (\mathbb{P}_{k-2})^2$
  where $f$ is a face of $K$ and $u \times \hat{n}$ denotes the normal trace of $u$;

  3 $\int_K \ u \cdot q \ \ dx; \forall q \in (\mathbb{P}_{k-3})^3$

Note that this does not say what the basis functions should be - this was left to future work. The Whitney functions are an example of lowest order Nédélec elements.
The $H(\text{div})$ finite element of degree $p$ from [158] is similar. Set

- $K$ to be a tetrahedron
- $P$ to be $\mathbb{P}_{p-1} \oplus \widetilde{\mathbb{P}}_p(x_1, x_2, x_3)$[1]
- $A$ to be the following linear functionals acting on an element $u \in \mathbb{P}p$

  Q$_\xi$1. $\int_f (u \cdot \hat{n}) \ \ q \ \ d\gamma, \forall q \in \mathbb{P}_{k-1}$
  where $f$ is a face of $K$;

  Q$_\xi$2. $\int_K \ u \cdot q \ \ dx; \forall q \in (\mathbb{P}_{k-2})^3$

Similarly, description of the basis functions was left to future work.

[1] The space $\widetilde{\mathbb{P}}_p(x_1, x_2, x_3)$ is called $\mathcal{D}^p$ in his work. It is the space of vectors formed by scaling $(x_1, x_2, x_3)$ by a homogeneous degree $p$ polynomial.

### 4.4.2 Whitney Elements

Let $\lambda_i$ be the **barycentric function** associated to vertex $\mathbf{v}_i$ in a primal mesh $K$. More precisely, $\lambda_i : K \to \mathbb{R}$ is the unique function which is linear on each simplex of $K$ satisfying $\lambda_i(\mathbf{v}_j) = \delta_{ij}$. The **Whitney function** $\mathcal{W}_{\sigma^k}$ associated to the $k$-simplex $\sigma^k := [\mathbf{v}_0, \ldots, \mathbf{v}_k]$ is given by

$$\mathcal{W}_{\sigma^k} := k! \sum_{i=0}^{k} (-1)^i \; \lambda_i \; d\lambda_0 \wedge \ldots \wedge \widehat{d\lambda_i} \wedge \ldots \wedge d\lambda_k \tag{4.2}$$

where $\widehat{d\lambda_i}$ indicates that $d\lambda_i$ is omitted. Note that $d\lambda$ should be interpreted as $d^{\,0}\lambda$ per Definition **??** or as ${}^1(\nabla\lambda)$ per Lemma **??**. ◊

We write out the Whitney functions explicitly for $n = 3$, our primary application context. Note that $\mathcal{W}_{\sigma^3}$ is the constant function with value $1/|\sigma^3|$. This is a consequence of the geometric identity

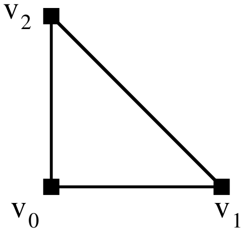
| $k$ | $\sigma^k$ | $\mathcal{W}_{\sigma^k}$ |
|---|---|---|
| 0 | $[\mathbf{v}_0]$ | $\lambda_0$ |
| 1 | $[\mathbf{v}_0, \mathbf{v}_1]$ | $\lambda_0 \nabla \lambda_1 - \lambda_1 \nabla \lambda_0$ |
| 2 | $[\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2]$ | $2(w_0 \nabla w_1 \times \nabla w_2 - w \nabla w_1 \times \nabla w02 + w \nabla w_2 \times \nabla w01)$ |
| 3 | $[\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$ | $1/\lvert\sigma^3\rvert$ |

Table 4.1: Whitney forms $W_{\sigma^k}$ for $n = 3$.

$$\nabla \lambda_i \cdot (\nabla \lambda_j \times \nabla \lambda_k) = \pm \frac{}{1} 3! |\sigma^3|$$

where the right side has sign $-1$ if an odd index was omitted from the scalar triple product and $+1$ otherwise. This reduces the sum in (4.2) to $(1/|\sigma^3|) \sum_i \lambda_i$, which is simply $1/|\sigma^3|$ due to the partition of unity formed by the barycentric functions.

**Whitney Element Example**



On the standard reference triangle shown above, the Whitney 0-forms are the barycentric functions:

$$\begin{aligned} \lambda_0 &= -x - y + 1 \\ \lambda_1 &= x \\ \lambda_2 &= y \end{aligned}$$

The Whitney 1-forms are formed by taking products of the form $\lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i$:

$$\begin{aligned} \mathcal{W}_{01} &= \lambda_0 \nabla \lambda_1 - \lambda_1 \nabla \lambda_0 = \begin{bmatrix} 1 - y \\ x \end{bmatrix} \\ \mathcal{W}_{02} &= \lambda_0 \nabla \lambda_2 - \lambda_2 \nabla \lambda_0 = \begin{bmatrix} y \\ 1 - x \end{bmatrix} \\ \mathcal{W}_{12} &= \lambda_1 \nabla \lambda_2 - \lambda_2 \nabla \lambda_1 = \begin{bmatrix} -y \\ x \end{bmatrix} \end{aligned}$$

## 4.5 Biological Applications

### 4.5.1 Solving Poisson's Equation and other PDEs from Biology

## Summary

## References and Further Reading

Although Whitney functions were developed out of theoretical considerations [220], it was recognized by Bossavit [34] that they provided a natural means for constructing stable bases for finite element methods, especially the edge elements and face elements that were gaining popularity at that time. Finite element exterior calculus (FEEC) [11] gives a full account of the analogies between spaces of Whitney functions and classical Nédélec [158, 159] and Raviart and Thomas [173] spaces.
Some work has explored the possibility of Whitney functions over non-simplicial elements as we do in this work. Gradinaru and Hiptmair defined Whitney-like functions on rectangular grids using Haar-wavelet approximations [104] and on square-base pyramids by considering the collapse of a cube to a pyramid [105]. Bossavit has given an approach to Whitney forms over standard finite element shapes (hexahedra, triangular prisms, etc.) based on extrusion and conation arguments [32].

## Exercises

# Chapter 5

# Numerical Integration, Linear Systems

## 5.1 Numerical Quadrature

### Quasi Monte Carlo method

The Monte Carlo numerical integration by sampling can be replaced by deterministic Quasi Monte Carlo (QMC) integration using Quasi random and low -discrepancy sampling.
Let $\mathbf{x} \in \{0,1\}^n$ be an $n$-bit string. Denote $\tilde{\mathbf{x}}$ as the binary value of the string $\mathbf{x}$ prepended by a radix point. For example, if $\mathbf{x} = \texttt{110011}$, then $\tilde{\mathbf{x}} = .110011_{\text{bin}} = 2^{-1} + 2^{-2} + 2^{-5} + 2^{-6} = \frac{51}{64}$.
Define $f(\mathbf{x}) = g(\tilde{\mathbf{x}})$. For all $0 \leq i \leq 2^n - 1$ and $y \in [\frac{i}{2^n}, \frac{i+1}{2^n}]$, we have

$$g\left(\frac{i}{2^n}\right) - \frac{c}{2^n} \leq g(y) \leq g\left(\frac{i}{2^n}\right) + \frac{c}{2^n}.$$

This implies that

$$\frac{1}{2^n} \sum_{\mathbf{x}\in\{0,1\}^n} \left(f(\mathbf{x}) - \frac{c}{2^n}\right) \leq \int_0^1 g(x)\ \mathrm{d}\,x \leq \frac{1}{2^n} \sum_{\mathbf{x}\in\{0,1\}^n} \left(f(\mathbf{x}) + \frac{c}{2^n}\right).$$

If $n$ is increasingly large, the estimate $\frac{1}{2^n}\sum_{\mathbf{x}\in\{0,1\}^n} f(\mathbf{x})$ converges to the true $\int_0^1 g(x)\ \mathrm{d}\,x = \mathrm{E}[g(X)]$. Furthermore, if we sample $\mathbf{x}$ either i.i.d. or pairwise independent, we can also bound the error (Figure 5.1(a)).
We can also bound the integration error by the discrepancy using the Koksma-Hlawka Inequality (Figure 5.1(b)). See [196] for a list of references.

**Theorem 5.1** (Koksma-Hlawka Inequality)**.** *Let $\Delta(t) = \frac{1}{N}\sum_{i=0}^{N-1} \mathbb{1}_{[0,t]}(x_i) - \frac{t}{1}$ be the local Star discrepancy of sampling. Then, for all $1 \leq p, q \leq \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$, we have*

$$\left| \int_0^1 g(x)\ \mathrm{d}\,x - \frac{1}{N}\sum_{i=0}^{N-1} g(x_i) \right| \leq \|\Delta\|_p \|g'\|_q.$$

Note that $\Delta$ is the difference of the fraction of points that fall in an interval $[0,t]$ and the fraction of the length of the interval $[0,t]$, and thereby a measure of distortion from uniformity.
Suppose $P = \{x_0, x_1, x_2, \ldots, x_{N-1}\}$. The *star discrepancy* is defined as

$$\Delta_P(t) = \frac{1}{N}\sum_{i=0}^{N-1} \mathbb{1}_{[0,t]}(x_i) - t,$$

where

$$\mathbb{1}_{[0,t]}(x) = \begin{cases} 1 & \text{if } x \in [0,t] \\ 0 & \text{otherwise} \end{cases}$$

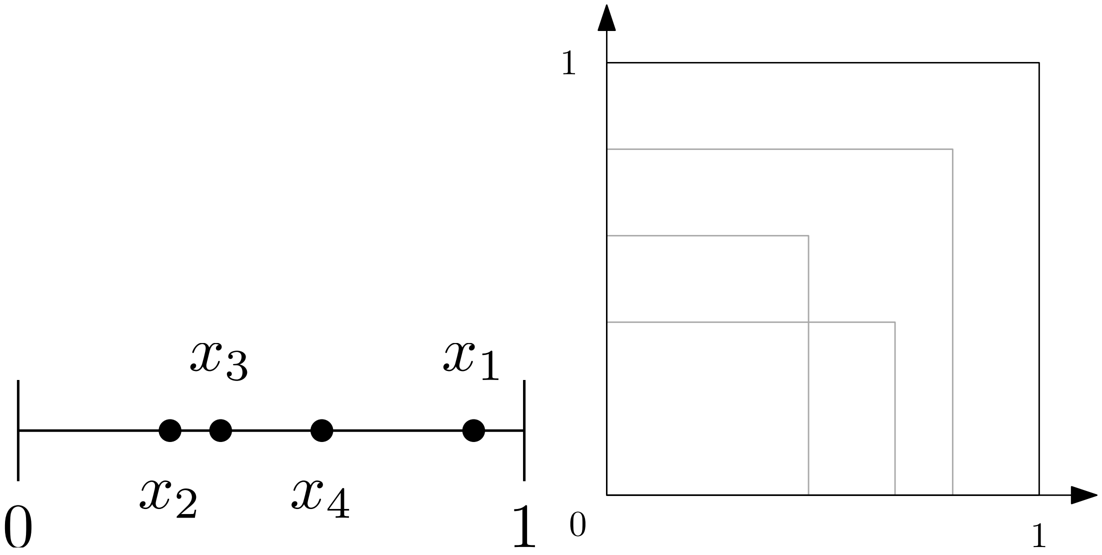


Figure 5.1: (a) Sample points in the unit interval used for integral estimation. (b) Axis-parallel rectangles anchored at the origin used to measure star discrepancy.

is the characteristic function. The star discrepany can be regarded as a test of randomness (i.e. how uniform is the distribution) using the family of all intervals with the left endpoint at the origin. It is related to the Kolmogorov-Smirnov test.

The Koksma-Hlawka inequality can be used the bound the integration error. See [196] for a list of references.

**Theorem 5.2** (Koksma-Hlawka inequality)**.** *For all* $1 \le p, q \le \infty$ *such that* $\frac{1}{p} + \frac{1}{q} = 1$,

$$\left| \int_0^1 f(x) \, \mathrm{d}\, x - \frac{1}{N} \sum_{i=0}^{N-1} f(x_i) \right| \le \|\Delta_P\|_p \|f'\|_q .$$

*Proof.* Observe that $f(x)$ and $t$ can be written as follows.

$$\begin{aligned} f(x) &= f(1) - \int_x^1 f'(t) \, \mathrm{d}\, t \\ &= f(1) - \int_0^1 \mathbb{1}_{[0,t]}(x) f'(t) \, \mathrm{d}\, t \\ t &= \int_0^1 \mathbb{1}_{[0,t]}(x) \, \mathrm{d}\, x \end{aligned}$$

We can rewrite the integral as follows.

$$\begin{aligned} \int_0^1 f(x) \, \mathrm{d}\, x &= \int_0^1 \left( f(1) - \int_0^1 \mathbb{1}_{[0,t]} f'(t) \, \mathrm{d}\, t \right) \mathrm{d}\, x \\ &= f(1) - \int_0^1 \int_0^1 \mathbb{1}_{[0,t]}(x) f'(t) \, \mathrm{d}\, t \, \mathrm{d}\, x \\ &= f(1) - \int_0^1 t f'(t) \, \mathrm{d}\, x. \end{aligned} \tag{5.1}$$

we can rewrite the average as follows.

$$\begin{aligned}\frac{1}{N}\sum_{i=0}^{N-1} f(x_i) &= \frac{1}{N}\sum_{i=0}^{N-1}\left(f(1) - \int_0^1 \mathbb{1}_{[0,t]}(x_i) f'(t)\ \mathrm{d}\,t\right)\\ &= f(1) - \int_0^1 \frac{1}{N}\sum_{i=0}^{N-1}\mathbb{1}_{[0,t]}(x_i) f'(t)\ \mathrm{d}\,t \qquad (5.2)\end{aligned}$$

The result follows by subtracting (8.24) from (8.25) and applying Hölder inequality. □

## 5.2 Collocation

## 5.3 Fast Multipole

## 5.4 Biological Applications

### 5.4.1 Efficient Computation of Molecular Energetics

### 5.4.2 PB and GB Energy Calculation

## Summary

## References and Further Reading

## Exercises

# Chapter 6

# Transforms

## 6.1 Radon Transform

and its inverse
discrete and continuous formulations

## 6.2 Fourier Transforms

and its inverse
discrete and continuous formulations
Fast Fourier Transforms (FFT) via wrapped convolutions

## 6.3 Fast Approximate Summations

using approximate FFT
(irregular / non-equispaced FFT)

## 6.4 Biological Applications

### 6.4.1 Fast Computation of Molecular Energetics

## Summary

## References and Further Reading

## Exercises

# Chapter 7

# Groups, Tilings, and Packings

## 7.1 Construction and Combinatorics of Families of *Almost-regular* Polyhedra

### 7.1.1 *TilingGen: an algorithm for generating all* almost-regular *polyhedra*

We present a simple algorithm for generating all *almost-regular* polyhedra (see Figure 7.1).

```
TILINGGEN($\mathcal{P}$, $\mathcal{L}$, $h$, $k$)
Constructs an almost-regular polyhedron using compatible mapping of polyhedron $\mathcal{P}$ onto lattice $\mathcal{L}$, such that the scaling and combinatorics are specified by $h$,$k$.

 $Q_\xi$1.  Assume that the lattice coordinate system is aligned with the Cartesian coordinate system such that the origings coincide and one of the axes is aligned to the X axis, and the other lies on the XY plane.
 $Q_\xi$2.  Place one point $A$ at the origin $(0, 0)$ of the lattice, a second point at $(h, k)$. Compute the other corners of the face $\mathcal{T}$. Note that we only need to know the number of vertices $n$ of $\mathcal{T}$.
 $Q_\xi$3.  Compute the location of the center $D$ of the face $\mathcal{T}$.
 $Q_\xi$4.  Let $\mathbb{T}_\mathbb{C}$ be the set of cyclic symmetry operations around $D$, such that $|\mathbb{T}_\mathbb{C}| = n$.
 $Q_\xi$5.  Initialize empty set $\mathbb{S}$
 $Q_\xi$6.  For each lattice point $p$ inside or on $\mathcal{T}$ do
 $Q_\xi$7.      Add $p$ to $\mathbb{S}$ if none of the transformations in $\mathbb{T}_\mathbb{C}$ applied to $p$ produces a point which is already in $\mathbb{S}$.
 $Q_\xi$8.  Compute the transformation $T_{map}$ which maps the face $\mathcal{T}$ to a face of the polyhedron $\mathcal{P}$. $T_{map}$ is composed of $T_{map_T} T_{map_S} T_{map_A}$ such that $T_{map_A}$ translates $\mathcal{T}$ along the lattice to take $D$ to the origin $O$, $T_{map_S}$ is a scaling that resizes $\mathcal{T}$ to the size of the faces in $\mathcal{P}$, then $T_{map_T}$ is a translation along Z-axis by an amount equal to the distance from the center of $\mathcal{P}$ to a face-center.
 $Q_\xi$9.  Let $\mathbb{T}_\mathbb{P}$ be the set of global symmetry operations (from the symmetry group of $\mathcal{P}$).
$Q_\xi$10.  Define a set of transformations $\mathbb{T}_{all} = \{T_2 T_{map} T_1 | T_2 \in \mathbb{T}_\mathbb{P} \& T_1 \in \mathbb{T}_\mathbb{S}\}$.
$Q_\xi$11.  All points on the almost-regular polyhedron is now generated by simply computing $\mathbb{T}_{all}(\mathbb{S})$.
```

Figure 7.1: TILINGGEN: Algorithm for constructing an *almost-regular* polyhedron using compatible mapping

### 7.1.2 Characterizing All Possible *Almost-Regular* Polyhedra and Completeness of TILINGGEN

Both Goldberg and Caspar-Klug constructions can be expressed as unfolding a regular polyhedron onto a 2D lattice and then refolding it with the lattice etched onto its faces. Pawley's wrapping idea is equivalent. We call this the *unfold-etch-refold* method. Here, we prove the conditions that must be satisfied to produce almost-regular polyhedra using the *unfold-etch-refold* mrthod for any regular solid, unfolded in any way, onto any 2D lattice.

Shepherd's conjecture [**?**] states that all convex polyhedra have a non self-overlapping planar unfolding with only edge-cuts. This conjecture is not proved or disproved yet for all possible convex polyhedra. However, for the set of special classes we are interested in, it is true. Hence, in principle it is possible to unfold one such polyhedra and lay it down on a 2D grid, use the grid to draw tiles of the unfolded polyhedron, and then fold it back up to get a tiled polyhdron. However, every polyhedron actually has many unfoldings. For example, icosahedron have 43380 unique unfoldings. Caspar and Klug's construction produced almost-regular polyhedra using 1 such unfolding, but it is not clear whether other unfoldings would also produce similar *almost-regular* polyhedra, or different types of *almost-regular* polyhedra, or not be *almost-regular*. To address this question, we have characterized [**?**] the relationship of the local and global symmetries of the almost-regular polyhedra, and the etched polyhedra (henceforth called *tiling*) produced using unfold-etch-refold construction. Here we only report the theoretical conclusion

**Theorem 7.1.** *The polyhedron generated by an* unfold-etch-refold *is* almost-regular *if and only if a compatible mapping of a regular polyhedron onto an* unfold-etch-refold *compatible lattice is performed.*

Please see [**?**] for detailed enumeration of possible compatible mapping of a regular polyhedron onto an *unfold-etch-refold* compatible lattice, and proof of the above theorem and associated lemmas.

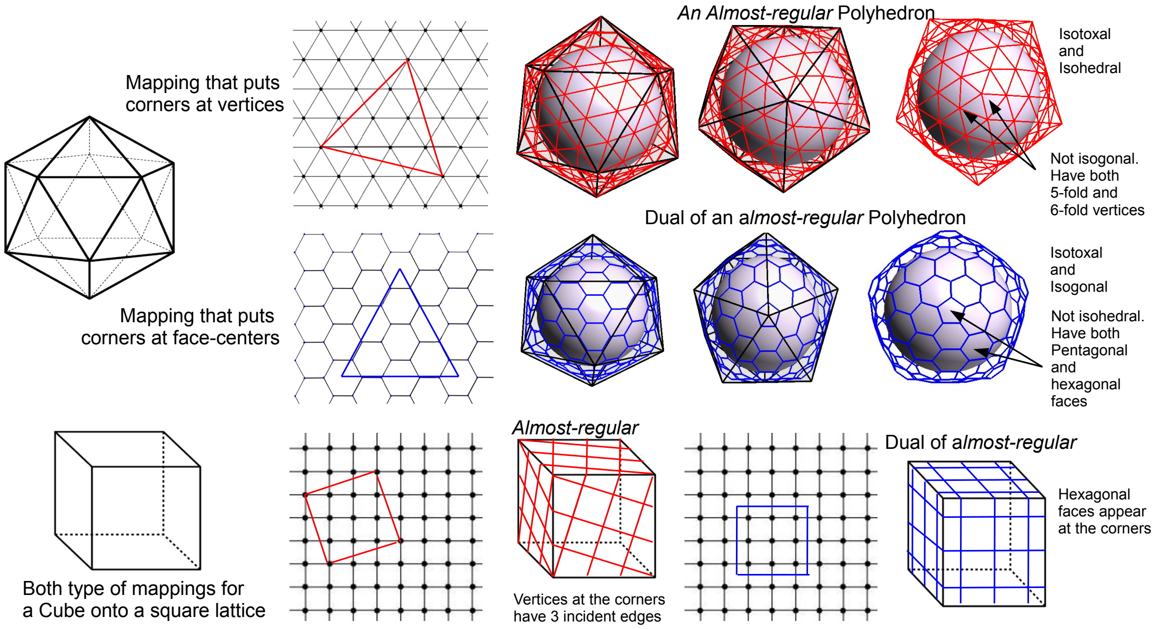

Mapping that puts corners at vertices
An *Almost-regular* Polyhedron
Isotoxal and Isohedral
Not isogonal. Have both 5-fold and 6-fold vertices
Dual of an *almost-regular* Polyhedron
Mapping that puts corners at face-centers
Isotoxal and Isogonal
Not isohedral. Have both Pentagonal and hexagonal faces
*Almost-regular*
Dual of *almost-regular*
Hexagonal faces appear at the corners
Both type of mappings for a Cube onto a square lattice
Vertices at the corners have 3 incident edges

Figure 7.2: **Illustration of the constructing *almost-regular* polyhedron and their duals**. Top row shows how placing corners of a polyhedral face on vertices of a compatible lattice produces an *almost-regular* polyhedron. The black lines show the original polyhedron, and the red lines show the etching/tiling induced by the lattice. The second row shows and example of placing the corners at face centers and producing duals of *almost-regular* polyhedron. Finally, the bottom row shows examples of both the primal and the dual constructions using a square lattice.

Figure 7.2 shows a few examples of constructing *almost-regular* polyhedron and their duals by compatible mapping of a face of a regular polyhedron on a *unfold-etch-refold* compatible lattice. Note that the etched-faces that cross an edge of $\mathcal{T}$ are geometrically not identical to the ones that do not. The ones crossing the boundary have a crease inside them, or if they are flattened, they are no longer regular. This is addressed in the Section **??**.

### 7.1.3 Parametrization and Geometric Aspects of TilingGen

The theoretical characterization of compatible mappings in the *unfold-etch-refold* protocol immediately lends itself to a simple parametrization of the *almost-regular* family. Furthermore, the symmetry at global and local levels lets us represent the geometry using a minimal set of points. Both of these insights are used in TilingGen.

For the sake of simplicity of presentation, the discussion in this section, in most cases, is focused solely on mapping icosahedron onto triangular lattices. Other compatible mappings can be discussed in the same manner with almost no difference in the theorems/lemmas presented here except for minor changes in counting. The choice to focus on the icosahedral case is primarily due to two reasons- first, it has the highest level of symmetry among the regular polyhedra which have a compatible mapping, and second, it has applications in modeling viruses, fullerenes etc.

### Parametrization

Let $\mathcal{L}$ be a lattice with origin $O$ and axes $H$ and $K$. Any point in the lattice is expressed using coordinates $(h, k)$ where both $h$ and $k$ are integers. Below we mention results (whose proofs can be found in [**?**]) which establish that a simple tuple $< \mathcal{P}, \mathcal{L}, h, k >$ is sufficient to represent the topology of a specific *almost-regular* polyhedron.

**Lemma 7.2.** *Assuming that one corner $A$ of the face $\mathcal{T}$ of the polyhedron is mapped to the origin $O$ of the lattice (or the nearest face-center for dual constructions). Then specifying the position of another compatibly placed point $B(h, k)$ is sufficient to parametrize the entire mapping.*

**Lemma 7.3.** *Topology of any* almost-regular *polyhedron or its dual can be expressed using a tuple $< \mathcal{P}, \mathcal{L}, h, k >$, where $\mathcal{P}$ is a regular polyhedron, $\mathcal{L}$ is a lattice represented using two axes, and $h$ and $k$ are integers.*

### Combinatorics and Symmetry

We consider the case where $\mathcal{P}$ is the icosahedron whose symmetry group will be denoted as $I$, and $\mathcal{L}$ is the triangular lattice which will be denoted as $\mathcal{L}^3$. Now, we discuss some properties of the lattice.

**Definition 7.4.** We define each triangle of the lattice $\mathcal{L}^3$ as a small triangle and use $t$ to denote such a triangle. Let us define a triple $< i, j, k >$ where $i$ and $j$ are integers and $k \in \{+, -\}$. Let the triangle produced by the intersections of the lines $h = i$, $k = j$ and $h + k = i + j + 1$ (having the vertices $(i, j), (i + 1, j)$ and $(i, j + 1)$) be denoted $t_{ij+}$. Similarly, the triangle denoted $t_{ij-}$ has vertices $(i, j), (i + 1, j - 1)$ and $(i + 1, j)$, and is produced by the intersections of the lines $h = i + 1$, $k = j$ and $h + k = i + j$.

The proof of the following lemma is immediate from this definition.

**Lemma 7.5.** *$t_{i_1 j_1 k_1}$ coincide with $t_{i_2 j_2 k_2}$ if and only if $i_1 = i_2$, $j_1 = j_2$ and $k_1 = k_2$. For any small triangle in $\mathcal{L}^3$, there exists a triple $< i, j, k >$ such that $t_{ijk}$ represents that small triangle.*

Through etching, the triangular lattice $\mathcal{L}^3$ produces a tiling of a face $\mathcal{T}$ (which will be called a large triangle in this section) of $\mathcal{P}$ where each tile is a small triangle. Now we consider some properties of this tiling.
Assuming $A$ is at $(0, 0)$, $B$ is $(h, k)$ such that $h$ and $k$ are integers, the tiling produced by $\mathcal{L}^3$ on $\mathcal{T}$ satisfies:

- The area of $\mathcal{T}$ is $\frac{\sqrt{3}}{4}(h^2 + hk + k^2)$, which is equal to the area of $h^2 + hk + k^2$ small triangles.
- In addition to $A, B$ and $C$, $\mathcal{T}$ includes exactly $\frac{h^2+hk+k^2-1}{2}$ more vertices of $\mathcal{L}^3$. Note that any vertex that lie on an edge of $\mathcal{T}$ is counted as half a vertex.
- Each edge of $\mathcal{T}$ is intersected by at most $2(h + k) - 3$ lines of the form $h = c$, $k = d$ and $h + k = e$, where $c, d$ and $e$ are integers.
- The number of small triangles intersected by any edge of $\mathcal{T}$ is at most $2(h + k - 1)$.

The following are some combinatorial properties of the overall tiled polyhedron-

- There are exactly $20(h^2 + hk + k^2)$ small triangles, and the same number of local 3-fold axes.
- The 12 gf-symmetry axes are surrounded by 5 small triangles.
- There are exactly $10(h^2 + hk + k^2 - 1)$ vertices (not lying on the gf-axes) with 6-fold local symmetry.

Similar properties can easily be derived for other mappings as well. The important point to note is that not only the topology, but also the symmetry and combinatorics are also parameterized by only $h$ and $k$.

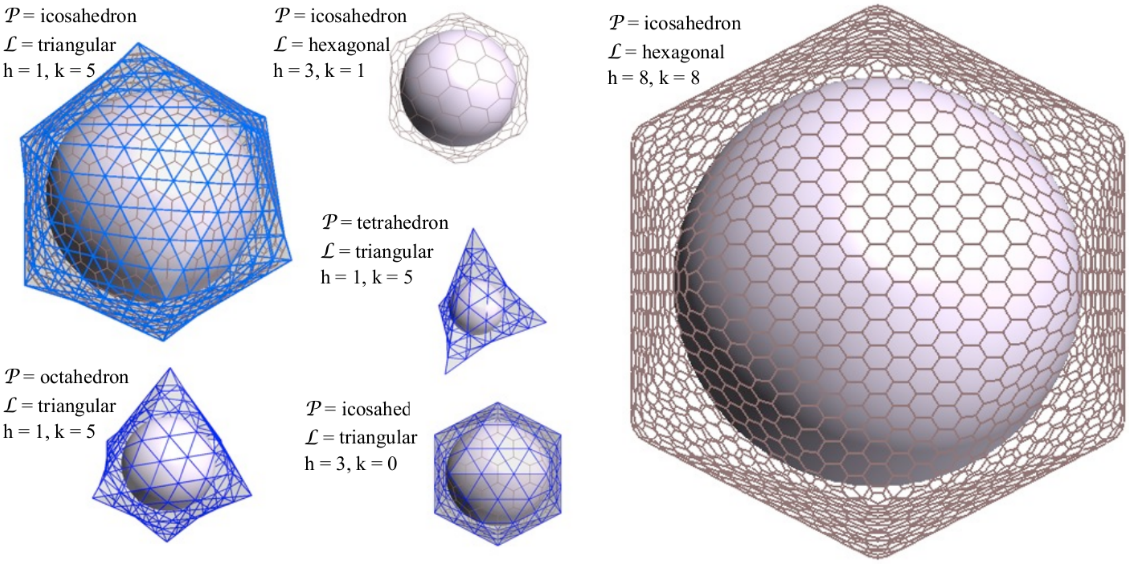

$\mathcal{P}$ = icosahedron
$\mathcal{L}$ = triangular
h = 1, k = 5
$\mathcal{P}$ = icosahedron
$\mathcal{L}$ = hexagonal
h = 3, k = 1
$\mathcal{P}$ = icosahedron
$\mathcal{L}$ = hexagonal
h = 8, k = 8
$\mathcal{P}$ = tetrahedron
$\mathcal{L}$ = triangular
h = 1, k = 5
$\mathcal{P}$ = octahedron
$\mathcal{L}$ = triangular
h = 1, k = 5
$\mathcal{P}$ = icosahed
$\mathcal{L}$ = triangular
h = 3, k = 0

Figure 7.3: **Some polyhedron generated by applying** TilingGen.

**Topology to Geometry**

Note that given a point $P$ with coordinate $(i, j)$ inside $\mathcal{T}$, there exists two other points $Q$ and $R$ such that $P$, $Q$ and $R$ are 3-fold symmetric around the center $D$ of $\mathcal{T}$. The two points $Q$ and $R$ have coordinates $(h - i - j, k + i)$ and $(h + k + j, -h - i)$ respectively. This can be seen by noticing that stepping along the $H$ and $K$ axis by $i$ and $j$ units from $A(0, 0)$ is $C^3$ symmetric (around $D$) to stepping in $-H + K$ and $-H$ directions by the same units from $B(h, k)$, and stepping in $-K$ and $H - K$ directions by the same units from $C$. We can further extend it to triangles and deduce the following.

**Lemma 7.6.** *If $A(h_1, k_1)$, $B(h_2, k_2)$ and $C(h_3, k_3)$ are three points in the $HK$ coordinate system such that $h_1, h_2, h_3, k_1, k_2, k_3$ are integers and $ABC$ is an equilateral triangle whose centroid is $O$, then the small triangles $t_{h_1+i,k_1+j,\pm}$, $t_{h_2-i-j-1,k_2+i,\pm}$ and $t_{h_3+j,k_3-i-j-1,\pm}$ are $C^3$ symmetric around $O$.*

Now, we define the minimal set of points or non-redundant set of points $\mathbb{S}$ such that no two points $s_i, s_j \in S$ are $C^3$ symmetric to each other around $D$, and all points in $\mathbb{S}$ lie inside or on $\mathcal{T}$. Clearly, $|\mathbb{S}| = \lceil \frac{h^2+hk+k^2}{3} \rceil$. Note that applying $C^3$ operations on $\mathbb{S}$ produces all points inside and on $\mathcal{T}$.

We conclude with the following theorem, whose proof is in [**?**]

**Theorem 7.7.** *The algorithm* TilingGen *constructs a minimal geometric representation of the* almost-regular *polyhedron in terms a set of points $\mathbb{S}$ embedded onto the XY plane and a set of 3D transformations $\mathbb{T}_{all}$.*

## 7.2 Crystal Symmetries

### 7.2.1 Symmetries

### 7.2.2 Quasi-symmetries

## 7.3 Hexagonal Tilings

### 7.3.1 Caspar-Klug coordinate system

### 7.3.2 T-numbers

### 7.3.3 P-numbers

## 7.4 Icosahedral Packings

## 7.5 Biological Applications

### 7.5.1 Crystal Structures

### 7.5.2 Viral Capsid Symmetry Detection and Classification

### 7.5.3 Characterization of Large Deformations in Molecules

## Summary

## References and Further Reading

## Exercises

# Chapter 8

# Motion Groups, Sampling

## 8.1 Rotation Group

## 8.2 Fourier Transforms

Solving optimization problems for protein structure interpretation reduces to a correlation based scoring and search over a space of relative transformations. The objective function in general is highly non-convex, possessing several local maxima and minima. The vast number of existing solutions to the rigid-body correlation problem can be distinguished by a few basic approaches.

Feature-based methods compute and correlate reduced representations of proteins. An early example of a feature-based approach is the method of vector quantization [227], in which sets of vectors are used to represent molecules. A similar approach is geometric hashing [139], whereby critical features are hashed into a table of values, and a score—related to the correlation score—measures the match between the participating proteins for a particular relative orientation. Feature-based approaches, used in docking [186] and fitting [224], result in improved performance due to the reduced search space, at the possible expense of poor resolution scaling.

Iterative approaches vary in sophistication, ranging from a simple version of steepest ascent [163] to more powerful techniques such as Powell optimization [225]. Most such approaches result in locally optimal solutions that, depending on the initial guess, may or may not be close to the globally optimal correlation. They are thus usually used in conjunction with an exhaustive approach that provides the requisite initial guess.

Exhaustive or Fourier-based approaches exploit the fact that it is beneficial if the computation of the objective function can be done relatively fast or if the search space is restricted. In these approaches the proteins are treated as rigid bodies. The automated search of all possible motions, i.e., translations and rotations to maximize the overlap between both structures is the main task of these programs. This is done by evaluating a correlation integral with respect to the motions using Fast Fourier transforms. Fourier based methods combine an accurate and exhaustive search with reduced computational cost, and have thus proven quite popular as a search scheme, see, e.g. [114, 176, 18, 52, 149, 164, 226, 236]. Many, if not most, existing Fourier-based docking algorithms use a regular discrete three-dimensional cartesian grid onto which the molecules are projected. The correlation score of these discretised and suitably weighted structures serves as the objective function for the optimization problem. The correlation score between pairs of grid cells is computed via fast Fourier transforms, thus implicitly searching over the three dimensional space of translations. The remaining rotational degrees of freedom however need to be incorporated into a global search. Such an approach has been first published by [129] in 1992. Since then, this approach has been adapted and improved many times. An overview of these translational grid-based FFT search schemes can be found in [75]. In recent approaches, the equispaced grid has been replaced with a non-equispaced Cartesian one, as in [18], or a polar one, as in [93, 177, 205, 77]. Fast translational matching exploits the fact that for each rotation, the objective function is a correlation integral, and can be computed by fast Fourier transforms. On the other hand, the space of rotations is still subject to exhaustive search.

While the optimal matching solution exists in a highly localized region of relative translations, the range of relative rotations varies widely for each translation. The disparity between size and sampling density of translational and rotational search spaces motivates this work. We present a fast rotational correlation matching that extends the methods of [177] and [205], which use spherical harmonic functions and classical orthogonal polynomials to model molecular shapes. Rotational speedups depend on representing the protein structures in a basis more amenable to rotational sampling. In Kovacs and Wriggers [136], Kovacs et. al [134], and Garccon et. al [90], that basis is the basis of functions on the unit sphere $\mathbb{S}^2$, i.e., the family of spherical harmonic functions $Y_l^m(\theta, \phi)$, whereas in the work of Ritchie [175, 177], a radial basis function $R_k^l(r)$ related to the Gaussian accompanies the spherical basis. Like their translational counterparts, rotational speedups compute a multiple exponential sum, or an FFT; unlike translational speedups, the FFT is computed on a uniformly spaced grid of z-y-z Euler angles.

We employ algorithms to compute the fast Fourier transform on the rotation group to solve the matching problem. In this work, instead of correlating functions defined on the unit cube, we use functions defined on $\mathbb{R}^3$ but split into $\mathbb{R}^+ \times \mathbb{S}^2$. We exploit the fact that correlations of functions defined on $\mathbb{S}^2$ can be computed by means of Fourier transforms on the rotation group. This enables efficient computation of the objective function over rotational degrees of freedom instead of translational degrees of freedom.

A three-dimensional translation can be expressed as a translation along the z-axis followed by two rotations, one about the y-axis and one about the z-axis. Hence, it has two rotational degrees of freedom and one translational. Combining this in a motion, we have five rotation angles that describe a motion and one absolute value of a translation along one axis. If we are able to speed up the computation for the rotations by correlating functions on the sphere, we get an improved complexity for five of the six degrees of freedom instead of the previous three. This approach has been suggested in [135] for protein fitting and can also be found in [77].

The essential mathematical tool used in this work for protein fitting is the fast calculation of the discrete Fourier transform on the rotation group $\mathrm{SO}(3)$. An implementation of such an algorithm can be found in [167]. For completeness, we also mention the related work of [113] and [51] . The paper of [113] uses representation theory of the rotation group for approaching optimization problems in the cryoEM setting, as we do. However the nature of our optimization problem is fundamentally different from theirs. The paper of [51] develops discrete Fourier transforms on the motion group $SE(3)$, and applies it to topics ranging from workspace density of robotic manipulators to conformational statistics of macromolecules. On the other hand, we use fast discrete Fourier transform on the rotation group $\mathrm{SO}(3)$ to provide an efficient solution to our optimization problem.

Current exhaustive techniques suffer from two main drawbacks. The first drawback relates to local refinement. Depending as they do on the equispaced FFT, exhaustive techniques cannot be gracefully used to refine existing solutions. Say we wish to improve a matching pose, obtained using a translational FFT speedup with a certain grid size. If we redo the experiment with half the grid length of the previous computation, the three dimensional FFT becomes eight times as expensive, but more importantly, it spends much of its time at points on the new grid already excluded by the initial experiment. A similar argument applies to rotational speedups; in both these approaches, the concept of a local refinement is largely absent.

A second drawback relates to the question of uniform sampling in rotational space. While sampling in translational space is straightforward, involving Cartesian grids with uniform, possibly differing grid-sizes in each independent direction, the notions of uniformity and direction do not translate easily to the rotational space $\mathrm{SO}(3)$. In particular, equispaced Euler angular grids do not result in equispaced $\mathrm{SO}(3)$ samples. Due to this, rotational FFT-based techniques are destined to oversample certain regions of $\mathrm{SO}(3)$ while leaving others wholly unexamined.

### 8.2.1 Proteins and flexibility

One main point of criticism of rigid-body fitting methods is that proteins undergo conformational changes during the induced fit, i.e., they not only move with respect to each other but also deform, shear or bend. Flexibility often involves movements between large rigid parts of the protein, called domains, between flexible loops on the molecular surface and between large side chain at active sites. Due to the vastness of the space of flexible motions, protein flexibility can be practically dealt with by (A) conducting all-atomistic local searches, as in the case of molecular dynamical algorithms [128, 142, 200, 195, 118], (B) Building a coarse-grained representation of the protein, also known as a domain decomposition [87, 3, 166, 190], or (C) A combination of the strategies in (A) and (B) [210, 211, 241].

Domain-based approaches have so far lacked a search scheme that takes advantage of the translational or rotational speedups that FFT-based approaches can afford. This has to do with the issue of focusing: in uniform FFT-based techniques, there is no way to restrict the search space to a small area of interest that can be occupied by a single domain rather than the entire protein. By contrast, searching over the entire space for each domain is both time-consuming and results in spurious and geometrically implausible false positives, and sifting through these grows rapidly inefficient as the number of domains increases. This is also why domain-based flexibility algorithms such as those in [209, 210, 211] prefer Monte-Carlo-based or steepest-ascent-based search schemes.

### 8.2.2 Our contributions

We address the drawbacks mentioned in Section 8.2 with a pair of rotationally exhaustive, non-equispaced techniques to compute rigid-body correlations. The resulting family of techniques, has the following properties:

- *Sampling robust.* The technique is capable of efficiently computing correlations over arbitrary samples of rigid body motions $\mathbb{R}^3 \times \mathrm{SO}(3)$.
- *Compatible.* It can be used along with existing equispaced FFT-based techniques.
- *General.* It unifies the rotationally-exhaustive paradigms in [136, 175, 90, 177].

Finally, this work also aims to be a self-contained overview of correlation techniques that depend on expressing the input scalar valued functions in terms of rotationally invariant bases. In particular, we prove all relevant properties inherent to our mathematical framework.

## 8.3 Background

In this Section we give some necessary definitions and background information for these algorithms. Note that we defer most multi-line proofs to the appendix.
Let $A, B : \mathbb{R}^3 \mapsto \mathbb{C}$ be a pair of scalar-valued functions. We define the rigid-body correlation problem as follows.

**Definition 8.1.** For two functions $A : \mathbb{R}^3 \mapsto \mathbb{C}$ and $B : \mathbb{R}^3 \mapsto \mathbb{C}$ we define

$$C(\mathbf{R}_i, \mathbf{t}_j) = \int_{\mathbb{R}^3} A(\mathbf{x})B(\mathbf{R}_i\mathbf{x} + \mathbf{t}_j)d\mathbf{x}, \qquad i \in \{1, \ldots, N_{rot}\}, j \in \{1, \ldots, N_{trans}\} \tag{8.1}$$

as the rigid-body correlation between $A$ and $B$ for a given set $S = \{(\mathbf{R}_i, \mathbf{t}_j)\}, R_i \in \mathrm{SO}(3), \mathbf{t}_j \in \mathbb{R}^3$ of rigid-body motions. The rigid-body correlation problem is to maximize $C(\mathbf{R}_i, \mathbf{t}_j)$ over the set $S$.

The rigid-body correlation problem is a non-convex geometric optimization problem. The several problem domains in computational biology to which it applies can be distinguished by their choice of $A$ and $B$. In protein-protein docking, for instance, $A$ and $B$ are affinity functions that represent a relevant property, such as shape or electrostatics, of the underlying protein; in protein-density map fitting, $A$ is a blurred representation of the atoms of the protein, while $B$ is the density map itself.
The objective function (8.1) can be efficiently calculated by expanding it in a series of orthogonal basis functions. The starting point of this approach is a coordinate transform of vectors $\overline{z}x \in \mathbb{R}^3$ from Cartesian to spherical coordinates. The inner product of two square-integrable functions $f, g : \mathbb{R}^3 \to \mathbb{C}$ parameterized in spherical coordinates is given by

$$\langle f, g \rangle \;\; = \;\; \int_{\mathbb{R}^+} \int_{\mathbb{S}^2} f(r\mathbf{u})\overline{g(r\mathbf{u})}r^2 \mathrm{d}\mathbf{u}\mathrm{d}r. \tag{8.2}$$

We now consider the orthogonal bases for the two components of the product space separately. Let $\xi \in \mathbb{S}^2$ and let $(\varphi, \theta) \in [0, 2\pi) \times [0, \pi]$ be its coordinates. For any $l \in \mathbb{N}_0$ and $m = -l, \ldots l$ the spherical harmonics of degree $l$ are defined as

$$Y_l^m(\xi) = \sqrt{\frac{2l+1}{4\pi}} P_l^{|m|}(\cos\theta)\mathrm{e}^{\mathrm{i}m\phi}$$

where $P_l^m : [-1, 1] \to \mathbb{R}$ are associated Legendre polynomials, cf. [206], that arise as the derivatives of ordinary Legendre polynomials $P_l(x)$.
The spherical harmonics satisfy the orthogonality relation

$$\int_{\mathbb{S}^2} Y_l^m(\xi)\overline{Y_{l'}^{m'}(\xi)}\,\mathrm{d}\xi = \delta_{ll'}\delta_{mm'}. \tag{8.3}$$

Secondly, we employ a weighted version of the Laguerre polynomials denoted by $R_k^l(r)$. These functions have been used to describe the radial part of the orbitals of hydrogenic atoms and are also known as radial wavefunctions, see [14, pp. 368 ff] for general informations. In [175] these functions have been employed in the context of six-dimensional rigid-body docking.

**Definition 8.2.** For $r \in \mathbb{R}_0^+,\ l,k \in \mathbb{N}_0,\ k > l$, the weighted Laguerre polynomials $R_k^l : \mathbb{R}^+ \to \mathbb{R}$ are given by

$$R_k^l(r) = \sqrt{\frac{2(k-l-1)!}{\Gamma(k+\frac{1}{2})}}\mathrm{e}^{-\frac{r^2}{2}} r^l L_{k-l-1}^{l+\frac{1}{2}}\left(r^2\right)$$

using the Laguerre polynomials $L_k^l$, see [206].

For $r \in \mathbb{R}_0^+,\ l,k \in \mathbb{N}_0,\ k > l$, the functions $R_k^l(r)$ satisfy

$$\int_0^\infty R_k^l(r) R_n^l(r) r^2 \mathrm{d}r = \delta_{k,n}. \tag{8.4}$$

Based on the previous orthogonality relations (8.3) and (8.4), we see that the functions $R_k^l(r) Y_l^m(\mathbf{u})$ for $k,l \in \mathbb{N},\ k > l \geq |m|$ are orthonormal with respect to the inner product from (8.2). This follows immediately by

$$\begin{aligned}\langle R_k^l(r) Y_l^m(\mathbf{u}), R_{k'}^{l'}(r) Y_{l'}^{m'}(\mathbf{u})\rangle &= \int_0^\infty R_k^l(r) R_{k'}^{l'}(r) r^2 \mathrm{d}r \int_{\mathbb{S}^2} Y_l^m(\mathbf{u}) \overline{Y_{l'}^{m'}(\mathbf{u})} \mathrm{d}\mathbf{u} \\ &= \delta_{k,k'} \delta_{l,l'} \delta_{m,m'}. \end{aligned} \tag{8.5}$$

Moreover, these products of functions constitute an orthogonal basis of the space of square-integrable functions on $\mathbb{R}^3$. Therefore, we find a unique series expansion of the two given functions $A(\overline{z}x)$ and $B(\overline{z}x)$ in terms of these functions as

$$A(\overline{z}x) = A(r\mathbf{u}) = \sum_{k=1}^{\infty} \sum_{l=0}^{k-1} \sum_{m=-l}^{l} \hat{a}_{klm} R_k^l(r) Y_l^m(\mathbf{u}) \tag{8.6}$$

with coefficients

$$\hat{a}_{klm} = \int_0^\infty \int_{\mathbb{S}^2} A(r\mathbf{u}) \overline{R_k^l(r) Y_l^m(\mathbf{u})} r^2 \mathrm{d}\mathbf{u} \mathrm{d}r, \tag{8.7}$$

and analogously for $B(\overline{z}x)$.

Typically the initial data for $A(\overline{z}x)$ and $B(\overline{z}x)$ will be obtained by an EM or read in from a database as an atomic structure in terms of a collection of atoms and charges. Either way the methods only provide a finite number of samples of the unknown functions $A$ and $B$. Hence the integral (8.7) will be approximated by a suitable quadrature rule. In PF*corr* we use a combination of the Clenshaw-Curtis formula for the spherical part cf. [62, pp. 86] and a Gauss-Legendre formula for the radial part cf. [62, pp. 222]. Alternatives to such deterministically-sampled quadrature schemes are quasi Monte-Carlo methods or Monte-Carlo methods. Since this is not the focus of this work we omit further details on quadrature and merely comment on the error induced by step.

**Lemma 8.3.** *Let $A : \mathbb{R}^3 \mapsto \mathbb{C}$ be a complex scalar-valued, 2-Lipschitz continuous function with finite support on the domain $\Omega \subset \mathbb{R}^3$. For a given spherical grid with maximum grid-diameter $h$*[1]*, for small $h$ the coefficients $\hat{a}_{klm}$ can be computed with an error $E = Ch|\Omega|$ for a constant $C \in \mathbb{R}$.*

### 8.3.1 Multi-basis framework

As a first step in solving the rigid-body correlation problem in Equation 8.1, $A$ and $B$ are represented in terms of orthogonal basis functions. PF*corr* offers two distinct choices of how to proceed.

$\mathrm{Q}_\xi 1$. **Mixed bases.** The standard approach uses the expansion (8.6) and approximates it by

$$A_L(\overline{z}x) = A_L(r\mathbf{u}) = \sum_{k=1}^{L} \sum_{l=0}^{k-1} \sum_{m=-l}^{l} \hat{a}_{klm} R_k^l(r) Y_l^m(\mathbf{u}), \tag{8.8}$$

and analogously for $B_L(\overline{z}x)$. For convenience, we shall omit the subscript $L$ from now on.

[1] The grid-diameter is the diameter of the smallest ball that the grid-cell can be enclosed in.

Q$_\xi$2. **Pure spherical basis.** This slightly modified approach following [136] and [90], divides the three dimensional space into discrete spherical slices and uses only a spherical basis expansion in terms of spherical harmonics on each slice with fixed radius $r$. The pure spherical representation of a scalar valued function $A : \mathbb{R}^3 \mapsto \mathbb{C}$ for a given radial coordinate $r$ is given by

$$A_r(\mathbf{u}) = \lim_{L\to\infty} \sum_{l=0}^{L} \sum_{m=-l}^{l} \hat{a}_{lm}(r) Y_l^m(\mathbf{u}) \tag{8.9}$$

with coefficients

$$\hat{a}_{lm}(r) = \int_{\mathbb{S}^2} A_r(\mathbf{u}) \overline{Y_l^m(\mathbf{u})} \mathrm{d}\mathbf{u}, \tag{8.10}$$

where $A_r(\mathbf{u}) = A(r, \mathbf{u})$.

The next step in solving the rigid-body correlation problem from Definition 8.1 involves applying a motion to the functions $A$ and $B$. We assume that $A$ and $B$ are rigid bodies, and restrict the motion to rotations and translations in three-dimensional space.

### 8.3.2 Rotating basis expansions of scalar-valued functions

We shall now examine how a function expanded as in (8.8) behaves under the application of a rotation. We conveniently employ the representation property of spherical harmonics stating for arbitrary rotations $\overline{z}R \in \mathrm{SO}(3)$:

$$Y_l^n(\overline{z}R^T\overline{z}u) = \sum_{m=-l}^{l} Y_l^m(\overline{z}u) D_l^{mn}(\overline{z}R), \qquad \text{for } |m| \leq l, \overline{z}u \in \mathbb{S}^2. \tag{8.11}$$

where $D_l^{mn}(\mathbf{R})$ is a Wigner-D function [222].
The Wigner-$D$ functions $D_l^{m,n}$ with degree $l$ and orders $m, n$ with $\max\{|m|, |n|\} \leq l$ are given by the explicit expression

$$D_l^{m,n}(\alpha, \beta, \gamma) = \mathrm{e}^{-\mathrm{i}m\alpha}\, d_l^{m,n}(\cos\beta)\, \mathrm{e}^{-\mathrm{i}n\gamma}$$

where $\alpha, \gamma \in [0, 2\pi)$ and $\beta \in [0, \pi]$ are the Euler angle decomposition of a rotation $\mathbf{R} \in \mathrm{SO}(3)$ and $d_l^{m,n}$ are the Wigner-$d$ functions

$$d_l^{m,n}(x) = \varepsilon \left( \frac{s!(s+\mu+\nu)!}{(s+\mu)!(s+\nu)!} \right)^{1/2} 2^{-\frac{\mu+\nu}{2}} (1-x)^{\frac{\mu}{2}} (1+x)^{\frac{\nu}{2}} P_{l-L_*}^{(\mu,\nu)}(x), \tag{8.12}$$

$P_{l-L_*}^{(\mu,\nu)}(x)$ are the Jacobi polynomials and

$$\begin{aligned} \mu &= |n-m|, & \nu &= |n+m|, \\ L_* &= \max\{|m|, |n|\}, & s &= l - L_*, \end{aligned} \qquad \varepsilon = \begin{cases} 1, & \text{if } m > n, \\ (-1)^{n-m}, & \text{if } m \leq n. \end{cases}$$

Note that $d_l^{m,n}$ is a polynomial of degree $l$ if $m+n$ is even. Otherwise, it is a polynomial of degree $l-1$ times a factor of $(1-x^2)^{1/2}$.
By virtue of (8.11), applying an arbitrary rotation $\overline{z}R \in \mathrm{SO}(3)$ to the given function $A(\overline{z}x)$ will yield

$$A(\overline{z}R^T\overline{z}x) = A(r\overline{z}R^T\overline{z}u) \quad = \quad \sum_{k=1}^{\infty} \sum_{l=0}^{k-1} \sum_{m,n=-l}^{l} \hat{a}_{kln} D_l^{mn}(\mathbf{R}) R_k^l(r) Y_l^m(\mathbf{u}).$$

Note that the rotation does not affect the radial parts of the function as a rotation preserves distance. Hence, a similar result holds for the radial-basis independent coefficients $\hat{a}_{lm}$.

**Lemma 8.4.** *Given two functions $A : \mathbb{R}^3 \mapsto \mathbb{C}$ and $B : \mathbb{R}^3 \mapsto \mathbb{C}$ expanded in terms of a mixed basis as given in* (8.8) *the pure rotational correlation can be obtained by evaluating*

$$C(\mathbf{R}) = \sum_{k=1}^{L} \sum_{l=0}^{k-1} \sum_{m=-l}^{l} \sum_{m'=-l}^{l} (-1)^m \hat{a}_{kl-m} \hat{b}_{klm'} D_l^{m,m'}(\mathbf{R}) \tag{8.13}$$

*for arbitrary choices of $\overline{z}R \in \mathrm{SO}(3)$.*

This is a direct result from using the orthogonality property (8.5) with the basis expansions of $A(\overline{z}x)$ and $B(\overline{z}R^T\overline{z}x)$ in

$$C(\mathbf{R}) = \int_{\mathbb{R}\times\mathbb{S}^2} \sum_{klm} \hat{a}_{klm} R_k^l(r) Y_l^m(\mathbf{u}) \sum_{k'l'm'm''} \overline{\hat{b}_{k'l'm'}} R_{k'}^{l'}(r) \overline{D_l^{m'',m'}(\mathbf{R}) Y_{l'}^{m''}(\mathbf{u})} r^2 dr d\mathbf{u}.$$

**Lemma 8.5.** *Given two functions $A : \mathbb{R}^3 \mapsto \mathbb{C}$ and $B : \mathbb{R}^3 \mapsto \mathbb{C}$ expanded in terms of a pure spherical basis as given in* (8.9) *the pure rotational correlation can be obtained by evaluating*

$$C(\mathbf{R}) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \sum_{m'=-l}^{l} (-1)^m (-1)^{m'} D_l^{m,m'}(\mathbf{R}) \int_{\mathbb{R}^+} \hat{a}_{l-m}(r) \overline{\hat{b}_{l-m'}}(r) r^2 dr \tag{8.14}$$

*for arbitrary choices of $\overline{z}R \in \mathrm{SO}(3)$.*

### 8.3.3 Fourier Transforms on the rotation group $\mathrm{SO}(3)$

To efficiently calculated the correlations (8.13), (8.14) , we will use the Fast SO(3) Fourier Transform. For details on the algorithm we refer the reader to [167]. Here we simply outline the basic idea and show how it can be applied to compute our scoring function.
The space of square integrable functions in SO(3) is denoted $\mathrm{L}^2\,(\mathrm{SO}(3))$ and defined via the standard inner product

$$\langle f, g \rangle = \int_0^{2\pi} \int_0^{\pi} \int_0^{2\pi} f(\alpha, \beta, \gamma) \overline{g(\alpha, \beta, \gamma)} \sin\beta \,\mathrm{d}\gamma \,\mathrm{d}\beta \,\mathrm{d}\alpha.$$

A convenient orthogonal basis for $\mathrm{L}^2(\mathrm{SO}(3))$ are the Wigner-D functions $D_l^{m,n}(\overline{z}R)$ which satisfy the orthogonality condition

$$\langle D_l^{m,n}, D_{l'}^{m',n'} \rangle = \frac{8\pi^2}{2l+1} \delta_{l,l'} \delta_{m,m'} \delta_{n,n'}.$$

**Definition 8.6** (NDSOFT)**.** The nonequispaced discrete SO(3) Fourier transform (NDSOFT) is defined as the evaluation of the sums

$$f(\alpha_q, \beta_q, \gamma_q) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \sum_{n=-l}^{l} \hat{f}_l^{m,n} \tilde{D}_l^{m,n}(\alpha_q, \beta_q, \gamma_q), \qquad q = 1, 2, \ldots, Q, \tag{8.15}$$

for given Fourier coefficients $\hat{f}_l^{m,n}$ and nodes $(\alpha_q, \beta_q, \gamma_q)$.

We outline our strategy for the fast approximate algorithm, a detailed description of this algorithm, called the nonequispaced fast Fourier transform (NFSOFT) can be found in [167]. We can rearrange (8.15) to

$$f(\alpha, \beta, \gamma) = \sum_{m=-L}^{L} \sum_{n=-L}^{L} \mathrm{e}^{-\mathrm{i}m\alpha}\, \mathrm{e}^{-\mathrm{i}n\gamma} \sum_{l=L_*}^{L} \hat{f}_l^{m,n} d_l^{m,n}(\cos\beta).$$

We can then calculate new coefficients $\bar{f}_l^{m,n}$ from the coefficients $\hat{f}_l^{m,n}$ in $\mathcal{O}(L^3 \log^2 L)$ arithmetic operations to rewrite the inner most sum for $m, n = -L, \ldots, L$ using the Chebyshev polynomials of first kind $T_l(x)$,

$$\sum_{l=L_*}^{L} \hat{f}_l^{m,n} d_l^{m,n}(\cos\beta) = \sum_{l=0}^{L-\chi} \tilde{f}_l^{m,n} (1-x^2)^{\chi/2} T_l(x), \tag{8.16}$$

where $\chi = [m + n \text{ odd}]$. We are now able to replace the Chebyshev polynomials of first kind with complex exponentials,

$$\sum_{l=0}^{L-\chi} \tilde{f}_l^{m,n} (1-x^2)^{\chi/2} T_l(x) = \sum_{l=-L}^{L} \hat{g}_l^{m,n} \mathrm{e}^{-\mathrm{i}l\beta}, \qquad m, n = -L, \ldots, L.$$

We can compute the coefficients $\hat{g}_l^{m,n}$ from the coefficients $\bar{f}_l^{m,n}$ with $\mathcal{O}(L^3)$ arithmetic operations. The obtained form is now ready to be inserted into (8.15) to become

$$f(\alpha,\beta,\gamma) = \sum_{l=-L}^{L} \sum_{m=-L}^{L} \sum_{n=-L}^{L} \hat{g}_l^{m,n} \mathrm{e}^{-\mathrm{i}m\alpha} \mathrm{e}^{-\mathrm{i}l\beta} \mathrm{e}^{-\mathrm{i}n\gamma}. \tag{8.17}$$

This is a plain three-dimensional Fourier sum and we can use the NFFT algorithm to evaluate it with $\mathcal{O}(L^3 \log L + Q)$ operations, where $Q$ is the number of nodes at which we evaluate the function; see [168]. Hence, the application of a NFSOFT results in $\mathcal{O}(L^3 \log^2 L + Q)$ operations.

## 8.4 Rigid-body correlations

Although not immediately apparent, the idea of exploiting the rotational invariance of the spherical harmonics that serve as basis functions in the Fourier expansion of a functions in $L^2(\mathbb{S}^2)$ has some advantages over translation-invariant Fourier expansion in [18, 52] .

The key idea is to first express the three-dimensional translation in terms of two rotations and a translation in one dimension. Hence, this translation will have two rotational degrees of freedom and one translational. A three dimensional translation $\overline{z}t \in \mathbb{R}^3$ of a object can be uniquely expressed as $\overline{z}t = r\,\overline{z}R_Z(\varphi)\overline{z}R_Y(\theta)\overline{z}e_z$ for $\varphi \in [0, 2\pi), \theta \in [0, \pi]$ and $r \in \mathbb{R}^+$ where $\overline{z}e_z = (0, 0, 1)^T$. Combining, this with the three independent rotation parameters of the object, we have five rotation angles that describe a motion and one absolute value of a translation along one axis. Consequently, are able to speed up the computation for the rotations by spherical Fourier transforms and obtain an improved complexity for five of the six degrees of freedom of rigid-body correlations instead of the previous three. For $\overline{z}U = \overline{z}R_Z(\varphi)\overline{z}R_Y(\theta)$, we have

$$\begin{aligned} C(\overline{z}R, \overline{z}t) = C(\tilde{\overline{z}}R, \overline{z}U z\overline{z}e_z) &= \int_{\mathbb{R}^3} A(\overline{z}x) B(\overline{z}R\overline{z}x - \overline{z}U z\overline{z}e_z)\,\mathrm{d}\overline{z}x \\ &= \int_{\mathbb{R}^3} A(\overline{z}U^T\overline{z}x) B(\tilde{\overline{z}}R\overline{z}x - z\overline{z}e_z)\,\mathrm{d}\overline{z}x, \end{aligned} \tag{8.18}$$

with $\tilde{\overline{z}}R = \overline{z}U^T\overline{z}R$. A similar approach has been previously suggested in [135] for protein matching. Here we will focus on its efficient computation. After having considered the effects of rotations, it remains for us to examine the effect of the single one dimensional translation, say along the z-axis. In spherical coordinates a translation of the vector $\overline{z}x$ about $z\overline{z}e_z$ is given by $\overline{z}x - z\overline{z}e_z = r_z\overline{z}u_z$ with $r_z = \sqrt{r^2 + 2rz\cos\theta + z^2}$ and $\overline{z}u_z = (\arccos\left(\genfrac{}{}{0pt}{}{-r_z\sin\theta}{r}\right), \varphi)$. We point out that the longitudal angle $\varphi$ does not change during a translation along the z-axis. The effect of a translation along the z-axis on the $\mathbb{R}^+ \times \mathbb{S}^2$ basis functions can be expressed in terms of translation matrix (T-matrix) elements $T^n_{jh,kl}(z)$ as described in [175] as

$$R_k^l(r_z) Y_l^n(\overline{z}u_z) = \sum_{k'=0}^{\infty} \sum_{l'=0}^{k'-1} T^n_{k'l',kl}(z) R_{k'}^{l'}(r) Y_{l'}^n(\overline{z}u). \tag{8.19}$$

Note that the T-Matrices apply only to mixed basis expansions (8.8); for pure spherical basis expansions, the coefficients $\hat{a}_{lm}$ for each radial slice with radius $r$ have to be recomputed after each translation $\mathbf{t} \in \mathbb{R}^3$.

These translation coefficients are expressed as

$$T^n_{k'l',kl}(z) = \mathrm{e}^{-z^2/4\lambda} \sum_{m=|l-l'|}^{l+l'} A_m^{ll'|n|} \sum_{j=0}^{k-l+k'-l'-2} C_j^{kl,k'l'} M!(z^2/4\lambda)^{m/2} L_M^{(m+1/2)}(z^2/4\lambda), \tag{8.20}$$

where

$$
\begin{aligned}
M &= j + \frac{l+l'-m}{2},\ C_j^{k'l',kl} = \sum_{j=0}^{k-l-1}\sum_{j'=0}^{k'-l'-1} \delta_{n,j+j'} X_{klj} X_{k'l'j'},\\
X_{klj} &= \left[\frac{(k-l-1)!(1/2)_k}{2}\right]^{1/2} \frac{(-1)^{k-l-j-1}}{j!(k-l-j-1)!(1/2)_{l+j+1}},\\
A_m^{l'l|n|} &= (-1)^{m+l'-l)/2+n}(2m+1)\left[(2l'+1)(2l+1)\right]^{1/2} \begin{pmatrix} l' & l & m\\ 0 & 0 & 0\end{pmatrix} \begin{pmatrix} l' & l & m\\ n & -n & 0\end{pmatrix}.
\end{aligned}
$$

Moreover $\begin{pmatrix} a & b & c\\ \alpha & \beta & \gamma\end{pmatrix}$ denotes the Wigner 3-j symbol and $(\cdot)_m$ is the Pochhammer symbol.

Directly computing T-Matrix entries in Equation 8.20 for fixed $k, l, k', l', m$ takes $\mathcal{O}(L^3 N_t)$ steps, where $N_t$ is the number of translations in one dimension. The overall complexity is thus $L^5 \cdot \mathcal{O}(L^3 N_t) = \mathcal{O}(L^8 N_t)$. An important contribution of the PF$corr$ algorithm is the fast and efficient computation of the T-Matrix entries in $\mathcal{O}(L^7 + L^6 N_t)$ steps. Details of this speedup can be found in the Appendix.

Having collected all the ingredients we state the following important Theorem.

**Theorem 8.7.** *For a fixed cut-off degree $L \in \mathbf{N}_0$ and two given functions*

$$
A(r\mathbf{u}) = \sum_{k=1}^{L}\sum_{l=0}^{k-1}\sum_{m=-l}^{l} \hat{a}_{klm} R_k^l(r) Y_l^m(\mathbf{u}),\ B(r\mathbf{u}) = \sum_{k=1}^{L}\sum_{l=0}^{k-1}\sum_{m=-l}^{l} \hat{b}_{klm} R_k^l(r) Y_l^m(\mathbf{u})
$$

*the objective function* (8.1) *can be evaluated by computing*

$$
\begin{aligned}
C(\overline{z}R, \overline{z}t) = C(\overline{z}R, \overline{z}U z\overline{z}e_z) \quad &= \quad \sum_{k,k'=1}^{L}\sum_{l=0}^{k-1}\sum_{l'=0}^{k'-1}\sum_{m'=-l'}^{l'}\sum_{m=-l}^{l}\sum_{n=-\min(l,l')}^{\min(l,l')} (-1)^n \hat{a}_{k'l'm'} \hat{b}_{klm}\\
&\times \quad D_h^{-nm'}(\mathbf{R}) D_l^{nm}(\mathbf{U}) T_{k'l',kl}^n(z)
\end{aligned}
$$

for arbitrary choices of $\overline{z}R \in \mathrm{SO}(3)$ and $\overline{z}t \in \mathbb{R}^3$. Its proof can be found in the Appendix.

The mixed basis expansions can be used to compute rigid-body correlations. Let $A$ and $B$ be scalar-valued functions, and let $B$ undergo rotations $\mathbf{R}$ relative to $A$. We are interested in the pure rotational correlation $C(\mathbf{R}) = \int_{\mathbb{R}^3} A(\mathbf{x})\overline{(B(\mathbf{R}\mathbf{x}))}d\mathbf{x}$, where the overbar represents complex conjugation[2]. The following two lemmas can be established, respectively, for mixed-basis coefficients $\hat{a}_{klm}, \hat{b}_{klm}$ and pure spherical basis coefficients $\hat{a}_{lm}, \hat{b}_{lm}$:

$$
C(\mathbf{R}) = \sum_{k=1}^{L}\sum_{l=0}^{k-1}\sum_{m=-l}^{l}\sum_{m'=-l}^{l} (-1)^m \hat{a}_{kl-m} (-1)^{m'} \overline{\hat{b}_{kl-m'}} D_l^{m,m'}(\alpha,\beta,\gamma). \tag{8.21}
$$

To derive the expression for general rigid-body correlations

$$
C(\mathbf{R},\mathbf{t}) = \int_{\mathbb{R}^3} A(\mathbf{x})\overline{B(\mathbf{R}\mathbf{x}+\mathbf{t})}d\mathbf{x}
$$

we can use Equation **??** along with an elementary fact: every rigid-body motion $(\mathbf{R},\mathbf{t})$ can be factored into a combination of five rotations and a single translation about the z-axis[3]. Let these five rotations be parametrized by z-y-z Euler angles $\mathbf{R}^A = (\alpha^A, \beta^A, \gamma^A)$ and $\mathbf{R}^B = (0, \beta^B, \gamma^B)$. Then we obtain, for the mixed-basis functions:

**Lemma 8.8.** *Given two functions $A : \mathbb{R}^3 \mapsto \mathbb{C}$ and $B : \mathbb{R}^3 \mapsto \mathbb{C}$ expanded in terms of a mixed basis as given in* (8.8) *the rigid-body correlation can be obtained by evaluating*

[2] The conjugation is used to simplify algebraic manipulations, and is otherwise redundant.

[3] It is enough to see that every translation $\mathbf{t}$ can be expressed as two rotations and a single translation along the z-axis. Starting at the origin, the point $\mathbf{t}$ can be reached by translating along the z-axis by $\|\mathbf{t}\|$, and then rotating about the $z$ and $y-$axes by $\theta$ and $\phi$, the spherical coordinates of $\mathbf{t}$.

$$C(\mathbf{R}, \mathbf{t}) = \sum_{klmn} \hat{a}_{klm} D_l^{n,m}(\mathbf{R}^A) \sum_{k'l'm'} (-1)^n \hat{b}_{k'l'm'} D_{l'}^{-n,m'}(\mathbf{R}^B) T_{kl,k'l'}^{|n|}(z) \tag{8.22}$$

*for arbitrary choices of* $\overline{z}R \in \mathrm{SO}(3)$ *and* $\overline{z}t \in \mathbb{R}^3$.

Following an observation in [90], it is not as efficient to use the pure spherical basis expansions to express a general rigid-body correlation. Instead, Equation **??** is used along with a scan of the translational degrees of freedom, in which the basis coefficients are recomputed for each distinct $\mathbf{t} \in \mathbb{R}^3$. Hence we omit mentioning the case of pure spherical expansions here.

We conclude this section with some notes on the complexity of the evaluation of the introduced expansions. If we use the NFSOFT [167] to compute Equations (8.21) then the pure rotational correlation in Lemma **??** can be computed in $\mathcal{O}(L^4 + N_{\mathbf{R}}^3)$ steps using the following recipe, where $N_{\mathbf{R}}$ is the number of distinct rotation angles per rotational degree of freedom, i.e., per Euler angle.

**Recipe 1.** Evaluate

$$C(\mathbf{R}) = \sum_{k=1}^{L} \sum_{l=0}^{k-1} \sum_{m=-l}^{l} \sum_{n=-l}^{l} (-1)^m \hat{a}_{kl-m} \hat{b}_{kln} D_l^{m,n}(\mathbf{R}) \tag{8.23}$$

for $N_{\mathbf{R}}^3$ different choices of Euler angles.

Q$_\xi$1. Rearrange the multiple summations such that the sum over $k$ becomes the innermost sum.

Q$_\xi$2. Compute

$$\hat{f}_{lmn} = \sum_{k=l+1}^{L} (-1)^m \hat{a}_{kl-m} \hat{b}_{kln}$$

in $\mathcal{O}(L^4)$ steps.

Q$_\xi$3. Use the $\mathrm{SO}(3)$ Fourier transform to compute the remaining sums

$$C(\mathbf{R}) = \sum_{l=0}^{L-1} \sum_{m=-l}^{l} \sum_{n=-l}^{l} \hat{f}_{lmn} D_l^{mn}(\overline{z}R)$$

in $\mathcal{O}(L^3 \log^2 L + N_{\mathbf{R}}^3)$ steps, where $N_{\mathbf{R}}$ is the number of unique Euler angles per rotation axis.

In a similar fashion, the pure rotational correlation in Lemma 8.5 can be computed in $\mathcal{O}(L^3 \log^2 L + N_{\mathbf{R}} + L^3 I)$ steps where $I$ is the complexity of computing the integral $\int_{R^+} \hat{a}(r)\hat{b}(r) r^2 dr$ for a given pair of scalar-valued functions $\hat{a}, \hat{b} : \mathbb{R}^+ \mapsto \mathbb{C}$. Since there are $\mathcal{O}(L^3)$ integrals $\int_{\mathbb{R}^+} \hat{a}(r)\hat{b}(r) r^2 dr$ we get the aforementioned complexity.

Let us now consider general rigid-body motion. The general rigid-body correlation in Theorem 8.22 can be computed in $\mathcal{O}(L^6 + L^4 N_{\mathbf{R}}^2 + N_{\mathbf{R}}^5) N_t$ steps using the outlined a way to speed up computations of the translation matrix entries (8.20) and the NFSOFT, where $N_{\mathbf{R}}^3$ and $N_{\mathbf{R}}^2$ are the number of rotations of $A$ and $B$ respectively, and $N_t$ is the number of one-dimensional translations. The computation is performed according to the following recipe.

**Recipe 2.** Evaluate

$$\begin{aligned} C(\overline{z}R, \overline{z}t) = C(\overline{z}R, \overline{z}U z \overline{z}e_z) \quad &= \quad \sum_{k,k'=1}^{L} \sum_{l=0}^{k-1} \sum_{l'=0}^{k'-1} \sum_{m'=-l'}^{l'} \sum_{m=-l}^{l} \sum_{n=-\min(l,l')}^{\min(l,l')} (-1)^{n} \hat{a}_{k'l'm'} \hat{b}_{klm} \\ &\times \quad D_h^{-nm'}(\mathbf{R}) D_l^{nm}(\mathbf{U}) T_{k'l',kl}^{n}(z) \end{aligned}$$

for $N_{\mathbf{R}}$ different choices of $\overline{z}R$ and $N_t$ different choices of one-dimensional translations $z \in \mathbb{R}$

Q$_\xi$1. Compute

$$\hat{c}_{kll'}^{m'n} = \sum_{k'=l'+1}^{L} \overline{z}1_{|n| \leq l'} \hat{b}_{k'l'm'} T_{k'l',kl}^{n}(z)$$

in $\mathcal{O}(L^6)$ steps.

Q$_\xi$2. Compute

$$\tilde{c}^{n}_{kl} = \sum_{l'=0}^{L-1} \sum_{m'=-l'}^{l'} \hat{c}^{m'n}_{kll'} D_{l'}^{-n,m'}(\mathbf{U})$$

using a modification of the NFSOFT in $\mathcal{O}(L^5 \log L + N_{\mathbf{R}}^2 L^3)$ steps.

Q$_\xi$3. Compute

$$c_l^{mn} = \sum_{k=l+1}^{L} (-1)^{n}\hat{a}_{klm}\tilde{c}^{n}_{kl}.$$

Q$_\xi$4. Compute

$$C(\overline{z}R, \overline{z}U z\overline{z}e_z) = \sum_{l=0}^{L-1} \sum_{m=-l}^{l} \sum_{n=-l}^{l} c_l^{mn} D_l^{n,m}(\mathbf{R}^A)$$

using the standard NFSOFT [167] in $\mathcal{O}(N_{\mathbf{R}}^2(L^4 + L^3 \log L + N_{\mathbf{R}}^3))$ steps.

Hence, the overall cost is $\mathcal{O}(L^6 + L^5 \log L + N_{\mathbf{R}}^2(L^3 + L^4) + N_{\mathbf{R}}^2 N_{\mathbf{R}}^3)N_t$, i.e., $\mathcal{O}(L^6 + L^4 N_{\mathbf{R}}^2 + N_{\mathbf{R}}^5)N_t$.
With these recipes established, we now outline algorithms to perform fast rigid-body correlations given a pair of scalar-valued functions as input. Algorithm 1 uses the mixed basis, while Algorithm 2 uses the pure spherical harmonic basis.

**Algorithm 1:** Fast Rotational Correlation with mixed radial/spherical basis functions

```
Input: $L$ : Expansion degree;
$G$ : Spherical grid with sizes $N_r$, $N_\theta$, $N_\phi$ in the radial, polar and azimuthal directions respectively. Let
$N = \max(N_r, N_\theta, N_\phi)$;
$A, B : \mathbb{R}^3 \mapsto \mathbb{C}$ : scalar-valued functions sampled on $G$ centered at $r = 0$;
$\mathcal{M} \subset \mathbb{R}^3 \times \mathrm{SO}(3)$ : a finite set of rigid-body motions;
foreach $(k, l, m)$ with $|m| \leq l \leq k \leq L$ do
    Calculate the coefficients $\hat{a}_{klm}$ and $\hat{b}_{klm}$ using Equation 8.7;
end
if $\mathbf{t} == \mathbf{0}\ \forall(\mathbf{R}, \mathbf{t}) \in \mathcal{M}$ then
Find the maximum value of $C(\mathbf{R}) = \int_{\mathbb{R}^3} A(\mathbf{x})B(\mathbf{R}\mathbf{x})d\mathbf{x}\ \forall\mathbf{R} \in \mathcal{M}$ using Recipe 1.;
else Find the maximum value of $C(\mathbf{R}, \mathbf{t}) = \int_{\mathbb{R}^3} A(\mathbf{x})B(\mathbf{R}\mathbf{x} + \mathbf{t})d\mathbf{x}\ \forall(\mathbf{R}, \mathbf{t}) \in \mathcal{M}$ using the steps from Recipe 2.;
;
Output: The maximum correlation $C \in \mathbb{C}$ between $A$ and $B$;
Complexity: $\mathcal{O}(\mathcal{C}_{\text{coeff}} + \mathcal{C}_{\text{PF}corr})$ flops, where $\mathcal{C}_{\text{coeff}} = \mathcal{O}(L^3 N^3)$ is the complexity of computing the coefficients $\hat{a}_{klm}$,
and $\mathcal{C}_{\text{PF}corr} = \mathcal{O}(L^4 + N_{\mathbf{R}})$ in the pure rotational case or $\mathcal{O}(L^6 + L^4 N_{\mathbf{R}}^2 + N_{\mathbf{R}}^5)N_t$ in the general case;
```

There are three sources of error . The first is the expansion error, i.e., the error induced by truncating the basis expansion at a finite value of $L$. The second is the representation error, i.e., the error induced in numerically integrating the coefficients in Equation (8.7). The third is the NFFT error, i.e., the error induced by approximating the exponential sums by the NFFT.
Following [175, 177, 90], the first two sources of error can be respectively mitigated by choosing an expansion degree between $20 \leq L \leq 25$, and using a single-point quadrature rule. We provide further evidence in Section 8.4.1 of the former assertion.
The NFFT approximates exponential sums with a kernel basis expansion, providing a choice of several kernels, and several parameters govern the actual error of the expansion. In our implementation, we choose the Gaussian kernel with an oversampling factor of 3, see [168], resulting in the errors in Table **??**. On more information about the error of the NFFT and the NFSOFT we refer to [168] and [167], respectively.
Note that, in solutions to the correlation problem, the absolute value of the correlation is less important than the value relative to other rigid-body rotations, i.e., the ability of the search scheme to discriminate between two different rigid-body motions. A measure of this ability is presented in Section 8.4.1 in the context of sampling arbitrary subsets of the motion group $SE(3)$.

**Algorithm 2:** Fast Rotational Correlations with pure spherical harmonic basis functions

```
Input: $L$ : Expansion degree;
  $G$ : Spherical grid with sizes $N_r$, $N_\theta$, $N_\phi$ in the radial, polar and azimuthal directions respectively. Let
  $N = \max(N_r, N_\theta, N_\phi)$;
  $A, B : \mathbb{R}^3 \mapsto \mathbb{C}$ : scalar-valued functions sampled on $G$ centered at $r = 0$;
  $\mathcal{T} \subset \mathbb{R}^3 \times \mathrm{SO}(3)$ : a finite set of pairs $\{(\mathbf{t}, \mathcal{R})\}$, where $\mathbf{t} \in \mathbb{R}^3$ is a translation and $\mathcal{R} \subset \mathrm{SO}(3)$ is a finite set of rotations
  corresponding to $\mathbf{t}$;
foreach $r \in G$ do
    foreach $(l, m)$ with $|m| \leq l \leq L$ do
        Compute $\hat{a}_{lm}(r)$ using Equation 8.10;
    end
end
foreach $(\mathbf{t}, \mathcal{R}) \in \mathcal{T}$ do
    Translate $B(\mathbf{x})$ by $\mathbf{t}$ ;
    foreach $(l, m)$ with $|m| \leq l \leq L$ do
        Compute $\hat{b}_{lm}(r)$ using Equation 8.10;
    end
    Compute $C(\mathbf{R}) = \int_{\mathbb{R}^3} A(\mathbf{x}) B(\mathbf{R}(\mathbf{x} + \mathbf{t})) d\mathbf{x} \; \forall \mathbf{R} \in \mathcal{R}$ as in Recipe 2.
end
Output: The maximum correlation $C \in \mathbb{C}$ between $A$ and $B$;
Complexity: $\mathcal{O}((\mathcal{C}_{\text{coeff}} + \mathcal{C}_{\text{PF}corr})|\mathcal{T}|)$ flops, where $\mathcal{C}_{\text{coeff}} = \mathcal{O}(N^2 L^2)$, and $\mathcal{C}_{\text{PF}corr} = \mathcal{O}(L^3 \log L + N_{\mathbf{R}}^3)$;
```

For the T-Matrix computation, a dramatic speedup with respect to the direct algorithm is observed in $L \geq 10$ regime, where the $L^7$ v/s $L^8$ scaling is apparent. However, for typical values of $L$ (see following paragraph), the computation times are still too slow to be usable in the inner loop of any Fourier-based correlation approach, including our own. Like prior work that uses the T-Matrix (See the introduction for an overview), we thus prefer to precompute and store T-Matrix entries for given values of $z$ and $\lambda$ (See Equation 8.20).

From a practical standpoint, our rigid-body correlation search is seen to be a viable, if somewhat slower, alternative to existing rigid-body correlation search techniques. Most of the degradation in performance is due to the NFFT, which uses, in its implementation, an oversampled FFT to enable the non-uniformity inherent to it. Following [177], we choose $L$ to typically lie between 20 and 25, in which case typical run times for an exhaustive correlation involving about $1.5 \cdot 10^7$ distinct rigid-body samples lie between 2 and 3 minutes. We also note that, other than the argument in Section 8.4.1, there is no reason to prefer the non-uniformity inherent to PF$corr$ and, if performance is a concern, each of the steps involving the NFFT can be replaced by the equispaced FFT.

### 8.4.1 Sampling arbitrary subsets of the motion group $SE(3)$; addressing the drawbacks of existing techniques

The main advantage of PF$corr$ is in sampling arbitrary (finite) subsets of the space of rigid body motion in three dimensions $SE(3) = \mathbb{R}^3 \times \mathrm{SO}(3)$. In our implementation, this is as simple as specifying a set of rigid-body motions on which correlations are to be performed. By contrast, all prior techniques *require* an equispaced angular grid for rotational search, a property that results in a highly non-uniform search of the space of rotations (See Drawback 2 in the introduction).

For exhaustive correlations between a pair of scalar-valued functions, one typically employs *uniform* sampling of the space of rotations $\mathrm{SO}(3)$. As we mention in the introduction, most of the uncertainty in the rigid-body correlation problem lies in the space of rigid-body rotations, and it is thus more important to sample this space exhaustively. There are several existing techniques that, given an angular sampling criterion, provide a set of samples that are uniform with respect to accepted metrics of uniformity [106, 155, 230]. We use the approach from [155], in which the metrics of *local separation* and *global coverage* compete to provide a set of highly uniform samples in $\mathrm{SO}(3)$.

The ability to sample and correlate over arbitrary subsets of $SE(3)$ is only useful if, at any expansion degree, the fineness of the rotational sampling does not exceed the accuracy with which $\hat{a}_{klm}$ and $\hat{b}_{klm}$ represent $A$ and $B$ respectively (See Equation

(8.7)). Such a scenario would give rise to correlations that are so close to each other as to be essentially indistinguishable, and would result in a set of correlations clustered around the average correlation. To measure this tendency, we compute the Z-score $z = \frac{x-\mu}{\sigma}$, a measure of the distance of each individual correlation from the average. The results, indicate that the top-ranking Z-score increases with increase in degree, as expected, leveling off at $L \geq 20$, where the error due to floating-point calculations begins to rival the error due to representation, and that even at very low expansion degrees, the top-ranking Z-score is 3 standard deviations from the mean, indicating a very high confidence. Another argument as to why the regime $20 \leq L \leq 25$ is best, as the latter provides a balance between the errors of representation and floating-point computation. For additional information on the Z-score measure see e.g. [165].

## 8.5 Flexible correlations: main results

We present an algorithm (Algorithm 3) for domain-based protein matching. This algorithm, given as input

Q$_\xi$1. A protein $\mathcal{P}$,

Q$_\xi$2. A hierarchical domain decomposition, defined in Section 8.5.1, of $\mathcal{P}$,

Q$_\xi$3. A scalar-valued function $B : \mathbb{R}^3 \mapsto \mathbb{R}$ representing a stationary target, and,

Q$_\xi$4. A scalar-valued representation $A$ of $\mathcal{P}$,

produces as output the optimal correlation between $A$ and $B$ under rigid-body motions of the domains of $\mathcal{P}$. Algorithm 3 makes use of the ability of PF$corr$ to uniformly sample arbitrary subsets of $\mathbb{R}^3 \times \mathrm{SO}(3)$.

### 8.5.1 Domain-based protein flexibility framework

We assume a generic framework for domain-based protein flexibility. This framework consists of ideas from domain-decomposition of proteins that have existed in various forms over the past decade (see especially [148]), as well as a set of techniques, described, for instance, in [19], to assign motions to each of these domains.
Let a protein crystal structure $\mathcal{P}$ comprise a set of atoms. Designate a subset of $\mathcal{P}$ as a domain $D$. A domain decomposition of $\mathcal{P}$ is a set $\mathcal{DD} = \{D_i\}, 1 \leq i \leq n_{\mathcal{DD}}$, where $D_i$ is a domain. A hierarchical domain decomposition $\mathcal{HD} = \{\mathcal{DD}_i\}, 1 \leq i \leq n_{\mathcal{HD}}$ is a set of domain decompositions $\mathcal{DD}_i$ such that each domain in $\mathcal{DD}_i$ is a subdomain of some domain in $\mathcal{DD}_{i-1}$ (See, for example, [29]). For each $\mathcal{DD}_i$ of the hierarchical domain decomposition $\mathcal{HD}$, a motion graph $MG$ specifying relative motions between domains of $\mathcal{DD}_i$ can be specified. The motion graph consists of a set of edges $F_{ij}$, called flexors, between pairs of domains $i, j$ that undergo relative motion. The geometric properties of each flexor imply a set of rigid-body transformations $(\mathbf{R}^k_{i,j}, \mathbf{t}^k_{i,j})$, $k \in \{1 \ldots N_\mathbf{T}\}$ applied to $D_j$ relative to $D_i$ [19].

### 8.5.2 Algorithm for flexible matching

Algorithm 3 applies to a particular domain decomposition of $\mathcal{P}$, i.e, it applies to a particular index in the hierarchical domain decomposition of $\mathcal{P}$. It uses the ability of PF$corr$ to sample arbitrary subsets of $SE(3)$ to match representations of domains $A_i \in A$ to a target scalar-valued function $B : \mathbb{R}^3 \mapsto \mathbb{R}$. Note by contrast that a classic equispaced Fourier-based correlation scheme would not be able to perform the tasks in Algorithm 3 without also producing several results that do not belong to the chosen subset of $SE(3)$. This focusing property enables PF$corr$ to combine the merits of both local and global optimization schemes in the following sense. The algorithm is *local* in that it is restricted to the chosen subset of $SE(3)$, but *global* in that it samples that subset exhaustively. It thus combines the speed of a local search without being sensitive, as local search algorithms are, to local optima.

## 8.6 Conclusion

## Appendix

Here we give additional details on the mathematical background of the used algorithms.

**Algorithm 3:** Greedy multi-domain matching

```
Input:
  Q_ξ1. $\mathcal{P}$ : Protein;
  Q_ξ2. $\mathcal{DD} = \{D_i, MG\}, i \in \{1 \dots N_D\}$ :  A domain decomposition of $\mathcal{P}$;
  Q_ξ3. $\mathcal{R}(\mathcal{DD}_i)$ :  A conversion from $D_i \in \mathcal{DD}$ into a function $A_i : \mathbb{R}^3 \mapsto \mathbb{R}$;
  Q_ξ4. $A : \mathbb{R}^3 \mapsto \mathbb{R}$ : Scalar-valued function representing $\mathcal{P}$;
  Q_ξ5. $B : \mathbb{R}^3 \mapsto \mathbb{R}$ :  Target scalar-valued function;
  Q_ξ6. $PQ$ : Priority queue with elements $(j, r)$, $j \in \mathbb{Z}^+, r \in \mathbb{R}$ ordered least-first w.r.t $r$;
Output: The optimal correlation between $A_i$ and $B$ under rigid-body transformations of $A_i$, $i \in \{1 \dots N_D\}$.
Use PF$corr$ to find the optimal rigid-body transformation $(\mathbf{R}, \mathbf{t})$ relating $A$ to $B$;
foreach $D_i \in \mathcal{DD}$ do
    Compute the correlation $C_i \leftarrow \int_{\mathbb{R}^3} A_i B d\mathbf{x}$ between each domain $D_i$ and the target $B$;
    Push $(i, C_i)$ to $PQ$;
end
$i \leftarrow 1$;
while $i \leq N_D$ do
    $k \leftarrow PQ[N_D - i - 1].j$;
    $D_i \leftarrow D_k$;
    $i \leftarrow i + 1$;
end
foreach $D_i \in \mathcal{DD}$, $i \neq 1$ do
    Using flexors $F_{i-1,i}$, compute the set of relative motions $T_{i-1,i} \leftarrow \{(\mathbf{R}^k_{i-1,i}, \mathbf{t}^k_{i-1,i})\}$, $k \in \{1 \dots N^i_{\mathbf{T}}\}$ of $D_i$ relative
      to $D_{i-1}$;
    Compute the set of absolute motions $T_i \leftarrow \{(\mathbf{R}^k_i, \mathbf{t}^k_i)\}$, $k \in \{1 \dots N^i_{\mathbf{T}}\}$ for each rigid-body transformation in the set
      $T_{i-1,i}$ relative to the stationary domain $D_1$;
end
foreach $(i, C_i) \in PQ$ do
    Use PF$corr$ to find the optimal rigid-body transformation $(\mathbf{R}_i, \mathbf{t}_i) \in T_i$ relating $A_i$ to $B$;
end

Complexity: $\mathcal{O}(C_{\text{PF}corr} N_D)$ flops, where $C_{\text{PF}corr}$ is the complexity of PF$corr$.
```

**T-Matrices Computation.**
The translation coefficients $T^{|m|}_{k'l',kl}(z) \cdot \exp(z^2/4\lambda)$ are polynomials of degree

$$\max(n+2M) = \max(n + 2(j + \frac{l+l'-k}{2})) = \max(2j + l + l') = 2k - l + 2k' - l' - 4.$$

Let $d = 2k - l + 2k' - l' - 4$, $n = \min(p, l+l') - s$ and $i = \frac{p-n}{2}$ Then Equation (8.20) can be arranged to obtain

$$T^{|m|}_{k'l',kl}(z)\exp(z^2/4\lambda) = \sum_{p=0}^{2k-l+2k'-l'-4} \alpha_p \cdot z^p$$

where

$$\alpha_p = \\ \sum_{s=0}^{\min(p,l+l')-|l-l'|} A_n^{ll'|m|} \sum_{j=\max(i-\frac{l+l'-n}{2},0)}^{k-l+k'-l'-2} C_j^{k,l,k'l'} M! \frac{(1/2)_{M+n+1}}{(M-i)!(1/2)_{n+i+1}} \cdot \frac{1}{(-4\lambda)^i i!},$$

and $s$ is even if and only if $d$ is even.

The coefficients $\alpha_p$ can be computed for all $p$ in $\mathcal{O}(L^3)$ steps. For fixed $k, l, k', l', m$, the T-Matrix polynomial can be computed in $\mathcal{O}(LN_t)$. The complexity for fixed $k, l, k', l', m$ is hence $\mathcal{O}(L^3 + LN_t)$, resulting in an overall complexity of $\mathcal{O}(L^8 + L^6 N_t)$. A polynomial can be evaluated at a set of equispaced arguments with $\mathcal{O}(L)$ multiplications. Applying Nuttall's update rule for polynomials [162] reduces these multiplications to additions without altering the number of operations required. This affords a small speedup.

**T-Matrices Computation Speedup.** If $A_n^{ll'|m|}$ is precomputed for all $m$, the other terms in Equation (8.20) have to be calculated only once for fixed $k, l, k', l'$. In the first step, we compute

$$b_n := \sum_{j=0}^{k-l+k'-l'-2} C_j^{kl,k'l'} M! \exp(-z^2/4\lambda)\ (z^2/4\lambda)^{n/2} L_M^{(n+1/2)}(z^2/4\lambda)$$

for all $m$ and fixed $n, l, n', l'$. The summation over $j$ and the computation of $L_M^{n+1/2}$ each takes $\mathcal{O}(L)$ steps, implying a complexity of $\mathcal{O}(L^2)$ for each $b_n$, and a complexity for all $m$ of $\mathcal{O}(L^3)$.

In the second step we compute the T-Matrix entries $T^{|m|}_{k'l',kl} = \sum\limits_{n=|l-l'|}^{l+l'} b_n \cdot A_n^{ll'|m|}$.

Since the above calculation has to be done for all $k, l, k', l'$ and for $N_t$ translations, the overall complexity for $T^{|m|}_{k'l',kl}$ is now $\mathcal{O}(L^7 N_t)$, instead of $\mathcal{O}(L^8)$.

Computation of the coefficients $C_j^{kl,k'l'}$ can also be sped up. Only these coefficients and the boundary of the innermost sum depend on $k$ and $k'$. If $k$ and $k'$ are switched, the boundary of the sum does not change, so for switched $k$ and $k'$ only the value $C_j^{kl,k'l'}$ changes. In the first step

$$l(z^2/4\lambda) := L_M^{(n+1/2)}(z^2/4\lambda)$$

is computed for all $j, n, l, l'$. In the second step

$$t_{kl,k'l'} := \sum_{j=0}^{k-l+k'-l'-2} C_j^{kl,k'l'} M! \exp(-z^2/4\lambda)\ (z^2/4\lambda)^{n/2} l(z^2/4\lambda)$$

and $t_{k'l,kl'}$ respectively are computed. In the third step

$$T^{|m|}_{k'l',kl} = \sum_{n=|l-l'|}^{l+l'} A_n^{ll'|m|} \cdot t_{kl,k'l'}$$

and $T^{|m|}_{kl',k'l}$ respectively are computed.

Moreover, the symmetry property [175] $T^{|m|}_{k'l',kl} = (-1)^{l-l'} T^{|m|}_{kl,k'l'}$ implies $T^{|m|}_{kl',k'l} = (-1)^{l-l'} T^{|m|}_{k'l,kl'}$. Hence, the dynamic programming approach above allows us to calculate $T^{|m|}_{kl,k'l'}$, $T^{|m|}_{kl',k'l}$ and $T^{|m|}_{k'l,kl'}$ by calculating $T^{|m|}_{k'l',kl}$.

The complexity of the approach of representing the $T$-coefficients as a polynomial can be reduced by using the speed-up by dynamic programming as explained above. To achieve the reduction in the complexity we consider the calculation of $\alpha_p$. Instead of computing $\alpha_p$ directly, first

$$b_s^p := \frac{1}{(-4\lambda)^i \cdot i!(1/2)_{n+i+1}} \sum_{j=\max(i-\frac{l+l'-n}{2},0)}^{k-l+k'-l'-2} C_j^{k,l,k'l'} M! \frac{(1/2)_{M+n+1}}{(M-i)!}$$

is precomputed. Due to the summation and the parameters $s$ and $p$, this computation has the complexity $\mathcal{O}(L^3)$ Afterwards the $\alpha_p$

$$\alpha_p = \sum_{s=0}^{\min(p,l+l')-|l-l'|} A_n^{ll'|m|} \cdot b_s^p.$$

are computed This has the complexity $\mathcal{O}(L^2)$, implying a complexity of $\mathcal{O}(L^3)$ for the precomputation of $\alpha_p$ for all $m$. The total computation of the $\alpha_p$ for all $m$ is hence $\mathcal{O}(L^3 + L^3) = \mathcal{O}(L^3)$.

The subsequent computation of

$$T^{|m|}_{kl',k'l} \exp(z^2/4\lambda) = \sum_{p=0}^{2k-l+2k'-l'-4} \alpha_p \cdot z^p$$

is for fixed $k, l, k', l', m$ and all $m$ is $\mathcal{O}(L^2 N_t)$. Therefore the overall complexity for fixed $k, l, k', l'$ and all $m$ is $\mathcal{O}(L^3 + L^2 N_t)$. Thus, for all $k, l, k', l'$ the complexity is $\mathcal{O}(L^4)\mathcal{O}(L^3 + L^2 N_t) = \mathcal{O}(L^7 + L^6 N_t)$.

**Proof of Lemma 8.3.** Let $\Omega$ be subdivided in $N$ grid-cells $\Omega_i$ with centers $\overline{z}x_i$, volume $V_i$ and diameter $d_i$. The approximation error in the $i$th grid-cell is given by

$$E_i = \left| \int_{\Omega_i} A(\mathbf{x}) d\mathbf{x} - A(\mathbf{x}_i) V_i \right| = \left| \int_{\Omega_i} (A(\mathbf{x}) - A(\mathbf{x}_i)) \, \mathrm{d}\mathbf{x} \right|.$$

Expanding $A(\mathbf{x})$ in a Taylor series about $\mathbf{x}_i$, we get

$$E_i = \left| \int_{\Omega_i} (\mathbf{x} - \mathbf{x}_i)^{\mathfrak{T}} \nabla A(\mathbf{x}_i) + O(||\overline{z}x - \overline{z}x_i||^2) \, \mathrm{d}\mathbf{x} \right| \leq \left| \int_{\Omega_i} c \cdot d_i \, \mathrm{d}\mathbf{x} \right| \leq c \cdot d_i \cdot V_i.$$

for some constant $c$, due to $A(\overline{z}x)$ being 2-Lipschitz continous on $\Omega$ . Thus the error across all grid cells is the sum $E = \sum_{\forall i} E_i \leq c \max_{\forall i} d_i |\Omega|$. Since the maximum diameter of the grid-cells is proportional to the grid fineness $h$, we have $E \leq Ch|\Omega|$ for a fixed constant $C$.

**Proof of Theorem 8.22.** Consider a rotation $\overline{z}R \in \mathrm{SO}(3)$ that is applied to the molecule $A$. By the representation property of spherical harmonics 8.11 the affinity function becomes

$$A(\overline{z}R^T \overline{z}x) = A(r\overline{z}R^T \overline{z}u) \quad = \quad \sum_{k=1}^{\infty} \sum_{l=0}^{k-1} \sum_{m,n=-l}^{l} \hat{a}_{klm} D_l^{nm}(\mathbf{R}) R_k^l(r) Y_l^n(\mathbf{u}).$$

The molecule $B$ will be rotated by $\overline{z}U = \overline{z}R_Z(\varphi)\overline{z}R_Y(\theta)$ and translated by the vector $(0,0,z)^{\mathrm{T}}$. Using (8.19), this yields the series expansion

$$\begin{aligned}B(\overline{z}U\overline{z}x - z\overline{z}e_z) = B(r_z(\overline{z}U^T\overline{z}u)_z) &= \sum_{k=1}^{\infty}\sum_{l=0}^{k-1}\sum_{m,n=-l}^{l} \hat{b}_{klm} D_l^{nm}(\mathbf{U}) R_k^l(r_z) Y_l^n(\mathbf{u}_z)\\ &= \sum_{k=1}^{\infty}\sum_{l=0}^{k-1}\sum_{m,n=-l}^{l}\sum_{k'=0}^{\infty}\sum_{l'=0}^{k'-1} \hat{b}_{klm} D_l^{nm}(\mathbf{U}) T^n_{k'l',kl}(z) R_{k'}^{l'}(r) Y_{l'}^n(\overline{z}u).\end{aligned}$$

After inserting both of the above expansions of the affinity functions into the correlation integral (8.18), we are now able to use the orthonormality property

$$\langle R_k^l(r)Y_l^m(\mathbf{u}), R_{k'}^{l'}(r)Y_{l'}^{m'}(\mathbf{u})\rangle = \delta_{k,k'}\delta_{l,l'}\delta_{m,m'}$$

to simplify the correlation integral to

$$\begin{aligned}C(\overline{z}R, \overline{z}U z\overline{z}e_z) &= \int_{\mathbb{R}^3} A(\overline{z}R^T\overline{z}x)B(\overline{z}U^T\overline{z}x - z\overline{z}e_z)\,\mathrm{d}\overline{z}x\\ &= \sum_{k,k',k''=1}^{\infty}\sum_{l''=0}^{k''-1}\sum_{m',n'=-l''}^{l''}\sum_{l=0}^{k-1}\sum_{m,n=-l}^{l}\sum_{l'=0}^{k'-1}(-1)^{n'}\hat{a}_{k''l''m'}D_{l''}^{n'm'}(\mathbf{R})\\ &\times\ \hat{b}_{klm}D_l^{nm}(\mathbf{U})T^n_{k'l',kl}(z)\delta_{k',k''}\delta_{l',l''}\delta_{n,-n'}\\ &= \sum_{k,k'=1}^{\infty}\sum_{l=0}^{k-1}\sum_{l'=0}^{k'-1}\sum_{m'=-l'}^{l'}\sum_{m=-l}^{l}\sum_{n=-\min(l,l')}^{\min(l,l')}(-1)^{n}\hat{a}_{k'l'm'}\hat{b}_{klm}\\ &\times\ D_{l'}^{-nm'}(\mathbf{R})D_l^{nm}(\mathbf{U})T^n_{k'l',kl}(z).\end{aligned}$$

If we now approximate the infinite sums by sums with a certain maximum degree $L$ we obtain the formula from Theorem 8.22.

## 8.7 Sampling

### 8.7.1 Monte Carlo method

Let $g\colon [0,1] \to [0,1]$ be integrable and $|g'(x)| \leq c$. We are interested in computing $\int_0^1 g(x)\ \mathrm{d}\,x$ using $m$ samples with some probability bound on the error.

Let $f\colon \{0,1\}^n \to [0,1]$ be a function mapping $n$-bit strings to reals such that $f(\mathbf{x}) = g(\widetilde{\mathbf{x}})$, where $\widetilde{\mathbf{x}}$ is the value of the binary string prepending by a radix point. Then, we have

$$\frac{1}{2^n}\sum_{\mathbf{x}\in\{0,1\}^n}\left(f(\mathbf{x}) - \frac{c}{2^n}\right) \leq \int_0^1 g(x)\ \mathrm{d}\,x \leq \frac{1}{2^n}\sum_{\mathbf{x}\in\{0,1\}^n}\left(f(\mathbf{x}) + \frac{c}{2^n}\right).$$

This implies that the average estimate $\widetilde{f} = \frac{1}{2^n}\sum_{\mathbf{x}\in\{0,1\}^n} f(\mathbf{x})$ of $f$ is a close approximation to the integral $\int_0^1 g(x)\ \mathrm{d}\,x$.

**Theorem 8.9** ([156, Theorem 13.5])**.** *Let $f$ and $\widetilde{f}$ be defined as above and let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be $m$ samples chosen i.i.d. uniform in $\{0,1\}^n$. If $m > \frac{1}{2\epsilon^2}\ln\frac{2}{\delta}$, then*

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i) - \widetilde{f}\right| \geq \epsilon\right] \leq \delta.$$

Suppose we want to obtain similar error with fewer number of purely random bits. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ be pairwise independent.

Let $Y = \frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i)$. Then, $\mathrm{E}[Y] = \mathrm{E}[f(\mathbf{x})]$. By Chebyshev inequality,

$$\begin{aligned}
\Pr\left[|Y - \mathrm{E}[f(\mathbf{x})]| \geq \epsilon\right] &\leq \frac{\mathrm{Var}[Y]}{\epsilon^2} \\
&= \frac{\mathrm{Var}[\frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i)]}{\epsilon^2} \\
&= \frac{\mathrm{Var}[\sum_{i=1}^{m} f(\mathbf{x}_i)]}{m^2\epsilon^2} \\
&\leq \frac{1}{m\epsilon^2}.
\end{aligned}$$

Thus, $\Pr[|Y - \widetilde{f}]| \geq \epsilon] \leq \delta$ when $m \geq \frac{1}{\delta\epsilon^2}$.

### 8.7.2 Quasi Monte Carlo method

Suppose $P = \{x_0, x_1, x_2, \ldots, x_{N-1}\}$. The *star discrepancy* is defined as

$$\Delta_P(t) = \frac{1}{N}\sum_{i=0}^{N-1} \mathbb{1}_{[0,t]}(x_i) - t,$$

where

$$\mathbb{1}_{[0,t]}(x) = \begin{cases} 1 & \text{if } x \in [0,t] \\ 0 & \text{otherwise} \end{cases}$$

is the characteristic function. The star discrepany can be regarded as a test of randomness (i.e. how uniform is the distribution) using the family of all intervals with the left endpoint at the origin. It is related to the Kolmogorov-Smirnov test.
The Koksma-Hlawka inequality can be used the bound the integration error. See [196] for a list of references.

**Theorem 8.10** (Koksma-Hlawka inequality)**.** *For all* $1 \leq p, q \leq \infty$ *such that* $\frac{1}{p} + \frac{1}{q} = 1$,

$$\left| \int_0^1 f(x)\ \mathrm{d}\,x - \frac{1}{N}\sum_{i=0}^{N-1} f(x_i) \right| \leq \|\Delta_P\|_p \|f'\|_q.$$

*Proof.* Observe that $f(x)$ and $t$ can be written as follows.

$$\begin{aligned}
f(x) &= f(1) - \int_x^1 f'(t)\ \mathrm{d}\,t \\
&= f(1) - \int_0^1 \mathbb{1}_{[0,t]}(x) f'(t)\ \mathrm{d}\,t \\
t &= \int_0^1 \mathbb{1}_{[0,t]}(x)\ \mathrm{d}\,x
\end{aligned}$$

We can rewrite the integral as follows.

$$\begin{aligned}
\int_0^1 f(x)\ \mathrm{d}\,x &= \int_0^1 \left( f(1) - \int_0^1 \mathbb{1}_{[0,t]} f'(t)\ \mathrm{d}\,t \right) \mathrm{d}\,x \\
&= f(1) - \int_0^1 \int_0^1 \mathbb{1}_{[0,t]}(x) f'(t)\ \mathrm{d}\,t\,\mathrm{d}\,x \\
&= f(1) - \int_0^1 t f'(t)\ \mathrm{d}\,x. \qquad (8.24)
\end{aligned}$$

we can rewrite the average as follows.

$$\begin{aligned}\frac{1}{N}\sum_{i=0}^{N-1} f(x_i) &= \frac{1}{N}\sum_{i=0}^{N-1}\left(f(1) - \int_0^1 \mathbb{1}_{[0,t]}(x_i) f'(t)\,\mathrm{d}\,t\right)\\ &= f(1) - \int_0^1 \frac{1}{N}\sum_{i=0}^{N-1} \mathbb{1}_{[0,t]}(x_i) f'(t)\,\mathrm{d}\,t \qquad (8.25)\end{aligned}$$

The result follows by subtracting (8.24) from (8.25) and applying Hölder inequality. □

### Hölder Inequality

If $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, then its $\ell_p$ norm is defined as $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{N} |x_i|^p\right)^{1/p}$.

**Theorem 8.11** (Hölder Inequality). *For all $1 \le p, q \le \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$,*

$$\sum_{i=1}^{N} |x_i| \cdot |y_i| \le \left(\sum_{i=1}^{N} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{N} |y_i|^q\right)^{1/q}.$$

*If $x(t)$ is a function, then its $L_p$ norm is defined as $\left(\int |x(t)|^p\,\mathrm{d}\,t\right)^{1/p}$. So, the error bound in the Koksma-Hlawka inequality reads*

$$\left(\int_0^1 |\Delta_P(t)|^p\,\mathrm{d}\,t\right)^{1/p} \left(\int_0^1 |f'(t)|^q\,\mathrm{d}\,t\right)^{1/q}.$$

### Monte Carlo and Quasi Monte Carlo Integration

Let $g\colon [0,1] \to [0,1]$ be an integrable function. Suppose that its derivative exists and is bounded by $|g'(x)| \le c$. We want to calculate $I(g) = \int_0^1 g(x)\,\mathrm{d}\,x$. Note that if $X$ is a random variable and $X_i \sim \mathrm{U}(0,1)$ are $n$ independent random values uniformly distributed in the interval $[0,1]$ then $\mathrm{E}[g(X)] = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \to \int_0^1 g(x)\,\mathrm{d}\,x$ and with probability 1 as $n \to \infty$. The error in the Monte Carlo (MC) estimate of the integral is supplied by the variance, i.e. an application of Chernoff bounds. $\mathrm{Var}[g(X)] = (\frac{1}{n}\sum_{i=1}^{m} g(X_i) - \int_0^1 g(x)\,\mathrm{d}\,x)^2 = \frac{\sigma^2}{n}$

Let $f\colon \{0,1\}^n \to [0,1]$ be a function mapping $n$-bit strings to reals such that $f(\mathbf{x}) = g(\tilde{\mathbf{x}})$, where $\tilde{\mathbf{x}}$ is the value of the binary string prepending by a radix point. Then, we have

$$\frac{1}{2^n}\sum_{\mathbf{x}\in\{0,1\}^n}\left(f(\mathbf{x}) - \frac{c}{2^n}\right) \le \int_0^1 g(x)\,\mathrm{d}\,x \le \frac{1}{2^n}\sum_{\mathbf{x}\in\{0,1\}^n}\left(f(\mathbf{x}) + \frac{c}{2^n}\right).$$

This implies that the average estimate $\tilde{f} = \frac{1}{2^n}\sum_{\mathbf{x}\in\{0,1\}^n} f(\mathbf{x})$ of $f$ is a close approximation to the integral $\int_0^1 g(x)\,\mathrm{d}\,x$.

**Theorem 8.12** ([156, Theorem 13.5]). *Let $f$ and $\tilde{f}$ be defined as above and let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be $m$ samples chosen i.i.d. uniform in $\{0,1\}^n$. If $m > \frac{1}{2\epsilon^2}\ln\frac{2}{\delta}$, then*

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i) - \tilde{f}\right| \ge \epsilon\right] \le \delta.$$

Suppose we want to obtain similar error with fewer number of purely random bits. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ be pairwise independent.

Let $Y = \frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i)$. Then, $\mathrm{E}[Y] = \mathrm{E}[f(\mathbf{x})]$. By Chebyshev inequality,

$$\begin{aligned}
\Pr\left[|Y - \mathrm{E}[f(\mathbf{x})]| \geq \epsilon\right] &\leq \frac{\mathrm{Var}[Y]}{\epsilon^2} \\
&= \frac{\mathrm{Var}[\frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i)]}{\epsilon^2} \\
&= \frac{\mathrm{Var}[\sum_{i=1}^{m} f(\mathbf{x}_i)]}{m^2\epsilon^2} \\
&\leq \frac{1}{m\epsilon^2}.
\end{aligned}$$

Thus, $\Pr[|Y - \tilde{f}]| \geq \epsilon] \leq \delta$ when $m \geq \frac{1}{\delta\epsilon^2}$.

## 8.8 Biological Applications

### 8.8.1 Docking Problem

### 8.8.2 Flexible Fitting

## Summary

## References and Further Reading

## Exercises

# Chapter 9

# Optimization

## 9.1 Convex and Non-Convex

### 9.1.1 Convex Set

- Convex Set $C$ Line joining all pair of points is always contained in the set C

$$\forall\, x_1, x_2 \in C \quad 0 \leq t \leq 1 \qquad \Longrightarrow \quad x_1 t + x_2(1-t) \in C$$

- Convex combination of $x_1, x_2, \ldots,\ x_k$ is given by

$$x = x_1 t_1 + x_2 t_2 + \ldots + x_k t_k = 1, \quad t_1 + t_2 + \ldots + t_k = 1, \quad t_k \geq 0$$

- For a point set S, **Convex hull**(s): Conv (s) $\overset{def}{=}$ set of all combinations of points in $S$ (This is called Partition of Unity)

- Conic combination of $x_1\ and\ x_2$ is any point $x = t_1 x_1 + t_2 x_2,\ \forall\ t_1 \geq 0,\ t_2 \geq 0$.Convex Cone: set of all conic combinations of points in the set.Note, it generalizes/sweeps out the line segment between $x_1\ and\ x_2$.

- Hyperplane: $\{x \mid w^T x = b, w \neq 01\}$, where $w$ is normal vector.

- Half-space: $\{x \mid w^T x \leq b, w \neq 0\}$.

- Polygons and Polyhedra: Intersection of Half-spaces

- Norm Balls and norm cones are convex. Norm unit ball at origin $\{x \mid \|x\| \leq 1\}$. Norm ball at center $x_c$ and radius $r$ is $\{x \mid \|x - x_c\| \leq\ r\}$, norm cone is defined as $\{(x,t) \mid \|x\|\ \leq\ t\}$

### 9.1.2 Convexity of functions

We say $f\ :\ \mathbb{R}^d \rightarrow \mathbb{R}$ is convex if dom $f$ is convex and

$$f(tx\ +\ (1-t)y)\ \leq\ tf(x)\ +\ (1-t)f(y),\ \ \forall x, y \in \text{dom}\, f,\ \ 0 \leq t \leq 1$$

.

If all inequalities are strict inequalities, i.e.,

$$f(tx\ +\ (1-t)y)\ <\ tf(x)\ +\ (1-t)f(y),\ \ \forall x, y \in \text{dom}\, f,\ \ 0 < t < 1$$

then this is strictly convex. Or, equivalently one can define a convex function as:

**Definition 9.1.** A continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is considered convex if for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

, where $\nabla f(\mathbf{x})$ is the gradient of $f$ at $\mathbf{x}$. Moreover, if $f$ is non-differentiable, one can replace $\nabla f(x)$ with the notion of subgradient.

We can say $f$ is concave if $-f$ is convex.
Here are several examples of convex functions.

- Affine transformation: $f(x) = W^T x \ + \ b$
- Exponential function: $e^{tx}$ for any $t \in R$
- Powers: $x^{\propto}$, $\forall \alpha \geq 1 \ or \ \alpha \leq 0$
- Negative entropy : $x \ln x$
- Norms: $f(X) = \|X\|_2 \ = \ \sigma_{max}(X) \ = \ (\lambda_{max}(X^T X))^{1/2}$, $X$ is a matrix.
- Quadratic Form: $f(x) = \frac{1}{2} x^T P x + q^T x + r$, $\nabla f(x) = \frac{P+P^T}{2} x + q = Px + q$, $H_f = \nabla^2(f(x)) = P$. This is convex if $P \succeq 0$ ($P$ is positive semi-definite)
- Sub-Level Set of $f : \mathbb{R}^d \to \mathbb{R}$ :
  $$C_\alpha = \{x \in dom \ f \mid f(x) \ \leq \ \alpha\}$$
  sub-level sets of convex functions are convex and of co-dimension 1
- Epigraph of $f : \mathbb{R}^d \to \mathbb{R}$. epi = $\{(x, t) \in \mathbb{R}^{n+1} \mid x \in dom \ f, \ f(x) \ \leq \ t\}$. Then $f$ is convex if and only if epi $f$ is a convex set.

## 9.2 Convex Optimization Problems

- Linear Programming
- Quadratic Programming
- Polynomial Optimization
- Geometric Programming
- Semi-Definite Programming

The general convex optimization problem can be formulated as:

$$\begin{aligned} \min \quad & f_0(x) \\ s.t. \quad & f_i(x) \leq 0, \ \forall i = 1, \ldots, m \\ & h_j(x) = 0, \ \forall j = 1, \ldots, p \end{aligned} \tag{9.1}$$

here $x \in R^d$ is an objective variable. $f_0 : \mathbb{R}^d \to \mathbb{R}$ is an objective or cost function. $f_i : \mathbb{R}^d \to \mathbb{R}, \forall i$ are inequality constrains while $h_i : \mathbb{R}^d \to \mathbb{R}$ are equality constraints.
We denote the optimal value of the optimization problem: $p^\star = \inf\{f_0(x) \mid f_i(x) \leq 0, h_j(x) = 0\}$
We say $p^\star = \infty$ if the problem is infeasible (no solution satisfies free constraints) and $p^\star = -\infty$ if the problem is unbounded below

- $x$ is feasible if $x \in dom f_0$ and satisfies all equality and inequality constraints

- a feasible point $x$ is optimal if $f_0(x) = p^\star$ ; $x_{opt}$ = set of all optimal points
- $x$ is **locally optimal** if $\exists\ R > 0$ such that $x$ is optimal for

$$\begin{aligned} \min_{z \in \mathbb{R}^n} \quad & f_0(z) \\ s.t. \quad & f_i(z) \leq 0,\ \forall i = 1, \ldots, m \\ & h_j(z) = 0,\ \forall j = 1, \ldots, p \\ & \|z - x\|_2 \leq R \end{aligned} \tag{9.2}$$

  Here are several examples:

  - $f_0(x) = \frac{1}{x}$, the domain of the function is $dom\ f_0 = \mathbb{R} \backslash \{0\}$, the optimal value is $p^\star = 0$ and no optimal point for $x \in (-\infty, \infty)$.
  - $f_0(x) = -\ln x$, the domain of the function is $dom\ f_0 = \mathbb{R}_+$, the optimal value is $p^\star = -\infty$, the problem is unbounded below.
  - $f_0(x) = x \ln x$, the domain of the function is $dom\ f_0 = \mathbb{R}_+$, the optimal value is $p^\star = -\frac{1}{e}, x = \frac{1}{e}$ is the optimal point.

- General Problem has implicit constraint $\quad x \in D = (\cap_{i=0}^{m}\ dom\ f_i) \cap (\cap_{i=0}^{p}\ dom\ h_i)$
- Unconstrained Problem (m = p = 0) is a special case where the optimization problem can be simplified since no explicit constraints are given. However, they might have implicit constraints. One example would be

$$\min f_0(x) = -\sum_{i=1}^{k} log(b_i - a_i^T x)$$

  Domain of the problem has no explicit constraint but $a_i^T x < b_i$ needs to be satisfied.

If $f_0$ is differentiable, then the optimality criterion for $f_0$ would be:

$$\nabla f_0(x)^T (y - x) \geq 0 \quad \forall \text{feasible } y$$

i.e. $\nabla f_0(x)$ is the tangent Hyperplane to feasible set at $x$.
As an example, let us review the least square problem. The least square problem aims at minimizing $\|A\mathbf{x} - \mathbf{b}\|_2^2$, where $A$ is a constant matrix and $b$ is a constant vector.
The solution of this problem is $x^\star = A^\dagger b$, where $A^\dagger$ is the pseudo inverse of $A$. The time complexity of the computation is $\Theta(n^2\, k)$, if $A \in \mathbb{R}^{k \times n}$

### 9.2.1 Linear Programming(LP)

The primal form of LP is:

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n} \quad & \mathbf{c}^T\mathbf{x} \quad \mathbf{c} \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m \\ s.t. \quad & A\mathbf{x} \leq \mathbf{b} \ \ (\text{or } A\mathbf{x} \geq \mathbf{b} \text{ or } A\mathbf{x} = \mathbf{b}) \\ & \mathbf{x} \geq 0 \end{aligned} \tag{9.3}$$

- Has many applications (industry)
- Proof technique for polynomialness
- Possibility of Polynomial time algorithms

Several algorithms has been made for solving LP problems:

- Simplex Method (Dantzig 1947)
- Ellipsoid Algorithm (Shor, Khachian 1979)
- Interior Point Methods (Karmarkar 1984)

Equivalent form:

| | | | |
|---|---|---|---|
| Max to Min | $\max\ \mathbf{c}^T\mathbf{x}$ | $\leftrightarrow$ | $\min\ -\mathbf{c}^T\mathbf{x}$ |
| Equality to Inequality | $a_i^T\mathbf{x} = b_i$ | $\leftrightarrow$ | $\{a_i^T\mathbf{x} \leq b_i, a_i^T\mathbf{x} \geq b_i\}$ |
| Inequality to Non-negativity | $a_i^T\mathbf{x} \leq b_i$ | $\leftrightarrow$ | $\begin{cases} a_i^T\mathbf{x} + s_i = b_i \\ \mathbf{s} \geq 0 \quad \mathbf{s} \in \mathbb{R}^n \end{cases}$ |
| Variables unrestricted in sign | $x_j$ unrestricted in sign | $\leftrightarrow$ | $\{x_j^+ \geq 0, x_j^- \geq 0\}$ |

### 9.2.2 Duality

Q: Given a solution $\mathbf{x}$ to an LP, how do we decide whether or not $\mathbf{x}$ is in fact an optimum solution?
A: Calculate a lower bound on $\min\ \mathbf{c}^T\mathbf{x}$, given $A\mathbf{x} = b, \mathbf{x} \geq 0$

Suppose $\exists\ \mathbf{y}$ such that $A^T\mathbf{y} \leq \mathbf{c}$, then $\mathbf{y}^T\mathbf{b} = \mathbf{y}^T A\mathbf{x} = (A^T\mathbf{y})^T\mathbf{x} \leq \mathbf{c}^T\mathbf{x}$. Hence $\mathbf{y}^T\mathbf{b}$ is a lower bound on LP; so to get best lower bound, which is the **duality form** of the original linear programming problem. The dual LP problem is defined as:

$$\begin{aligned} \max \quad & \mathbf{b}^T\mathbf{y} \\ s.t. \quad & A^T\mathbf{y} \leq \mathbf{c} \end{aligned} \tag{9.4}$$

The **Weak Duality** condition is related with the following theorem:

**Theorem 9.2.** *(Lower Bound Theorem)*
*If the primal LP problem has an optimum value* $\mathbf{z}$*, then it has a dual LP with optimum value* $\mathbf{w}$ *and* $\mathbf{z} \geq \mathbf{w}$*. Moreover, for infeasible LP problem:*

$$\textit{infeasible min prob} \leftrightarrow value = +\infty$$
$$\textit{unbounded min prob} \leftrightarrow value = -\infty$$
$$\textit{infeasible max prob} \leftrightarrow value = -\infty$$
$$\textit{unbounded max prob} \leftrightarrow value = +\infty$$

There is also a **Strong Duality** condition for the LP problem.

**Theorem 9.3.** *If Primal or Dual is feasible, then* $\mathbf{z} = \mathbf{w}$

In general optimization problem, one can consider the Lagrangian of the optimization problem:

$$L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R} \text{ with } dom\ L = D \times \mathbb{R}^m \times \mathbb{R}^p$$

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) = \sum_{j=1}^{p} \nu_j h_j(x)$$

where $\lambda$ and $\mu$ are Lagrange Multipliers.
The **Lagrange Dual Function** of the optimization problem can be defined as $g : \mathbb{R}^m \times \mathbb{R}^p \rightarrow R$

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$$

The function satisfies Lower Bound Property automatically: if $\lambda \geq 0$ then $g(\lambda, \nu) \leq \mathbf{b}^*$

Let us consider the following optimization as an example:

$$\begin{aligned} \min \quad & \mathbf{x}^T\mathbf{x} \\ s.t. \quad & A\mathbf{x} = \mathbf{b} \end{aligned} \tag{9.5}$$

The Lagrangian is $L(\mathbf{x}, \nu) = \mathbf{x}^T\mathbf{x} + \nu^T(A\mathbf{x} - b)$. Thus, the Lagrangian will reach its minimum if

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \nu) = 2\mathbf{x} + A^T\nu = 0$$

Hence we solved out for $\mathbf{x}$: $\mathbf{x} = -\frac{1}{2}A^T\nu$. If we substitute the equation into $L$, then we could obtain the Lagrangian Dual Function.

$$g(\nu) = L(-\frac{1}{2}A^T\nu, \nu) = -\frac{1}{4}\nu^T AA^T\nu - b^T\nu, \ \forall\ \nu$$

The Lower Bound Property tells us that the optimal value $p^\star$ satisfies:

$$p^\star \geq -\frac{1}{4}\nu^T AA^T\nu - b^T\nu, \ \forall\ \nu$$

### 9.2.3 Non-convex Problems

## 9.3 Combinatorial and Geometric

## 9.4 Biological Applications

### 9.4.1 Fast Computation Methods

## Summary

## References and Further Reading

## Exercises

# Chapter 10

# Statistics

## 10.1 Probability Primer

### 10.1.1 Probability Definitions

Data in bioinformatics is noisy. It can be due to measurement noise or errors in computation. For instance, sometimes we need to take Fourier transform and error may propagate. If we want to prove the effective rate of a drug, we would like to say that the drug has maximum binding affinity. This translates to solving an optimization problem, and we would like to be able to tell how close our solution is to the true maximum. So, we can regard input as random variables with certain mean and variance. For example, each data pixel in an image is a random variable, with some mean and variance. We can then track the propagation of uncertainly in the algorithm. Useful techniques include spectral properties and inequalities for vectors, matrices and tensors.

**Random Variable**

A *random variable* (r.v.) $X$ is values as result of an outcome. A *sample space* is a set of all possible outcomes. $\Pr[x] \in [0, 1]$ is the probability of occurrence of each outcome. A function that assigns probabilities is called a *probability distribution function*. For example, the uniform distributions, the Gaussian distributions (which have nice concentration properties), and the Poisson distributions (which often appear in image pixels because they count the number of hits over time).
Sometimes, probability mass function does not exists. For example, if $X$ has uniform probability in $[0, 1]$, then $\Pr[X = 0.527] = \frac{1}{\infty} = 0$. We can define *probability density function* (pdf) $p$ such that

$$\Pr[a \leq X \leq b] = \int_a^b p(x)\ \mathrm{d}\,x.$$

Notice that the integral is a linear operator on $p$. The *cummulative distribution function* (cdf) is

$$P(a) = \Pr[X \leq a] = \int_{-\infty}^a p(x)\ \mathrm{d}\,x.$$

Figure 10.1(a) shows the integrals as areas under the probability density function.
An *event* is a subset of the sample space. For example, if we have $n$ unbiased coin flips giving random variables $x_1, x_2, \ldots, x_n \in \{0, 1\}$, then the sample space is $\{0, 1\}^n$. The event of an odd number of ones occurring in the sequence consists of elements in $\{0, 1\}^n$.

**Independence**

For two events $A$ and $B$, we define the *conditional probability* as

$$\Pr[A|B] = \Pr[A \cap B] / \Pr[B],$$

where $\Pr[A \cap B]$ is the probability of the joint occurrence of the events.

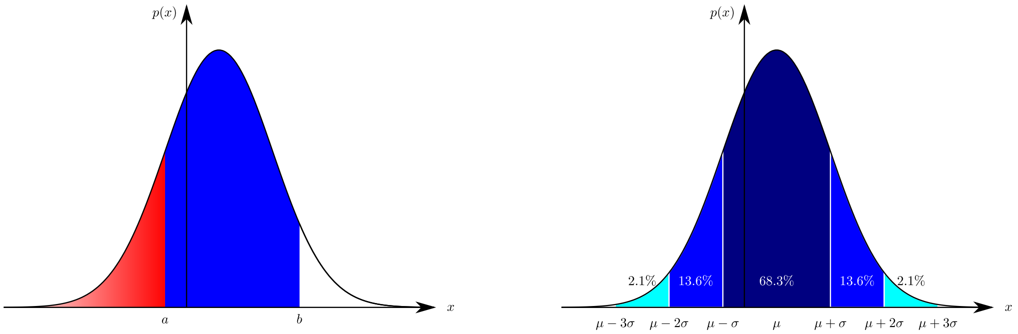


Figure 10.1: (a) The probability $\Pr[a \leq X \leq b] = \int_a^b p(x)\,\mathrm{d}\,x$ and the cummulative probability $P(a) = \Pr[X \leq a] = \int_{-\infty}^a p(x)\,\mathrm{d}\,x$ as areas under the probability density function $p(x)$. (b) Approximately 68.3%, 95.4% and 99.7% of probability mass of a Gaussian within $\sigma$, $2\sigma$ and $3\sigma$ from the mean $\mu$.

We say that two events $A$ and $B$ are *independent* if $\Pr[A \cap B] = \Pr[A]\Pr[B]$, or equivalently $\Pr[A|B] = \Pr[A]$. A sequence of $n$ random variables $x_1, x_2, \ldots, x_n$ are *mutually independent* if for all possible $A_1, A_2, \ldots, A_n$, of values of $x_1, x_2, \ldots, x_n$,

$$\Pr[x_1 \in A_1, x_2 \in A_2, \ldots, x_n \in A_n] = \Pr[x_1 \in A_1]\Pr[x_2 \in A_2]\ldots\Pr[x_n \in A_n].$$

Notice that pairwise independence (or even $k$-wise independence) is weaker than mutual independence.
Sometime, our sample points are not mutually independent. Suppose we want to generate sample points from the $d$-dimensional cube leaving no large gap. A naive way would be to use a regular lattice (Figure 10.2(a)). A better way used in Quasi Monte Carlo method is to generate a sample of bounded discrepancy (Figure 10.2(b)).

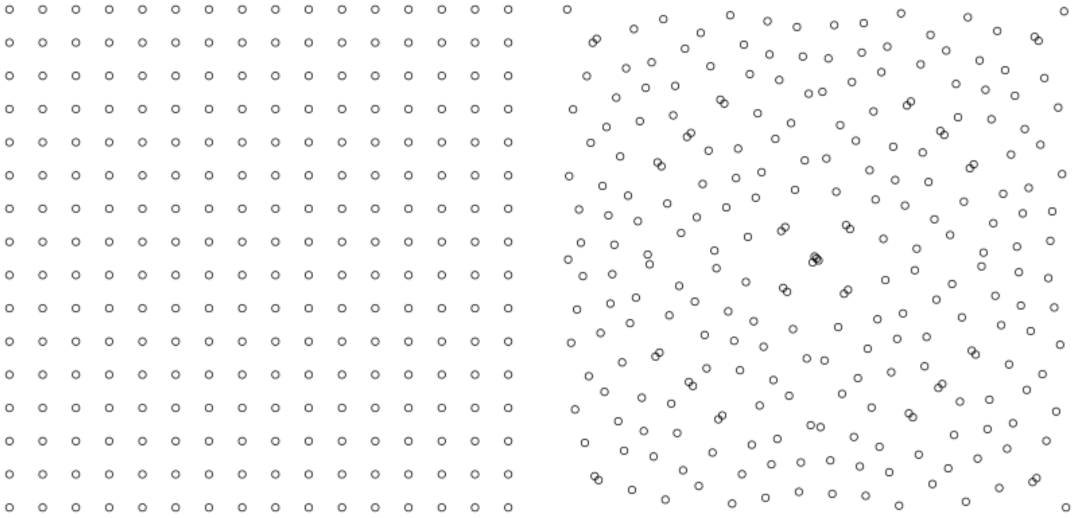


Figure 10.2: (a) Samples from a regular lattice. (b) Low discrepancy samples from the Sobol sequence.

Notice that normalization kills independence. If $x, y \in \mathbb{R}$ are independent, then

$$\text{normalize}(x, y) = \left( \frac{x}{\sqrt{x^2+y^2}}, \frac{y}{\sqrt{x^2+y^2}} \right)$$

may no longer be independent.

### Expectation

The *expectation* of a random variable $X$ with pdf $p$ is defined as

$$E[X] = \int_{-\infty}^{+\infty} xp(x) \,\mathrm{d}\, x.$$

The linearity of expectation

$$E[X_1 + X_2 + \ldots + X_n] = E[X_1] + E[X_2] + \ldots + E[X_n]$$

holds even without independence.
The union bound

$$\Pr[A_1 \cup A_2 \cup \ldots \cup A_n] \le \sum_{i=1}^{n} \Pr[A_i]$$

is an upper bound of the unions of events.
The inclusion-exclusion principle says that

$$\Pr[A_1 \cup A_2 \cup \ldots \cup A_n] = \sum_{i=1}^{n} \Pr[A_i] - \sum_{i<j} \Pr[A_i \cap A_j] + \sum_{i<j<k} \Pr[A_i \cap A_j \cap A_k] - \ldots.$$

One application of the inclusion-exclusion principle is volume calculation of molecules represented as a union of atoms. An atom consists of a nucleus in its center surrounded by a electron cloud, which can be represented by a ball with radius equal to the range of its van der Waals force. The atoms are bonded together, forming a geometry of union of balls. Examples include NaCl salt, protein, and water molecule ($H_2O$) which polarizes like a magnet with Hydrogen (H) positively charged and Oxygen (O) negatively charged. Two (or a small number of) balls may overlap each other. Since the volume is proportional to finding electrons in certain region, we can apply the inclusion-exclusion principle.

### Variance

The *variance* of a random variable $X \in \mathbb{R}$ is given by

$$\begin{aligned} Var(X) = \sigma^2(X) &= E[X - E^2[X]]^2 \\ &= E[X^2] - 2E[X]E[X] + E^2[X] \\ &= E[X^2] - E^2[X]. \end{aligned}$$

For a Gaussian random variable, its standard deviation $\sigma$ tells that more than $68\%$ of the probability mass is with $\sigma$ from its mean. For $2\sigma$ and $3\sigma$ from the mean, the probability masses are more than $95\%$ and $99\%$ respectively. (Figure 10.1(b))
In general, $Var(X_1 + X_2) \neq Var(X_1) + Var(X_2)$. However, equality holds if $X_1$ and $X_2$ are pairwise independent. In fact, if $X_1, X_2, \ldots, X_n$ are pairwise independent, then

$$Var(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} Var(X_i).$$

## 10.1.2 Probability Distributions

### Gaussian Distributions

The *Gaussian distribution* is related to Central Limit Theorem.

**Theorem 10.1** (Central Limit Theorem [31, Theorem 12.2])**.** *If $X_1, \ldots, X_n \in \mathbb{R}$ is a sequence of independent identically distributed (i.i.d.) random variables each with mean $\mu$ and variance $\sigma^2$, then*

$$X = \frac{1}{\sqrt{n}} \left( \sum_{i=1}^{n} X_i - n\mu \right)$$

*converges to the distribution $N(0, \sigma^2)$.*

The univariate Gaussian distribution $N(\mu, \sigma^2)$ is given by the pdf

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

For $d$-variate Gaussian distribution $N(\mu, \Sigma)$ where $\mu \in \mathbb{R}^d$ is the mean vector and $\Sigma \in \mathbb{R}^{d\times d}$ is the covariance matrix, the pdf is given by

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} Exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^{\mathfrak{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right].$$

When $d = 3$, an *isotropic* Gaussian has $4$ degrees of freedom, corresponding to the number of parameters necessary to define a sphere in $\mathbb{R}^3$. Meanwhile, an *anisotropic* Gaussian would have $9 = \binom{2+3}{2} - 1$ degrees of freedom, corresponding to the number of parameters necessary to define an ellipsoid. If the ellipsoid is *isothetic*, then the degree of freedom reduces to $6$.

**Binomial Distributions**

A *Bernoulli distribution* is a stochastic process with two outcomes

$$X = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob. } 1-p \end{cases}.$$

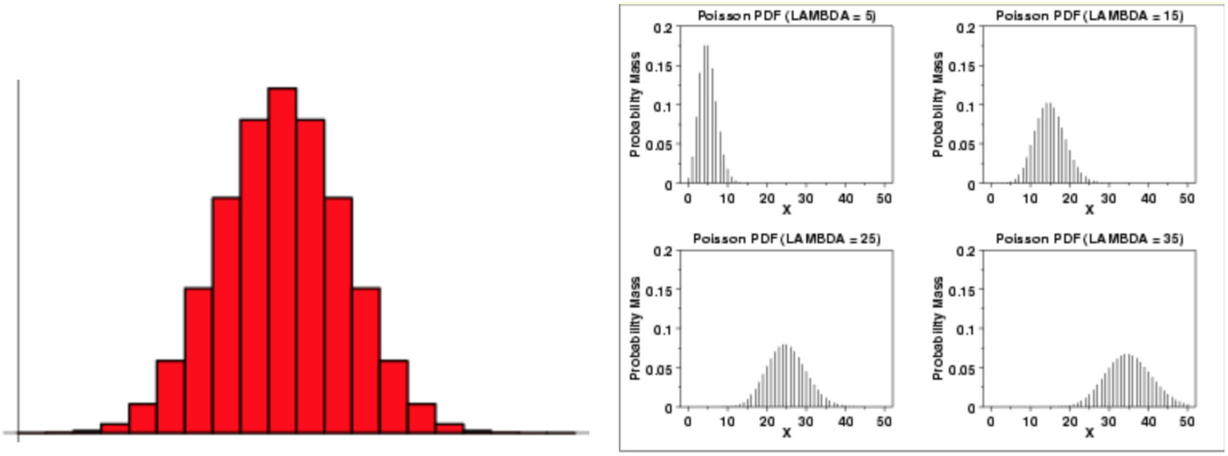


Figure 10.3: (a) The above plot shows the Binomial distribution of exactly $n$ successes out of N = 20 trials with p = q = $\frac{1}{2}$ . (b) The above plot shows the Poisson distribution for four different values of $\lambda$.

The *binomial distribution* $Bin(n, p)$ counts the number $X$ of ones in $n$ independent Bernoulli trials.

$$\Pr[X = k] = \Pr[\text{(total number of ones)} = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

It has mean $np$ and variance $np(1-p)$. It also satisfies the property that if $X \sim Bin(n_1, p)$ and $Y \sim Bin(n_2, p)$ are independent, then $X + Y \sim Bin(n_1 + n_2, p)$.

**Poisson Distribution**

Let $\lambda$ be the average rate per unit of time and $n$ be the number of division of a unit time interval into segments, where the probability of two events occurring in the same segment is negligible. The Poisson distribution counts the number $X$ of events occurring in a unit of time as $n \to \infty$. It is the limit of $Bin(n, p = \lambda/n)$. For the Poisson distribution both the mean and variance equal $\lambda$.

$$\Pr[X = k] = \Pr[k \text{ events occurs in a unit of time}] = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

### 10.1.3 Pairwise independence

A set of events $E_1, E_2, \ldots, E_n$ are *mutually independent* if for any subset $I \subseteq \{1, 2, \ldots, n\}$,

$$\Pr\left[\bigcap_{i\in I} E_i\right] = \prod_{i\in I} \Pr[E_i].$$

A set of discrete random variables $X_1, X_2, \ldots, X_n$ are *mutually independent* if for any subset $I \subseteq \{1, 2, \ldots, n\}$,

$$\Pr\left[\bigwedge_{i\in I} X_i = x_i\right] = \prod_{i\in I} \Pr[X_i = x_i].$$

For example, if $X = (X_1, X_2, \ldots, X_n) \sim \mathrm{N}(\mathbf{0}, I)$, then $X_i \sim \mathrm{N}(0, 1)$ are mutually independent. If $X = (X_1, X_2, \ldots, X_n)$ is uniformly distributed on the hypercube $[0, 1]^n$, then $X_i \sim \mathrm{U}(0, 1)$ are mutually independent.
A set of events $E_1, E_2, \ldots, E_n$ are *$k$-wise independent* if for any subset $I \subseteq \{1, 2, \ldots, n\}$ such that $|I| \leq k$,

$$\Pr\left[\bigcap_{i\in I} E_i\right] = \prod_{i\in I} \Pr[E_i].$$

A set of discrete random variables $X_1, X_2, \ldots, X_n$ are *$k$-wise independent* if for any subset $I \subseteq \{1, 2, \ldots, n\}$ such that $|I| \leq k$,

$$\Pr\left[\bigwedge_{i\in I} X_i = x_i\right] = \prod_{i\in I} \Pr[X_i = x_i].$$

When $k = 2$, it is also called *pairwise independence*. In other words, $X_1, X_2, \ldots, X_n$ are pairwise independent if for all $i \neq j$ and any pair of values $(a, b)$,

$$\Pr[X_1 = a \wedge X_2 = b] = \Pr[X_1 = a] \Pr[X_2 = b].$$

**Lemma 10.2** ([156, Lemma 13.1])**.** *We can generate $m = 2^n - 1$ uniform pairwise independent bits from $n$ uniform mutually independent bits. (A random bit is uniform if it assumes values $0$ or $1$ with equal probability $1/2$).*

*Idea of proof.* Generate $n$ uniform random bits $X_1, X_2, \ldots, X_n \in \{0, 1\}$. Enumerate all $(2^n - 1)$ non-empty subsets of $\{1, 2, \ldots, n\}$. Let $S_j$ be the $j^{\text{th}}$ subset of the enumeration. Then, set

$$\begin{aligned} Y_j &= \left(\sum_{i\in S_j} X_i\right) \mod 2 \\ &= \bigoplus_{i\in S_j} X_i, \end{aligned}$$

where $\oplus$ denotes XOR. □

**Lemma 10.3** ([156, Lemma 13.2])**.** *Let $X_1$ and $X_2$ be independent and uniform over $\mathrm{GF}(p)$. Generate $Y_i = (X_1 + iX_2) \bmod p$ for $i \in \{0, 1, 2, p - 1\}$. Then, $Y_0, Y_1, Y_2, \ldots, Y_{p-1}$ are pairwise independent.*

*Proof.* For a given $X_2$, we know that $Y_i$ is uniform over $\mathrm{GF}(p)$ as $X_1$ is uniform over $\mathrm{GF}(p)$.
Consider $Y_i$ and $Y_j$, where $i \neq j$. For any $a, b \in \{0, 1, 2, \ldots, p - 1\}$,

$$\begin{aligned} &Y_i = X_1 + iX_2 = a \text{ and } Y_j = X_1 + jX_2 = b \\ \iff &X_2 = \frac{b - a}{j - i} \text{ and } X_1 = a - \frac{i(b - a)}{j - i}. \end{aligned}$$

Hence, $\Pr[Y_i = a \wedge Y_j = b] = \frac{1}{p^2}$. □

**Finite Field**

Finite fields appear in Rijndael — the AES cryptographic system. The Reed Solomon code uses the Galois field $\mathrm{GF}(2^n)$, which is a field of characteristic 2.
The field $\mathrm{GF}(2^n)$ consists over polynomials

$$P(x) = \sum_{i=0}^{n-1} c_i x^i \qquad\qquad (c_i \in \mathrm{GF}(2) = \{0,1\})$$

of degree less than $n$ over the field $\mathrm{GF}(2)$ modulo an irreducible polynomial over $\mathrm{GF}(2)$. (For example, the polynomial $x^2 + 2x + 1 = (x+1)^2$ is reducible over $\mathbb{R}$. The polynomial $x^2 + 1 = (x+i)(x-i)$ is irreducible over $\mathbb{R}$, but reducible over $\mathbb{C}$.)
The $2^n$ polynomials can be encoded by $2^n$ strings of bits. For example, $x^7 + x^6 + x^4 + 1$ can be encoded by the bit string `11010001` of length $8$.
In algebraic geometry, we can study polynomials modulo over the sphere. The famous result of Bézout Theorem says that a curve of degree $d$ and a curve of degree $e$, with some caveats, intersect at $(d \cdot e)$ points.

### 10.1.4 Transformation of Random Variables

Consider next a transformation of random variables. We shall revisit the Box-Muller method from lecture 3, and prove that it indeed uniformly samples the Gaussian.
Suppose we generate a sample point $(X_1, X_2)$ according to the joint pdf $\mu(x_1, x_2)$, and then apply an injective function $f$ to obtain $(Y_1, Y_2) = f(X_1, X_2)$. It can be shown that the sample $(Y_1, Y_2)$ follows the joint pdf

$$\rho(y_1, y_1) = \mu(f^{-1}(y_1, y_2)) \cdot \left|\det\left(\frac{\partial\,\mathbf{x}}{\partial\,\mathbf{y}}\right)\right|,$$

where the Jacobian matrix is defined as

$$\frac{\partial\,\mathbf{x}}{\partial\,\mathbf{y}} = \begin{pmatrix} \frac{\partial\,x_1}{\partial\,y_1} & \frac{\partial\,x_1}{\partial\,y_2} \\ \frac{\partial\,x_2}{\partial\,y_1} & \frac{\partial\,x_2}{\partial\,y_2} \end{pmatrix}.$$

Similar techniques also work in $\mathbb{R}^d$.
The Box-Muller method transforms a sample point $(X_1, X_2)$ generated uniformly random from $(0,1)^2$, and transforms it to

$$(Y_1, Y_2) = \Bigg( \sqrt{-2\ln X_1}\cos(2\pi X_2)\,,\ \sqrt{-2\ln X_1}\sin(2\pi X_2) \Bigg).$$

We now calculate the partial derivatives and verify that the resulting $(Y_1, Y_2)$ distributes according to $N(\mathbf{0}, I)$. First, we express $x_1$ and $x_2$ in terms of $y_1$ and $y_2$.

$$\begin{aligned} x_1 &= \exp(-(y_1^2 + y_2^2)/2) \\ x_2 &= \arctan(y_2/y_1)/(2\pi) \end{aligned}$$

Then, we calculate the partial derivatives.

$$\begin{aligned} \frac{\partial\,x_1}{\partial\,y_1} &= -\,y_1\exp(-(y_1^2 + y_2^2)/2) \\ \frac{\partial\,x_1}{\partial\,y_2} &= -\,y_2\exp(-(y_1^2 + y_2^2)/2) \\ \frac{\partial\,x_2}{\partial\,y_1} &= \frac{-y_2/y_1^2}{(2\pi)(1 + y_2^2/y_1^2)} \\ \frac{\partial\,x_2}{\partial\,y_2} &= \frac{1/y_1}{(2\pi)(1 + y_2^2/y_1^2)} \end{aligned}$$

Hence, the Jacobian determinant is

$$\begin{aligned}
\left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right| &= \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} \\
&= \begin{vmatrix} -y_1 \exp(-(y_1^2+y_2^2)/2) & -y_2 \exp(-(y_1^2+y_2^2)/2) \\ \frac{-y_2/y_1^2}{(2\pi)(1+y_2^2/y_1^2)} & \frac{1/y_1}{(2\pi)(1+y_2^2/y_1^2)} \end{vmatrix} \\
&= \frac{\exp(-(y_1^2+y_2^2)/2)}{(2\pi)(1+y_2^2/y_1^2)} \begin{vmatrix} -y_1 & -y_2 \\ -y_2/y_1^2 & 1/y_1 \end{vmatrix} \\
&= \frac{\exp(-(y_1^2+y_2^2)/2)}{(2\pi)(1+y_2^2/y_1^2)} \cdot (-1 - y_2^2/y_1^2) \\
&= -\exp(-(y_1^2+y_2^2)/2)/(2\pi)
\end{aligned}$$

Therefore, the pdf of $(Y_1, Y_2)$ is

$$\mu(f^{-1}(\mathbf{y})) \cdot \left|\det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right)\right| = 1 \cdot \frac{1}{2\pi} e^{-(y_1^2+y_2^2)/2} = \frac{1}{2\pi} e^{-(y_1^2+y_2^2)/2},$$

which shows that $(Y_1, Y_2) \sim \mathrm{N}(\mathbf{0}, I)$.

### 10.1.5 Annular Concentration of Gaussian

A one-dimensional Gaussian has its mass close to its mean (Figure 10.1(b)). However, for large $d$, a $d$-dimensional Gaussian $N(\mathbf{0}, \sigma^2 I)$ with pdf

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sigma^d} e^{-\|\mathbf{x}\|^2/(2\sigma^2)}$$

has very little mass close to the origin, even though its maximum probability density is at the origin. In fact,

$$\int_0^1 p(r)\ \mathrm{d}\,r$$

is vanishing small, where $r$ is the $Ell_2$ distance from the center (Figure 10.4(b)) and $p(r)$ is the marginal probability density (Figure 10.4(c)). For $N(\mathbf{0}, I)$, we know that

$$p(r) \propto r^{d-1} e^{r^2/2}.$$

So, where is the maximum mass? We can set the derivative to zero.

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\,r} r^{d-1} e^{r^2/2} = (d-1) r^{d-2} e^{-r^2/2} - r^d e^{-r^2/2} &= 0 \\
r^2 &= d-1 \\
r &= \sqrt{d-1}
\end{aligned}$$

So, we need $r \approx \sqrt{d}$ to see significant probability mass. For $r \ll \sqrt{d}$, the mass is non-significant. For $r \gg \sqrt{d}$, mass also disappears.

**Theorem 10.4** ([31, Theorem 2.9])**.** *Let* $X = X_1 + X_2 + \ldots + X_n$*, where* $X_i$ *are mutually independent with mean* 0 *and variance at most* $\sigma^2$*. Let* $0 \le a \le \sqrt{2}n\sigma^2$*. Suppose* $|E[X_i^s]| \le \sigma^2 s!$ *for all* $3 \le s \le (a^2/(4n\sigma^2))$*, then*

$$\Pr[|X| \ge a] \le 3e^{-a^2/(12n\sigma^2)}.$$

We will prove the above theorem in the next lecture, and using Markov inequality. Here we use this theorem to prove the multivariate spherical Gaussian Annulus Theorem. Recall the intimate relationship of the level sets of spherical Gaussian in $R^d$ and balls in $R^d$ . So similar to the Theorem 2.8 of BHK for unit balls, we discussed in lecture 2, we have

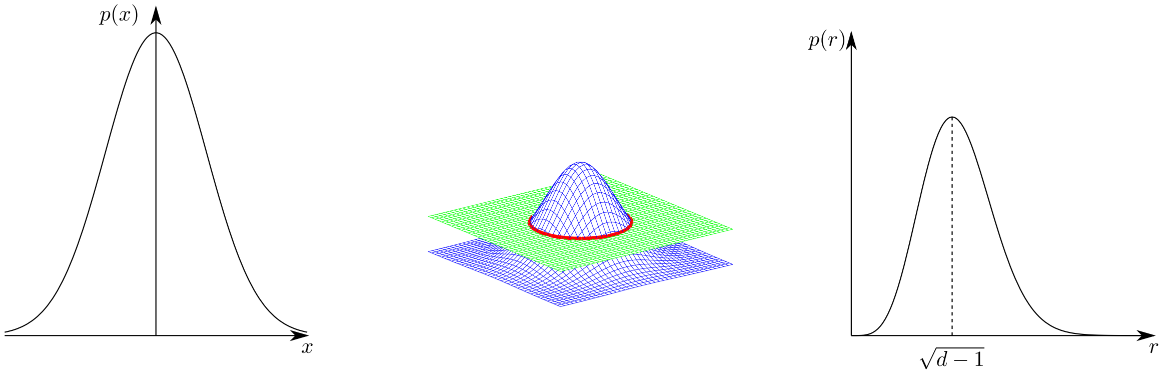


Figure 10.4: (a) A univariate Gaussian concentrated around its mean. (b) Spherical level sets of $Ell_2$ distance. (c) The marginal distribution $p(r)$ with peak at $r = \sqrt{d-1}$.

**Theorem 10.5** (Gaussian Annulus Theorem [31, Theorem 2.8])**.** *For a $d$-dimensional spherical Gaussian $N(\mathbf{0}, I)$ and $c \le \sqrt{d}$, all but $3e^{-c^2/96}$ of the probability mass lies within an annulus of $\sqrt{d} - c \le r \le \sqrt{d} + c$.*

*Proof.* For a point $\mathbf{x} = (x_1, x_2, \ldots, x_d) \sim N(\mathbf{0}, I)$, we have

$$r^2 = \|\mathbf{x}\|^2 = x_1^2 + x_2^2 + \ldots + x_d^2.$$

Consider

$$\begin{aligned} & |r - \sqrt{d}| \ge c \\ \implies & r^2 - d = |r - \sqrt{d}| \cdot |r + \sqrt{d}| \ge c\sqrt{d} && \text{(since } |r+\sqrt{d}| \ge \sqrt{d}) \\ \implies & |y_1 + y_2 + \ldots + y_d| \ge c\sqrt{d} && (y_i = x_i^2 - 1) \\ \implies & |z_1 + z_2 + \ldots + z_d| \ge c\sqrt{d}/2. && (z_i = y_i/2) \end{aligned}$$

To use the above theorem, we bound the moments of $z_i$ using the Gamma integral.

$$\begin{aligned} |E[z_i^s]| &= 2^{-s} E[|y_i|^s] \\ &\le 2^{-s} E[1 + x_i^2] \\ &= 2^{-s} + 2^{-s}\sqrt{\frac{2}{\pi}} \int_0^\infty x^{2s} e^{-x^2/2} \,\mathrm{d}\,x \\ &= 2^{-s} + \frac{1}{\sqrt{\pi}} \int_0^\infty u^{s-0.5} e^{-u} \,\mathrm{d}\,u && \text{(substitue } x = \sqrt{2u}) \\ &\le s! \end{aligned}$$

Hence, we can apply the above theorem with $\sigma^2 = E[z_i^2] \le 4$ and $|E[z_i^s]| \le 4(s!)$.

$$\Pr[|z_1 + z_2 + \ldots + z_d| \ge c\sqrt{d}/2] \le 3e^{-c^2/96}. \qquad \square$$

### 10.1.6 Distribution Sampling

How can we sample points from a given distribution, for example $N(\mathbf{0}, I)$? If $x \in \mathbb{R}$ has pdf $p(x)$, then we can define use its cdf

$$\begin{aligned} P\colon \mathbb{R} &\to \mathbb{R} \\ x &\mapsto P(x) = \int_{-\infty}^{x} p(t) \,\mathrm{d}\,t. \end{aligned}$$

If $u$ is uniformly sampled from $[0,1]$, then $x = P^{-1}(u)$ will be distributed according to pdf $p(x)$. So, for $d$-dimensional Gaussian $N(\mathbf{0}, I)$, we can draw $\mathbf{u} = (u_1, u_2, \ldots, u_d)$ uniformly from $[0,1]^d$ and take inverse of the cdf componentwise $\mathbf{x} = (x_1, x_2, \ldots, x_d) = (P^{-1}(u_1), P^{-1}(u_2), \ldots, P^{-1}(u_d))$.
Another example is the Cauchy distribution with pdf

$$p(x) = \frac{1}{\pi(1+x^2)}.$$

Its cdf is given by

$$P(x) = \int_{-\infty}^{x} \frac{1}{\pi(1+t^2)}\,\mathrm{d}\,t = \frac{1}{\pi}\arctan x + \frac{1}{2}.$$

**Box-Muller Method**

Independent samples $X_1, X_2 \sim U(0,1)$ can be used to generate samples $(Y_1, Y_2)$ of a bivariate Gaussian distribution $N(\mathbf{0}, I)$ using the Box-Muller method as follows.

$$\begin{aligned} Y_1 &= \sqrt{-2\ln X_1}\cos(2\pi X_2) \\ Y_2 &= \sqrt{-2\ln X_1}\sin(2\pi X_2) \end{aligned}$$

Then, we have the following.

$$\begin{aligned} X_1 &= e^{-(Y_1^2+Y_2^2)/2} \\ X_2 &= \arctan(Y_2/Y_1) \end{aligned}$$

The Jacobian determinant equals

$$J = \left(\frac{1}{\sqrt{2\pi}}e^{-y_1^2/2}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-y_2^2/2}\right).$$

### 10.1.7 Mixture of Gaussians

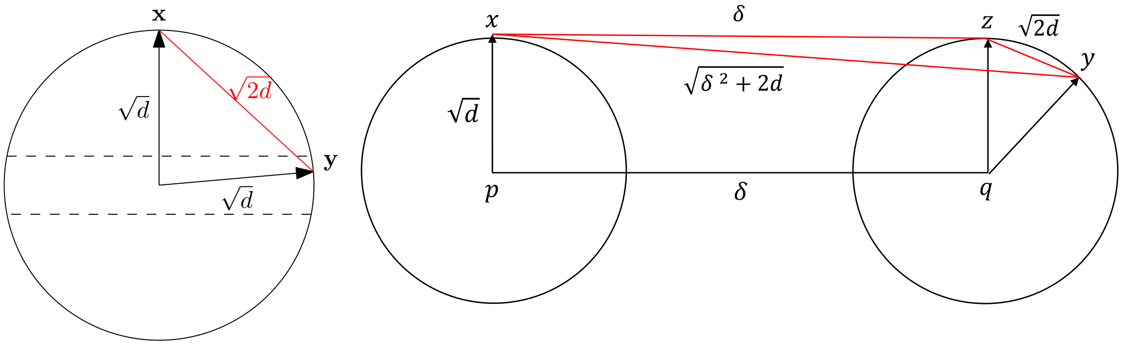


Figure 10.5: (a) Nearly orthogonal Gaussian samples $\mathbf{x}$ and $\mathbf{y}$. (b) Samples $\mathbf{x}$ and $\mathbf{y}$ from two Gaussians centered at $\mathbf{p}$ and $\mathbf{q}$ respectively.

Given a mixture of two Gaussian densities

$$p(\mathbf{x}) = w_1 p_1(\mathbf{x}) + w_2 p_2(\mathbf{x}),$$

where $w_1 + w_2 = 1$ is a convex combination. It can be shown that if the means of the $d$-dimension spherical unit-variance Gaussians are separated by $\Omega(d^{1/4})$, then they are separable. The idea is that with high probability, points in the same cluster belong to the same Gaussian because most of the points are concentrated. More formally, with high probability $\|\mathbf{x} - \mathbf{y}\|^2 =$

$2d \pm O(\sqrt{d})$ if they come from the same Gaussian, and $\|\mathbf{x}-\mathbf{y}\|^2 = \delta^2 + 2d \pm O(\sqrt{d})$ if they come from different Gaussian separated by $\delta$.

Suppose $\mathbf{x}, \mathbf{y} \sim N(\mu, I)$ come from the same Gaussian (Figure 10.5(a)). Observe that most probability mass lies in an annulus of width $O(1)$ and radius $\sqrt{d-1}$. Rotate the coordinate system so that $\mathbf{x}$ is at the North pole. With high probability, $\mathbf{y}$ is in the slab $\{(y_1, y_2, \ldots, y_d) \colon -c \leq y_1 \leq c\}$ for some $c \in O(1)$. So, $\mathbf{y}$ is nearly orthogonal to $\mathbf{x}$ and hence $\|\mathbf{x}-\mathbf{y}\| \approx \sqrt{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2}$. More precisely, we can further rotate the coordinate system so that

$$\mathbf{x} = (\sqrt{d} \pm O(1), 0, 0, \ldots, 0) \qquad \text{and} \qquad \mathbf{y} = (\pm O(1), \sqrt{d} \pm O(1), 0, \ldots, 0).$$

Hence, $\|\mathbf{x}-\mathbf{y}\|^2 = (d \pm O(\sqrt{d})) + (d \pm O(\sqrt{d})) = 2d \pm O(\sqrt{d})$. See also (Figure 10.5(b))

Suppose $\mathbf{x} \sim N(\mathbf{p}, I)$ and $\mathbf{y} \sim N(\mathbf{q}, I)$ come from different Gaussians . With high probability, $\mathbf{x}$ and $\mathbf{y}$ lies in an annulus of of width $O(1)$ and radius $\sqrt{d-1}$ centered at $\mathbf{p}$ and $\mathbf{q}$ respectively. Also, $(\mathbf{x}-\mathbf{p}), (\mathbf{p}-\mathbf{q}), (\mathbf{q}-\mathbf{y})$ are nearly mutually perpendicular. Hence, $\|\mathbf{x}-\mathbf{y}\|^2 \approx \|\mathbf{x}-\mathbf{p}\|^2 + \delta^2 + \|\mathbf{q}-\mathbf{y}\|^2 = \delta^2 + 2d \pm O(\sqrt{d})$.

Thus, if $\delta = \Omega(d^{1/4})$, we can separable constant number of points with constant failure probability.

In general, we can ask the question of separating many Gaussians. It can be solved if they are well-separated. One application of separating a mixture of Gaussians is in locating the position of atoms by sampling, which is possible if the separation of atoms in the lattice is large enough.

### 10.1.8 Concentration Theorems

The following probability inequalities by Markov and Chebyshev are used to prove the **Law of Large Numbers**.

**Theorem 10.6** (Markov Inequality [31, Theorem 2.1] [156, Theorem 3.1])**.** *Let $X \geq 0$ be a random variable and $a > 0$.*

$$\Pr[X \geq a] \leq \mathrm{E}[X]/a.$$

*Proof.* For a continuous non-negative random variable $x$ with probability density $p(x)$

$$\begin{aligned}
\mathrm{E}[X] &= \int_0^\infty xp(x)\,\mathrm{d}\,x \\
&= \int_0^a xp(x)\,\mathrm{d}\,x + \int_a^\infty xp(x)\,\mathrm{d}\,x \\
&\geq \int_a^\infty xp(x)\,\mathrm{d}\,x \\
&\geq \int_a^\infty ap(x)\,\mathrm{d}\,x \\
&= a\int_a^\infty p(x)\,\mathrm{d}\,x \\
&= a\Pr[X \geq a]
\end{aligned}$$

□

Note the proof works for discrete probability distributions. Replace summations for all the integrals.

**Corollary 10.7.** *Let $X$ be a non-negative random variable and $c > 0$. Then,* $\Pr[X \geq c\,\mathrm{E}[X]] \leq 1/c$.

The above says that the value of $X$ is not far from the mean $E(X)$. The Chebyshev's inequality (below) can be proved by applying Markov's inequality on the variance.

**Theorem 10.8** (Chebyshev Inequality [31, Theorem 2.3] [156, Corollary 3.7])**.** *Let $X$ be a random variable with mean $m$ and variance $\sigma^2$. Then, for all $a > 0$,*

$$\Pr[\,|X-m| \geq a\sigma\,] \leq 1/a^2.$$

Using the Chebyshev inequality, we can now prove the Law of Large Numbers.

**Theorem 10.9** (**Law of Large Numbers**)**.** *Let $S$ be the sample mean of $n$ independent random variables $X_1, X_2, \ldots, X_n$ with means* $\mathrm{E}[X_i] = m$ *and variances* $\mathrm{Var}[X_i] = \sigma^2$. *Then, for all $\epsilon > 0$,*

$$\Pr[|S-m| \geq \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}.$$

*Proof.* Applying Chebyshev Inequality on $S$ with $a = \epsilon/\sigma(S)$, we get

$$\begin{aligned}\Pr\left[|S-m| \geq \epsilon\right] &\leq \sigma^2(S)/\epsilon^2 \\ &= \frac{1}{\epsilon^2}\sigma^2\left(\frac{X_1+X_2+\ldots+X_n}{n}\right) \\ &= \frac{1}{n^2\epsilon^2}\sigma^2(X_1+X_2+\ldots+X_n) \\ &= \frac{\sigma^2}{n\epsilon^2}.\end{aligned}$$

□

### 10.1.9 Application of Markov and Chebyshev Inequalities

**Theorem 10.10** ([31, Theorem 2.9]). *Let $X = X_1 + X_2 + \ldots + X_n$, where $X_i$ are mutually independent with mean 0 and variance at most $\sigma^2$. Let $0 \leq a \leq \sqrt{2}n\sigma^2$. Suppose $|\mathrm{E}[X_i^s]| \leq \sigma^2 s!$ for all $3 \leq s \leq (a^2/(4n\sigma^2)$, then*

$$\Pr[|X| \geq a] \leq 3e^{-a^2/(12n\sigma^2)}.$$

*Proof.* We will bound $|\mathrm{E}[X^r]|$ and use Markov inequality.
Consider the expansion

$$(X_1+X_2+\ldots+X_n)^r = \sum_{\sum r_i = r}\binom{r}{r_1\ r_2\ \ldots\ r_n}X_1^{r_1}X_2^{r_2}\ldots X_n^{r_n},$$

where $\binom{r}{r_1\ r_2\ \ldots\ r_n} = \frac{r!}{r_1!\cdot r_2!\cdot\ldots\cdot r_n!}$.
Since $X_i$ are independent,

$$\mathrm{E}\left[(X_1+X_2+\ldots+X_n)^r\right] = \sum_{\sum r_i = r}\binom{r}{r_1\ r_2\ \ldots\ r_n}\mathrm{E}[X_1^{r_1}]\,\mathrm{E}[X_2^{r_2}]\ldots\mathrm{E}[X_n^{r_n}].$$

Note that $\mathrm{E}[X_i] = 0$. So, all terms with $r_i = 1$ are zero. So, all non-zero terms have $r_i \geq 2$ or $r_i = 0$. Since $\sum_i r_i = r$, this implies that in a non-zero term, there are at most $r/2$ non-zero indices $r_i$.
The number of non-zero terms with exactly $t$ indices $r_i \geq 2$ equals

$$\binom{n}{t}\binom{r-t-1}{t-1},$$

because there are $\binom{n}{t}$ ways to choose a subset of $\{1, 2, \ldots, n\}$ of cardinality $t$ corresponding to the $t$ indices with weight at least 2, and there are $\binom{(r-2t)+(t-1)}{t-1} = \binom{r-t-1}{t-1}$ ways to allocate the remaining $(n-2t)$ weights. (This is analogous to the fact that the number of monic monomials in a polynomial of degree at most $e = r - 2t$ in $n = t - 1$ variables equals $\binom{e+n}{n} = \binom{e+n}{e}$.)
Also using $|\mathrm{E}[X_i]^{r_i}| \leq \sigma^2 r_i!$, we get

$$\begin{aligned}\mathrm{E}[X^r] &\leq r!\sum_{t=1}^{r/2}\binom{n}{t}\binom{r-t-1}{t-1}\sigma^{2t} \\ &\leq r!\sum_{t=1}^{r/2}\frac{(n\sigma^2)^t}{t!}2^{r-t-1}.\end{aligned}$$

Let $h(t) = \frac{(n\sigma^2)^t}{t!}2^{r-t-1}$. Since $a \leq \sqrt{2}n\sigma^2$ and $s \leq a^2/(4n\sigma^2)$, we have $r \leq s \leq n\sigma^2/2$. For $t \leq r/2$, increasing $t$ by 1 increases $h(t)$ by $n\sigma^2/(2t) \geq 2$. Thus,

$$\begin{aligned}\mathrm{E}[X^r] &\leq r!\sum_{t=1}^{r/2}h(t) \\ &\leq r!\cdot h\left(\frac{r}{2}\right)\left(1+\frac{1}{2}+\frac{1}{4}+\ldots\right) \\ &\leq r!\cdot\frac{(n\sigma^2)^{r/2}}{(r/2)!}2^{r/2}.\end{aligned}$$

Applying the Markov Inequality, we get

$$\Pr[|X| \geq a] = \Pr[X^r \geq a^r] \leq \frac{r! \cdot (n\sigma^2)^{r/2} \cdot 2^{r/2}}{(r/2)! \cdot a^r} \leq \left(\frac{2nr\sigma^2}{a^2}\right)^{r/2}.$$

Setting $r$ to be the largest even integer less than $a^2/(6n\sigma^2)$ completes the proof. □

**Lemma 10.11.** *For any integer $s > 0$, the $s^{th}$ moment of $X \sim \mathrm{N}(0,1)$ is at most $(s!)$.*

*Proof.* It follows from the following integral.

$$\mathrm{E}[X^s] = \int_{-\infty}^{+\infty} \frac{x^s}{\sqrt{2\pi}} e^{-x^2/2} \,\mathrm{d}\, x = \begin{cases} 0 & \text{if } s \text{ is odd} \\ (s-1)!! & \text{if } s \text{ is even} \end{cases}$$

□

### 10.1.10 Chernoff Bounds

Recall the *binomial distribution* $\mathrm{Bin}(n,p)$ counts the number $X$ of ones in $n$ independent Bernoulli trials.

$$\Pr[X = k] = \Pr[(\text{total number of ones}) = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

It can be written as a sum $X = \sum X_i$ of Bernoulli random variables $X_i$ with parameter $p$, in other words,

$$X_i = \begin{cases} 0 & \text{with prob. } (1-p) \\ 1 & \text{with prob. } p \text{ the } i^{th} \text{ trial is a success} \end{cases}.$$

It has expectation $\mathrm{E}[X] = np$ and variance $\mathrm{Var}[X] = np(1-p)$. What we desire is a bound on the probability that the sum random variable $X$, does not deviate too far from this expected value.

What we next present is a general technique to compute such probability bounds.

The Chernoff bounds [31, Theorem 12.3 & Theorem 12.4] [156, Theorem 4.4 & Theorem 4.5] say that for all $\delta > 0$,

$$\Pr[X > (1+\delta)m] \leq \left[\frac{e^\delta}{(1+\delta)^{1+\delta}}\right]^m,$$

and for all $0 \leq \gamma \leq 1$,

$$\Pr[X < (1-\gamma)m] \leq e^{-\gamma^2 m/2},$$

where $X = \sum_i X_i$ and $m = \mathrm{E}[X]$.

For several NP-hard problems, we can use input or data sampling to obtain a probabilistic approximation algorithm. We can then apply Markov, Chebyshev and Chernoff tail bounds. Sometimes, we can also weaken the assumption of mutually independent sample to $k$-wise independence (e.g. $k = 2$). More on this in the next lecture.

In Monte Carlo methods, we use random sampling. In Quasi Monte Carlo, we use deterministic methods and measure the distortion from uniformity by discrepancy (for example, by considering the family of isothetic rectangles). See for e.g. [**?**]

There are also methods known as Markov Chain Monte Carlo (MCMC) or Markov Chain Quasi Monte Carlo (MCQMC).

#### Estimating $\pi$

We can use the Monte Carlo method to estimate $\pi$. Sample $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m$ independently and uniformly from $[0,1]^2$. Define indicator random variables

$$Z_i = \begin{cases} 1 & \text{if } \|\mathbf{z}_i\|_2 \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

Then,

$$\Pr[Z_i = 1] = \frac{(\text{area of the unit circle inside } [0,1]^2)}{(\text{area of the unit square})} = \frac{\pi}{4}.$$

Let $W = \sum_{i=1}^{m} Z_i$. Then,

$$\mathrm{E}[W] = \sum_{i=1}^{m} \mathrm{E}[Z_i] = \frac{m\pi}{4}.$$

Let $W' = \frac{4}{m}W$. Then, $\mathrm{E}[W'] = \pi$. Hence, $W'$ gives an estimate for $\pi$.
In fact, it improves as $m$ gets larger. By the Chernoff inequality bounds, we get

$$\Pr[|W - \mathrm{E}[W]| \geq \epsilon\, \mathrm{E}[W]] \leq 2e^{-m\pi\epsilon^2/12}.$$

For $\epsilon < 1$, we can choose $m \geq \frac{12}{\pi\epsilon^2} \ln \frac{2}{\delta}$. Then, the above algorithm is an $(\epsilon, \delta)$-approximation.

**Probabilistic Approximation Algorithm**

A randomized algorithm gives an $(\epsilon, \delta)$-*approximation* for a value $v$ if the output $X$ of the algorithm satisfies

$$\Pr[|X - v| \leq \epsilon v] \geq 1 - \delta.$$

**Theorem 10.12** ([156, Theorem 10.1])**.** *Let $X_1, X_2, \ldots, X_m$ be i.i.d. indicator random variables with $\mu = \mathrm{E}[X_i]$. If $m \geq \frac{3}{\epsilon^2\mu} \ln \frac{2}{\delta}$, then $\Pr\left[\left|\left(\frac{1}{m}\sum_{i=1}^{m} X_i\right) - \mu\right| \geq \epsilon\mu\right] \leq \delta$. That is, an $m$-sampling provides an $(\epsilon, \delta)$-approximation.*

*Proof.* This can be proved using the above stated Chernoff bound. □

## 10.2 Bayesian

### 10.2.1 Bayes Rule

We will learn to use probability theory to use sampling for parameter estimation. Consider the Bayes rule.

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]}$$

This follows from $\Pr[A|B]\Pr[B] = \Pr[B|A]\Pr[A]$. We can regard $B$ as the measurement samples that we have. Using this data, we try to estimate $A$. In the numerator, $\Pr[B|A]$ is the likelihood of $A$ and $\Pr[A]$ is the prior probability. The normalization appears in the denominator $\Pr[B]$. The left hand side is the posterior probability $\Pr[A|B]$.
For example, suppose that a product is defective $0.1\%$ of the time, and a test fails $1\%$ of the time to detect a defective product. Also, assuming that a product is not defective, a test says a product is defective $2\%$ of the time.
Let $A$ be the event that a product is defective. Let $B$ be the event that a test says a product is defective. Then, we have the followings.

$$\begin{aligned}
\Pr[B|A] &= 0.99 \\
\Pr[A] &= 0.001 \\
\Pr[B|\overline{A}] &= 0.02 \\
\Pr[B] &= \Pr[B|A]\Pr[A] + \Pr[B|\overline{A}]\Pr[\overline{A}] = 0.99 \times 0.001 + 0.02 \times 0.999 = 0.02097
\end{aligned}$$

So, using the Bayes rule, we can estimate

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]} = \frac{0.99 \times 0.001}{0.0297} \approx 0.047,$$

which is surprising.
The Bayes rule can be applied to molecule reconstruction from projection images. In such applications, the molecule imaging is used to reconstruct the structure of the molecule. This is analogous to using measurements to make estimations according to the Bayes rule.

### 10.2.2 Maximum Likelihood Estimator

Suppose a probability distribution of a random variable $X$ depends on parameter $r$. So, $\Pr[X|r]$ denotes the probability of observing $X$ if parameter value is $r$. If $r$ is also random, after observing the value of $X$, one can find the best $r$ maximizing the posterior probability

$$\Pr[r|X] = \frac{\Pr[X|r]\Pr[r]}{\Pr[X]}.$$

Assume $\Pr[r]$ is the same for all $r$. Since the unconditional probability of $X$ in the denominator is independent of $r$, it reduces to finding the *maximum likelihood estimator* (MLE)

$$\arg\max_r \mathrm{L}(r|X) = \arg\max_r \Pr[X|r].$$

**Example**

Consider the example of flipping a biased coin in $n$ trials with unknown probability $r$ of getting head. The probability of getting $k$ heads follows the binomial disbtribution $\mathrm{Bin}(n, r)$ such that

$$\Pr[k|r] = \binom{n}{k} r^k (1-r)^{n-k}.$$

If we get 62 heads and 38 tails in 100 trails, the maximum likelihood estimator gives $r = 0.62$ when $\Pr[62|r]$ is maximized. One can see this by setting the derivative (with respect to $r$) to 0.

We can study a single particle using cryo-electron microscopy. To do so, we build a specimen grid of millions of in-vitro molecules and take a snapshot by shooting X-ray and measuring its projection. We can reconstruct the locations of the molecules by solving a least square optimization with regularizer, but it is unstable. How many samples would we need? We want to show that the solution converges as the sample size increase. We can recast the problem by regarding the data as a random variable with certain mean and variance. We can then solve for the maximum a-posterior estimator. We would also like to output a confidence level of our estimation.

### 10.2.3 Unbiased Estimator

Let $X = (X_1, X_2, \ldots, X_n)$ be samples or observations from a distribution having parameter $\theta$. (For example, the Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$ has parameters mean $\mu$ and variance $\sigma^2$, while the binomial distribution $\mathrm{Bin}(n, p)$ has parameters $n$ and success probability $p$.)

Let $D(X)$ be an estimator of some function $h(\theta)$. The *bias* is defined as

$$\mathrm{E}[D(X) - h(\theta)].$$

It is called an *unbiased estimator* when the bias equals zero.

The quality of the estimator can be measured by the *mean squared error* (MSE)

$$\mathrm{E}[(D(X) - h(\theta))^2] = \mathrm{Var}(D(X)) + \mathrm{Bias}^2.$$

**Theorem 10.13.** *Let $X_1, X_2, \ldots, X_n$ be independent samples, each with mean $\mu$ and variance $\sigma^2$.*

*$Q_\xi$1. [**?**, Example 14.3] $D(X) = \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimator of $\mu$.*

*$Q_\xi$2. If $\mu$ is known, then $D(X) = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ is an unbiased estimator of $\sigma^2$.*

*$Q_\xi$3. [**?**, Example 14.5] If $\mu$ is not known, then $D(X) = \frac{1}{n-1}\sum_{i=1}^n (X_i - m)^2$ is an unbiased estimator of $\sigma^2$, where $m = \frac{1}{n}\sum_{i=1}^n X_i$.*

*Proof of 3.* Let $S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - m)^2$. Observe that

$$\begin{aligned}\sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n \Big[(X_i - m) + (m - \mu)\Big]^2 \\ &= \sum_{i=1}^n (X_i - m)^2 + n(m - \mu)^2\end{aligned}$$

Hence,

$$
\begin{aligned}
S^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (m - \mu)^2 \\
\mathrm{E}[S^2] &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(X_i) - \mathrm{Var}(m) \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

Thus, $\left(\frac{n}{n-1}S^2\right)$ is an unbiased estimator of $\sigma^2$. □

## 10.3 Biological Applications

## Summary

## References and Further Reading

ExerSection

# Biology Appendix

# Conclusions

Something to conclude

# Bibliography

[1] S. S. Abhyankar and C. L. Bajaj. Automatic parameterization of rational curves and surfaces III: algebraic plane curves. *Comput. Aided Geom. Des.*, 5(4):309–321, 1988.

[2] R. Abraham, J. E. Marsden, and T. Ratiu. *Manifolds, tensor analysis, and applications*, volume 75 of *Applied Mathematical Sciences*. Springer-Verlag, New York, second edition, 1988.

[3] A. Abyzov, R. Bjornson, M. Felipe, and M. Gerstein. RigidFinder: A fast and sensitive method to detect rigid blocks in large macromolecular complexes. *PROTEINS: Structure, Function, and Bioinformatics*, 78:309–324, 2010.

[4] N. Akkiraju and H. Edelsbrunner. Triangulating the surface of a molecule. *Discrete Applied Mathematics*, 71(1-3):5–22, 1996.

[5] G. Albers and T. Roos. Voronoi diagrams of moving points in higher dimensional spaces. In *Proc. 3rd Scand. Workshop Algorithm Theory*, volume 621 of *Lecture Notes in Computer Science*, pages 399–409. Springer-Verlag, 1992.

[6] D. Apprato, R. Arcanceli, and J. L. Gout. Rational interpolation of Wachspress error estimates. *Comput. Math. Appl.*, 5(4):329–336, 1979.

[7] D. Apprato, R. Arcangeli, and J. L. Gout. Sur les elements finis rationnels de Wachspress. *Numer. Math.*, 32(3):247–270, 1979.

[8] M. A. Armstrong. *Basic Topolgy*. Springer, New York, 1983.

[9] D. Arnold, R. Falk, and R. Winther. Finite element exterior calculus, homological techniques, and applications. *Acta Numerica*, pages 1–155, 2006.

[10] D. Arnold, R. Falk, and R. Winther. Geometric decompositions and local bases for spaces of finite element differential forms. *Comput. Methods Appl. Mech. Engrg.*, 198(21-26):1660–1672, 2009.

[11] D. Arnold, R. Falk, and R. Winther. Finite element exterior calculus: from Hodge theory to numerical stability. *Bulletin of the American Mathematical Society*, 47(2):281–354, 2010.

[12] D. N. Arnold and G. Awanou. The serendipity family of finite elements. *Found. Comput. Math.*, 11(3):337–344, 2011.

[13] D. N. Arnold, D. Boffi, and R. S. Falk. Approximation by quadrilateral finite elements. *Math. Comput.*, 71(239):909–922, 2002.

[14] P. Atkins and J. de Paula. *Physical Chemistry for the Life Sciences*. Oxford University Press, 2006.

[15] F. Aurenhammer. Voronoi diagrams: A survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23:345–405, 1991.

[16] C. Bajaj, F. Bernardini, and G. Xu. Automatic reconstruction of surfaces and scalar fields from 3D scans. In *ACM SIGGRAPH*, pages 109–118, 1995.

[17] C. Bajaj and W. J. Bouma. Dynamic Voronoi diagrams and Delaunay triangulations. In *Proceedings of the 2nd Canadian Conference on Computational Geometry*, pages 273–277, 1990.

[18] C. Bajaj, R. Chowdhury, and V. Siddahanavalli. F2Dock: Fast fourier protein-protein docking. *IEEE/ACM Trans. Comput. Biol. Bioinf*, 8(1):45–58, 2011.

[19] C. Bajaj, R. A. Chowdhury, and V. Siddavanahalli. F3dock: A fast, flexible and fourier-based approach to protein-protein docking. Technical report, The University of Texas, 2007.

[20] C. Bajaj and S. Goswami. Automatic fold and structural motif elucidation from 3d em maps of macromolecules. In *ICVGIP 2006*, pages 264–275, 2006.

[21] C. Bajaj, H. Y. Lee, R. Merkert, and V. Pascucci. NURBS based B-rep models for macromolecules and their properties. In *Proceedings of the 4th ACM Symposium on Solid Modeling and Applications*, pages 217–228, New York, NY, USA, 1997. ACM.

[22] C. Bajaj, V. Pascucci, A. Shamir, R. Holt, and A. Netravali. Multiresolution molecular shapes. Technical report, TICAM, Univ. of Texas at Austin, Dec. 1999.

[23] C. Bajaj, V. Pascucci, A. Shamir, R. Holt, and A. Netravali. Dynamic maintenance and visualization of molecular surfaces. *Discrete Applied Mathematics*, 127(1):23–51, 2003.

[24] C. Bajaj, G. Xu, and Q. Zhang. A fast variational method for the construction of resolution adaptive $c^2$-smooth molecular surfaces. *Computer Methods in Applied Mechanics and Engineering*, 198(21-26):1684–1690, 2009.

[25] N. Basdevant, D. Borgis, and T. Ha-Duong. A coarse-grained protein-protein potential derived from an all-atom force field. *Journal of Physical Chemistry B*, 111(31):9390–9399, 2007.

[26] M. Behzad and G. Chartrand. *Introduction to the theory of graphs*. Allyn and Bacon Inc., Boston, Mass., 1971.

[27] L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L. D. Marini, and A. Russo. Basic principles of virtual element methods. *Math. Models Methods Appl. Sci.*, 23(1):199–214, 2013.

[28] L. Beirão da Veiga, F. Brezzi, L. Marini, and A. Russo. $H(\mathrm{div})$ and $H(\mathrm{curl})$ conforming VEM. *arXiv:1407.6822*, 2014.

[29] R. Bettadapura, A. Vollrath, and C. Bajaj. Pfflexfit: Hierarchical flexible fitting in 3d em. Technical Report 12-29, University of Texas at Austin, July 2012.

[30] J. F. Blinn. A generalization of algebraic surface drawing. *ACM Transactions on Graphics*, 1(3):235–256, July 1982.

[31] A. Blum, J. Hopcroft, and R. Kannan. Foundations of data science. *Vorabversion eines Lehrbuchs*, 2016.

[32] A. I. Bobenko. Delaunay triangulations of polyhedral surfaces, a discrete laplace-beltrami operator and applications. In *SCG '08: Proceedings of the twenty-fourth annual symposium on Computational geometry*, pages 38–38, New York, NY, USA, 2008.

[33] P. Bochev and J. Hyman. Principles of mimetic discretizations of differential operators. In *Compatible Spatial Discretizations*, pages 89–119. Springer, 2006.

[34] A. Bossavit. Whitney forms: a class of finite elements for three-dimensional computations in electromagnetism. *Proc. of the IEEE*, 135(8):493–500, 1988.

[35] Y. Bourne, J. Grassi, P. E. Bougis, and P. Marchot. Conformational flexibility of the acetylcholinesterase tetramer suggested by x-ray crystallography. *Journal of Biological Chemistry*, 274(43):30370–30376, 1999.

[36] Y. Bourne, P. Taylor, and P. Marchot. Acetylcholinesterase inhibition by fasciculin: crystal structure of the complex. *Cell*, 83:503, 1995.

[37] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[38] F. Brezzi, A. Buffa, and G. Manzini. Mimetic scalar products of discrete differential forms. *J. Comput. Phys.*, in press, 2013.

[39] F. Brezzi, J. Douglas Jr, R. Durán, and M. Fortin. Mixed finite elements for second order elliptic problems in three variables. *Numer. Math.*, 51(2):237–250, 1987.

[40] F. Brezzi, J. Douglas Jr, and L. D. Marini. Two families of mixed finite elements for second order elliptic problems. *Numer. Math.*, 47(2):217–235, 1985.

[41] F. Brezzi, K. Lipnikov, and V. Simoncini. A family of mimetic finite difference methods on polygonal and polyhedral meshes. *Math. Models Meth. Appl. Sci.*, 15(10):1533–1551, 2005.

[42] J. Canny. Generalised characteristic polynomials. *Journal of Symbolic Computation*, 9(3):241–250, 1990.

[43] F. Cazals, F. Chazal, and T. Lewiner. Molecular shape analysis based upon the morse-smale complex and the connolly function. In *19th Ann. ACM Sympos. Comp. Geom.*, pages 351–360, 2003.

[44] R. Chaine. A geometric convection approach of 3D reconstruction. In *Proc. Eurographics Sympos. on Geometry Processing*, pages 218–229, 2003.

[45] G. Chartrand and L. Lesniak. *Graphs & digraphs*. Chapman & Hall/CRC, Boca Raton, FL, fourth edition, 2005.

[46] D. Chavey. Tilings by regular polygons II: A catalog of tilings. *Comput. Math. Appl.*, 17(1–3):147–165, 1989.

[47] F. Chazal and A. Lieutier. Stability and homotopy of a subset of the medial axis. In *Proc. 9th ACM Sympos. Solid Modeling and Applications*, pages 243–248, 2004.

[48] W. Chen and Y. Wang. Minimal degree H(curl) and H(div) conforming finite elements on polytopal meshes. *arXiv:1502.01553*, 2015.

[49] L. P. Chew. Constrained Delaunay triangulations. In *Proc. 3rd Annual ACM Symposium Computational Geometry*, pages 215–222, 1987.

[50] P. Chew. Near-quadratic bounds for the $L_1$ Voronoi diagram of moving points. *Computational Geometry Theory and Applications*, 7, 1997.

[51] G. Chirikjian and I. Ebert-Uphoff. Numerical convolution on the euclidean group with applications to workspace generation. *IEEE Trans. on Robotics and Automation*, 14(1):123–136, 1998.

[52] R. Chowdhury, M. Rasheed, D. Keidel, M. Moussalem, A. Olson, and C. Bajaj. Protein-protein docking with f2dock 2.0 and gb-rerank. *PLoS ONE, doi:10.1371/journal.pone.0051307*, 8(3: e51307), 2013.

[53] S. H. Christiansen. A construction of spaces of compatible differential forms on cellular complexes. *Math. Models Methods Appl. Sci.*, 18(5):739–757, 2008.

[54] S. H. Christiansen and R. Winther. Smoothed projections into finite element exterior calculus. *Math. Comput.*, 77(262):813–829, 2008.

[55] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*, volume 40 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, second edition, 2002.

[56] P. Clément. Approximation by finite element functions using local regularization. *RAIRO Anal. Numér.*, 9(R-2):77–84, 1975.

[57] Cocone. Tight Cocone Software for surface reconstruction and medial axis approximation. *http://www.cse.ohio-state.edu/∼tamaldey/cocone.html*, 2001.

[58] M. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612):709–713, 19 August 1983.

[59] M. L. Connolly. Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16:548–558, 1983.

[60] E. Cueto, N. Sukumar, B. Calvo, M. A. Martínez, J. Cegoñino, and M. Doblaré. Overview and recent advances in natural neighbour Galerkin methods. *Arch. Comput. Methods Engrg.*, 10(4):307–384, 2003.

[61] CVC, UT Austin. Volrover. *http://cvcweb.ices.utexas.edu/software/guides.php*, 2011.

[62] P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration (Second Edition)*. Academic Press Inc., 1984.

[63] M. de Berg and K. T. G. Dobrindt. On levels of detail in terrains. *Graphical Models and Image Processing*, 60:1–12, 1998.

[64] S. Dekel and D. Leviatan. The Bramble-Hilbert lemma for convex domains. *SIAM J. Math. Anal.*, 35(5):1203–1212, 2004.

[65] M. Desbrun and M. Gascuel. Animating soft substances with implicit surfaces. In R. Cook, editor, *SIGGRAPH 95 Conference Proceedings*, Annual Conference Series, pages 287–290, Los Angeles, August 1995. Addison Wesley.

[66] T. K. Dey, J. Giesen, and S. Goswami. Shape segmentation and matching with flow discretization. In F. Dehne, J.-R. Sack, and M. Smid, editors, *Proc. Workshop Algorithms Data Strucutres (WADS 03)*, LNCS 2748, pages 25–36, Berlin, Germany, 2003.

[67] M. Eck, T. DeRose, T. Duchamp, T. Hoppe, H. Lounsbery, and W. Stuetzle. Multiresolution analysis of arbitrary meshes. In *ACM Computer Graphics Proceedings, SIGGRAPH'95*, pages 173–180, 1995.

[68] H. Edelsbrunner. Weighted alpha shapes. Technical Report 1760, University of Illinois at Urbana-Champaign, 1992.

[69] H. Edelsbrunner. Surface reconstruction by wrapping finite point sets in space. In B. Aronov, S. Basu, J. Pach, and M. Sharir, editors, *Ricky Pollack and Eli Goodman Festschrift*, pages 379–404. Springer-Verlag, 2002.

[70] H. Edelsbrunner, M. Facello, and J. Liang. On the definition and the construction of pockets in macromolecules. Tech Report UIUCDCS-R-95-1935, University of Illinois Urbana-Champaign, 1995.

[71] H. Edelsbrunner, J. Harer, V. Natarajan, and A. Zomorodian. Morse-smale complexes for piecewise linear 3-manifolds. In *19th Ann. Sympos. Comp. Geom.*, pages 361–370, 2003.

[72] H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical morse-smale complexes for piecewise linear 2-manifolds. *Discrete Computational Geometry*, 30:87–107, 2003.

[73] H. Edelsbrunner and E. P. Mücke. Three-dimensional alpha shapes. *ACM Transactions on Graphics*, 13(1):43–72, 1994.

[74] H. Edelsbrunner and N. R. Shah. Incremental topological flipping works for regular triangulations. *Algorithmica*, 15:223–241, 1996.

[75] M. Eisenstein and E. Katchalski-Katzir. On proteins, grids, correlations, and docking. *C R Biol*, 327:409 – 420, 2004.

[76] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[77] J. Esquivel-Rodriguez and D. Kihara. Fitting multimeric protein complexes into electron microscopy maps using 3d zernike descriptors. *J. Phys. Chem. B*, 116(6854-6861), 2012.

[78] T. Euler, R. Schuhmann, and T. Weiland. Polygonal finite elements. *Magnetics, IEEE Transactions on*, 42(4):675–678, 2006.

[79] E. Eyal and D. Halperin. Dynamic maintenance of molecular surfaces under conformational changes. In *SCG '05: Proceedings of the 21st Annual Symposium on Computational Geometry*, pages 45–54, 2005.

[80] E. Eyal and D. Halperin. Improved maintenance of molecular surfaces using dynamic graph connectivity. *Algorithms in Bioinformatics*, pages 401–413, 2005.

[81] M. A. Facello. *Geometric Techniques for Molecular Shape Analysis*. PhD thesis, University of Illinois, 1996. Department of Computer Science Technical Report # 1967.

[82] G. Farin. Surfaces over Dirichlet tessellations. *Computer Aided Geometric Design*, 7(1-4):281–292, 1990.

[83] M. Floater. Mean value coordinates. *Computer Aided Geometric Design*, 20(1):19–27, 2003.

[84] M. Floater, A. Gillette, and N. Sukumar. Gradient bounds for Wachspress coordinates on polytopes. *SIAM J. Numer. Anal.*, 52(1):515–532, 2014.

[85] M. Floater, K. Hormann, and G. Kós. A general construction of barycentric coordinates over convex polygons. *Adv. Comput. Math.*, 24(1):311–331, 2006.

[86] M. S. Floater, G. Kós, and M. Reimers. Mean value coordinates in 3D. *Computer Aided Geometric Design*, 22(7):623–631, 2005.

[87] S. Flores, L. Lu, J. Yang, N. Carriero, and M. Gerstein. Hinge atlas: relating protein sequence to sites of structural flexibility. *BMC Bioinformatics*, 8:167–186, 2007.

[88] J. J. Fu and R. C. T. Lee. Voronoi diagrams of moving points in the plane. *International Journal of Computational Geometry Applications*, 1(1):23–32, 1991.

[89] T. Fujita, K. Hirota, and K. Murakami. Representation of splashing water using metaball model. *Fujitsu*, 41(2):159–165, 1990. in Japanese.

[90] J. I. Garccon, J. A. Kovacs, and R. Abagyan. Adp_em: Fast exhaustive multi-resolution docking with high-throughput coverage. *Bioinformatics*, 23(4):427–433, 2007.

[91] M. Garland. QSlim. *http://graphics.cs.uiuc.edu/ garland/software/qslim.html*, 2004.

[92] M. Garland and P. Heckbert. Simplifying surfaces with color and texture using quadric error metrics. In *IEEE Visualization*, pages 263–270, 1998.

[93] J. Garzon, J. Lopéz-Blanco, C. Pons, J. A. Kovacs, R. Abagyan, J. Fernandez-Recio, and P. Chacon. FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics*, 25:2544–2551, 2009.

[94] P. Giblin. *Graphs, surfaces and homology*. Cambridge University Press, Cambridge, third edition, 2010.

[95] T. S. Gieng, B. Hamann, K. I. Joy, G. L. Schussman, and I. J. Trotts. Constructing hierarchies for triangle meshes. *IEEE Transactions on Visualization and Computer Graphics*, 4(2):145–161, 1998.

[96] J. Giesen and M. John. The flow complex: a data structure for geometric modeling. In *Proc. 14th ACM-SIAM Sympos. Discrete Algorithms*, pages 285–294, 2003.

[97] A. Gillette. *Stability of Dual Discretization Methods for Partial Differential Equations*. PhD thesis, University of Texas at Austin, 2011.

[98] A. Gillette and C. Bajaj. A generalization for stable mixed finite elements. In *Proc. 14th ACM Symp. Solid Phys. Modeling*, pages 41–50, 2010.

[99] A. Gillette and C. Bajaj. Dual formulations of mixed finite element methods with applications. *Computer Aided Design*, 43(10):1213–1221, 2011.

[100] A. Gillette, A. Rand, and C. Bajaj. Error estimates for generalized barycentric coordinates. *Advances in Computational Mathematics*, 37(3):417–439, 2012.

[101] S. Goswami, T. K. Dey, and C. L. Bajaj. Identifying flat and tubular regions of a shape by unstable manifolds. In *Proc. 11th ACM Sympos. Solid and Phys. Modeling*, pages 27–37, 2006.

[102] J. L. Gout. Construction of a Hermite rational “Wachspress type” finite element. *Comput. Math. Appl.*, 5(4):337–347, 1979.

[103] J. L. Gout. Rational Wachspress-type finite elements on regular hexagons. *IMA J. Numer. Anal.*, 5(1):59, 1985.

[104] V. Gradinaru. *Whitney elements on sparse grids*. PhD thesis, Universität Tübingen, 2002.

[105] V. Gradinaru and R. Hiptmair. Whitney elements on pyramids. *Electronic Transactions on Numerical Analysis*, 8:154–168, 1999.

[106] M. Gräf and D. Potts. Sampling sets and quadrature formulae on the rotation group. *Numer. Funct. Anal. Optim.*, 30:665 – 688, 2009.

[107] R. L. Graham and P. Hell. On the history of the minimum spanning tree problem. *Ann. Hist. Comput.*, 7(1):43–57, 1985.

[108] J. Grant and B. Pickup. A gaussian description of molecular shape. *Journal of Physical Chemistry*, 99:3503–3510, 1995.

[109] G. Graves. The magic of metaballs. *Computer Graphics World*, 16(5):27–32, 1993.

[110] X. Gu and S.-T. Yau. Global conformal surface parameterization. In *SGP '03: Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 127–137, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.

[111] L. Guibas, J. S. B. Mitchell, and T. Roos. Voronoi diagrams of moving points in the plane. In *Proceedings of the 17th Internatational Workshop on Graph-Theoretical Concepts in Computer Science*, volume 570 of *Lecture Notes in Computer Science*, pages 113–125. Springer-Verlag, 1991.

[112] V. Guillemin and A. Pollack. *Differential Topology*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1974.

[113] R. Hadani and A. Singer. Representation theoretic patterns in three dimensional cryo-electron microscopy I - the intrinsic reconstitution algorithm. *Annals of Mathematics*, 2011.

[114] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *PROTEINS: Struct. Funct. Genet.*, 47:409 – 443, 2002.

[115] R. Hiptmair. Finite elements in computational electromagnetism. *Acta Numerica*, pages 237–339, 2002.

[116] A. N. Hirani. *Discrete Exterior Calculus*. PhD thesis, California Institute of Technology, 2003.

[117] C. Ho and C. Yap. Polynomial Remainder Sequences and Theory of Subresultants. Technical report, Technical Report, 1987.

[118] L. Holm and C. Sander. Parser for protein folding units. *Proteins*, 19(3):256–268, July 1994.

[119] M. Holst, N. Baker, and F. Wang. Adaptive multilevel finite element solution of the poisson-boltzmann equation i: Algorithms and examples. *Journal of Compututational Chemistry*, 21:1319–1342, 2000.

[120] H. Hoppe. Progressive meshes. In *ACM Computer Graphics Proceedings, SIGGRAPH'96*, pages 99–108, 1996.

[121] P. H. Hubbard. *Collision Detection for Interactive Graphics Applications*. PhD thesis, Department of Computer Science, Brown University, Providence, Rhode Island, April 1995.

[122] P. M. Hubbard. Approximating polyhedra with spheres for time-critical collision detection. *ACM Transactions on Graphics*, 15(3), 1996.

[123] T. Hughes. *The finite element method*. Prentice Hall Inc., Englewood Cliffs, NJ, 1987. Linear static and dynamic finite element analysis, With the collaboration of Robert M. Ferencz and Arthur M. Raefsky.

[124] H. Imai, M. Iri, and K. Murota. Voronoi diagram in the laguerre geometry and its applications. *SIAM J. Comput.*, 14:93–10, 1985.

[125] P. Joshi, M. Meyer, T. DeRose, B. Green, and T. Sanocki. Harmonic coordinates for character articulation. *ACM Transactions on Graphics*, 26:71, 2007.

[126] T. Ju, F. Losasso, S. Schaefer, and J. Warren. Dual contouring of hermite data. In *SIGGRAPH*, pages 339–346, 2002.

[127] T. Ju, S. Schaefer, J. D. Warren, and M. Desbrun. A geometric construction of coordinates for convex polyhedra using polar duals. In *Symposium on Geometry Processing*, pages 181–186, 2005.

[128] L. Kale, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten. Namd2: Greater scalability for parallel molecular dynamics. *J. Comput. Phys.*, 151(1):283–312, 1999.

[129] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. Friesem, C. Aflalo, and I. Vakser. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Nat. Acad. Sci. USA*, 89:2195 – 2199, 1992.

[130] A. Khodakovsky, N. Litke, and P. Schroder. Globally smooth parameterizations with low distortion. *ACM Transactions on Graphics*, 22(3):350–357, 2003.

[131] F. Kikuchi, M. Okabe, and H. Fujio. Modification of the 8-node serendipity element. *Comput. Methods Appl. Mech. Engrg.*, 179(1-2):91–109, 1999.

[132] R. Klausen, A. Rasmussen, and A. Stephansen. Velocity interpolation and streamline tracing on irregular geometries. *Computational Geosciences*, pages 1–16, 2011.

[133] R. Klein and J. Kramer. Multiresolution representations for surface meshes. In *Proceedings of the SCCG*, 1997.

[134] J. A. Kovacs, P. Chacón, Y. Cong, E. Metwally, and W. Wriggers. Fast rotational matching of rigid bodies by fast fourier transform acceleration of five degrees of freedom. *Acta Crystallographica Section D*, D59:1371–1376, 2003.

[135] J. A. Kovacs, P. Chacón, Y. Cong, E. Metwally, and W. Wriggers. Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom. *Acta Crystallogr. Sect. D*, 59:1371 – 1376, 2003.

[136] J. A. Kovacs and W. Wriggers. Fast rotational matching. *Acta Crystallographica Section D*, 58(8):1282–1286, Aug 2002.

[137] J. B. Kruskal, Jr. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.*, 7:48–50, 1956.

[138] S. Kumar, D. Manocha, and A. Lastra. Interactive display of large-scale NURBS models. *IEEE Transactions on Visualization and Computer Graphics*, 2(4):323–336, 1996.

[139] Y. Lamdan and H. Wolfson. Geometric hashing: a general and efficient model-based recognition scheme. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 238–249, 1988.

[140] T. Langer and H. Seidel. Higher order barycentric coordinates. In *Comput. Graphics Forum*, volume 27, pages 459–466. Wiley Online Library, 2008.

[141] M. Lee, M. Feig, F. Salsbury, and C. Brooks. New analytic approximation to standard molecular volume definition and its application to generalized born calculation. *Journal of Computational Chemistry*, 24:1348–1356, 2003.

[142] J. Leech, J. Prins, and J. Hermans. Smd: visual steering of molecular dynamics for protein design. *IEEE Computational Science and Engineering*, 3(4):38–45, 1996.

[143] M. Levitt, C. Sander, and P. S. Stern. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *Journal of Molecular Biology*, 181:423 – 447, 1985.

[144] R. Loos. Generalized Polynomial Remainder Sequences, 1983.

[145] F. Macaulay. Some formulae in elimination. *Proceedings of the London Mathematical Society*, 1(1):3, 1902.

[146] R. H. MacNeal and R. L. Harder. Eight nodes or nine? *Int. J. Numer. Methods Eng.*, 33(5):1049–1058, 1992.

[147] A. Maheshwari, P. Morin, and J. R. Sack. Progressive tins: Algorithms and applications. In *Proceedings 5th ACM Workshop on Advances in Geographic Information Systems*, Las Vegas, 1997.

[148] V. N. Maiorov and R. A. Abagyan. A new method for modeling large-scale rearrangements of protein domains. *Proteins*, 27:410–424, 1997.

[149] J. G. Mandell, V. A. Roberts, M. E. Pique, V. Kotlovyi, J. C. Mitchell, E. Nelson, I. Tsigelny, and L. F. T. Eyck. Protein docking using continuum electrostatics and geometric fit. *Protein Engineering Design and Selection*, 14(2):105–113, 2000.

[150] G. Manzini, A. Russo, and N. Sukumar. New perspectives on polygonal and polyhedral finite element methods. *Math. Models Methods Appl. Sci.*, 24(08):1665–1699, 2014.

[151] S. Martin, P. Kaufmann, M. Botsch, M. Wicke, and M. Gross. Polyhedral finite elements using harmonic basis functions. In *Proc. Symp. Geom. Proc.*, pages 1521–1529, 2008.

[152] N. L. Max. Computer representation of molecular surfaces. *IEEE Computer Graphics and Applications*, 3(5):21–29, August 1983.

[153] P. Milbradt and T. Pick. Polytope finite elements. *Int. J. Numer. Methods Eng.*, 73(12):1811–1835, 2008.

[154] J. Milnor. *Morse Theory*. Princeton University Press, New Jersey, 1963.

[155] J. C. Mitchell. Discrete uniform sampling of rotation groups using orthogonal images. *SIAM Journal of Scientific Computing*, 30(1):525–547, 2007.

[156] M. Mitzenmacher and E. Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press, 2005.

[157] J. R. Munkres. *Elements of algebraic topology*. Addison-Wesley Publishing Company, Menlo Park, CA, 1984.

[158] J.-C. Nédélec. Mixed finite elements in $\mathbf{R}^3$. *Numer. Math.*, 35(3):315–341, 1980.

[159] J.-C. Nédélec. A new family of mixed finite elements in $\mathbf{R}^3$. *Numer. Math.*, 50(1):57–81, 1986.

[160] H. Nishimura, M. Hirai, T. Kawai, T. Kawata, I. Shirakawa, and K. Omura. Object modeling by distribution function and a method of image generation. *Transactions IECE Japan, Part D*, J68-D(4):718–725, 1985.

[161] T. Nishita and E. Nakamae. A method for displaying metaballs by using Bézier clipping. In *Computer Graphics Forum*, volume 13, pages 271–280. Eurographics, Basil Blackwell Ltd, 1994. Eurographics '94 Conference issue.

[162] A. H. Nuttall. Efficient evaluation of polynomials and exponentials of polynomials for equispaced arguments. *IEEE Transactions On Acoustics, Speech And Signal Processing*, 35(10):1486–1487, 1987.

[163] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. Ucsf chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–12, 2004.

[164] B. G. Pierce, Y. Hourai, and Z. Weng. Accelerating protein docking in zdock using an advanced 3d convolution library. *PLoS One*, 6(9), 2011.

[165] G. D. Pintilie, J. Zhang, T. D. Goddard, W. Chiu, and D. C. Gossard. Quantitative analysis of cryo-em density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to quantitative analysis of cryo-em density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J. Struct. Biol.*, 170:427–438, 2010.

[166] G. Poornam, A. Matsumoto, H. Ishida, and S. Hayward. A method for the analysis of domain movements in large biomolecular complexes. *PROTEINS: Structure, Function, and Bioinformatics*, 76:201–212, 2009.

[167] D. Potts, J. Prestin, and A. Vollrath. A fast algorithm for nonequispaced fourier transforms on the rotation group. *Numerical Algorithms*, 52(3):355–384, 2009.

[168] D. Potts, G. Steidl, and M. Tasche. Fast Fourier transforms for nonequispaced data: A tutorial. In J. J. Benedetto and P. J. S. G. Ferreira, editors, *Modern Sampling Theory: Mathematics and Applications*, chapter 12, pages 247 – 270. Birkhäuser, Boston, 2001.

[169] A. Rand. Average interpolation under the maximum angle condition. *SIAM J. Numer. Anal.*, 50(5):2538–2559, 2012. arxiv.org:1112.4100.

[170] A. Rand, A. Gillette, and C. Bajaj. Interpolation error estimates for mean value coordinates. *Advances in Computational Mathematics*, in press:1–18, 2011.

[171] A. Rand, A. Gillette, and C. Bajaj. Quadratic serendipity finite elements on polygons using generalized barycentric coordinates. *arXiv:1109.3259*, 2011.

[172] M. M. Rashid and M. Selimotic. A three-dimensional finite element method with arbitrary polyhedral elements. *Int. J. Numer. Methods Eng.*, 67(2):226–252, 2006.

[173] 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.

[174] J. Renegar. On the computational complexity and geometry of the first-order theory of the reals. Part I: Introduction. Preliminaries. The geometry of semi-algebraic sets. The decision problem for the existential theory of the reals. *Journal of symbolic computation*, 13(3):255–299, 1992.

[175] D. W. Ritchie. High order analytic translation matrix elements for real six-dimensional polar Fourier correlations. *J. Appl. Cryst.*, 38:808 – 818, 2005.

[176] D. W. Ritchie. Recent progress and future directions in protein-protein docking. *Curr. Prot. Pep. Sci.*, 9:1 – 15, 2008.

[177] D. W. Ritchie, D. Kozakov, and S. Vajda. Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational fft generating functions. *Bioinformatics*, 24:1865–1873, 2008.

[178] T. Roos. Voronoĭ diagrams over dynamic scenes. *Discrete Applied Mathematics. Combinatorial Algorithms, Optimization and Computer Science*, 43(3):243–259, 1993.

[179] T. Roos. New upper bounds on Voronoi diagrams of moving points. *Nordic Journal of Computing*, 4(2):167–171, 1997.

[180] K. Rosen. *Discrete mathematics and its applications*. McGraw-Hill, 2003.

[181] J. Rossignac and P. Borrel. Multi-resolution 3d approximation for rendering complex scenes. In B. Falcidieno and T. Kunii, editors, *Geometric Modeling in Computer Graphics*, pages 455–465. Springer-Verlag, 1993.

[182] G. Salmon. *Lessons introductory to the modern higher algebra*. Hodges, Figgis, and co., 1885.

[183] M. Sanner, A. Olson, and J. Spehner. Fast and robust computation of molecular surfaces. In *Proceedings of the eleventh annual symposium on Computational geometry*, pages 406–407. ACM Press, 1995.

[184] M. Sanner, A. Olson, and J. Spehner. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, March 1996.

[185] M. F. Sanner and A. J. Olson. Real time surface reconstruction for moving molecular fragments. In *Proceedings of the Pacific Symposium on Biocomputing '97*, Maui, Hawaii, January 1997.

[186] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson. Geometry-based flexible and symmetric protein docking. *Proteins: Structure, Function, and Bioinformatics*, 60(2):224–231, 2005.

[187] J. Schreiner, A. Asirvatham, E. Praun, and H. Hoppe. Inter-surface mapping. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 870–877. ACM, 2004.

[188] W. J. Schroeder. A topology modifying progressive decimation algorithm. In R. Yagel and H. Hagen, editors, *IEEE Visualization '97*, pages 205–212. IEEE, nov 1997.

[189] L. R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comput.*, 54(190):483–493, 1990.

[190] M. Shatsky, R. Nussinov, and H. Wolfson. Flexible protein alignment and hinge detection. *PROTEINS: Structure, Function, and Genetics*, 48:242–256, 2002.

[191] A. Sheffer, E. Praun, and K. Rose. Mesh parameterization methods and their applications. *Foundations and Trends® in Computer Graphics and Vision*, 2(2):105–171, 2006.

[192] R. Sibson. A vector identity for the Dirichlet tessellation. *Math. Proc. Cambridge Philos. Soc.*, 87(1):151–155, 1980.

[193] D. Sieger, P. Alliez, and M. Botsch. Optimizing voronoi diagrams for polygonal finite element computations. *Proc. 19th Int. Meshing Roundtable*, pages 335–350, 2010.

[194] D. Siersma. Voronoi diagrams and morse theory of the distance function. In O. E. Barndorff and E. B. V. Jensen, editors, *Geometry in Present Day Science*, pages 187–208. World Scientific, 1999.

[195] R. D. Skeel, I. Tezcan, and D. J. Hardy. Multiple grid methods for classical molecular dynamics. *Journal of Computatioanl Chemistry*, 23(6):673–684, 2002.

[196] I. Sloan. Integration and approximation in high dimensions–a tutorial. *Uncertainty Quantification, Edinburgh*, 2010.

[197] Y. Song, Y. Zhang, C. Bajaj, and N. Baker. Continuum diffusion reaction rate calculations of wild type and mutant mouse acetylcholinesterase: Adaptive finite element analysis. *Biophysical Journal*, 87(3):1558–1566, 2004.

[198] Y. Song, Y. Zhang, T. Shen, C. Bajaj, J. McCammon, and N. Baker. Finite element solution of the steady-state smoluchowski equation for rate constant calculations. *Biophysical Journal*, 86(4):2017–2029, 2004.

[199] O. G. Staadt and M. H. Gross. Progressive tetrahedralizations. In *Proceddings of the IEEE Visualization Conference*, pages 397–402, 1998.

[200] J. E. Stone, J. Gullingsrud, and K. Schulten. A system for interactive molecular dynamics simulation. In *Proceedings of the 2001 symposium on Interactive 3D graphics*, pages 191–194. ACM Press, 2001.

[201] G. Strang and G. J. Fix. *An analysis of the finite element method*. Prentice-Hall Inc., Englewood Cliffs, N. J., 1973. Prentice-Hall Series in Automatic Computation.

[202] N. Sukumar. Quadratic maximum-entropy serendipity shape functions for arbitrary planar polygons. *Comput. Methods Appl. Mech. Engrg.*, 263:27–41, 2013.

[203] N. Sukumar and E. A. Malsch. Recent advances in the construction of polygonal finite element interpolants. *Archives Comput. Methods. Eng.*, 13(1):129–163, 2006.

[204] N. Sukumar and A. Tabarraei. Conforming polygonal finite elements. *Int. J. Numer. Methods Eng.*, 61(12):2045–2066, 2004.

[205] K. Sumikoshi, T. Terada, S. Nakamura, and K. Shimizu. A fast protein-protein docking algorithm using series expansions in terms of spherical basis functions. *Genome Informatics*, 16:161 – 193, 2005.

[206] G. Szegő. *Orthogonal Polynomials*. Amer. Math. Soc., Providence, 4th edition, 1975.

[207] A. Tabarraei and N. Sukumar. Application of polygonal finite elements in linear elasticity. *International Journal of Computational Methods*, 3(4):503–520, 2006.

[208] F. Tama. Normal mode analysis with simplified models to investigate the global dynamics of biological systems. *Protein and Peptide Letters*, 10(2):119 – 132, 2003.

[209] M. Topf, M. L. Baker, B. John, W. Chiu, and A. Sali. Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. *Journal of Structural Biology*, 149:191–203, 2005.

[210] M. Topf, K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali. Protein structure fitting and refinement guided by cryoem density. *Structure*, 16(2):295–307, 2008.

[211] L. G. Trabuco, E. Villa, K. Mitra, J. Frank, and K. Schulten. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure*, 2008.

[212] A. Varshney and F. Brooks. Fast analytical computation of richards's smooth molecular surface. In *VIS '93: Proceedings of the 4th conference on Visualization '93*, pages 300–307, 1993.

[213] R. Verfürth. A note on polynomial approximation in Sobolev spaces. *Math. Modelling Numer. Anal.*, 33(4):715–719, 1999.

[214] R. Voorintholt, M. T. Kosters, G. Vegter, G. Vriend, and W. G. Hol. A very fast program for visualizing protein surfaces, channels and cavities. *Journal of Molecular Graphics*, 7(4):243–245, December 1989.

[215] E. Wachspress. Barycentric coordinates for polytopes. *Comput. Math. Appl.*, 61(11):3319–3321, 2011.

[216] E. L. Wachspress. *A Rational Finite Element Basis*, volume 114 of *Mathematics in Science and Engineering*. Academic Press, New York, 1975.

[217] R. Walker. *Algebraic Curves*. Springer Verlag, New York, 1978.

[218] J. Warren. Barycentric coordinates for convex polytopes. *Adv. Comput. Math.*, 6(1):97–108, 1996.

[219] J. Warren, S. Schaefer, A. N. Hirani, and M. Desbrun. Barycentric coordinates for convex sets. *Adv. Comput. Math.*, 27(3):319–338, 2007.

[220] H. Whitney. *Geometric Integration Theory*. Princeton University Press, 1957.

[221] M. Wicke, M. Botsch, and M. Gross. A finite element method on convex polyhedra. *Comput. Graphics Forum*, 26(3):355–364, 2007.

[222] E. P. Wigner and J. J. Griffin. *Group theory and its application to the quantum mechanics of atomic spectra*, volume 4. Academic Press New York, 1959.

[223] J. B. with C. Bajaj, J. Blinn, M.-P. Cani-Gascuel, A. Rockwood, B. Wyvill, and G. Wyvill, editors. *Introduction to Implicit Surfaces*. Morgan Kaufmann Publishers, San Francisco, 1997.

[224] N. Woetzel, S. Lindert, P. Stewart, and J. Meiler. Bcl::em-fit: Rigid body fitting of atomic structures into density maps using geometric hashing and real space refinement. *Journal of Structural Biology*, 3(264-76), 2011.

[225] W. Wriggers. Using situs for the integration of multi-resolution structures. *Biophysical Reviews*, 2(1):21–27, 2010.

[226] W. Wriggers and P. Chacon. Multi-resolution contour-based fitting of macromolecular structures. *Journal of Molecular Biology*, 317:375–384, 2002.

[227] W. Wriggers, R. A. Milligan, K. Schulten, and J. A. McCammon. Self-organizing neural networks bridge the biomolecular resolution gap. *Journal of Molecular Biology*, 287(1247-1254), 1998.

[228] B. Wyvill, C. McPheeters, and G. Wyvill. Animating soft objects. *The Visual Computer*, 2(4):235–242, 1986.

[229] B. Wyvill, C. McPheeters, and G. Wyvill. Data structure for soft objects. *The Visual Computer*, 2(4):227–234, 1986.

[230] A. Yershova and S. M. LaValle. Deterministic sampling methods for spheres and SO(3). In *Proceedings. IEEE International Conference on Robotics and Automation.*, pages 3974 – 3980, 2004.

[231] S. Yoshimoto. Ballerinas generated by a personal computer. *Journal of Visualization and Computer Animation*, 3(2):55–90, 1992.

[232] T. You and D. Bashford. An analytical algorithm for the rapid determination of the solvent accessibility of points in a three-dimensional lattice around a solute molecule. *Journal of Computational Chemistry*, 16(6):743–757, 1995.

[233] Z. Yu and C. Bajaj. A segmentation-free approach for skeletonization of gray-scale images via anisotropic vector diffusion. *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 1:I–415–I–420 Vol.1, 2004.

[234] D. Zhang, J. Suen, Y. Zhang, Y. Song, Z. Radic, P. Taylor, M. Holst, C. Bajaj, N. Baker, and J. McCammon. Tetrameric mouse acetylcholinesterase: Continuum diffusion rate calculations by solving the steady-state smoluchowski equation using finite element methods. *Biophysical Journal*, 88(3):1659–1665, 2004.

[235] J. Zhang and F. Kikuchi. Interpolation error estimates of a modified 8-node serendipity finite element. *Numer. Math.*, 85(3):503–524, 2000.

[236] Q. Zhang, R. Bettadapura, and C. Bajaj. Macromolecular structure modeling from 3dem using volrover 2.0. *Biopolymers*, 97(9):709–731, 2012.

[237] Y. Zhang, C. Bajaj, and B.-S. Sohn. Adaptive and quality 3d meshing from imaging data. In *ACM Solid Modeling and Applications*, pages 286–291, 2003.

[238] Y. Zhang, C. Bajaj, and B.-S. Sohn. 3d finite element meshing from imaging data. *Special issue of Computer Methods in Applied Mechanics and Engineering on Unstructured Mesh Generation, in press*, 2004.

[239] W. Zhao, G. Xu, and C. Bajaj. An algebraic spline model of molecular surfaces. In *SPM '07: Proceedings of the 2007 ACM symposium on Solid and physical modeling*, pages 297–302, 2007.

[240] W. Zhao, G. Xu, and C. Bajaj. An algebraic spline model of molecular surfaces. In *Proceedings of the 2007 ACM symposium on Solid and physical modeling*, pages 297–302. ACM, 2007.

[241] W. Zheng. Accurate flexible fitting of high-resolution protein structures into cryo-electron microscopy maps using coarse-grained pseudo-energy minimization. *Biophysical Journal*, 100:478–488, 2011.

[242] O. Zienkiewicz and R. Taylor. *The Finite Element Method*. Butterworth-Heinemann, London, fifth edition, 2000.

[243] A. J. Zomorodian. *Topology for computing*, volume 16 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2009.