A Mathematical Primer for Computational Data Sciences

C. Bajaj

February 18, 2018
2.5 Polynomials, Piecewise Polynomials, Splines .................................................. 43
  2.5.1 Univariate case ......................................................................................... 43
  2.5.2 Bivariate case .......................................................................................... 44
  2.5.3 Multivariate case ....................................................................................... 44
2.6 Parametric and Implicit Representation .............................................................. 45
  2.6.1 Curves ....................................................................................................... 45
  2.6.2 Surface ..................................................................................................... 47
  2.6.3 Examples .................................................................................................. 47
2.7 Finite Elements and Error Estimation ................................................................. 50
  2.7.1 Tensor Product Over The Domain: Irregular Triangular Prism ................. 50
  2.7.2 Error Estimation ....................................................................................... 53
2.8 Biological Applications ....................................................................................... 53
  2.8.1 Tertiary Motif Detection ......................................................................... 53
  2.8.2 Ion channel models .................................................................................. 55
  2.8.3 Ribosome models ...................................................................................... 55
  2.8.4 Topological Agreement of Reduced Models ............................................. 56
  2.8.5 Dynamic Deformation Visualization ....................................................... 57
3 Differential Geometry, Operators ........................................................................... 59
  3.1 Shape Operators, First and Second Fundamental Forms ................................ 59
    3.1.1 Curvature: Gaussian, Mean .................................................................... 59
    3.1.2 The Shape of Space: convex, planar, hyperbolic ..................................... 59
    3.1.3 Laplacian Eigenfunctions ....................................................................... 59
  3.2 Finite Element Basis, Functional Spaces, Inner Products ................................. 59
    3.2.1 Hilbert Complexes .................................................................................. 59
  3.3 Topology of Function Spaces .......................................................................... 59
  3.4 Differential Operators and their Discretization formulas .................................. 59
  3.5 Conformal Mappings from Intrinsic Curvature ............................................... 60
  3.6 Biological Applications .................................................................................. 61
    3.6.1 Molecular Surface Analysis ................................................................... 61
    3.6.2 Solving PDEs in Biology ....................................................................... 61
3.7 Summary ......................................................................................................... 61
3.8 References and Further Reading ........................................................................ 61
3.9 Exercises ........................................................................................................ 61

4 Differential Forms and Homology of Discrete Functions ....................................... 63
  4.1 Exterior Calculus ............................................................................................ 63
  4.2 deRham Cohomology ...................................................................................... 64
  4.3 k-forms and k-cochains .................................................................................. 64
    4.3.1 Discrete Differential Forms .................................................................... 64
    4.3.2 Discrete Exterior Derivative ................................................................... 65
  4.4 Types of k-form Finite Elements ..................................................................... 66
    4.4.1 Nédelec Elements .................................................................................. 66
    4.4.2 Whitney Elements .................................................................................. 67
  4.5 Biological Applications ................................................................................... 68
    4.5.1 Solving Poisson’s Equation and other PDEs from Biology ....................... 68
4.6 Summary ......................................................................................................... 68
4.7 References and Further Reading ....................................................................... 68
4.8 Exercises ......................................................................................................... 68
## Contents

9 Optimization .................................................. 79
  9.1 Convex and Non-Convex .................................. 79
    9.1.1 Geometry of Convexity ............................ 79
    9.1.2 Convexity of Functions ............................ 79
    9.1.3 Convex Optimization Problems ..................... 80
    9.1.4 Non-convex Problems ............................... 84
  9.2 Combinatorial and Geometric ............................. 84
  9.3 Biological Applications ................................ 84
    9.3.1 Fast Computation Methods ....................... 84
Summary ......................................................... 84
References and Further Reading .............................. 84
Exercises ......................................................... 84

10 Statistics .................................................... 85
  10.1 Probability Primer ....................................... 85
    10.1.1 Probability Definitions ......................... 85
    10.1.2 Probability Distributions ....................... 87
    10.1.3 Pairwise independence ........................... 89
    10.1.4 Transformation of Random Variables ............. 90
    10.1.5 Annular Concentration of Gaussian ............... 91
    10.1.6 Distribution Sampling ............................ 92
    10.1.7 Mixture of Gaussians ............................. 93
    10.1.8 Concentration Theorems .......................... 94
    10.1.9 Application of Markov and Chebyshev Inequalities 95
    10.1.10 Chernoff Bounds ................................ 96
  10.2 Bayesian .................................................. 97
    10.2.1 Bayes Rule ......................................... 97
    10.2.2 Maximum Likelihood Estimator ..................... 98
    10.2.3 Unbiased Estimator ............................... 98
  10.3 Biological Applications ................................ 99
Summary ......................................................... 99
References and Further Reading .............................. 99

Biology Appendix ................................................. 101

Conclusion ...................................................... 103
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.
These sub-areas include sampling theory, matrix equations, numerical solution of differential equations, and optimization. Book recommendation:
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 dimension.
- 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.

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 directed 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 $\mathbb{R}^n$. 
1.3. TOPOLOGICAL COMPLEXES

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**

Triangulations 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 meaningful topological invariant. We will discuss this further in Section 1.4.1.

1.3.3 Morse Functions and the Morse-Smale Complex

Morse theory 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 [84] 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.

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. [42, 41]. Cazals, Chazal and Lewiner used the Morse complex for molecular shape analysis in [24]. 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 of 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 origin and destination (a minimum and a maximum). The yellow curve is a 1-cell (also called an arc): the intersection of an ascending and descending 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 [8, 25, 39, 54, 119]. Chazal and Lieutier [27] have used it for stable medial axis construction. Dey, Giesen and Goswami have used distance function for object segmentation and matching [36]. Goswami, Dey and Bajaj have used it for detailed feature analysis of shape via an annotation of flat and matching [36]. Goswami, Dey and Bajaj have used it for stable medial axis construction. Dey, Giesen and Goswami have used distance function for object segmentation and matching [36].

Cazals, Chazal and Lewiner used the Morse complex for molecular shape analysis in [24]. As described in Section 3, our work uses the Morse complex to aid in the curation of molecular surfaces.

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 $\Sigma$.

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

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 $\Sigma$.

$$h_\Sigma : \mathbb{R}^3 \to \mathbb{R}, \quad x \mapsto \inf_{p \in \Sigma} \|x - p\|$$
1.3. TOPOLOGICAL COMPLEXES

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 tessellation of space into cells induced by the gradient of 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 origin and destination (a minimum and a maximum). The yellow curve is a 1-cell (also called an arc): the intersection of an ascending 2-manifold (blue region) and a descending 1-manifolds (green+yellow curves, originating from the same saddle point).
CHAPTER 1. GRAPHS, TRIANGULATIONS AND COMPLEXES

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 [54, 55]. See also [39].

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 [104] 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 [36, 54] 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 finite portion of \( \mathbb{R}^3 \), 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:

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

We include a point at infinity denoted \( (m_\infty) \) in the set \( C_{3,E} \) to compactify the complementary 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).](image)

### 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 [54] 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 [36]. This produces a tetrahedral decomposition of the features captured.
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 [33] 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 8rA.

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)

We have so far 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 [30]. 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 manifolds using the algorithm described in [55]. \( U^2 \) 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 \( \text{Bd}(\sigma^k) \). The interior of \( \sigma^k \) is \( \text{Int}(\sigma^k) = \sigma^k \setminus \text{Bd}(\sigma^k) \). Note that \( \text{Int}(\sigma^0) = \sigma^0 \). The volume of \( \sigma^k \) is denoted \( |\sigma^k| \).

**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 \( |K| \).

**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.
1.5. PRIMAL AND DUAL COMPLEXES

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.

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.

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.

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 $\tau^k$.

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.

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.

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 $*\sigma^{n-k}$.
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 $\ast$ to indicate dual cells, i.e.

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

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 $\ast K$.

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 $s[c(\sigma^0), \ldots, c(\sigma^n)]$ 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 $\text{Vor} P$ of the point set $P$. 

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 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):

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

$$ (d - r)^4 = d^4 - 4d^3r + 6d^2r^2 - 4dr^3 + r^4$$

$$\leq d^4 - 2d^2r^2 + r^4 = (d^2 - r^2)^2$$

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

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

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

\[
\begin{align*}
    d & \leq r \\
    d - r & \leq 0 \\
    2d^2 - 2rd & \leq 0 \\
    2d^2 - 2rd + r^2 & \leq r^2 \\
    d^2 - 2rd + r^2 & \leq r^2 - d^2 \\
    (d - r)^2 & \leq r^2 - d^2 \\
    (d - r) & \leq \sqrt{r^2 - d^2} \\
    E(p, B) & \leq P(p, B)
\end{align*}
\]

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 $B = \{B_1, \ldots, B_n\}$ the distance functions are extended as follows:

\[
E(p, \mathcal{B}) = \min_{1 \leq i \leq n} |d_i - r_i| \tag{1.2}
\]
1.6. VORONOI AND DELAUNAY DECOMPOSITIONS

\[ P(p, B) = \sqrt{\min_{1 \leq i \leq n} |d_i^2 - r_i^2|} \] (1.3)

The problem in comparing \( E(p, B) \) with \( P(p, B) \) 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 \( B = \{B_1, B_2\} \) is achieved at \( i = 1 \) for \( P(p, B) \) and at \( i = 2 \) for \( E(p, B) \):

\[ P(p, B) = P(p, B_1) \]
\[ E(p, B) = E(p, B_2). \]

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

\[ P(p, B) = P(p, B_{i_P}) \]
\[ E(p, B) = E(p, B_{i_E}). \]

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

\[ E(p, B) = E(p, B_{i_E}) \leq E(p, B_{i_P}) \leq P(p, B_{i_P}) = P(p, B). \]

![Power diagram and Voronoi diagram of two circles](image1.png)

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)).
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 \( p \in \mathbb{R}^n \) and \( w \in \mathbb{R} \), the power distance from a point \( x \in \mathbb{R}^n \) 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 \( B \) 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 \( B \):

\[
V_B = \{x \in \mathbb{R}^2 | \pi_B(x) \leq \pi_C(x) \forall C \in B \}.
\]

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

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 stereographic 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):

\[
QEF[x] = \sum_i |n_i \cdot (x - p_i)|^2,
\]

(1.4)
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 [121, 122], 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 \( S_0 \).

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

(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, $S_0(0) = \{x|g(x) = 0\}$. All the grid points inside the interval volume $I_0(0) = \{x|g(x) \leq 0\}$ have negative function values, and all the grid points outside it have positive values.

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 [121, 122], 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 [121, 122] 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](image)

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.

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

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 [32, 98, 99, 100, 110, 111]. Other work on surface representations features the use of metaballs, molecular surfaces, and blobby models [3, 17, 114, 35, 49, 60, 70, 83, 89, 90, 115, 116, 117].

In previous work on dynamic triangulations the focus has been mostly on the simpler Delaunay/Voronoi structures (unweighted case) [9, 71, 29, 48, 62, 4, 93, 94]. 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
1.7. BIOLOGICAL APPLICATIONS

surface construction and second, an enumeration of the patches that make up the analytical surface (See [32], [31] and his PhD thesis). In [111], 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 [43], including parallel algorithms in [110] and a triangulation scheme in [3]. Fast computations of SES is described in [99] and [98], 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 [12], [11] and [13]. You and Bashford in [118] 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., [32, 31, 111, 43, 110, 3, 99, 98, 118, 59, 12, 11, 123, 14]), not much work has been done on neighborhood data structures for the dynamic maintenance of molecular surfaces as needed in MD. In [13] Bajaj et al. considered limited dynamic maintenance of molecular surfaces based on Non Uniform Rational BSplines ( NURBS ) descriptions for the patches. Eyal and Halperin [45, 46] presented a algorithm based on dynamic graph connectivity that updates the union of balls molecular surface after a conformational change in $O(\log^2 n)$ 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 [60, 90], blobby models [117]) or for collision detection [70]. Direct application of previous approaches for the decimation and multiresolution representation of the surfaces themselves [99, 77] 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 [108] 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 [96, 82, 75]. They vary in both the simplification scheme like vertex removal [34], edge contraction [68], triangle contraction [53], vertex clustering [102], wavelet analysis [37], 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 [28] 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 [69] 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 [70] 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 [95].
Useful references for graph theory include Behzad and Chartrand [16], Chartrand and Lesniak [26], and Giblin [52].

Minimal spanning trees are discussed in Graham and Hell [58] and algorithms for finding them in Kruskal [76].

Useful algebraic topology texts include Armstrong [5] and Hatcher [64]. 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 [65].

For practical computational topology, see Zomorodian [125].

For more on CW complexes, see Munkres [86].

The power distance metric is described in [7].

More on regular triangulations in [40, 44] and on weighted alpha shapes in [38, 47].

Some application references: [121, 122, 21, 22, 106, 107, 120, 73, 51, 67, 78].

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. \( f : V \to \mathbb{R} \times \cdots \times \mathbb{R} \) \( k \) 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 \( T : V^p \to \mathbb{R} \) that 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 \( \text{Alt} T \), 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

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

\( \diamond \)

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 $F$ together with two operations called addition and multiplication, $F$ with addition forms an Abelian group with identity element "0" while $F$ with multiplication forms an Abelian group with identity element "1"

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

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

where $\alpha, \beta \in F, a, b \in X$

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)
- $\bigcup_{i \in I} X_i \in \mathcal{X}$ (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:

- $d(x, y) \geq 0$
2.2. INNER PRODUCTS AND NORMS

- \( d(x, y) = 0 \iff x = y \)
- \( d(x, y) = d(y, x) \)
- \( d(x, z) \leq d(x, y) + d(y, z) \)

**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:

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

### 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: \( u + v \in V \text{ if } u, v \in V \)
  - scalar multiplication is continuous: \( \alpha u \in V \text{ if } \alpha \in F, u \in V \)

### 2.2.5 Normed Space

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

- \( N(v) \geq 0 \), and \( N(v) = 0 \text{ if and only if } v = 0 \)
- \( N(\alpha v) = |\alpha|N(v) \)
- \( N(u + v) \leq N(u) + N(v) \)

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:

  \[ \|(u_n + v_n) - (u + v)\| \leq \|u - u_n\| + \|v - v_n\| \to 0 \text{ as } n \to 0 \]

- Suppose \( \alpha_n \to \alpha \) in \( F \), then:

  \[ \|\alpha_n u_n - \alpha u\| \leq |\alpha - \alpha_n||u_n|| + |\alpha||u - u_n| \to 0 \text{ as } n \to \infty \]

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**: $U \rightarrow V$, $U, V$ 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 $V$:
  \[ \| u \|_U \leq C_1 \implies \exists C_2 \text{ such that } \| Au \|_V \leq C_2 \]
- **A** is **continuous** if and only if
  \[ \forall \varepsilon > 0, \exists \delta > 0 \text{ such that: } \| u - v \|_U < \delta \implies \| Au - Av \|_V < \varepsilon \]
  or if and only if, whenever $u_n \rightarrow u$ ($\| u - u_n \|_V \rightarrow 0$ as $n \rightarrow \infty$), we have:
  \[ \| Au - Av \|_V \rightarrow 0 \text{ as } n \rightarrow \infty \]

Theorem 2.17. Let $(U, \| \cdot \|_U), (V, \| \cdot \|_V)$ be normed spaces over the same field. Let $A : U \rightarrow V$ 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:
   \[ \| Au \|_V \leq C \| u \|_U \quad \forall u \in U \]

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

4)⇒ 1):
If $u_n \to u$, then:

$$\|A u - A u_n\|_V \leq C \|u - u_n\|_V \to 0 \text{ as } n \to \infty$$

\[\square\]

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:

- $\forall u \in V, p(u, u) \geq 0; p(u, u) = 0 \iff u = 0$
- $\forall u, v \in V, p(u, v) = \overline{p(v, u)}$ (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:

$$p(u, \beta_1 v_1 + \beta_2 v_2) = \overline{\beta_1} p(u, v_1) + \beta_2 p(u, v_2)$$

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 Cauchy-Schwarz Inequality.

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

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

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

$$0 \leq (u - \alpha v, u - \alpha v)$$

$$= (u, u) - \alpha(v, u) - \alpha(u, v) + \alpha \overline{v}(v, v)$$

$$= (u, u) - \frac{(v, v)}{(v, v)}(v, v) - \frac{(u, v)}{(v, v)}(u, v) + \frac{(v, u)(u, v)}{(v, v)^2}(v, v)$$

$$= \frac{1}{(v, v)} [(u, u)(v, v) - 2(u, v)^2 + |(u, v)|^2]$$

$$\text{Therefore } |(u, v)|^2 \leq (u, u)(v, v).$$

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

- $\|u\| \geq 0$ and $\|u\| = 0 \iff u = 0$
\[ \|u + v\| \leq \|u\| + \|v\| \]
\[ \forall \alpha \in \mathbb{F}, \|\alpha u\| = |\alpha|\|u\| \]

**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:

\[
\begin{aligned}
| (v_m, u_m) - (v, u) | &= | (v_m, u_m) + (v_m, u) - (v_m, u) - (v, u) | \\
&= | (v_m, u_m - u) + (v_m - v, u) | \\
&\leq \|v_m\|\|u_m - u\| + \|v_m - v\|\|u\| \\
&\to 0 \text{ as } m \to \infty
\end{aligned}
\]

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

- \( \cos \theta \overset{\text{def}}{=} \frac{(u, v)}{\|u\|\|v\|} \quad (\mathbb{F} = \mathbb{R}) \)
- 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 \( \mathbb{CP}^n \) 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 \( \mathbb{CA}^n \) 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 \( Z_d(f) \) of \( \mathbb{CA}^n \) 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.
of degree $d$. The hypersurface $Z^d_1(f)$ is also known as a flat or a hyperplane, a $Z^2_2(f)$ is known as a quadric hypersurface, and a $Z^3_3(f)$ is known as a cubic hypersurface. The hypersurface $Z^d_2$ is a plane curve of degree $d$, a $Z^2_3$ 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 $Z^d_n(f_i)$.

An irreducible rational hypersurface $Z^d_3(f)$, can additionally be defined by rational parametric equations which are given as $(x_1 = G_1(u_1, u_2, ..., u_{n-1}), x_2 = G_2(u_1, u_2, ..., u_{n-1}), ..., x_n = G_n(u_1, u_2, ..., u_{n-1}))$, where $G_1, G_2, ..., G_n$ are rational functions of degree $d$ in $u = (u_1, u_2, ..., u_{n-1})$, i.e., each is a quotient of polynomials in $u$ 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 [81]. 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[97]. 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 [81] 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

\[
\mathcal{X}^d = \{ X^\alpha | \alpha_0 + \ldots + \alpha_{m-1} = d \} \]

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 \( F \) using monomials in \( \mathcal{X}_i^d \). Specifically, let

\[
F_i^d = \frac{\mathcal{X}_i^d}{X_i^d} f_i.
\]

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 \( F_i^d \). 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 \( P \).

\[
A = \begin{pmatrix} X_0^d \\ \\ \\ X_m^d \end{pmatrix} = \begin{pmatrix} \ldots & a_{i1} & a_{i2} & a_{i3} & \ldots \\ \ldots & b_{j1} & b_{j2} & b_{j3} & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ X_m^d \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ X_m^d \end{pmatrix}
\]

Now, if the \( F_i \) have a common root \((\hat{X}_0, \ldots, \hat{X}_{m-1})\), then this root must satisfy all of the polynomial equations in the \( F_i^d \). This fact implies that the nontrivial vector \((\hat{X}_0, \ldots, \hat{X}_{m-1})\) 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

\[
R = \frac{\text{det}(A)}{\text{det}(B)}
\]

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 \( \text{det}(A) \) and \( \text{det}(B) \) may evaluate to zero even though \( R \) is not identically zero. Techniques for computing \( R \) by specializing before division has recently been considered in [23, 92].

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 \( \text{deg} f_1(f_2) \leq \text{deg} f_1(f_1) \), is the result of one step of pseudo-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 \( \alpha, \beta \in C \) and \( \text{deg} f_1(g) < \text{deg} f_1(f_2) \). Repeating the pseudo-division with \( f_2 \) and \( g \) and ensuring that the factors \( \alpha \) and \( \beta \) are ‘primitive’, one can compute a subresultant polynomial remainder sequence (p.r.s):

\[
f_1, f_2, g = S_{k-1}, \ldots, S_1, S_0
\]

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 [66, 80]. 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 \( X_1 \).)
2.4. HOMOGENEOUS AND BARYCENTRIC COORDINATES

For the set of polynomial equations (2.2), 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 \( S^{m-1} \):

\[
\begin{align*}
\hat{f}_1(x_2, \ldots, x_n) &= 0 \\
\ldots \\
\hat{f}_{m-1}(x_2, \ldots, x_n) &= 0
\end{align*}
\]

(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

\[
S = S^m, S^{m-1}, S^{m-2}, \ldots, S^1, S^0
\]

(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 \( \text{generalized barycentric coordinates} \)) are any set of functions satisfying certain key properties.

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

1. **Non-negative:** \( \lambda_i(x) \geq 0 \) on \( \Omega \).
2. **Linear Completeness:** For any linear function \( L : \Omega \to \mathbb{R}, L = \sum_{i=1}^{n} L(v_i)\lambda_i \).

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 \( T\Omega \).

1. **Invariance:** \( \lambda_i(x) = \lambda_i^T(T(x)) \).

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, \( |u|_{H^1(\Omega)} = \sqrt{\int |\nabla u(x)|^2 \, dx} \) 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, \( T(x) := hx \). 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:

\[
|\lambda_i^T|_{H^1(T\Omega)}^2 = \int_{T\Omega} |\nabla \lambda_i^T(x)|^2 \, dx = \int_{T\Omega} \left| \frac{1}{h} \nabla \left( \lambda_i^T(hx) \right) \right|^2 \, dx
\]

\[
= h^{d-2} \int_\Omega |\nabla \lambda_i(y)|^2 \, dy = h^{d-2} |\lambda_i|_{H^1(\Omega)}^2.
\]

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 \( L^p \) seminorm (\( p = 2 \), 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).
To describe more complicated curves through

Thus, barycentric coordinates give us an easy way to define and control the shape of lines through a triangle.

\[ \gamma \]

such that

\[ \gamma \]

real coefficients

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

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

\[
\begin{pmatrix}
  x \\
  y \\
  1
\end{pmatrix} =
\begin{pmatrix}
  x_1 & x_2 & x_3 \\
  y_1 & y_2 & y_3 \\
  1 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
  \lambda_1 \\
  \lambda_2 \\
  \lambda_3
\end{pmatrix}
\]

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

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

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

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

where

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

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 $\{ p_i \}_{i=0}^n$:

$$P_n = \sum_{i=0}^{m} p_i N^n_i(x)$$

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

$$N^n_i(x) = \begin{cases} 1 & \text{for } u_i \leq x \leq u_{i+1} \\ 0 & \text{otherwise.} \end{cases}$$
\[ N_n^l(x) = \frac{x - u_l - 1}{u_{l+n-1} - u_l - 1} N_n^{l-1}(x) + \frac{u_l + n - x}{u_{l+n} - u_l} N_n^{l+1}(x) \]

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

**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**

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

where

\[ B_i^m(x, y) = \binom{m}{i} x^i y^{m-i} (1 - x - y)^{m-i-j} \]

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

\[ \frac{x - a_i}{b_i - a_i}, \quad i = 1, 2, \ldots, d \]

\[ B_i^n(t) = \frac{n!}{(n-i)!} t^i (1-t)^{n-i} \]

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

where

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

**2.5.3 Multivariate case**

**Tensor Product**

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

**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:

\[ P(x, y, z) = \sum_{i=0}^{m} \sum_{j=0}^{m-i} \sum_{k=0}^{m-i-j} b_{ijk} B_i^n(x, y)B_j^n(y)B_k^n(z) \]
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

\[
p(x_1, \ldots, x_d) = \sum_{i_1 + \ldots + i_{d_1} \leq m} \sum_{j_1 = 0}^{n_1} \ldots \sum_{j_{d_2} = 0}^{n_{d_2}} b_{i_1 \ldots i_{d_1} j_1 \ldots j_{d_2}} \hat{P}_{i_1 \ldots i_{d_1}}^{m_{j_1}}(t_1) \ldots B_{j_{d_2}}^{n_{d_2}}(t_{d_2})
\]

where

\[
(x_1, \ldots, x_d)^T \in [\mathbf{p}_0, \ldots, \mathbf{p}_{d_1}] \times [a_{11}, b_{11}] \times \ldots \times [a_{d_2}, b_{d_2}]
\]

and

\[
t_i = \frac{x_{d_1 + i} - a_i}{b_i - a_i}, \quad i = 1, 2, \ldots, d_2
\]

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 \( [f_1(s)x - f_1(s) = 0, f_3(s)y - f_2(s) = 0] \) 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 set 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 \([112]\). 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.
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 \( O(d^3) \). 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 \( O(d^3) \), 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 \( O(d^5) \).

Let \( C : (f_1(x, y, z) = 0, f_2(x, y, z) = 0) \) 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 \( C : (x = G_1(t), y = G_2(t), z = G_3(t)) \) 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:

1. Substitute \( (x = G_1(t), y = G_2(t), z = G_3(t)) \) into the equation \( f(x, y, z) = 0 \).
2. 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 coefficients 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 \( C \).

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. \( \square \)

### 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 \((x, y)\) 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 \( c^2 \geq 4ab \), 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 \( c^2 \geq 4ab \), 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 \( 4af - e^2 \) 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 \( a_1 \)’s, \( c_1 \)’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 \( c_1 \)’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 some 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 conoid is implemented on a VAX-780 using VAXIMA.)

Geometrically speaking, a conoid 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 conoid meet the conoid 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 \((x, y)\) 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 conoid 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 conoid. 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 conoid becomes important and may be made appropriately, when the parameterization is desired only for a specific patch of the conoid.

**Step (1)** If \( C(x, y, z) \) has a real root at infinity, an *alinear transformation* of the type \( x' = a_1 x + b_1 y + c_1 z + d_1 \), \( y' = a_2 x + b_2 y + c_2 z + d_2 \) and \( z' = a_3 x + b_3 y + c_3 z + d_3 \) will suffice. If \( C(x, y, z) \) has no real root at infinity, we must use a *linear transformation* of the type \( x' = (a_1 x + b_1 y + c_1 z + d_1) h(a_4 x + b_4 y + c_4 z + d_4), y' = a_2 x + b_2 y + c_2 z + d_2) h(a_4 x + b_4 y + c_4 z + d_4) \) and \( z' = (a_3 x + b_3 y + c_3 z + d_3) h(a_4 x + b_4 y + c_4 z + d_4) \). This is equivalent to a *homogeneous linear transformation* of the type \( X' = a_1 X + b_1 Y + c_1 Z + d_1 H, Y' = a_2 X + b_2 Y + c_2 Z + d_2 H, Z' = a_3 X + b_3 Y + c_3 Z + d_3 H \) and \( H' = a_4 X + b_4 Y + c_4 Z + d_4 H \) applied to the homogeneous conoid \( C(X, Y, Z, H) = a X^2 + b Y^2 + c Z^2 + d X Y + e X Z + f Y Z + g X H + h Y H + i Z H + j H^2 = 0 \).

**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 conoid this corresponds to the roots of the homogeneous conic equation \( C(x, y, z) = a x^2 + b y^2 + c x z + f y z + c_2 = 0 \). Also, here the simultaneous truth of \( d^2 = 4 a b c, e^2 = 4 a c \) and \( f^2 = 4 b c \) corresponds to the existence of a polynomial parameterization for the conoid, 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 \( C(x', y', z') = I(a_1 x + b_1 y + c_1 z + d_1, a_2 x + b_2 y + c_2 z + d_2, a_3 x + b_3 y + c_3 z + d_3) \). 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 \( C(X', Y', Z', H') = I(a_1 X + b_1 Y + c_1 Z + d_1 H, a_2 X + b_2 Y + c_2 Z + d_2 H, a_3 X + b_3 Y + c_3 Z + d_3 H, a_4 X + b_4 Y + c_4 Z + d_4 H) \). To eliminate the \( Y^2 \) term we choose \( b_4 = 1 \) and \((b_1, b_2)\) to be a point on either the conic \( a x^2 + b y^2 + d x y + e x z + f y z + j z^2 = 0 \) with \( b_3 = 0 \) or a point on the conic \( a x^2 + b y^2 + d x y + (e + g) x z + (f + h) y z + (e + i + j) z^2 = 0 \) 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 \( a_2 = 1, c_1 = 1 \) 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 \( T \), let \([v_i, v_j, v_k]\) be one of the triangles where \( v_i, v_j, v_k \) are the vertices of the triangle. Suppose the unit normals of the surface at the vertices are also known, denoted as \( n_l \), \( (l = i, j, k) \). Let \( v_l(\lambda) = v_l + \lambda n_l \). First we define a prism (Figure 2.4) \( D_{ijk} := \{ p : p = b_1v_i(\lambda) + b_2v_j(\lambda) + b_3v_k(\lambda), \lambda \in I_{ijk} \} \), where \( (b_1, b_2, b_3) \) are the barycentric coordinates of points in \([v_i, v_j, v_k]\), and \( I_{ijk} \) is a maximal open interval containing 0 and for any \( \lambda \in I_{ijk}, v_i(\lambda), v_j(\lambda), v_k(\lambda) \) are not collinear and \( n_i, n_j, n_k \) point to the same side of the plane \( P_{ijk}(\lambda) := \{ p : p = b_1v_i(\lambda) + b_2v_j(\lambda) + b_3v_k(\lambda) \} \). Next we define a function in the Bernstein-Bezier (BB) basis over the prism \( D_{ijk} \):

\[
F(b_1, b_2, b_3, \lambda) = \sum_{i+j+k=n} b_{ijk}(\lambda) B_{n}^{ijk}(b_1, b_2, b_3),
\]

where \( B_{n}^{ijk}(b_1, b_2, b_3) \) is the Bezier basis

\[
B_{n}^{ijk}(b_1, b_2, b_3) = \frac{n!}{i!j!k!} b_1^i b_2^j b_3^k.
\]
2.7. **FINITE ELEMENTS AND ERROR ESTIMATION**

Non-degeneracy

(yiwang): I am not sure if the subsection title is correct. Let \( p_{ijk}^{(t)}(\lambda) = \text{det}\{n_i, v_j(\lambda) - v_i(\lambda), v_k(\lambda) - v_i(\lambda)\}, \ l = i, j, k. \)

Assume

\[
p_{ijk}^{(t)}(\lambda) > 0, \quad \forall \lambda \in [0, 1], \quad l = i, j, k.
\]

(2.10)

Consider the real numbers \( \lambda_1, \cdots, \lambda_s \) \( (s \leq 6) \) that solve one of these three equations of degree 2: \( p_{ijk}^{(t)}(\lambda) = 0, \ l = i, j, k, \) and define \( a = \max(-\infty, \{\lambda_i : \lambda_i < 0\}) \), \( b = \min(\infty, \{\lambda_i : \lambda_i > 1\}) \), 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

\[
n_i^T [v_j(\lambda) - v_i(\lambda)] \times [v_k(\lambda) - v_i(\lambda)] = p_{ijk}^{(t)}(\lambda) > 0,
\]

(2.11)

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

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 \( v_i, v_j, v_k, \) we define

\[
b_{300} = b_{030} = b_{003} = \lambda.
\]

(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 \( v_i, \) we require that the directional derivatives of \( F \) at \( v_i \) in the direction of \( b_2 \) \( b_3 \) are equal to \( \nabla F \cdot (v_j - v_i) \) \( \nabla F \cdot (v_k - v_i), \) respectively. Noticing that \( F \) has the form of (2.9) and \( (b_1, b_2, b_3) = (1, 0, 0) \) at \( v_i, \) one can derive that \( b_210 - b_{300} = \frac{1}{3} \nabla F(v_i) \cdot (v_j(\lambda) - v_i(\lambda)), \) where \( \nabla F(v_i) = n_i. \) Therefore

\[
b_{210} = \lambda + \frac{1}{3} n_i \cdot (v_j(\lambda) - v_i(\lambda)).
\]

(2.13)

\( b_{120}, b_{201}, b_{021}, b_{032}, b_{032} \) are defined similarly.

To obtain the \( C^1 \) continuity at the midpoints of the edges of \( T, \) we define \( b_{111} \) by using the side-vertex scheme [2]:

\[
b_{111} = w_1 b_{111}^{(1)} + w_2 b_{111}^{(2)} + w_3 b_{111}^{(3)},
\]

(2.14)

where

\[
w_i = \frac{b_i^2 b_k^2}{b_2^2 b_3^2 + b_1^2 b_3^2 + b_1^2 b_2^2}, \quad i = 1, 2, 3, \ i \neq j \neq k.
\]

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 \( v_j v_k, v_i v_k \) and \( v_i v_j. \) Consider the edge \( v_i v_j. \) Recall that any point \( p = (x, y, z) \) in \( D_{ijk} \) can be represented by

\[
(x, y, z)^T = b_1 v_i(\lambda) + b_2 v_j(\lambda) + b_3 v_k(\lambda).
\]

(2.15)

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

\[
I_3 = \begin{pmatrix}
\frac{\partial b_1}{\partial \lambda} & \frac{\partial b_2}{\partial \lambda} & \frac{\partial b_3}{\partial \lambda} \\
\frac{\partial b_1}{\partial \alpha} & \frac{\partial b_2}{\partial \alpha} & \frac{\partial b_3}{\partial \alpha} \\
\frac{\partial b_1}{\partial \beta} & \frac{\partial b_2}{\partial \beta} & \frac{\partial b_3}{\partial \beta}
\end{pmatrix}
\begin{pmatrix}
(v_i(\lambda) - v_k(\lambda))^T \\
(v_j(\lambda) - v_k(\lambda))^T \\
(b_1 n_i + b_2 n_j + b_3 n_k)^T
\end{pmatrix},
\]

(2.16)
CHAPTER 2. SETS, FUNCTIONS AND MAPPINGS

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

$$M := \begin{pmatrix} (v_i(\lambda) - v_k(\lambda))^T \\ (v_j(\lambda) - v_k(\lambda))^T \\ (b_1 n_i + b_2 n_j + b_3 n_k)^T \end{pmatrix}.$$  

(2.17)

and let $A = v_i(\lambda) - v_k(\lambda)$, $B = v_j(\lambda) - v_k(\lambda)$ and $C = b_1 n_i + b_2 n_j + b_3 n_k$, then $M = (A B C)^T$. From (2.16) we have

$$\begin{pmatrix} \frac{\partial b_1}{\partial x} & \frac{\partial b_2}{\partial x} & \frac{\partial b_3}{\partial x} \\ \frac{\partial b_1}{\partial y} & \frac{\partial b_2}{\partial y} & \frac{\partial b_3}{\partial y} \\ \frac{\partial b_1}{\partial z} & \frac{\partial b_2}{\partial z} & \frac{\partial b_3}{\partial z} \end{pmatrix} = M^{-1} = \frac{1}{\det(M)} (B \times C, C \times A, A \times B).$$  

(2.18)

According to (2.9), at the midpoint of $v_i\mathcal{V}v_j$, $(b_1, b_2, b_3) = (\frac{1}{2}, \frac{1}{2}, 0)$, we have

$$\begin{pmatrix} \frac{\partial F}{\partial b_1} \\ \frac{\partial F}{\partial b_2} \\ \frac{\partial F}{\partial b_3} \end{pmatrix} = \begin{pmatrix} (v_i(\lambda) - v_k(\lambda))^T \\ (v_j(\lambda) - v_k(\lambda))^T \\ (n_i + n_j)^T/2 \end{pmatrix} \begin{pmatrix} n \end{pmatrix} + \begin{pmatrix} \frac{3}{2}(b_{210} - b_{111}) \\ \frac{3}{2}(b_{120} - b_{111}) \end{pmatrix}.$$  

(2.19)

Define vectors

$$\begin{align*}
d_1(\lambda) &= v_j(\lambda) - v_i(\lambda) = B - A, \\
d_2(b_1, b_2, b_3) &= b_1 n_i + b_2 n_j + b_3 n_k = C, \\
d_3(b_1, b_2, b_3, \lambda) &= d_1 \times d_2 = B \times C + C \times A. 
\end{align*}$$

(2.20)

Let

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

(2.21)

$$d_3(\lambda) = d_3(\lambda) = C(\frac{1}{2}, \frac{1}{2}, 0, \lambda) = B = c \times c \times A.$$  

(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 $\nabla F \cdot d_3(\lambda) = 0$. Therefore, by (2.19) and (2.22), we have

$$b_{111} = \frac{d_3(\lambda)^T (3b_{210} B \times c + 3b_{120} c \times A + A \times B)}{3 \|d_3(\lambda)\|^2}.$$  

(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 $[v_i,v_j,v_k]$, 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

$$(x, y, z)^T = b_1 v_i(\lambda) + b_2 v_j(\lambda) + b_3 v_k(\lambda).$$  

(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 $v_i v_j$ of $\mathcal{T}$ satisfies $n_i \cdot (v_i - v_j) = n_j \cdot (v_j - v_i)$.

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 [124].
2.7.2 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 \( I : H^2(\Omega) \rightarrow \text{span}\{\lambda_i\} \subset H^1(\Omega) \) is given by

\[
Iu := \sum_i u(v_i)\lambda_i. 
\] (2.25)

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 \( x = (x, y) \), define \( x^\alpha := x^{\alpha_1}y^{\alpha_2} \), \( |\alpha| := \alpha_1 + \alpha_2 \), and \( D^\alpha u := \partial^{\alpha_1}u/\partial x^{\alpha_1}\partial y^{\alpha_2} \). The Sobolev semi-norms and norms over an open set \( \Omega \) are defined by

\[
|u|^2_{H^m(\Omega)} := \int_{\Omega} \sum_{|\alpha| = m} |D^\alpha u(x)|^2 \partial x \quad \text{and} \quad ||u||^2_{H^m(\Omega)} := \sum_{0 \leq k \leq m} |u|^2_{H^m(\Omega)}.
\]

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

\[
||u - Iu||_{H^1(\Omega)} \leq C \text{diam}(\Omega) ||u||_{H^2(\Omega)}, \quad \forall u \in H^2(\Omega). 
\] (2.26)

To prove estimate (2.26) 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

\[
||Iu||_{H^1(\Omega)} \leq C_I ||u||_{H^2(\Omega)}, \quad \forall u \in H^2(\Omega). 
\] (2.27)

We call equation (2.27) the \( H^1 \)-interpolant estimate associated to the barycentric coordinates \( \lambda_i \) used to define \( I \).

The optimal convergence estimate (2.26) 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).

Geometric Criteria

TBD

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’.
Figure 2.6: 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.
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 2.6a, 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 2.6. 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 2.6 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
Figure 2.8: 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 [50], 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.

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 2.7 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 [15]. 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 2.8. We use the software QSLim [50] 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) [79, 109]. 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, 63].

Exercises
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 $d_k : W^k \rightarrow W^{k+1}$ 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

$$(u, v)^{V_k} := (u, v)_{W^k} + (d_k u, d_k v)_{W^{k+1}}.$$ 

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:

$$0 \xrightarrow{d_0} H \Lambda^0 \xrightarrow{d_1} H \Lambda^1 \xrightarrow{d_2} \cdots \xrightarrow{d_{n-1}} H \Lambda^n \xrightarrow{d_n} 0.$$ 

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

A conformal map \( f : X \to Y \) is a 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 infinitesimal sense. Conformal maps are useful when parameterizing molecular surfaces as a triangulation with good angle bounds will preserve the angle bounds when passed through the parametrization to a surface triangulation.

To create a global conformal parameterization of a surface \( \Omega \) with arbitrary topology, one must solve 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 \( \omega + \sqrt{-1} \ast \omega \) where \( \omega \) and \( \ast \omega \) are real gradient fields on \( \Omega \). Further, \( \omega \) and \( \ast \omega \) 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:

\[
\begin{align*}
\frac{d\omega}{d} &= 0 \\
\Delta \omega &= 0 \\
\int_{e_i} \omega &= c_i
\end{align*}
\]

Given a solution to the above, one can construct an appropriate \( \ast \omega \) as well. The approach below is based on that of Gu and Yau [61] Section 3.2.

**Approach:** Let \( M \) be a mesh of \( \Omega \) with the correct topology. The computation of the homology basis is as follows.

B1. 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.

B2. Find a minimal spanning tree \( \bar{T} \) of the vertices of \( \bar{M} \).

B3. 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.

B4. 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 \( G \).
B5. 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 $\{L_1, \ldots, L_{2g}\}$ 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

$$L_i = \sum_{j=1}^{n_i} [u_{j-1}^i, u_j^i], \quad u_0 = u_{n_i}.$$

Then the PDE is discretized as

$$\begin{align*}
\sum_{j=1}^3 \omega([u_{j-1}, u_j]) &= 0 \quad \forall [u_0, u_1, u_2] \in M, u_0 = u_3 \\
\sum_{[u,v] \in M} k_{u,v} \omega([u,v]) &= 0 \quad \forall u \in M \\
\sum_{j=1}^{n_i} \omega([u_{j-1}, u_j]) &= c_i \quad \forall L_i
\end{align*}$$

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 [101, 61, 74, 103]

**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 \( \omega : \Omega \to \Lambda^k(T_x(\Omega)^*) \), \( \omega(x) : \bigoplus_{i=1}^k T_x(\Omega) \to \mathbb{R} \), 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 \( \Lambda^k(\Omega) \).

Definition 4.2. A differential \( dx_i \) is a 1-form whose action at \( x \in M \) is to assign the \( i \)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

\[
\{dx_I : I = \{i_1, \ldots, i_k\}, \ 1 \leq i_1 < \cdots < i_k \leq n\}
\]

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

\[
L^2\Lambda^k(\Omega) := \left\{ \sum_I a_I dx_I \in \Lambda^k(\Omega) : a_I \in L^2(\Omega) \ \forall I \right\}
\]
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 $I := \{i_1, \ldots, i_k\}$ 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 $\Omega$ ($0 \leq k < n$). 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 $C_k$. It has dimension $|C_k|$, equal to the number of elements of $K_k$. A $k$-chain $c$ is represented as a column vector of length $|C_k|$.

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 $\overline{C}_k$. 

Definition 4.10. A primal $k$-cochain $w$ is a linear functional on primal $k$-chains, i.e.

$$w : C_k \rightarrow \mathbb{R} \quad \text{via} \quad c \mapsto w(c),$$

where $w$ is a linear mapping. It is represented as a column vector of length $|C_k|$ so that the action of $w$ on a $k$-chain $c$ is the matrix multiplication $w^T c$, yielding the scalar $w(c)$. The space of primal cochains is denoted $C^k$.

A dual $k$-cochain $\mathcal{W}$ is a linear functional on dual $k$-chains, i.e.

$$\mathcal{W} : \mathcal{C}_k \rightarrow \mathbb{R} \quad \text{via} \quad c \mapsto \mathcal{W}(c),$$

where $\mathcal{W}$ is a linear mapping. The space of dual cochains is denoted $\mathcal{C}^k$.

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 $w$ over a primal $k$-chain $c$ is defined to be

$$\int_c w := w^T c = w(c).$$

Hence, the integration of $w$ over $c$ is exactly the same as the evaluation of $w$ 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:

$$\partial_k[v_0, v_1, \ldots, v_k] := \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k]$$

where $\hat{v}_i$ indicates that $v_i$ is omitted. The primal boundary operator is represented as a matrix of size $|C_{k-1}| \times |C_k|$ 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 $w$ 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 $|C_{k+1}| \times |C_k|$ so that the action of $\mathbb{D}_k$ on a primal $k$-cochain $w$ is the usual matrix multiplication $\mathbb{D}_k w := \partial_{k+1}^T w$.

The discrete exterior derivative satisfies the discrete version of Stokes’ theorem.

Lemma 4.14. Let $w$ be a primal $k$-cochain and $c \in C_{k+1}$ any primal $(k + 1)$-chain. Then

$$\int_c \mathbb{D}_k w = \int_{\partial_{k+1} c} w.$$ 

Proof. By Definition 4.11 we see that

$$\int_{\partial_{k+1} c} w = w^T \partial_{k+1} c = (\partial_{k+1}^T w)^T c = (D_k w)^T c = \int_c D_k w.$$ 

We now consider the analogous constructions for dual cochains. Observe that mesh duality allows us to view a dual $k$-chain $\mathcal{T}$ as a primal $(n - k)$-chain $c$. Hence $\partial_{n-k+1}$ 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 $\mathcal{W}$ is $\mathbb{D}_{n-k-1}^T$, which is equal to $\partial_{n-k}$.

It is represented as a matrix of size $|\mathcal{C}_{n-k-1}| \times |\mathcal{C}_k|$.
4.4 Types of \(k\)-form Finite Elements

4.4.1 Nédélec Elements

The original Nédélec paper [87] 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 \(S_p \subset \widetilde{P}_p\) is defined for domains embedded in \(\mathbb{R}^3\) by

\[
S_p := \{ f \in \widetilde{P}_p : (x_1, x_2, x_3) \cdot f = 0 \}.
\]

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 [87] is defined as follows. Set

- \(K\) to be a tetrahedron
- \(P\) to be \(P_{p-1} \oplus S_p\)
- \(A\) to be the following linear functionals acting on an element \(u \in \widetilde{P}_p\)

\[
\begin{align*}
1 & \int_{\partial K} (u \cdot \hat{e}) \cdot q \, ds, \forall q \in P_{k-1} \\
& \text{where } \hat{e} \text{ is the unit vector directed along the edge } e \text{ of } K; \\
2 & \int_{\partial K} (u \times \hat{n} \cdot q) \, d\gamma, \forall q \in (P_{k-2})^2 \\
& \text{where } f \text{ is a face of } K \text{ and } u \times \hat{n} \text{ denotes the normal trace of } u; \\
3 & \int_K u \cdot q \, dx; \forall q \in (P_{k-3})^3
\end{align*}
\]

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 [87] is similar. Set

- \(K\) to be a tetrahedron
- \(P\) to be \(P_{p-1} \oplus \widetilde{P}_p (x_1, x_2, x_3)\)
- \(A\) to be the following linear functionals acting on an element \(u \in \widetilde{P}_p\)

\[
\begin{align*}
\text{B1.} & \int_{\partial K} (u \cdot \hat{n}) \cdot q \, d\gamma, \forall q \in P_{k-1} \\
& \text{where } f \text{ is a face of } K; \\
\text{B2.} & \int_K u \cdot q \, dx; \forall q \in (P_{k-2})^3
\end{align*}
\]

Similarly, description of the basis functions was left to future work.

\footnote{The space \(\widetilde{P}_p (x_1, x_2, x_3)\) is called \(D^p\) 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 $v_i$ 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 $\lambda_i(v_j) = \delta_{ij}$. The Whitney function $W_{\sigma^k}$ associated to the $k$-simplex $\sigma^k := [v_0, \ldots, v_k]$ is given by

$$W_{\sigma^k} := k! \sum_{i=0}^{k} (-1)^i \lambda_i \, d\lambda_0 \wedge \ldots \wedge \hat{d\lambda_i} \wedge \ldots \wedge d\lambda_k$$

(4.2)

where $\hat{d\lambda_i}$ 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 $W_{\sigma^3}$ is the constant function with value $1/|\sigma^3|$. This is a consequence of the geometric identity

$$\nabla \lambda_i \cdot (\nabla \lambda_j \times \nabla \lambda_k) = \pm \frac{1}{3!} |\sigma^3|$$

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 $(1/|\sigma^3|) \sum \lambda_i$, which is simply $1/|\sigma^3|$ due to the partition of unity formed by the barycentric functions.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\sigma^k$</th>
<th>$W_{\sigma^k}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$[v_0]$</td>
<td>$\lambda_0$</td>
</tr>
<tr>
<td>1</td>
<td>$[v_0, v_1]$</td>
<td>$\lambda_0 \nabla \lambda_1 - \lambda_1 \nabla \lambda_0$</td>
</tr>
<tr>
<td>2</td>
<td>$[v_0, v_1, v_2]$</td>
<td>$2(w_0 \nabla w_1 \times \nabla w_2 - w \nabla w_1 \times \nabla w_02 + w \nabla w_2 \times \nabla w_01)$</td>
</tr>
<tr>
<td>3</td>
<td>$[v_0, v_1, v_2, v_3]$</td>
<td>$1/</td>
</tr>
</tbody>
</table>

Table 4.1: Whitney forms $W_{\sigma^k}$ for $n = 3$.

**Whitney Element Example**

On the standard reference triangle shown above, the Whitney 0-forms are the barycentric functions:

- $\lambda_0 = -x - y + 1$
- $\lambda_1 = x$
- $\lambda_2 = y$

The Whitney 1-forms are formed by taking products of the form $\lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i$:

- $W_{01} = \lambda_0 \nabla \lambda_1 - \lambda_1 \nabla \lambda_0 = \begin{bmatrix} 1 - y \\ x \end{bmatrix}$
- $W_{02} = \lambda_0 \nabla \lambda_2 - \lambda_2 \nabla \lambda_0 = \begin{bmatrix} y \\ 1 - x \end{bmatrix}$
- $W_{12} = \lambda_1 \nabla \lambda_2 - \lambda_2 \nabla \lambda_1 = \begin{bmatrix} -y \\ x \end{bmatrix}$
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 [113], it was recognized by Bossavit [20] 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) [6] gives a full account of the analogies between spaces of Whitney functions and classical Nédélec [87, 88] and Raviart and Thomas [91] 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 [56] and on square-base pyramids by considering the collapse of a cube to a pyramid [57]. Bossavit has given an approach to Whitney forms over standard finite element shapes (hexahedra, triangular prisms, etc.) based on extrusion and conation arguments [19].

Exercises
Chapter 5

Numerical Integration, Linear Systems

5.1 Numerical Quadrature

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 Crystal Symmetries
  7.1.1 Symmetries
  7.1.2 Quasi-symmetries

7.2 Hexagonal Tilings
  7.2.1 Caspar-Klug coordinate system
  7.2.2 T-numbers
  7.2.3 P-numbers

7.3 Icosahedral Packings

7.4 Biological Applications
  7.4.1 Crystal Structures
  7.4.2 Viral Capsid Symmetry Detection and Classification
  7.4.3 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

8.3 Sampling

8.3.1 Monte Carlo and Quasi Monte Carlo Integration

Let \( g: [0, 1] \rightarrow [0, 1] \) be an integrable function. Suppose that its derivative exists and is bounded by \( |g'(x)| \leq c \). We want to calculate \( I(g) = \int_0^1 g(x) \, dx \). Note that if \( X \) is a random variable and \( X_i \sim U(0, 1) \) are \( n \) independent random values uniformly distributed in the interval \( [0, 1] \) then \( E[g(X)] = \frac{1}{n} \sum_{i=1}^n g(X_i) \rightarrow \int_0^1 g(x) \, dx \) 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.

\[
\text{Var}[g(X)] = \left( \frac{1}{n} \sum_{i=1}^n g(X_i) - \int_0^1 g(x) \, dx \right)^2 = \frac{\sigma^2}{n}
\]

Let \( f: \{0, 1\}^n \rightarrow [0, 1] \) be a function mapping \( n \)-bit strings to reals such that \( f(x) = g(\tilde{x}) \), where \( \tilde{x} \) is the value of the binary string prepending by a radix point. Then, we have

\[
\frac{1}{2^n} \sum_{x \in \{0, 1\}^n} \left( f(x) - \frac{c}{2^n} \right) \leq \int_0^1 g(x) \, dx \leq \frac{1}{2^n} \sum_{x \in \{0, 1\}^n} \left( f(x) + \frac{c}{2^n} \right).
\]

This implies that the average estimate \( \tilde{f} = \frac{1}{2^n} \sum_{x \in \{0, 1\}^n} f(x) \) of \( f \) is a close approximation to the integral \( \int_0^1 g(x) \, dx \).

**Theorem 8.1** ([85, Theorem 13.5]). Let \( f \) and \( \tilde{f} \) be defined as above and let \( x_1, \ldots, x_m \) be \( m \) samples chosen i.i.d. uniform in \( \{0, 1\}^n \). If \( m > \frac{1}{\epsilon^2} \ln \frac{2}{\delta} \), then

\[
\Pr \left[ \left| \frac{1}{m} \sum_{i=1}^m f(x_i) - \tilde{f} \right| \geq \epsilon \right] \leq \delta.
\]

Suppose we want to obtain similar error with fewer number of purely random bits. Let \( x_1, x_2, \ldots, x_m \) be pairwise independent. Let \( Y = \frac{1}{m} \sum_{i=1}^m f(x_i) \). Then, \( E[Y] = E[f(x)] \). By Chebyshev inequality,

\[
\Pr \left[ |Y - E[f(x)]| \geq \epsilon \right] \leq \frac{\text{Var}[Y]}{\epsilon^2} \leq \frac{\text{Var}[\frac{1}{m} \sum_{i=1}^m f(x_i)]}{\epsilon^2} = \frac{\text{Var}[\sum_{i=1}^m f(x_i)]}{m^2 \epsilon^2} \leq \frac{1}{m \epsilon^2}.
\]
Thus, \( \Pr[|Y - \hat{f}| \geq \epsilon] \leq \delta \) when \( m \geq \frac{1}{\delta^2} \).

8.3.2 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 \( x \in \{0, 1\}^n \) be an \( n \)-bit string. Denote \( \tilde{x} \) as the binary value of the string \( x \) prepended by a radix point. For example, if \( x = 110011 \), then \( \tilde{x} = 0.110011 \)\text{bin} = \( \frac{51}{64} \).

Define \( f(x) = g(\tilde{x}) \). 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

\[
\frac{1}{2^n} \sum_{x \in \{0,1\}^n} \left( f(x) - \frac{c}{2^n} \right) \leq \int_0^1 g(x) \, dx \leq \frac{1}{2^n} \sum_{x \in \{0,1\}^n} \left( f(x) + \frac{c}{2^n} \right).
\]

If \( n \) is increasingly large, the estimate \( \frac{1}{2^n} \sum_{x \in \{0,1\}^n} f(x) \) converges to the true \( \int_0^1 g(x) \, dx = \mathbb{E}[g(X)] \). Furthermore, if we sample \( x \) either i.i.d. or pairwise independent, we can also bound the error (Figure 8.1(a)).

![Figure 8.1: (a) Sample points in the unit interval used for integral estimation. (b) Axis-parallel rectangles anchored at the origin used to measure star discrepancy.](image)

We can also bound the integration error by the discrepancy using the Koksma-Hlawka Inequality (Figure 8.1(b)). See [105] for a list of references.

**Theorem 8.2** (Koksma-Hlawka Inequality). Let \( \Delta(t) = \frac{1}{N} \sum_{i=0}^{N-1} \mathbb{1}_{[0,t]}(x_i) - \frac{t}{\mathbb{1}} \) 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

\[
\left| \int_0^1 g(x) \, dx - \frac{1}{N} \sum_{i=0}^{N-1} g(x_i) \right| \leq \|\Delta\|_p \|g'\|_q.
\]

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
\[
\mathbb{1}_{[0,t]}(x) = \begin{cases} 
1 & \text{if } x \in [0,t] \\
0 & \text{otherwise}
\end{cases}
\]
is the characteristic function. The star discrepancy 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 to bound the integration error. See [105] for a list of references.

**Theorem 8.3 (Koksma-Hlawka inequality).** For all \( 1 \leq p, q \leq \infty \) such that \( \frac{1}{p} + \frac{1}{q} = 1 \),
\[
\left| \int_0^1 f(x) \, dx - \frac{1}{N} \sum_{i=0}^{N-1} f(x_i) \right| \leq \|\Delta_P\|_p \|f'\|_q.
\]

**Proof.** Observe that \( f(x) \) and \( t \) can be written as follows.
\[
f(x) = f(1) - \int_x^1 f'(t) \, dt
\]
\[
= f(1) - \int_0^1 \mathbb{1}_{[0,t]}(x)f'(t) \, dt
\]
\[
t = \int_0^1 \mathbb{1}_{[0,t]} \, dx
\]
We can rewrite the integral as follows.
\[
\int_0^1 f(x) \, dx = \int_0^1 \left( f(1) - \int_0^1 \mathbb{1}_{[0,t]}f'(t) \, dt \right) \, dx
\]
\[
= f(1) - \int_0^1 \int_0^1 \mathbb{1}_{[0,t]}(x)f'(t) \, dt \, dx
\]
\[
= f(1) - \int_0^1 tf'(t) \, dx.
\]  
(8.1)

we can rewrite the average as follows.
\[
\frac{1}{N} \sum_{i=0}^{N-1} f(x_i) = \frac{1}{N} \sum_{i=0}^{N-1} \left( f(1) - \int_0^1 \mathbb{1}_{[0,t]}(x_i)f'(t) \, dt \right)
\]
\[
= f(1) - \frac{1}{N} \sum_{i=0}^{N-1} \mathbb{1}_{[0,t]}(x_i)f'(t) \, dt
\]  
(8.2)

The result follows by subtracting (8.1) from (8.2) and applying Hölder inequality. \( \square \)
8.4 Biological Applications

8.4.1 Docking Problem

8.4.2 Flexible Fitting

Summary

References and Further Reading

Exercises
Chapter 9

Optimization

9.1 Convex and Non-Convex

9.1.1 Geometry of Convexity

Convex Set $C$

Line joining all pair of points is always contained in the set $C$

$\forall x_1, x_2 \in C \quad 0 \leq t \leq 1 \rightarrow x_1 t + x_2 (1 - t) \in C$

Convex combination of $x_1, x_2, \ldots, x_k$ is given by $x = t_1 + t_2 + \ldots + t_k = 1 \quad t_k \geq 0$

For a point set $S$, Convex hull(s): set of all combinations of points in $S$

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 that it generalizes/sweeps out the line segment between $x_1$ and $x_2$

Hyperplane: $\{ x \mid w^T x = b, w \neq 0 \}$

$w$ is normal vector, then we can define the Half-space $\{ x \mid w^T x \leq b, w \neq 0 \}$ which is the region determined by the hyperplane.

Polygons and Polyhedra = Intersection of Half-spaces

Norm Balls and Norm Cones

Norm Unit Ball at origin $\{ x \mid \| x \| \leq 1 \}$ Norm Ball at center $x_c$ and radius $r$ $\{ x \mid \| x - x_c \| \leq r \}$ Norm Cone $\{ (x, t) \mid \| x \| \leq t \}$

In general, balls with respect to $L_q$ norms are denoted as $B_q(r) := \{ x \in \mathbb{R}^p, \| x \|_q \leq r \}$. As a special case, the notation $B_0(s)$ is used to denote the set of $s$-sparse vectors, which might not be a convex set.

9.1.2 Convexity of Functions

Definition

A. $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex if $\text{dom} f$ is convex and

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$$

(Jensen’s Inequality)

Strictly convex: strict inequality, $0 \leq t \leq 1; 0 < t < 1$

B. $f$ is concave if $-f$ is convex.
CHAPTER 9. OPTIMIZATION

Figure 9.1: Examples for convex sets and non-convex sets. Adapted from [72]

Example

- Affine $f(x)^F = W^T x + b$
- Exponential $e^{tx}$ for any $t \in R$
- Powers $x^\alpha \ \forall \ \alpha \geq 1 \ or \ \alpha \leq 0$
- Negative Entropy $x \log x$
- Norms $f(x) = \| x \|_2 = G_{\text{max}}(x) = (\lambda_{\text{max}}(x^T x))^{1/2}$ spectral norm: $f(x) = \| x \|_2$ max singular value:
  - Quadratic Form $f(x) = 1/2 x^T P x + q^T x + r$

Convex of $P \geq 0$ (positive semi-definite)

- Sub-Level Set of $f : R^d \to R \quad C_\infty = \{ x \in \text{dom} f | f(x) \leq \infty \}$
  - sub-level sets of convex functions are convex and of co-dim = 1
- Epigraph of $f : R^d \to R$
  - $\text{epi} f = \{ (x,t) \in R^{n+1} | x \in \text{dom} f, f(x) \leq t \}$
  - $f$ is convex if $\text{epi} f$ is a convex set

9.1.3 Convex Optimization Problems

A Linear Programming
B Quadratic Programming
C Polynomial optimization
D Geometric Programming
E Semi-Definite Programming

**optimal value**

\[ p^* = \inf \{ f \circ (x) | f_i(x) \leq 0, h_j(x) = 0 \} \]

\[ p^* = \infty \text{ if problem is infeasible (no solution salesfree constraints)} \]

\[ p^* = -\infty \text{ if problem is unbounded below} \]

- \( x \) if feasible if \( x \in \text{dom} f \circ \) and satisfies constraints
- feasible \( x \) is optimal if \( f \circ (x) = p^* ; x_{opt} = \text{set of optimal points} \)
- \( x \) is locally optimal if \( \exists R > 0 \) such that \( x \) is optimal for \( \min f \circ (z) \)

\[ z \in \mathbb{R}^n \]

\[ \text{s.t. } f_i(z) \geq 0, h_j(z) \leq 0, \|z - x\|_2 \leq r \] \hspace{1cm} (9.1)

**Examples**

(i) \( f_0(x) = 1/x \) \hspace{1cm} \( \text{dom } f_0 = \mathbb{R} + \text{ortho} \) : \( p^* = 0 \) ; no optimal point

(ii) \( f_0(x) = -\log x \) \hspace{1cm} \( \text{dom } f_0 = \mathbb{R} + \text{ortho} \) : \( p^* = -\infty \)

(iii) \( f_0(x) = x \log x \) \hspace{1cm} \( \text{dom } f_0 = \mathbb{R} + \text{ortho} \) : \( p^* = -1/e, x = 1/e \) is optimal

- General Problem has implicit constraint \( x \in D = \cap \text{dom } f_i \cap \text{dom } h_i \) \hspace{1cm} \((i = 0, d = 1)\)

- Unconstrained Problem (\( m = p = 0 \))

Minimize \( f_0(x) = -\sum_{i=1}^k \log(b_i - a_i^T x) \)

Domain of problem has no explicit constraint but implicit \( a_i^T x < b_i \)

- Optimality criterion for differentiable \( f_0 \)

\[ \nabla f_0(x)^T(y - x) \geq 0 \] for all feasible \( y \)

i.e. \( \nabla f_0(x) \) is te tangent Hyperplane to feasible set \( x \) at \( x \)

- Least Squares Problem \hspace{1cm} (see lecture II)

minimize \( ||Ax - b||_2^2 \)

**Solution** (Analytic) \( x^* = (A^T A)^{-1} A^T b \)

Comp. Time \( O(n^2 k) \) when \( A \in \mathbb{R}^{k \times n} \)
Linear Programming (LP)

Primal LP: minimize $C^T x$  $C \in \mathbb{R}^n$, $X \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$

subject to $A x \{\leq, =, \geq\} b$, $x \geq 0$

- Has many applications (industry)
- Proof technique for polynomial
- Hardest of Polynomial time algorithms

Algorithms
- Simplex Method (Dantzig 1947)
- Ellipsoid Algorithm (Shor, Khachian 1979)
- Interior Point Methods (Karmarkar 1984)

Equivalent Forms

Max to Min  $\max c^T x \iff \min \{ -c^T x \}$

Equality to Inequality  $a_i^T x = b_i \iff \{a_i^T x \leq b_i, a_i^T x \geq b_i\}$

Inequality to Non-negativity  $a_i^T x \leq b_i \iff \{a_i^T x + s = b_i \mid s \geq 0, s \in \mathbb{R}^n\}$

Variables unrestricted $x_j$ unrest $\iff \{x_j^+ \geq 0, x_j^- \geq 0\}$

Duality

Given a solution $x$ to an LP, how do we decide whether or not $x$ is in fact an optimum solution?

Calculate a lower bound on $\min C^T x$ given $Ax = b$, $x \geq 0$

Suppose $\exists y$ such that $A^T y \leq c$

Then $y^T b = y^T Ax = (A^T y)^T x \leq C^T x$

Hence $y^T b$ is a lower bound on LP; so to get best lower bound

Dual LP  $\max b^T y$ subject to $A^T y \leq c$

Weak Duality

If the primal LP has an optimum value $z$, then it has a dual LP with optimum value $w$ and $z \geq w$
9.1. CONVEX AND NON-CONVEX

Also for infeasible LP

- infeasible min prob \( \iff \) value = +\( \infty \)
- unbounded min prob \( \iff \) value = +\( \infty \)
- infeasible max prob \( \iff \) value = -\( \infty \)
- unbounded max prob \( \iff \) value = -\( \infty \)

Strong Duality Thm

If Primal or Dual is feasible, then \( z = w \)

Lagrangian

\[
L : R^n \times R^m \times R^p \to R \quad \text{with} \quad \text{dom} L = D \times R^m \times R^p
\]

\[L(x, \lambda, \nu) = f \circ (x) + \sum_{i=1}^{m} \lambda_i f_i(x) = \sum_{j=1}^{p} \nu_j h_j(x)\]

Lagrange Multipliers

Lagrange Dual Function

\[
g : R^m \times R^p \to R \quad g(\lambda, \nu) = \inf L(x, \lambda, \nu)
\]

Lower Bound Property if \( \lambda \geq 0 \) then \( g(\lambda, \nu) \leq b^* \)

Example

Least norm sol of Linear Equations

\[
\begin{align*}
\min & \quad x^T x \\
\text{subject to} & \quad Ax = b
\end{align*}
\]

Lagrangia \( L(x, \nu) = x^T x + \nu^T (Ax - b) \)

Minimize \( \iff \) \( \nabla_x L(x, \nu) = 2x + A^T \nu = 0 \quad \to \quad x = \frac{1}{2} A^T \nu \)

Substitute into \( L \) to obtain \( g(\nu) = L(-1/2A^T \nu, \nu) = -1/4x^T AA^T \nu - b^T \nu \)

Lower Bound Property \( p^* \geq -1/4 \nu^T AA^T \nu - b^T \nu, \forall \nu \)
9.1.4 Non-convex Problems

9.2 Combinatorial and Geometric

9.3 Biological Applications

9.3.1 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 uncertainty 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. 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 $\Pr[X = 0.527] = \frac{0}{\infty} = 0$. We can define probability density function (pdf) $p$ such that

$$\Pr[a \leq X \leq b] = \int_a^b p(x) \, dx.$$

Notice that the integral is a linear operator on $p$. The cumulative distribution function (cdf) is

$$P(a) = \Pr[X \leq a] = \int_{-\infty}^a p(x) \, dx.$$

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 \( \Pr[a \leq X \leq b] = \int_a^b p(x) \, dx \) and the cumulative probability \( P(a) = \Pr[X \leq a] = \int_{-\infty}^a p(x) \, dx \) 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.
10.1. PROBABILITY PRIMER

Expectation

The *expectation* of a random variable $X$ with pdf $p$ is defined as

$$E[X] = \int_{-\infty}^{+\infty} xp(x) \, dx.$$ 

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

$$\Pr[A_1 \cup A_2 \cup \ldots \cup A_n] \leq \sum_{i=1}^{n} \Pr[A_i]$$

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 an 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$_2$O) 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

$$\text{Var}(X) = \sigma^2(X) = E[X - E^2[X]]^2$$

$$= E[X^2] - 2E[X]E[X] + E^2[X]$$

$$= E[X^2] - E^2[X].$$

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, $\text{Var}(X_1 + X_2) \neq \text{Var}(X_1) + \text{Var}(X_2)$. 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

$$\text{Var}(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i).$$

10.1.2 Probability Distributions

Gaussian Distributions

The *Gaussian distribution* is related to Central Limit Theorem.

**Theorem 10.1** (Central Limit Theorem [18, 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

$$\phi(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \text{Exp} \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right].$$

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 $9 = \left(\frac{2+3}{2}\right) - 1$ 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

$$X = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob. } 1 - p \end{cases}.$$  

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 discrete random variables $X_1, X_2, \ldots, X_n$ are mutually independent if for any subset $I \subseteq \{1, 2, \ldots, n\}$,

$$\Pr \left[ \bigcap_{i \in I} X_i = x_i \right] = \prod_{i \in I} \Pr[X_i = x_i].$$

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 $|I| \leq k$,

$$\Pr \left[ \bigcap_{i \in I} X_i = x_i \right] = \prod_{i \in I} \Pr[X_i = x_i].$$

For example, if $X = (X_1, X_2, \ldots, X_n) \sim N(0, I)$, then $X_i \sim N(0, 1)$ are mutually independent. If $X = (X_1, X_2, \ldots, X_n)$ is uniformly distributed on the hypercube $[0, 1]^n$, then $X_i \sim U(0, 1)$ 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 $|I| \leq k$,

$$\Pr \left[ \bigcap_{i \in I} X_i = x_i \right] = \prod_{i \in I} \Pr[X_i = x_i].$$

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 $(a, b)$,

$$\Pr[X_1 = a \land X_2 = b] = \Pr[X_1 = a] \Pr[X_2 = b].$$

Lemma 10.2 ([85, 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 $j$th subset of the enumeration. Then, set

$$Y_j = \left( \sum_{i \in S_j} X_i \right) \mod 2 = \bigoplus_{i \in S_j} X_i,$$

where $\oplus$ denotes XOR.

Lemma 10.3 ([85, Lemma 13.2]). Let $X_1$ and $X_2$ be independent and uniform over $GF(p)$. Generate $Y_i = (X_1 + iX_2) \mod p$ 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 $GF(p)$ as $X_1$ is uniform over $GF(p)$.

Consider $Y_i$ and $Y_j$, where $i \neq j$. For any $a, b \in \{0, 1, 2, \ldots, p - 1\}$,

$$Y_i = X_1 + iX_2 = a \land Y_j = X_1 + jX_2 = b \iff X_2 = \frac{b - a}{j - i} \text{ and } X_1 = a - \frac{i(b - a)}{j - i}.$$ 

Hence, $\Pr[Y_i = a \land Y_j = b] = \frac{1}{p^2}$.
CHAPTER 10. STATISTICS

Finite Field

Finite fields appear in Rijndael — the AES cryptographic system. The Reed Solomon code uses the Galois field GF(2^n), which is a field of characteristic 2.

The field GF(2^n) consists over polynomials

\[ P(x) = \sum_{i=0}^{n-1} c_i x^i \quad (c_i \in GF(2) = \{0, 1\}) \]

of degree less than \( n \) over the field GF(2) modulo an irreducible polynomial over GF(2). (For example, the polynomial \( x^2 + 2x + 1 = (x + 1)^2 \) is reducible over IR. The polynomial \( x^2 + 1 = (x + i)(x - i) \) is irreducible over IR, but reducible over 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

\[ \rho(y_1, y_2) = \mu(f^{-1}(y_1, y_2)) \cdot \left| \det \left( \frac{\partial x}{\partial y} \right) \right|, \]

where the Jacobian matrix is defined as

\[ \frac{\partial x}{\partial y} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix}. \]

Similar techniques also work in IR^d.

The Box-Muller method transforms a sample point \((X_1, X_2)\) generated uniformly random from \((0, 1)^2\), and transforms it to

\[ (Y_1, Y_2) = \left( \sqrt{-2 \ln X_1 \cos(2\pi X_2)}, \sqrt{-2 \ln X_1 \sin(2\pi X_2)} \right). \]

We now calculate the partial derivatives and verify that the resulting \((Y_1, Y_2)\) distributes according to \( N(0, I) \). First, we express \( x_1 \) and \( x_2 \) in terms of \( y_1 \) and \( y_2 \).

\[ x_1 = \exp(-(y_1^2 + y_2^2)/2) \]
\[ x_2 = \arctan(y_2/y_1)/(2\pi) \]

Then, we calculate the partial derivatives.

\[ \frac{\partial x_1}{\partial y_1} = -y_1 \exp(-(y_1^2 + y_2^2)/2) \]
\[ \frac{\partial x_1}{\partial y_2} = -y_2 \exp(-(y_1^2 + y_2^2)/2) \]
\[ \frac{\partial x_2}{\partial y_1} = -y_2/y_1^2 \]
\[ \frac{\partial x_2}{\partial y_2} = (2\pi)(1 + y_2^2/y_1^2) \]
\[ \frac{\partial x_2}{\partial y_1} = 1/y_1 \]
\[ \frac{\partial x_2}{\partial y_2} = (2\pi)(1 + y_2^2/y_1^2) \]
Hence, the Jacobian determinant is

\[
\left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right| = \begin{vmatrix}
\frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \frac{\partial x_1}{\partial y_3} \\
\frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \frac{\partial x_2}{\partial y_3} \\
\frac{\partial x_3}{\partial y_1} & \frac{\partial x_3}{\partial y_2} & \frac{\partial x_3}{\partial y_3}
\end{vmatrix}
= \begin{vmatrix}
-y_1 \exp(-(y_1^2 + y_2^2)/2) & -y_2 \exp(-(y_1^2 + y_2^2)/2) \\
\frac{1/y_1}{(2\pi)^{3/2} / y_1^{3/2}} & \frac{1/y_2}{(2\pi)^{3/2} / y_2^{3/2}} \\
-\frac{1}{y_1^2} & -1
\end{vmatrix}
= \frac{\exp(-(y_1^2 + y_2^2)/2)}{(2\pi)(1 + y_2^2/y_1^2)} (-y_1 - y_2)
= \exp(-(y_1^2 + y_2^2)/2)/(2\pi)
\]

Therefore, the pdf of \((Y_1, Y_2)\) is

\[
\mu(f^{-1}(y)) \left| \det \left( \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right) \right| = 1 \cdot \frac{1}{2\pi} e^{-(y_1^2+y_2^2)/2} = \frac{1}{2\pi} e^{-(y_1^2+y_2^2)/2},
\]

which shows that \((Y_1, Y_2) \sim N(0, I)\).

### 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 \(N(0, \sigma^2 I)\) with pdf

\[
p(x) = \frac{1}{(2\pi)^{d/2} \sigma^d} e^{-\|x\|^2/(2\sigma^2)}
\]

has very little mass close to the origin, even though its maximum probability density is at the origin. In fact,

\[
\int_0^1 \! p(r) \, dr
\]

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 \(N(0, I)\), 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.

\[
\frac{d}{dr} r^{d-1} e^{-r^2/2} = (d-1) r^{d-2} e^{-r^2/2} - r^d e^{-r^2/2} = 0
\]

\[
r^2 = d - 1
\]

\[
r = \sqrt{d - 1}
\]

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 ([18, 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 \(0 \leq a \leq \sqrt{2n}\sigma^2\). Suppose \(|E[X_i^n]| \leq \sigma^2 s!\) for all \(3 \leq s \leq (a^2/(4n\sigma^2))\), then

\[
\Pr[|X| \geq a] \leq 3e^{-a^2/(12n\sigma^2)}.
\]

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 $\mathit{Ell}_2$ distance. (c) The marginal distribution $p(r)$ with peak at $r = \sqrt{d-1}$.

**Theorem 10.5** (Gaussian Annulus Theorem [18, Theorem 2.8]). For a $d$-dimensional spherical Gaussian $N(0, I)$ and $c \leq \sqrt{d}$, all but $3e^{-c^2/96}$ of the probability mass lies within an annulus of $\sqrt{d} - c \leq r \leq \sqrt{d} + c$.

**Proof.** For a point $x = (x_1, x_2, \ldots, x_d) \sim N(0, I)$, we have

$$r^2 = \|x\|^2 = x_1^2 + x_2^2 + \ldots + x_d^2.$$

Consider

$$|r - \sqrt{d}| \geq c \implies r^2 - d = |r - \sqrt{d}| \cdot |r + \sqrt{d}| \geq c\sqrt{d}$$  

(since $|r + \sqrt{d}| \geq \sqrt{d}$)

$$|y_1 + y_2 + \ldots + y_d| \geq c\sqrt{d}$$  

($y_i = x_i^2 - 1$)

$$|z_1 + z_2 + \ldots + z_d| \geq c\sqrt{d}/2.$$  

($z_i = y_i/2$)

To use the above theorem, we bound the moments of $z_i$ using the Gamma integral.

$$|E[z_i^s]| = 2^{-s} E[|y_i|^s] \leq 2^{-s} E[1 + x_i^2] = 2^{-s} + 2^{-s} \frac{2}{\pi} \int_0^\infty x^{2s} e^{-x^2/2} \, dx \leq 2^{-s} + \frac{1}{\sqrt{\pi}} \int_0^\infty u^{s-0.5} e^{-u} \, du \leq s!$$

(substitute $x = \sqrt{2u}$)

Hence, we can apply the above theorem with $\sigma^2 = E[z_i^2] \leq 4$ and $|E[z_i^s]| \leq 4(s!)$.

$$\Pr[|z_1 + z_2 + \ldots + z_d| \geq c\sqrt{d}/2] \leq 3e^{-c^2/96}.$$  

**10.1.6 Distribution Sampling**

How can we sample points from a given distribution, for example $N(0, I)$? If $x \in \mathbb{R}$ has pdf $p(x)$, then we can define use its cdf

$$P: \mathbb{R} \to \mathbb{R}$$  

$$x \mapsto P(x) = \int_{-\infty}^x p(t) \, dt.$$
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 $N(0, I)$, we can draw $u = (u_1, u_2, \ldots, u_d)$ uniformly from $[0, 1]^d$ and take inverse of the cdf componentwise $x = (x_1, x_2, \ldots, x_d) = (P^{-1}(u_1), P^{-1}(u_2), \ldots, P^{-1}(u_d))$.

Another example is the Cauchy distribution with pdf $p(x) = \frac{1}{\pi(1 + x^2)}$.

Its cdf is given by

$$P(x) = \int_{-\infty}^{x} \frac{1}{\pi(1 + t^2)} \, dt = \frac{1}{\pi} \arctan x + \frac{1}{2}.$$  

**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 $N(0, I)$ using the Box-Muller method as follows.

$$Y_1 = \sqrt{-2 \ln X_1} \cos(2\pi X_2)$$
$$Y_2 = \sqrt{-2 \ln X_1} \sin(2\pi X_2)$$

Then, we have the following.

$$X_1 = e^{-(Y_1^2 + Y_2^2)/2}$$
$$X_2 = \arctan(Y_2/Y_1)$$

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

Given a mixture of two Gaussian densities

$$p(x) = w_1 p_1(x) + w_2 p_2(x),$$

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 $\|x - y\|^2 = \left(\sqrt{d} - 2\delta - \sqrt{2d}\right)^2$.
2d ± O(\sqrt{d}) if they come from the same Gaussian, and \|x - y\|^2 = \delta^2 + 2d ± O(\sqrt{d}) if they come from different Gaussian separated by \delta.

Suppose \( x, y \sim N(\mu, I) \) 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 \( x \) is at the North pole. With high probability, \( y \) is in the slab \( \{(y_1, y_2, \ldots, y_d): -c \leq y_1 \leq c\} \) for some \( c \in O(1) \). So, \( y \) is nearly orthogonal to \( x \) and hence \( \|x - y\| \approx \sqrt{\|x\|^2 + \|y\|^2} \).

More precisely, we can further rotate the coordinate system so that

\[
x = (\sqrt{d} \pm O(1), 0, 0, \ldots, 0)
\quad \text{and} \quad
y = (\pm O(1), \sqrt{d} \pm O(1), 0, \ldots, 0).
\]

Hence, \( \|x - y\|^2 = (d \pm O(\sqrt{d})) + (d \pm O(\sqrt{d})) = 2d \pm O(\sqrt{d}) \). See also (Figure 10.5(b))

Suppose \( x \sim N(p, I) \) and \( y \sim N(q, I) \) come from different Gaussians. With high probability, \( x \) and \( y \) lies in an annulus of width \( O(1) \) and radius \( \sqrt{d-1} \) centered at \( p \) and \( q \) respectively. Also, \( (x - p), (p - q), (q - y) \) are nearly mutually perpendicular. Hence, \( \|x - y\|^2 \approx \|x - p\|^2 + \delta^2 + \|q - y\|^2 = \delta^2 + 2d \pm O(\sqrt{d}) \).

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 [18, Theorem 2.1] [85, Theorem 3.1]). Let \( X \geq 0 \) be a random variable and \( a > 0 \).

\[
\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}.
\]

**Proof.** For a continuous non-negative random variable \( x \) with probability density \( p(x) \)

\[
\mathbb{E}[X] = \int_0^\infty xp(x) \, dx = \int_0^a xp(x) \, dx + \int_a^\infty xp(x) \, dx \\
\geq \int_a^\infty xp(x) \, dx \\
\geq \int_a^\infty ap(x) \, dx \\
= a \int_a^\infty p(x) \, dx \\
= a \Pr[X \geq a] \quad \Box
\]

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, \( \Pr[X \geq c \mathbb{E}[X]] \leq 1/c \).

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 [18, Theorem 2.3] [85, Corollary 3.7]). Let \( X \) be a random variable with mean \( m \) and variance \( \sigma^2 \). Then, for all \( a > 0 \),

\[
\Pr[|X - m| \geq a\sigma] \leq \frac{1}{a^2}.
\]

Using the Chebyshev inequality, we can now prove the Law of Large Numbers.
10.1.9 Application of Markov and Chebyshev Inequalities

**Theorem 10.10 ([18, 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 $0 \leq a \leq \sqrt{2n}\sigma^2$. Suppose $|E[X_i]| \leq \sigma^2!$ 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 $|E[X^r]|$ 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 $\binom{r}{r_1 r_2 \ldots r_n} = \frac{r!}{r_1! r_2! \ldots r_n!}$.

Since $X_i$ are independent,

$$E\left[(X_1 + X_2 + \ldots + X_n)^r\right] = \sum_{\sum r_i = r} \binom{r}{r_1 r_2 \ldots r_n} E[X_1^{r_1}] E[X_2^{r_2}] \ldots E[X_n^{r_n}].$$

Note that $E[X_i] = 0$. So, all terms with $r_i = 1$ are zero. So, all non-zero terms have $r_i \geq 2$ or $r_i = 0$. Since $\sum r_i = r$, 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 $\binom{r - t - 1}{t - 1}$ ways to allocate the remaining $(n - t) = t - 1$ 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 $\binom{e + n}{n} = \binom{r + n}{e}$.)

Also using $|E[X_i]|^s \leq \sigma^2 r_i!$, we get

$$E[X^r] \leq r! \sum_{t=1}^{r/2} \binom{n}{t} \binom{r - t - 1}{t - 1} \sigma^{2t}$$

$$\leq r! \sum_{t=1}^{r/2} \frac{(n\sigma^2)^t}{t!} 2^{r-t-1}.$$
Let \( h(t) = \frac{(n\sigma^2)^r}{t} 2^{r-t-1} \). Since \( a \leq \sqrt{2n}\sigma^2 \) 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,

\[
E[X^r] \leq r! \sum_{t=1}^{r/2} h(t) \\
\leq r! \cdot \frac{(n\sigma^2)^{r/2}}{(r/2)!} \cdot 2^{r/2}.
\]

Applying the Markov Inequality, we get

\[
\Pr[|X| \geq a] = \Pr[X^r \geq a^r] \leq \frac{r! \cdot (n\sigma^2)^{r/2}}{(r/2)!} \cdot \frac{2^{r/2}}{a^r}.
\]

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 \( X \sim N(0, 1) \) is at most \( (s!) \).

**Proof.** It follows from the following integral.

\[
E[X^s] = \int_{-\infty}^{+\infty} \frac{x^s}{\sqrt{2\pi}} e^{-x^2/2} \, dx = \begin{cases} 0 & \text{if } s \text{ is odd} \\ (s-1)!! & \text{if } s \text{ is even} \end{cases}
\]

\[\square\]

### 10.1.10 Chernoff Bounds

Recall the binomial distribution \( \text{Bin}(n, p) \) 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 \( E[X] = np \) and variance \( \text{Var}[X] = np(1-p) \). 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 [18, Theorem 12.3 & Theorem 12.4] [85, 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 \( 0 \leq \gamma \leq 1 \),

\[
\Pr[X < (1 - \gamma)m] \leq e^{-\gamma^2m/2},
\]

where \( X = \sum X_i \) and \( m = E[X] \).

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. [2]

There are also methods known as Markov Chain Monte Carlo (MCMC) or Markov Chain Quasi Monte Carlo (MCQMC).
10.2. BAYESIAN

Estimating $\pi$

We can use the Monte Carlo method to estimate $\pi$. Sample $z_1, z_2, \ldots, z_m$ independently and uniformly from $[0, 1]^2$. Define indicator random variables

$$Z_i = \begin{cases} 1 & \text{if } \|z_i\|_2 \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$ 

Then,

$$\operatorname{Pr}[Z_i = 1] = \frac{\text{(area of the unit circle inside } [0, 1]^2)}{\text{(area of the unit square)}} = \frac{\pi}{4}.$$

Let $W = \sum_{i=1}^m Z_i$. Then,

$$\operatorname{E}[W] = \sum_{i=1}^m \operatorname{E}[Z_i] = \frac{m \pi}{4}.$$ 

Let $W' = \frac{4}{m} W$. Then, $\operatorname{E}[W'] = \pi$. Hence, $W'$ gives an estimate for $\pi$. In fact, it improves as $m$ gets larger. By the Chernoff inequality bounds, we get

$$\operatorname{Pr}[|W - \operatorname{E}[W]| \geq \epsilon \operatorname{E}[W]] \leq 2e^{-m \pi \epsilon^2/12}.$$ 

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

$$\operatorname{Pr}[|X - v| \leq \epsilon v] \geq 1 - \delta.$$ 

Theorem 10.12 ([85, Theorem 10.1]). Let $X_1, X_2, \ldots, X_m$ be i.i.d. indicator random variables with $\mu = \operatorname{E}[X_i]$. If $m \geq \frac{1}{\pi \mu^2} \ln \frac{2}{\delta}$, then $\operatorname{Pr}[|\frac{1}{m} \sum_{i=1}^m X_i - \mu| \geq \epsilon \mu] \leq \delta$. That is, an $m$-sampling provides an $(\epsilon, \delta)$-approximation.

Proof. This can be proved using the above stated Chernoff bound. \hfill $\Box$

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.

$$\operatorname{Pr}[A|B] = \frac{\operatorname{Pr}[B|A] \operatorname{Pr}[A]}{\operatorname{Pr}[B]}.$$ 

This follows from $\operatorname{Pr}[A|B] \operatorname{Pr}[B] = \operatorname{Pr}[B|A] \operatorname{Pr}[A]$. We can regard $B$ as the measurement samples that we have. Using this data, we try to estimate $A$. In the numerator, $\operatorname{Pr}[B|A]$ is the likelihood of $A$ and $\operatorname{Pr}[A]$ is the prior probability. The normalization appears in the denominator $\operatorname{Pr}[B]$. The left hand side is the posterior probability $\operatorname{Pr}[A|B]$.

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.

$$\operatorname{Pr}[B|A] = 0.99$$

$$\operatorname{Pr}[A] = 0.001$$

$$\operatorname{Pr}[B|\overline{A}] = 0.02$$

$$\operatorname{Pr}[B] = \operatorname{Pr}[B|A] \operatorname{Pr}[A] + \operatorname{Pr}[B|\overline{A}] \operatorname{Pr}[\overline{A}] = 0.99 \times 0.001 + 0.02 \times 0.999 = 0.02097$$
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)
\[
\arg \max_r L(r|X) = \arg \max_r \Pr[X|r].
\]

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 distribution \(\text{Bin}(n, r)\) such that
\[
\Pr[k|r] = \binom{n}{k} r^k (1 - r)^{n-k}.
\]
If we get 62 heads and 38 tails in 100 trials, 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 \(N(\mu, \sigma^2)\) has parameters mean \(\mu\) and variance \(\sigma^2\), while the binomial distribution \(\text{Bin}(n, p)\) has parameters \(n\) and success probability \(p\).) Let \(D(X)\) be an estimator of some function \(h(\theta)\). The bias is defined as
\[
\text{E}[D(X) - h(\theta)].
\]
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)
\[
\text{E}[(D(X) - h(\theta))^2] = \text{Var}(D(X)) + \text{Bias}^2.
\]

Theorem 10.13. Let \(X_1, X_2, \ldots, X_n\) be independent samples, each with mean \(\mu\) and variance \(\sigma^2\).

\(B1.\) \([?], \text{Example 14.3}\] \(D(X) = \frac{1}{n} \sum_{i=1}^{n} X_i\) is an unbiased estimator of \(\mu\).
B2. If $\mu$ is known, then $D(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ is an unbiased estimator of $\sigma^2$.

B3. [?, 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

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} \left[ (X_i - m) + (m - \mu) \right]^2$$

$$= \sum_{i=1}^{n} (X_i - m)^2 + n(m - \mu)^2$$

Hence,

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - m)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 - (m - \mu)^2$$

$$E[S^2] = \frac{1}{n} \sum_{i=1}^{n} \text{Var}(X_i) - \text{Var}(m)$$

$$= \frac{n-1}{n} \sigma^2$$

Thus, $\left( \frac{n}{n-1} S^2 \right)$ is an unbiased estimator of $\sigma^2$. \qed

10.3 Biological Applications

Summary

References and Further Reading

ExerSection
Biology Appendix
Conclusions

Something to conclude
Bibliography


