COMPUTATIONAL TOPOLOGY

AN INTRODUCTION

Herbert Edelsbrunner and John Harer

Departments of Computer Science and Mathematics
Duke University

To our families and friends

Table of Contents

Part A

Part B

Part C

Part D

Preface

The last ten years have witnessed that geometry, topology, and algorithms form a potent mix of disciplines with many applications inside and outside academia. We aim at bringing these developments to a larger audience. This book has been written to be taught, and it is based on notes developed during courses delivered at Duke University and at the Berlin Mathematical School, primarily to students of computer science and mathematics. The organization into chapters, sections, exercises, and open problems reflects the teaching style we practice. Each chapter develops a major topic and is worth about two weeks of teaching. The chapters are divided into sections, each a lecture of one and a quarter hours. To convey a feeling for the boundary of the current knowledge, we complement the material with descriptions of open problems. An interesting challenge is the mixed background of the audience. How do we teach topology to students with limited background in mathematics, and how do we convey algorithms to students with limited background in computer science? Assuming no prior knowledge and appealing to the intelligence of the listener is a good first step. Motivating the material by relating it to situations in different walks of life is helpful in building up intuition that can cut through otherwise necessary formalism. Exposing central ideas with simple means helps and so does minimizing the necessary amount of detail.

The material in this book is a combination of topics in geometry, topology, and algorithms. Far from getting diluted, we find that the fields benefit from each other. Geometry gives a concrete face to topological structures and algorithms offer a means to construct them at a level of complexity that passes the threshold necessary for practical applications. As always, algorithms have to be fast because time is the one fundamental resource human kind did not yet learn to manipulate for its selfish purposes. Beyond these obvious relationships, there is a symbiotic affinity between algorithms and the algebra used to capture topological information. It is telling that both fields trace their name back to the writing of the same Persian mathematician, al-Khwarizmi, work-

ing in Baghdad during the ninth century after Christ. Besides living in the triangle spanned by geometry, topology, and algorithms, we find it useful to contemplate the place of the material in the tension between extremes such as

$$\begin{aligned} \text{local} &\quad \text{vs.} \quad \text{global;} \\ \text{discrete} &\quad \text{vs.} \quad \text{continuous;} \\ \text{abstract} &\quad \text{vs.} \quad \text{concrete;} \\ \text{intrinsic} &\quad \text{vs.} \quad \text{extrinsic.} \end{aligned}$$

Global insights are often obtained by a meaningful integration of local information. This is how we proceed in many fields, taking on bigger challenges after mastering the small. But small things are big from up close and big ones small from afar. Indeed, the question of scale lurking behind this thought is the driving force for much of the development described in this book. The dichotomy between discrete and continuous structures is driven by opposing goals, machine computation and human understanding. Both are illusions that are useful to have but should not be confused with anything intangible like reality. The tension between the abstract and the concrete as well as between the intrinsic and the extrinsic have everything to do with human approach to knowledge. An example close to home is the step from geometry to topology in which we remove the burdens of size to focus on the phenomenon of connectivity. The more abstract the context the more general the insight. Now, generality is good but it is not a substitute for the concrete steps that have to be taken to build bridges to applications. Zooming in and out of generality leads to unifying viewpoints and suggests meaningful integrations where they exist.

While these thoughts have certainly influenced us in the selection of the material and in its presentation, there is a long way to the concrete instantiation we call this book. The hardest part is to land, and we do in four parts decomposed into a total of ten chapters. Part A is a gentle introduction to topological thought. Discussing Graphs in Chapter I, Surfaces in Chapter II, and Complexes in Chapter III, we gradually build up topological sophistication, always in combination with geometric and algorithmic ideas. Part B presents classical material from topology. We focus on what we deem useful and efficiently computable. The material on Homology in Chapter IV and on Duality in Chapter V is exclusively algebraic. In the discussion of Morse Theory in Chapter VI, we build a bridge to differential concepts in topology. Part C is mostly novel and indeed the main reason we write this book. The main new concept is Persistence introduced in Chapter VII and its Stability discussed in Chapter VIII. Finally, we address connections to Singularities in Chapter IX. Part D concludes the book with a small collection of open problems in computational

topology. It is our hope that they point in the right direction, leading a new generation of researchers far and beyond what we currently imagine.

In a project like writing this book there are many who contribute, directly or indirectly. We want to thank all and we don't know where to begin. Above all, we thank our colleagues in academia and industry, our students, and our postdoctoral fellows for their ideas, criticism, and encouragement. And most of all for the sense of purpose they provide. We thank Duke University for providing the facilities and intellectual environment that allowed us to engage in the line of research leading to this book. We thank the Computer Science and the Mathematics Departments at Duke University and the Berlin Mathematical School for the opportunity to teach computational topology to their students. These courses provided the motivation to develop the notes that turned into this book. Last not least, we thank the funding agencies, in particular DARPA but also NSF and NIH, for nurturing this research and for opening up numerous connections to topics that lie well beyond the scope of this book.

Herbert Edelsbrunner and John Harer
Durham, North Carolina, 2008

Chapter I

Graphs

In topology we think of a graph as a 1-dimensional geometric object, vertices being points and edges being curves connecting these points in pairs. This view is different but compatible with the interpretation of a graph common in discrete mathematics where the vertices are abstract elements and the edges are pairs of these elements. In more than one way, this book lives in the tension between the discrete and the continuous and graphs are just one example of this phenomenon. We begin with the discussion of an intrinsic property, namely whether a graph is connected or not. Indeed, this does not depend on where we draw the graph, on paper or in the air. Following are extrinsic considerations about curves and graphs in the plane and in three-dimensional space. While the extrinsic questions are natural to people, the mathematician usually favors the intrinsic point of view since it tends to lead to more fundamental insights of more general validity.

I.1 Connected Components

A theme that goes through this entire book is the exchange between discrete and continuous models of reality. In this first section, we compare the notion of connectedness in discrete graphs and continuous spaces.

Simple graphs. An abstract *graph* is a pair $G = (V, E)$ consisting of a set of *vertices*, V, and a set of *edges*, E, each a pair of vertices. We draw the vertices as points or little circles and edges as line segments or curves connecting the points. For now, crossings between the curves are allowed. The graph is *simple* if the edge set is a subset of the set of unordered pairs, $E \subseteq \binom{V}{2}$, which means that no two edges connect the same two vertices and no edge joins a vertex to itself. For $n = \operatorname{card} V$ vertices, the number of edges is $m = \operatorname{card} E \leq \binom{n}{2}$. Every simple graph with n vertices is a subgraph of the *complete graph*, K_n, that contains an edge for every pair of vertices; see Figure I.1.

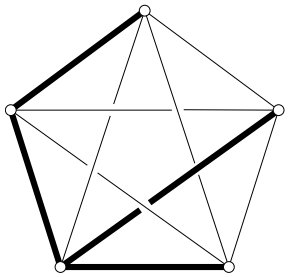

Figure I.1: The complete graph with five vertices, K_5. It has ten edges which form five crossings if drawn as sides and diagonals of a convex pentagon. The four thick edges connect the same five vertices and form a spanning tree of the complete graph.

In a simple graph, a *path* between vertices u and v can be described by a sequence of vertices, $u = u_0, u_1, u_2, \ldots, u_k = v$, with an edge between u_i and u_{i+1} for each $0 \leq i \leq k-1$. The *length* of this path is its number of edges, k. Vertices can repeat, allowing the path to cross itself or backtrack. The path is *simple* if the vertices in the sequence are distinct, that is, $u_i \neq u_j$ whenever $i \neq j$.

DEFINITION A. A simple graph is *connected* if there is a path between every pair of vertices.

A *(connected) component* is a maximal subgraph that is connected. The smallest connected graphs are the *trees*, which are characterized by having a unique simple path between every pair of vertices. Removing any one edge disconnects the tree. A *spanning tree* of $G = (V, E)$ is a tree (V, T) with $T \subseteq E$; see Figure I.1. It has the same vertex set as the graph and uses a minimal set of edges necessary to be connected. This requires that the graph is connected to begin with. Indeed, a graph is connected iff it has a spanning tree. An alternative characterization of a connected graph can be based on the impossibility to cut it in two.

DEFINITION B. A *separation* is a non-trivial partition of the vertices, that is, $V = U \dot\cup W$ with $U, W \neq \emptyset$, such that no edge connects a vertex in U with a vertex in W. A simple graph is *connected* if it has no separation.

Topological spaces. We now switch to a continuous model of reality, the topological space. There are similarities to graphs, in particular if our interest is limited to questions of connectedness. Starting with a point set, we consider a topology, which is a way to define which points are near without specifying how near they are from each other. Think of it as an abstraction of Euclidean space in which neighborhoods are open balls around points. Concretely, a *topology* on a point set $\mathbb{X}$ is a collection $\mathcal{U}$ of subsets of $\mathbb{X}$, called *open sets*, such that

(i) $\mathbb{X}$ is open and the empty set $\emptyset$ is open;

(ii) if U_1 and U_2 are open, then $U_1 \cap U_2$ is open;

(iii) if U_i is open for all i in some possibly infinite, possible uncountable, index set, then the union of all U_i is open.

The pair $(\mathbb{X}, \mathcal{U})$ is a called a *topological space*, but we will usually tacitly assume that $\mathcal{U}$ is understood and refer to $\mathbb{X}$ a topological space. Since we can repeat the pairwise intersection, Condition (ii) is equivalent to requiring that common intersections of finitely many open sets are open.

To build interesting topologies, we start with some initial notion of which sets might be open and then form appropriate combinations of these until the three conditions are satisfied. A *basis* of a topology on a point set $\mathbb{X}$ is a collection $\mathcal{B}$ of subsets of $\mathbb{X}$, called *basis elements*, such that each $x \in \mathbb{X}$ is contained in at least one $B \in \mathcal{B}$ and $x \in B_1 \cap B_2$ implies there is a third basis element with $x \in B_3 \subseteq B_1 \cap B_2$. The topology $\mathcal{U}$ *generated* by $\mathcal{B}$ consists of all sets $U \subseteq \mathbb{X}$ for which $x \in U$ implies there is a basis element $x \in B \subseteq U$. This topology can be constructed explicitly by taking all possible unions of all possible finite

intersections of basis sets. As an example consider the real line, $\mathbb{X} = \mathbb{R}$, and let $\mathcal{B}$ be the collection of open intervals. This gives the usual topology of the real line. Note that the intersection of the intervals $(-\frac{1}{k}, +\frac{1}{k})$, for the infinitely many integers $k \geq 1$, is the point 0. This is not an open set which illustrates the need for the restriction to finite intersections.

We often encounter sets inside other sets, $\mathbb{Y} \subseteq \mathbb{X}$, and in these cases we can borrow the topology of the latter for the former. Specifically, if $\mathcal{U}$ is a topology of $\mathbb{X}$ then the collection of sets $\mathbb{Y} \cap U$, for $U \in \mathcal{U}$, is the *subspace topology* of $\mathbb{Y}$. As an example consider the closed interval $[0, 1] \subseteq \mathbb{R}$. We have seen that the open intervals form the basis for a topology of the real line. The intersections of open intervals with $[0, 1]$ form the basis of the subspace topology of the closed interval. Note that the interval $(1/2, 1]$ is considered an open set in $[0, 1]$, but isn't open when considered as a set in $\mathbb{R}$.

Continuity, paths and connectedness. A function from one topological space to another is *continuous* if the preimage of every open set is open. This is derived from the concept of continuity familar from calculus; for example the function $f : \mathbb{R} \to \mathbb{R}$ that maps $(-\infty, 0]$ to 0 and $(0, \infty)$ to 1 is not continuous because for any $0 < \varepsilon < 1$, $(-\varepsilon, \varepsilon)$ is open, but $f^{-1}((-\varepsilon, \varepsilon))$ is not.

A *path* is a continuous function from the unit interval, $\gamma : [0, 1] \to \mathbb{X}$. It *connects* the point $\gamma(0)$ to the point $\gamma(1)$ in $\mathbb{X}$. Similar to paths in graphs we allow for self-intersections, that is, values $s \neq t$ in the unit interval for which $\gamma(s) = \gamma(t)$. If there are no self-intersections then the function is injective and the path is *simple*. Now we are ready to adapt our first definition of connectedness to topological spaces.

Definition A. A topological space is *path-connected* if every pair of points is connected by a path.

There is also a counterpart of our second definition of connectedness. We formulate it using open sets and there is an equivalent formulation in terms of *closed sets* which, by definition, are complements of open sets.

Definition B. A *separation* of a topological space $\mathbb{X}$ is a partition $\mathbb{X} = U \dot{\cup} W$ into two non-empty, open subsets. A topological space is *connected* if it has no separation.

It turns out connectedness is strictly weaker than path-connectedness, although for most spaces we will encounter they are the same. An example of a space that

is connected but not path-connected is the comb with a single tooth deleted. It is constructed by gluing vertical teeth to a horizontal bar and finally deleting the interior of the last tooth: taking the union of $[0,1] \times 0$ with $\frac{1}{k} \times [0,1]$, for all positive integers k, we finally delete $0 \times (0,1)$. To construct a topology, we take the collection of open disks as the basis of a topology on $\mathbb{R}^2$ and we use the subspace topology for the comb. This space is connected because it is the union of a path-connected set and a limit point. It is not path-connected because no path from anywhere else can reach 0×1.

Disjoint set systems. We return to graphs and consider the algorithmic question of deciding connectedness. There are many approaches and we present a solution based on maintaining a disjoint set system. This particular algorithm has various other applications, some of which will be discussed in later chapters of this book. Using the integers from 1 to n as the names of the vertices, we store each component of the graph as a subset of $[n] = \{1, 2, \ldots, n\}$. Adding the edges one at a time and maintaining the system of sets representing the components, we find that the graph is connected iff in the end there is only one set left, namely $[n]$. Formulated as an abstract data type, we have two operations manipulating the system:

FIND(i)**:** return the name of the set that contains i;

UNION(i, j)**:** assuming i and j belong to different sets in the system, replace the two sets by their union.

We need the find operation to test whether i and j indeed belong to different sets. Each successful union operation reduces the number of sets in the system by one. Starting with n singleton sets, it therefore takes $n-1$ union operations to get to a single set. Since trees connecting the n vertices can be generated this way, we thus have a proof that every tree with n vertices has $m = n - 1$ edges.

A standard data structure implementing a disjoint set system stores each set as a tree embedded in a linear array, $V[1..n]$. Each node in the tree is equipped with a pointer to its *parent*, except for the *root* which has no parent; see Figure I.2. Who is parent of whom is not important as long as the vertices are connected in a single tree. We implement the find operation by traversing the tree upward until we reach the root, reporting the root as the name of the set.

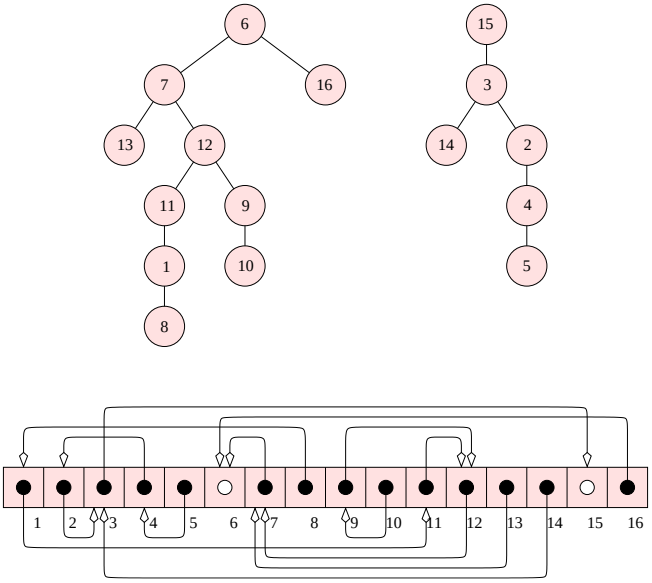

Figure I.2: Top: two trees representing two disjoint sets. Bottom: storing the two trees in a linear array using an arbitrary ordering of the nodes.

```
int FIND(i)
  if V[i].parent ≠ null then return FIND(V[i].parent)
                         else return i
  endif.
```

If i is not the root then we find the root recursively and finally return it. Otherwise, we return i as the root. We implement the union operation by linking one root to the other.

```
void UNION(i, j)
  x = FIND(i);  y = FIND(j);
  if x ≠ y then V[x].parent = y endif.
```

After making sure that the two sets are different, we assign one root as the parent of the other.

Improving the running time. The above implementation is not very efficient, in particular if we have long paths that are repeatedly traversed. To

prevent this from happening we always link the smaller to the larger tree.

```
void UNION(i, j)
  x = FIND(i);  y = FIND(j);
  if x ≠ y then if V[x].size > V[y].size then x ↔ y endif;
                V[x].parent = y
  endif.
```

Now a tree of k nodes cannot have paths longer than $\log_2 k$ edges since the size of the subtree grows by at least a factor of two each time we pass to the parent. To further improve the efficiency, we compress paths whenever we traverse them. Here it is convenient to assume that roots are identified by pointing to themselves.

```
int FIND(i)
  if V[i].parent ≠ i then
    return V[i].parent = FIND(V[i].parent)
  endif;
  return i.
```

If i is not the root then the function recursively finds the root, makes the root the parent of i, reports the root, and exits. Otherwise, the function reports i as the root and exits.

In analyzing the algorithm, we are interested in the running-time for executing a sequence of m union and find operations. Finding tight bounds turns out to be a difficult problem and we limit ourselves to stating the result. Specifically, if n is the number of vertices then the running-time is never more than $O(m\alpha(n))$, where $\alpha(n)$ is the notoriously slow growing inverse of the Ackermann function. Eventually, $\alpha(n)$ goes to infinity but to reach even beyond five we need an astronomically large number of vertices, more than the estimated number of electrons in our Universe. In other words, for all practical purposes the algorithm takes constant average time per operation but theoretically this is not a true statement.

Bibliographic notes. Graphs are ubiquitous objects and appear in most disciplines. Within mathematics, the theory of graphs is considered part of combinatorics. There are many good books on the subject, including the one by Tutte [3]. The basic topological notions of connectedness are treated in books on point-set or general topology, including the text by Munkres [1]. The computational problem of maintaining a system of disjoint sets, often

referred to as the union-find problem, is a classic topic in the field of algorithms. Solutions to it are known as union-find data structures and the most efficient of all is the up-tree representation maintained through weighted union and path-compression as explained in this section. A complete description of the highly non-trivial analysis of the algorithm can be found in the text by Tarjan [2].

[1] J. R. MUNKRES. *Topology. A First Course.* Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[2] R. E. TARJAN. *Data Structures and Network Algorithms.* SIAM, Philadelphia, Pennsylvania, 1983.

[3] W. T. TUTTE. *Graph Theory.* Addison-Wesley, Reading, Massachusetts, 1984.

I.2 Curves in the Plane

In the previous section, we used paths to merge points into connected components. To capture aspects of connectivity that go beyond components, we need different maps.

Closed curves. We distinguish primarily between two kinds of (connected) curves, *paths* and *closed curves*. As defined in the previous section, paths are continuous maps from $[0, 1]$ to $\mathbb{X}$. Sometimes, a closed curve is defined as a path in which 0 and 1 map to the same point. Usually, we will define a closed curve to be a map from the unit circle, $\gamma : \mathbb{S}^1 \to \mathbb{X}$, where $\mathbb{S}^1 = \{x \in \mathbb{R}^2 \mid \|x\| = 1\}$. This second version emphasizes the important fact that paths and closed curves capture different properties of topological spaces, since the interval and the circle are different topological spaces. To make this precise, we call two topological spaces *homeomorphic* or *topologically equivalent* if there exists a continuous bijection from one space to the other whose inverse is also continuous. A map with these properties is called a *homeomorphism*. Notice that a homeomorphism between two spaces gives a bijection between their open sets. The unit interval and the unit circle are not homeomorphic. Indeed, removing the midpoint decomposes the interval into two components while removing its image leaves the circle connected. This contradicts the existence of a bijection that is continuous in both directions.

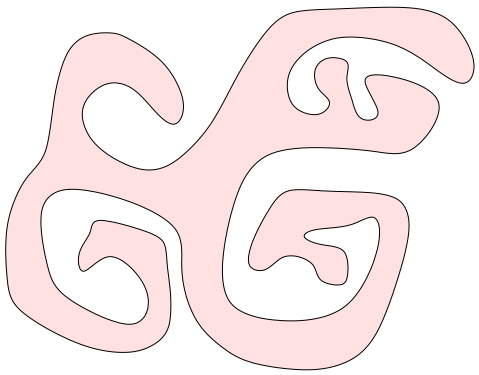

Figure I.3: The shaded inside and the white outside of a simple closed curve in the plane.

Considering maps into the Euclidean plane, $\mathbb{X} = \mathbb{R}^2$, it makes sense to distinguish curves with and without self-intersections. A *simple closed curve* is

a curve without self-intersections, that is, a continuous injection $\gamma : \mathbb{S}^1 \to \mathbb{R}^2$. Interestingly, every such curve decomposes the plane into two pieces, one on each side of the curve, as in Figure I.3.

Jordan Curve Theorem. Removing the image of a simple closed curve from $\mathbb{R}^2$ leaves two connected components, the bounded *inside* and the unbounded *outside*. The inside together with the image of the curve is homeomorphic to a closed disk.

This may seem obvious but proving it is challenged by the generality of the claim which is formulated for all and not just smooth or piecewise linear simple closed curves. There are reasons to believe that there is no simple proof for such a general claim. The fact that the inside together with the curve is homeomorphic to the closed disk, $\mathbb{B}^2 = \{x \in \mathbb{R}^2 \mid \|x\| \leq 1\}$, is known as the Schönflies Theorem. The Jordan Curve Theorem remains valid if we replace the plane by the sphere, $\mathbb{S}^2 = \{x \in \mathbb{R}^3 \mid \|x\| = 1\}$, but not if we replace it by the torus.

Parity algorithm. Given a simple closed curve in the plane, a fundamental computational question asks whether a given *query point* $x \in \mathbb{R}^2$ lies inside, on, or outside the curve. To write an algorithm answering this question, we assume a finite approximation of the curve. For example, we may specify γ at a finite number of points and interpolate linearly between them. The result is a *closed polygon*; see Figure I.4. It is *simple* if it is a closed curve itself. To decide whether the point x lies inside such a simple closed polygon, we draw a half-line emanating from x and count how often it crosses the polygon. Assuming x does not lie on the polygon, it lies inside if the number of crossings is odd and outside if that number is even. Hence, the name Parity Algorithm. In the implementation of this idea, we let $x = (x_1, x_2)$ be the query point and $a = (a_1, a_2)$, $b = (b_1, b_2)$ the endpoints of an edge of the polygon. We assume the generic case in which no three points are collinear and no two lie on a common vertical or horizontal line. To simplify the code, we choose the horizontal half-line leaving x toward the right and we assume that a is below b, that is, $a_2 < b_2$. We first make sure that the entire horizontal line crosses the edge, which we do by testing $a_2 < x_2 < b_2$. If it does then we test whether the crossing lies to the left or the right of the query point. To this end we compute the determinant of the matrix

$$\Delta(x, a, b) \quad = \quad \begin{bmatrix} 1 & x_1 & x_2 \\ 1 & a_1 & a_2 \\ 1 & b_1 & b_2 \end{bmatrix},$$

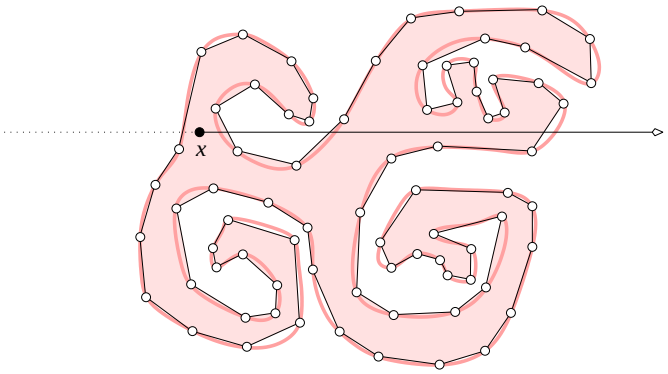

Figure I.4: Approximation of the simple closed curve in Figure I.3 by a simple closed polygon. The point x lies inside the polygon and the half-line crosses the polygon an odd number of times.

which is positive iff the sequence of points x, a, b forms a left-turn. To see this, we verify the claim for $x = (0, 0)$, $a = (1, 0)$, $b = (0, 1)$ and then notice that the sign of the determinant switches exactly when the three points become collinear. We use this fact to decide whether the half-line crosses the edge:

```
boolean doesCross(x, a, b)
  if not a_2 < x_2 < b_2 then return false endif;
  return det Δ(x, a, b) > 0.
```

Now we run this test for all edges and this way count the crossings. The trouble with this implementation are the non-generic cases. We finesse them using two infinitesimally small, positive numbers $0 < \varepsilon_1 \ll \varepsilon_2$ and substituting $x' = (x_1 + \varepsilon_1, x_2 + \varepsilon_2)$ for x. A generic case for x is generic for x' and we get the same decision for both points. A non-generic case for x is generic for x' and we use the decision for x'.

Polygon triangulation. Sometimes it is useful to have a more structured representation of the inside of the polygon, for example for navigation to find the exit out of a maze. The most common such structural representation is a *triangulation* which is a decomposition into triangles. Here we require that the triangles use the vertices of the polygon but do not introduce new ones. Furthermore, they use the edges of the polygon together with *diagonals*, which are new edges that connect non-adjacent vertices of the polygon. The diagonals

are required to pass through the inside and not cross any other diagonals and any polygon edges; see Figure I.5.

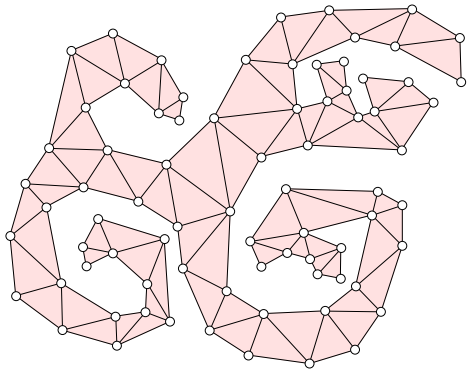

Figure I.5: A triangulation of the polygon in Figure I.4. Each diagonal passes from one side of the inside to the other.

To prove that a triangulation always exists we just need to show that there is at least one diagonal, unless the number of edges in the polygon is $n = 3$. Indeed, we may consider the leftmost vertex, b, of the polygon. Either we can connect its two neighbors, a and c, or we can connect b to the leftmost vertex u that lies inside the triangular region abc. Drawing this diagonal decomposes the n-gon into two, an n_1-gon and an n_2-gon. We have $n_1+n_2 = n+2$ and since both are at least three, we also have $n_1, n_2 < n$. We can thus use induction to complete the proof. The same inductive argument shows that there are $n-3$ diagonals and $n-2$ triangles, no matter how we triangulate. This is suggestive. Indeed, we can think of the triangles as the nodes and the diagonals as the arcs of a tree. Since every tree with $n-2 \geq 2$ nodes has at least one leaf, that is, a node with only one neighbor, every triangulation has an *ear*, that is, a triangle formed by one diagonal and two polygon edges. Incidentally, this property does not generalize to tetrahedral decompositions in $\mathbb{R}^3$.

Winding number. We return to a general, not necessarily simple, closed curve $\gamma : \mathbb{S}^1 \to \mathbb{R}^2$. Let x be a point not in the image of the curve. Suppose we traverse γ and view the moving point from x. Specifically, we let s go once around the circle and observe the unit vector $(\gamma(s)-x)/\|\gamma(s)-x\|$ rotate about the origin. When the vector completes a full turn we count $+1$ or -1 depending on whether this turn is counterclockwise or clockwise. The sum of these numbers is the *winding number* of γ and x, denoted as $W(\gamma, x)$. It

is necessarily an integer and gives the net number of counterclockwise turns we observe. If γ is simple then the points inside the curve all have the same winding number, -1 or $+1$. To reduce this to one case we may reorient the curve, e.g. by reflecting the unit circle along the horizontal coordinate axis, and get

$$W(\gamma, x) \quad = \quad \begin{cases} +1 & \text{if } x \text{ is inside;} \\ 0 & \text{if } x \text{ is outside.} \end{cases}$$

However, for non-simple curves we can get winding numbers of absolute value larger than one; see Figure I.6. Suppose we move x in the plane. As long as

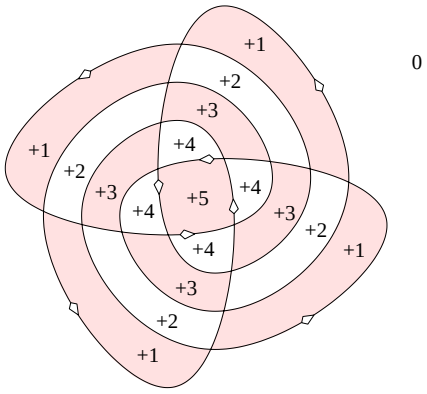

Figure I.6: An oriented non-simple closed curve with regions distinguished by the winding number of their points.

it does not cross the curve, the winding number does not change. Crossing the curve changes the winding number, namely by -1 if we cross from left to right and by $+1$ if we cross from right to left. But this implies that at least two regions in the decomposition defined by γ have their boundary arcs consistently oriented. Specifically, the neighbors of a region with locally maximum winding number all have winding number one less so the region lies to the left of all its boundary arcs. Similarly, a region with locally minimal winding number lies to the right of all its boundary arcs.

Bibliographic notes. The Jordan Curve Theorem is well known also beyond topology, in part because it seems so obvious but at the same time is difficult to prove. We refer to [4] for a deeper discussion. The difficulties encountered in the implementation of the parity algorithm have been voiced in

[3]. A provably correct implementation can be achieved with exact arithmetic and symbolic perturbation as described in [2]. Triangulations of simple closed polygons in the plane have been studied in computational geometry. Constructing such a triangulation in time proportional to the number of vertices seems rather difficult and the algorithm by Chazelle [1] that achieves this feat is not recommended for implementation.

[1] B. Chazelle. Triangulating a simple polygon in linear time. *Discrete Comput. Geom.* **6** (1991), 485–524.

[2] H. Edelsbrunner and E. P. Mücke. Simulation of Simplicity: a technique to cope with degenerate cases in geometric algorithms. *ACM Trans. Graphics* **9** (1990), 86–104.

[3] A. R. Forest. Computational geometry in practice. In *Fundamental Algorithms for Computer Graphics*, E. A. Earnshaw (ed.), Springer-Verlag, Berlin, Germany, 1985, 707–724.

[4] C. T. C. Wall. *A Geometric Introduction to Topology.* Addison-Wesley, 1971.

I.3 Knots and Links

In this section, we study closed curves in three-dimensional Euclidean space and questions about how they relate to each other or to themselves.

Knots. A closed curve embedded in $\mathbb{R}^3$ does not decompose the space but it can be tangled up in inescapable ways. The field of mathematics that studies such tangles is knot theory. Its prime subject is a *knot* which is an *embedding* $\kappa : \mathbb{S}^1 \to \mathbb{R}^3$, that is, an injective, continuous function that is a homeomorphism onto its image. It turns out that any injective, continuous function from the S^1 to $\mathbb{R}^3$ is an embedding, but this is not true for general domains. Another knot is *equivalent* to κ if it can be continuously deformed into κ without crossing itself during this process. Equivalent knots are considered the same. The simplest

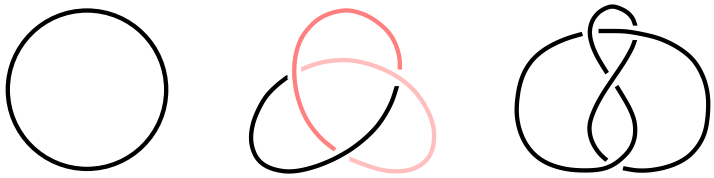

Figure I.7: From left to right: the unknot, the trefoil knot, and the figure-eight knot. The trefoil knot is tricolored.

knot is the *unknot*, also known as the *trivial knot*, which can be deformed to a geometric circle in $\mathbb{R}^3$. Two other and only slightly tangled up knots are the *trefoil* and the *figure-eight knots*, both illustrated in Figure I.7. A subtlety in the definition of equivalence is that deformations in which knotted parts disappear in the limit are not allowed. It is therefore useful to think of knots as curves with small but positive thickness, similar to shoelaces and ropes.

Reidemeister moves. Let us follow a deformation of a knot by drawing its projections to a plane, keeping track of the under- and over-passes at crossings. We are primarily interested in generic projections defined by the absence of any violations to injectivity, other than a discrete collection of double-points where the curve crosses itself in the plane. In a generic deformation, we observe three types of non-generic projections that transition between generic projections, which are illustrated in Figure I.8. It is plausible and also true that any two generic projections of the same knot can be transformed into each other by Reidemeister moves.

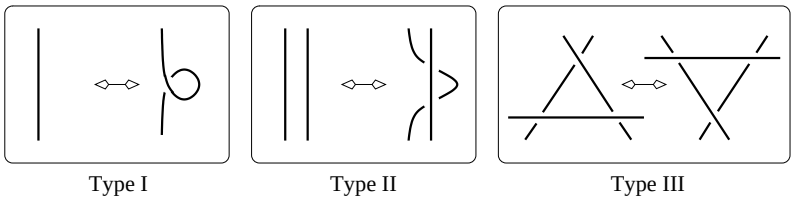

Figure I.8: The three types of Reidemeister moves.

It seems clear that the trefoil knot is not equivalent to the unknot, and there is indeed an elementary proof using Reidemeister moves. Call a piece of the knot from one under-pass to the next a *strand*. A *tricoloring* of a generic projection colors each strand with one of three colors such that

(i) at each crossing either three colors or only one color come together;

(ii) at least two colors are used.

Figure I.7 shows that the standard projection of the trefoil knot is tricolorable. A useful property of Reidemeister moves is that they preserve tricolorability, that is, the projection before the move is tricolorable iff the projection after the move is tricolorable.

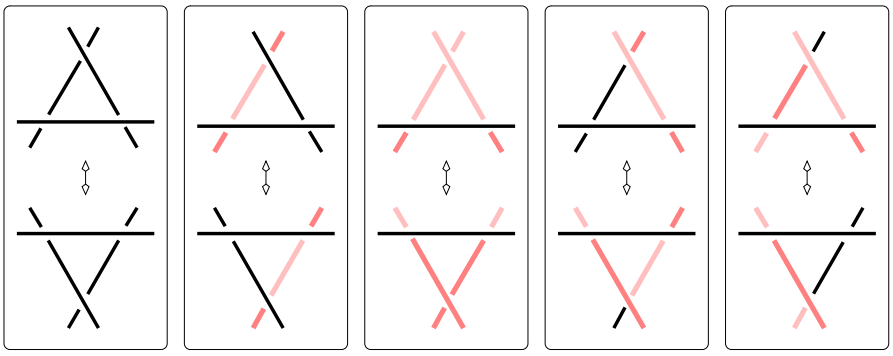

Figure I.9: The different cases in the proof that the Type III Reidemeister move preserves tricolorability. In each case there is only one new strand whose color can be chosen anew.

Type I. Going from left to right in Figure I.8, we use the same one color, and

going from right to left we observe that we have only one color coming together at the crossing.

Type II. From left to right we have two possibilities, either using only one color or going from two to three colors. The reverse direction is symmetric.

Type III. There are five cases to be checked, all shown in Figure I.9.

The trefoil knot is tricolorable and the unknot is not tricolorable. It follows that the two are not equivalent. It is not difficult to see that the figure-eight knot is not tricolorable. This implies that the trefoil knot and the figure-eight knot are different but the method is not powerful enough to distinguish the figure-eight from the unknot.

Links. A *link* is a collection of two or more disjoint knots. Equivalence between links is defined the same way as between knots, and Reidemeister moves again suffice to go from one generic projection to another. A disjoint plane *splits* a link if there are knots on both sides. A link is *splittable* if an equivalent link has a splitting plane. The *unlink* or *trivial link* of size two consists of two unknots that can be split, like the two circles in Figure I.10 on the left. The easiest non-splittable link consisting of two unknots is the *Hopf link*, which is shown in Figure I.10 in the middle. We can again

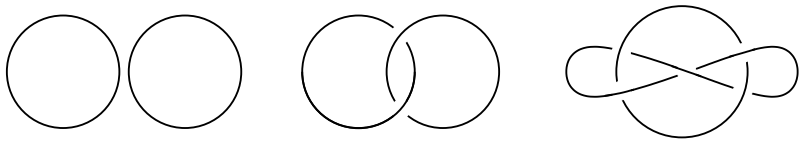

Figure I.10: From left to right: the unlink, the Hopf link, and the Whitehead link.

use tricolorability to prove that the Hopf link is different from the unlink. Alternatively, we may count the crossings between the two knots, κ and λ, counting with a sign. Specifically, we orient each knot arbitrarily and we look at each crossing locally. If the under-pass goes from the left of the over-pass to its right then we count $+1$ and otherwise we count -1. Letting x be a crossing and sign$x$ be plus or minus one as explained, the *linking number* is half the sum of these numbers over all crossings,

$$Lk(\kappa, \lambda) \quad = \quad \frac{1}{2} \sum_{x} \text{sign} x.$$

Changing the orientation of one knot but not the other has the effect of reversing the sign of the linking number. Clearly, Reidemeister moves do not affect

the linking number. Since the linking number of the unlink is zero and that of the Hopf link is plus or minus one, we have another proof that the two links are different. An easy link that is not splittable but has vanishing linking number is the Whitehead link in Figure I.10. It consists of two unknots but cannot be tricolored, which implies that it is not splittable.

Writhing number. Next we introduce a number that measures how contorted the curve is in space. Let $\kappa : \mathbb{R}^1 \to \mathbb{R}^3$ be a knot and assume that it is smooth and its tangent vector $\dot{\kappa}(s)$ is non-zero for every s. Projecting along a direction $u \in \mathbb{S}^2$ we get a closed curve in the plane. Assuming the projection is generic, we distinguish under- from over-passes and count each crossing plus or minus one time, as before. However, different from the case of the linking number, we count crossings the curve makes with itself and we do not divide by two. The sum of these numbers is the *directional writhing number*, $DWr(\kappa, u)$. The *writhing number* is the average over all directions. This is the integral of the directional writhing number over all directions divided by the area of the unit sphere,

$$Wr(\kappa) \quad = \quad \frac{1}{4\pi} \int_{u \in \mathbb{S}^2} DWr(\kappa, u) \, \mathrm{d}u.$$

The directions with non-generic projections form only a measure zero subset of the sphere. We therefore make no mistake when we average only over all generic projections. In contrast to the linking number, the writhing number is not necessarily an integer and it depends on the exact shape of the curve. Besides the shape it also captures topological information as we will see shortly.

The main motivation for studying the writhing number comes from molecular biology and, more specifically, the shape of DNA within the cell. Modeling its double-helix structure with a constant width ribbon, we are interested in the writhing number of the center axis, κ. The boundary of the ribbon consists of two closed curves. We need only one, $\lambda : \mathbb{S}^1 \to \mathbb{R}^3$. In the case of DNA, λ twists and turns around κ. Intuitively, the *twisting number* is the average motion of λ relative to κ. To formalize this idea, we assume that the center axis and the boundary curve are one unit of length apart and parametrized such that $\lambda(s) - \kappa(s)$ has unit length and is normal to the center axis. We construct a frame of mutually orthogonal unit vectors consisting of the tangent vector at s, $T(s) = \dot{\kappa}(s)/\|\dot{\kappa}(s)\|$, the normal vector connecting the two curves, $N(s) = \lambda(s) - \kappa(s)$, and the binormal vector, $B(s) = T(s) \times N(s)$. Using this frame, the twisting number is the average length of the projection of the

derivative of the normal vector onto the binormal vector,

$$Tw(\kappa, \lambda) \quad = \quad \frac{1}{2\pi} \int_{s \in \mathbb{S}^1} \langle \dot{N}(s), B(s) \rangle \, \mathrm{d}s.$$

This number may be interpreted as the number of local crossings between κ and λ, counted with a sign and averaged over all directions $u \in \mathbb{S}^2$. To make sense of the idea of a *local* crossing we use a limit process in which the distance between κ and λ goes to zero. Details are omitted. Similarly, the writhing number of κ may be interpreted as the number of global crossings between κ and λ, again counted with a sign, averaged over all directions, and in the limit when the separation between the knots goes to zero. Since the linking number counts all crossings, we get a relationship between the three measures, which we state without formal proof.

Călugăreanu-White Formula. Let κ be smooth closed curve in $\mathbb{R}^3$ and λ one of the two boundary curves of a ribbon centered along κ. Then $Lk(\kappa, \lambda) = Tw(\kappa, \lambda) + Wr(\kappa)$.

Relation to winding number. The writhing number of a is related to the winding number of the curve of critical directions. It is defined such that the directional writhing number remains unchanged as long as we move u on the sphere of directions without crossing the curve and it changes as soon as we cross the curve. The only Reidemeister move that affects the directional writhing number is Type I. The curve of critical directions is therefore traced out by the unit tangent vector and its negative, $T, -T : \mathbb{S}^1 \to \mathbb{S}^2$. In other words, we have two curves decomposing the sphere into maximal faces of invariant directional writhing number. It will be convenient to identify antipodal points on the sphere and think of a direction as a pair $(u, -u)$. Formally, this means we replace the sphere by the two-dimensional projective plane but we don't have to be this formal yet. The pair $(u, -u)$ crosses the curve T iff u crosses T or $-T$.

Recall that the winding number is defined for a closed curve and a point in the plane. Here we have a closed curve and an antipodal point pair on the sphere. Assuming u and $-u$ are not on the curve, we let the *winding number* be the net number of counterclockwise turns formed by T around the directed line defined by u. We use the same notation as in the plane denoting this number by $W(T, u)$. Here we define counterclockwise as seen by looking in the direction u. Figure I.11 illustrates the situation in which $-u$ crosses T from its left to its right. The winding number of T and $(u, -u)$ thus decreases by one, same as the directional writhing number. Indeed, the two change in synchrony

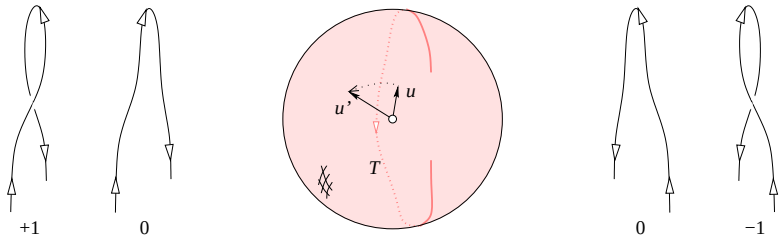

Figure I.11: The change of the viewpoint from u to u' is indicated on the sphere of directions. On the left, this removes a positive crossing and on the right, this adds a negative crossing. The effect is the same, namely a decrease in the directional writhing number by one. It remains the same even if the curves change their orientation.

in all cases and we have $DWr(\kappa, u_0) - DWr(\kappa, u) = W(T, u_0) - W(T, u)$ for all $u_0, u \in \mathbb{S}^2$. As a consequence, the average winding number differs from the average directional writhing number by an integer. Integrating the above relation over all directions of the sphere gives $DWr(\kappa, u_0)$ minus $Wr(\kappa)$ on the left and $W(\kappa, u_0)$ minus the average winding number on the right. Hence,

$$Wr(\kappa) \quad = \quad DWr(\kappa, u_0) - W(\kappa, u_0) + \frac{1}{4\pi} \int_{u \in \mathbb{S}^2} W(\kappa, u) \, \mathrm{d}u.$$

Bibliographic notes. Knots and links have been studied for centuries and there are a number of excellent books on the subject, including the text by Adams [1]. Motivation for studying the writhing number of a space curve and the twisting number of a ribbon is derived from the double-helix structure of DNA whose discovery is comparably recent [7]. These numbers measure how wound up, locally and globally, DNA is within the cell [3]. The noteworthy relation between writhing, twisting, and linking numbers has been discovered independently by Călugăreanu [4], Fuller [5], Pohl [6], and White [8]. The relationship to the winding number has been described in [2] and used to give an algorithm that computes the writhing number of a closed space polygon in subquadratic time.

[1] C. C. ADAMS. *The Knot Book.* W. H. Freeman, New York, 1994.

[2] P. K. AGARWAL, H. EDELSBRUNNER AND Y. WANG. Computing the writhing number of a polygonal knot. *Discrete Comput. Geom.* **32** (2004), 37–53.

[3] W. R. BAUER, F. H. C. CRICK, AND J. H. WHITE. Supercoiled DNA. *Scientific American* **243** (1980), 118–133.

[4] G. Călugăreanu. Sur les classes d'isotopie des noeuds tridimensionnels et leurs invariants. *Czech. Math. J.* **11** (1961), 588–625.

[5] F. B. Fuller. The writhing number of a space curve. *Proc. Natl. Acad. Sci. USA* **68** (1971), 815–819.

[6] W. F. Pohl. The self-linking number of a closed space curve. *J. Math. Mech.* **17** (1968), 975–985.

[7] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acid. A structure for deoxyribose nucleic acid. *Nature* **171** (1953), 737–738.

[8] J. H. White. Self-linking and the Gauss integral in higher dimensions. *Amer. J. Math.* **XCI** (1969), 693–728.

I.4 Planar Graphs

Only graphs with relatively few edges can be drawn without crossings in the plane. We consider properties that distinguish such graphs from others. We also prove Tutte's Theorem which implies that every graph that can be drawn without crossing can also be drawn this way with straight edges.

Embeddings. Let $G = (V, E)$ be a simple, undirected graph. A *drawing* maps every vertex $u \in V$ to a point $f(u)$ in $\mathbb{R}^2$, and it maps every edge $uv \in E$ to a path with endpoints $f(u)$ and $f(v)$. The drawing is an *embedding* if the points are distinct, the paths are simple and do not cross each other, and incidences are limited to endpoints. Not every graph can be drawn without crossings. The graph is *planar* if it has an embedding in the plane. As illustrated in Figure I.12 for the complete graph of four vertices, there are many drawings of a planar graph, some with and some without crossings. A *face* of

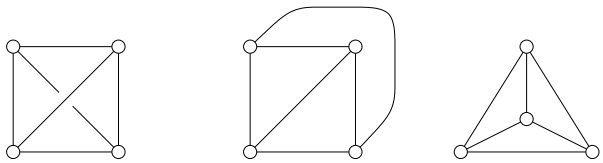

Figure I.12: Three drawings of K_4. From left to right: a drawing that is not an embedding, and embedding with one curved edge, and a straight-line embedding.

an embedding is a component in the defined decomposition of the plane. We write $n = \operatorname{card} V$, $m = \operatorname{card} E$, and ℓ for the number of faces. Euler's formula is a linear relation between these numbers.

EULER RELATION FOR PLANAR GRAPHS. Every embedding of a connected graph in the plane satisfies $n - m + \ell = 2$.

PROOF. Choose a spanning tree of $G = (V, E)$. It has n vertices, $n - 1$ edges, and one face. We have $n - (n - 1) + 1 = 2$, which proves the formula if G is a tree. Otherwise, draw the remaining edges, one at a time. Each edge decomposes one face into two, thus maintaining the relation by increasing both the number of edges and the number of faces by one. ◻

If the graph has more than one connected component then the right hand side of the equation is replaced by one plus that number. Note that the Euler Relation implies that the number of faces is the same for all embeddings and is

therefore a property of the graph. We get bounds on the number of edges and faces, in terms of the number of vertices, by considering *maximally connected* graphs for which adding any one edge would violate planarity. Every face of a maximally connected planar graph with three or more vertices is necessarily a triangle, for if there is a face with more than three edges we can add a path that crosses none of the earlier paths. Let $n \geq 3$ be the number of vertices, as before. Since every face has three edges and every edge belong to two triangles, we have $3\ell = 2m$. We use this relation to rewrite the Euler Relation: $n - m + \frac{2m}{3} = 2$ and $n - \frac{3\ell}{2} + \ell = 2$ and hence $m = 3n - 6$ and $\ell = 2n - 4$. Every planar graph can be completed to a maximally connected planar graph, which implies that it has at most these numbers of edges and faces.

Non-planarity. We can use the Euler Relation to prove that the complete graph of five vertices and the complete bipartite graph of three plus three vertices are not planar. Consider first K_5, which is drawn in Figure I.13, left. It

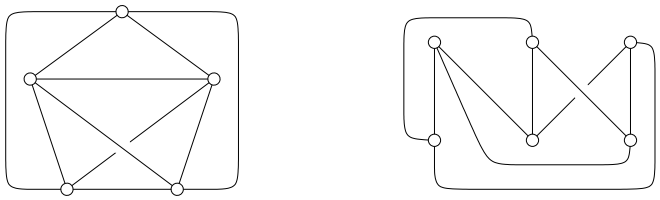

Figure I.13: K_5 on the left and $K_{3,3}$ on the right, each drawn with the unavoidable one crossing.

has $n = 5$ vertices and $m = 10$ edges, contradicting the upper bound of at most $3n-6 = 9$ edges for maximally connected planar graphs. Consider second $K_{3,3}$, which is drawn in Figure I.13, right. It has $n = 6$ vertices and $m = 9$ edges. Each cycle has even length, which implies that each face of a hypothetical embedding has four or more edges. We get $4\ell \leq 2m$ and $m \leq 2n - 4 = 8$ after plugging the inequality into the Euler Relation, again a contradiction.

In a sense, K_5 and $K_{3,3}$ are the quintessential non-planar graphs. Two graphs are *homeomorphic* if one can be obtained from the other by a sequence of operations, each deleting a degree-2 vertex and merging their two edges into one or doing the inverse.

Kuratowski Theorem. A simple graph is planar iff no subgraph is homeomorphic to K_5 or to $K_{3,3}$.

The proof of this result is omitted. The remainder of this section focuses on straight-line embeddings of planar graphs.

Convex combinations. Two points $a_0 \neq a_1$ define a unique line that passes through both. Each point on this line can be written as $x = (1-t)a_0 + ta_1$, for some $t \in \mathbb{R}$. For $t = 0$ we get $x = a_0$, for $t = 1$ we get $x = a_1$, and for $0 < t < 1$ we get a point in between. If we have more that two points, we repeat the construction by adding all points $y = (1-t)x + ta_2$ for which $0 \leq t \leq 1$, and so on, as illustrated in Figure I.14. Given $k+1$ points $a_0, a_1, \ldots, a_k$, we can

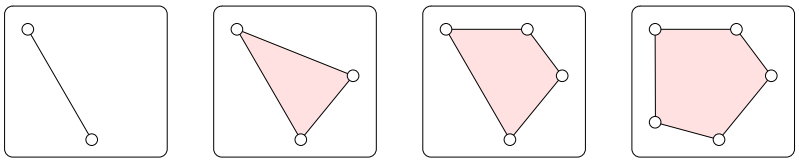

Figure I.14: From left to right: the construction of the convex hull of five points by adding one point at a time.

do the same construction in one step, calling a point $x = \sum_{i=0}^{k} t_i a_i$ a *convex combination* of the a_i if $\sum_{i=0}^{k} t_i = 1$ and $t_i \geq 0$ for all $0 \leq i \leq k$. The set of convex combinations is the *convex hull* of the a_i.

We are interested in graphs that arise as edge-skeletons of triangulations of the disk, like the one in Figure I.15. Letting $G = (V, E)$ be such a graph, we distinguish edges and vertices on the boundary from the ones in the interior of the disk. When we embed G in $\mathbb{R}^2$, we make sure that the boundary edges and vertices map to the boundary of the outer face. Since we only consider straight-line embeddings, it suffices to study mappings of the vertex set into the plane. We call $f : V \to \mathbb{R}^2$ a *strictly convex combination mapping* if for every interior vertex $u \in V$ there are real numbers $t_{uv} > 0$ with $\sum_v t_{uv} = 1$ and $f(u) = \sum_v t_{uv} f(v)$, where both sums are over all neighbors v of u. In words, every interior vertex maps to a point in the interior of the convex hull of the images of its neighbors. We will repeatedly use this mapping in combination with a linear function $h : \mathbb{R}^2 \to \mathbb{R}$ defined by $h(x) = \langle x, p \rangle + c$, where $p \in \mathbb{R}^2$ is a non-zero vector and c is a real number. Composing f with h we get $h(f(u)) = \sum_v t_{uv} h(f(v))$. In words, the value we get for u is the same strictly convex combination of the values for its neighbors.

Straight-line embeddings. Suppose we have a straight-line embedding of G in which the boundary edges map to the boundary of the outer face. Then every interior vertex lies inside the cycle connecting its neighbors. It follows that this embedding defines a strictly convex combination mapping. We now show that the reverse is also true provided the boundary vertices map to the corners of a strictly convex polygon.

TUTTE'S THEOREM. Let $G = (V, E)$ be the edge-skeleton of a triangulation of the disk and $f : V \to \mathbb{R}^2$ a strictly convex combination mapping that maps the boundary vertices to the corners of a strictly convex polygon. Then drawing straight edges between the image points gives a straight-line embedding.

We will give the proof in three steps, which we now prepare with two observations. A *separating edge* of G is an interior edge that connects two boundary vertices. It is convenient to assume that G has no separating edge, but if it does we can split the graph into two and do the argument for each piece. Call a path in G *interior* if all its vertices are interior except possibly the first and the last. Under the assumption of no separating edge, every interior vertex u can be connected to every boundary vertex by an interior path. Indeed, we can find an interior path that connects u to a first boundary vertex w. Let w_0 and w_1 be the neighboring boundary vertices. Since none of the edges separate, the neighbors of w form a unique interior path connecting w_0 to w_1. It follows that there is an interior path connecting u to w_0. By repeating the argument substituting w_0 for w we eventually see that u has interior paths to all boundary vertices.

Now suppose that $h \circ f$ takes its maximum at an interior vertex, u. Since $h \circ f(u)$ is a strictly convex combination of the values the the neighbors, we conclude that $h \circ f(v) = h \circ f(u)$ for all neighbors v of u. We can iterate and because of the mentioned interior path property we eventually reach every vertex. It thus follows that $h \circ f$ has the same value at all vertices of G. We refer to this observation as the *maximum principle* and its symmetric version as the *minimum principle*.

Proof of Tutte's Theorem. We now present the proof in three steps. First, all interior vertices u of V map to the interior of the strictly convex polygon whose corners are the images of the boundary vertices. To see this, choose $p \in \mathbb{R}^2$ and $c \in \mathbb{R}$ such that the line $h^{-1}(0)$ defined by $h(x) = \langle x, p\rangle + c$ passes through a boundary edge and $h(f(w)) > 0$ for all boundary vertices other than the endpoints of that edge. Then $h(f(u)) > 0$ else the maximum principle would imply $h(f(v)) = 0$ for all vertices. Repeating this argument for

all edges of the convex polygon implies that all interior vertices u have $f(u)$ in the interior of the polygon. This implies in particular that each triangle incident to a boundary edge is non-degenerate, that is, its three vertices are not collinear.

Second, letting yuv and zuv be the two triangles sharing the interior edge uv in G, the points $f(y)$ and $f(z)$ lie on opposite sides of the line $h^{-1}(0)$ that passes through $f(u)$ and $f(v)$. To see this, assume $h(f(y)) > 0$ and find a strictly rising path connecting y to the boundary. It exists because $h(f(y)) > h(f(u))$ so one of the neighbors of y has strictly larger function value, and the same is true for the next vertex on the path and so on. Similarly, find a strictly falling path connecting u to the boundary and the same for v, as illustrated in Figure I.15. The rising path does not cross the falling paths, but the two falling

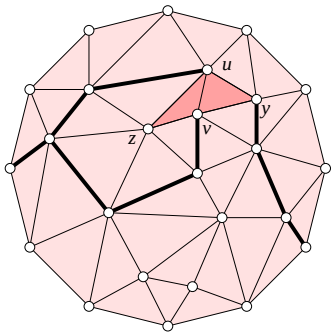

Figure I.15: One strictly rising and two strictly falling paths connecting y, u, and v to the boundary.

paths may share a vertex, as in Figure I.15. In either case, we get a piece of the triangulation bounded by vertices with non-positive function values. Other than u and v all other vertices in this boundary have strictly negative function values. If z belongs to the boundary of this piece then it has strictly negative function value simply because it differs from u and v. Else it belongs to the interior of the piece and we have $h(f(z)) < 0$ by the maximum principle. We note that this argument uses $h(f(y)) > 0$ in an essential manner. To show that this assumption is justified, we connect yuv by a sequence of triangles to one incident to a boundary edge. In this sequence, any two contiguous triangles share an edge. As observed in the first step, the image of the last triangle is non-degenerate. Going backward, this implies that the image of the second to the last triangle is non-degenerate and so on. Finally, the image of yuv is non-degenerate, as required.

Third, no two of the edges cross. To get a contradiction, assume x is a point in the common interiors of two edges, uv and $u'v'$. Choose a half-line that emanates from x and avoids the images of all vertices. Since the vertices y and z that form triangles with uv map to opposite sides of the line passing through $f(u)$ and $f(v)$, the half-line intersects exactly one of the edges yu, yv, zu, zv. Continuing this way we get a sequence of edges starting with uv and ending with a boundary edge. Similarly, the half-lines defines another sequence of edges starting with $u'v'$ and ending with the same boundary edge. Going back in both sequences, we pass from one edge to an unambiguously define preceding edge. Since we start with the same boundary edge we get $uv = u'v'$. This completes the proof of Tutte's Theorem.

Constructing straight-line embeddings. Tutte's Theorem leads to a simple algorithm for constructing a straight-line embedding of a planar graph. For simplicity, we assume that it is the edge-skeleton of a triangulation of the disk and that none of its edges separates. We reindex such that u_1 to u_k are ordered along the boundary of the outer face and u_{k+1} to u_n are the interior vertices of the graph. First, we set $f(u_i) = (\cos(2i\pi/k), \sin(2i\pi/k))$, for $1 \leq i \leq k$, to place the boundary vertices in order on the unit circle in the plane. They form the corners of a strictly convex polygon, as required. Expressing the image of each interior vertex as a strictly convex combination of the images of its neighbors, we write

$$f(u_j) \quad = \quad \frac{1}{d_j} \sum f(v),$$

for each $k+1 \leq j \leq n$, where d_j is the degree of u_i and the sum is over all neighbors v of u_j in the graph. We get a system of $n-k$ linear equations in $n-k$ unknowns, the images of the interior vertices. Writing the system in matrix form, we get one non-zero coefficient for each interior vertex and two more for each edge connecting two interior vertices. By Euler's relation, the number of edges is less than $3n$. If follows that the system is sparse with fewer than $7n$ non-zero coefficients. It thus permits efficient methods to find the solution, which by Tutte's Theorem corresponds to a straight-line embedding of the graph.

Bibliographic notes. Graphs that can be drawn in the plane without crossings arise in a number of applications, including geometric modeling, geographic information systems, and others. We refer to [3] for a collection of mathematical and algorithmic results specific to planar graphs. The fact that all planar

graphs have straight-line embeddings has been known long before Tutte's Theorem. Early last century, Steinitz showed that every 3-connected planar graph is the edge-skeleton of a convex polytope in $\mathbb{R}^3$ [4]. This skeleton can be projected to $\mathbb{R}^2$ to give a straight-line embedding. In the 1930s, Koebe proved that every planar graph is the intersection graph of a collection of possibly touching but not otherwise overlapping closed disks in $\mathbb{R}^2$ [2]. We get a straight-line embedding by connecting the centers of the touching disks. The original theorem by Tutte is for coefficients t_{uv} equal to one over the degree of u [6]. The more general version and the proof presented in this section are taken from the more recent paper by Floater [1]. Efficient numerical methods for solving systems of linear equations can be found in the linear algebra text by Strang [5].

[1] M. S. FLOATER. One-to-one piecewise linear mappings over triangulations. *Math. Comput.* **72** (2003), 685–696.

[2] P. KOEBE. Kontaktprobleme der konformen Abbildung. *Ber. Sächs. Akad. Wiss. Leipzig, Math.-Phys. Kl.* **88** (1936), 141–164.

[3] T. NISHIZEKI AND N. CHIBA. *Planar Graphs: Theory and Algorithms.* North-Holland, Amsterdam, the Netherlands, 1988.

[4] E. STEINITZ. *Polyeder und Raumeinteilung.* In *Enzykl. Math. Wiss., Part 3AB12* (1922), 1–139.

[5] G. STRANG. *Introduction to Linear Algebra.* Wellesley-Cambridge Press, Wellesley, Massachusetts, 1993.

[6] W. T. TUTTE. How to draw a graph. *Proc. London Math. Soc.* **13** (1963), 743–768.

Exercises

The credit assignment reflects a subjective assessment of difficulty. A typical question can be answered using knowledge of the material combined with some thought and analysis.

1. **Deciding connectivity** (two credits). Given a simple graph with n vertices and m edges, the disjoint set system takes time proportional to $(n+m)\alpha(n)$ to decide whether or not the graph is connected.

   (i) Describe a different an algorithm that makes the same decision in time proportional to $n+m$.

   (ii) Modify the algorithm so it computes the connected components in time proportional to $n+m$.

2. **Shelling disks** (three credits). Consider a triangulation of a simple closed polygon in the plane, but one that may have interior vertices inside the polygon. A *shelling* is a total order of the triangles such that the union of the triangles in any initial sequence is homeomorphic to a closed disk. Prove that every such triangulation has a shelling.

3. **Jordan curve** (one credit). Recall the Jordan Curve Theorem which says that every simple closed curve in the plane decomposes $\mathbb{R}^2$ into two.

   (i) Show the same is true for a simple closed curve on the sphere, $\mathbb{S}^2 = \{x \in \mathbb{R}^2 \mid \|x\| = 1\}$.

   (ii) Give an example that shows the result does not hold for simple closed curves on the torus.

4. **Homeomorphisms** (two credits). Give explicit homeomorphisms to show that the following spaces with topologies inherited from the respective containing Euclidean spaces are homeomorphic:

   - $\mathbb{R}^1 = \mathbb{R}$, the real line;
   - $(0,1)$, the open interval;
   - $\mathbb{S}^1 - \{(0,1)\}$, the circle with one point removed.

   Generalize your homeomorphisms to show the same for the Euclidean plane, the open disk, and the sphere with one point removed.

5. **Splitting a link** (two credits). Prove that the Borromean rings shown in Figure I.16 on the left are not splittable.

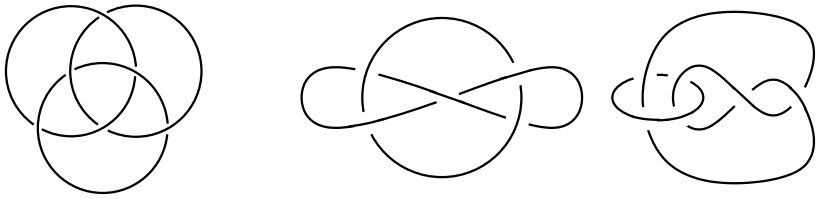

Figure I.16: Left: Any two of the three knots of the Borromean rings can be split but are held together by the third knot. Right: Two generic projections of the Whitehead link.

6. **Deforming a link** (two credits). Use Reidemeister moves to demonstrate that the two links in Figure I.16 in the middle and on the right are equivalent.

7. **Planar graph coloring** (two credits). Recall that every planar graph has a vertex of degree at most five. We can use this fact to show that every planar graph has a vertex 6-coloring, that is, a coloring of each vertex with one of six colors such that any two adjacent vertices have different colors.

   Indeed, after removing a vertex with fewer than six neighbors we use induction to 6-color the remaining graph and when we put the vertex back we choose a color that differs from the colors of its neighbors. Refine the argument to prove that every planar graph has a vertex 5-coloring.

8. **Edge coloring** (three credits). We color each edge of a maximally connected planar graph with one of three colors such that each face (triangle) has all three colors in its boundary.

   (i) Show that a 4-coloring of the vertices implies a 3-coloring of the edges.
   (ii) Show that a 3-coloring of the edges implies a 4-coloring of the vertices.

   In other words, proving that every planar graph has a vertex 4-coloring is equivalent to proving that every triangulation in the plane has an edge 3-coloring.

Chapter II

Surfaces

The most common two-dimensional spaces are 2-manifolds, or surfaces, which come in two varieties: with and without boundary. We usually envision them put into three-dimensional space, sometimes with and preferably without self-intersections. Not all surfaces can be embedded in three-dimensional Euclidean space and self-intersections are unavoidable, but often they are accidental. Indeed, choosing a nice embedding of a surface in space is an interesting computational problem. We address this question for surfaces made out of triangles.

II.1 Two-dimensional Manifolds

In our physical world, the use of the term surface usually implies a 3-dimensional, solid shape of which this surface is the boundary. In mathematics, the solid shape is not assumed and we discuss surfaces in their own right. Indeed, there are closed surfaces that are not the boundary of any solid shape. They are non-orientable and do not embed into three-dimensional Euclidean space, which is why our intuition for them is lacking.

Topological 2-manifolds. Consider the open disk of points at distance less than one from the origin, $D = \{x \in \mathbb{R}^2 \mid \|x\| < 1\}$. It is homeomorphic to $\mathbb{R}^2$, as for example established by the homeomorphism $f : D \to \mathbb{R}^2$ defined by $f(x) = x/(1 - \|x\|)$. We will call any subset of a topological space that is homemorphic to D an open disk. A 2-*manifold (without boundary)* is a topological space $\mathbb{M}$ whose points all lie in open disks. Intuitively, this means that $\mathbb{M}$ looks locally like the plane.

$\mathbb{M}$ is *compact* if for every covering of $\mathbb{M}$ by open sets, called an *open cover*, we can find a finite number of the sets that cover $\mathbb{M}$. We say that the open cover always has *finite subcover*. Examples of non-compact 2-manifolds are $\mathbb{R}^2$ itself and open subsets of $\mathbb{R}^2$. Examples of compact 2-manifolds are shown in Figure II.1, top row. We get 2-*manifolds with boundary* by removing open disks from

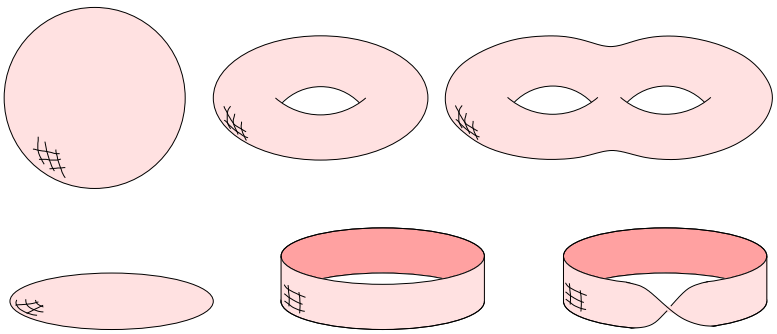

Figure II.1: Top from left to right: the sphere, $\mathbb{S}^2$, the torus, $\mathbb{T}^2$, the double torus, $\mathbb{T}^2\#\mathbb{T}^2$. Bottom from left to right: the disk, the cylinder, the Möbius strip.

2-manifolds without boundary. Alternatively, we could require that each point has a neighborhood homeomorphic to either D or to D_+, the half disk obtained by removing all points with negative second coordinate from D. The *boundary*

of a 2-manifold with boundary consists of all points x whose neighborhoods are homemorphic. Within the boundary, the neighborhood of every point x is an open interval, which is the defining property of a 1-*manifold*, or *curve*. There is only one type of connected, compact 1-manifold, namely the closed curve. Following the practice of considering topologically equivalent spaces the same, we will therefore often refer to it as a circle. If $\mathbb{M}$ is compact, this implies that its boundary is a collection of circles. Examples of 2-manifolds with boundary are the (closed) disk, the cylinder, and the Möbius strip, all illustrated in Figure II.1, bottom row.

We get new 2-manifolds from old ones by gluing them to each other. Specifically, remove an open disk each from two 2-manifolds, $\mathbb{M}$ and $\mathbb{N}$, find a homeomorphism between the two boundary circles, and identify corresponding points. The result is the *connected sum* of the two manifolds, denoted as $\mathbb{M}\#\mathbb{N}$. Forming the connected sum with the sphere does not change the manifold since it just means replacing one disk by another. Adding the torus is the same as attaching the cylinder at both boundary circles after removing two open disks.

Orientability. Of the examples we have seen so far, the Möbius strip has the curious property that it seems to have two sides locally at every interior point but there is only one side globally. To express this property intrinsically, without reference to the embedding in $\mathbb{R}^3$, we consider a small, oriented circle inside the strip. We move it around without altering its orientation, like a clock whose fingers keep turning in the same direction. However, if we slide the clock

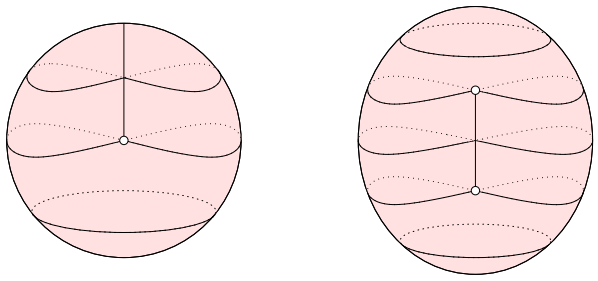

Figure II.2: Left: the projective plane, $\mathbb{P}^2$, obtained by gluing a disk to a Möbius strip. Right: the Klein bottle obtained by gluing two Möbius strips together. The vertical lines are self-intersections that are topologically not important.

once around the strip its orientation is the reverse of what it used to be and we call the path of its center an *orientation-reversing* closed curve. There are also *orientation-preserving* closed curves in the Möbius strip, such as the one

that goes around the strip twice following along close to the boundary. If all closed curves in a 2-manifold are orientation-preserving then the 2-manifold is *orientable*, else it is *non-orientable*. The curves drawn on the projective plane and the Klein bottle in Figure II.2 are all orientation-preserving. We leave finding orientation reversing curves on the same two surfaces as an instructive exercise to the reader.

Note that the boundary of the Möbius strip is a single circle. We can therefore glue the strip to a sphere or a torus after removing an open disk from the latter. This operation is often referred to as adding a *cross-cap* to the sphere or torus. In the first case we get the *projective plane*, the sphere with one cross-cap, and in the second case we get the *Klein bottle*, the sphere with two cross-caps. Both cannot be embedded in $\mathbb{R}^3$, so we have to draw them with self-intersections, but these should be ignored when we think about these surfaces.

Classification. As it turns out, we have seen examples of each major kind of compact 2-manifold. They have been completely classified about a century ago by cutting and gluing to arrive at a unique representation for each type. This representation is a convex polygon whose edges are glued in pairs, called a

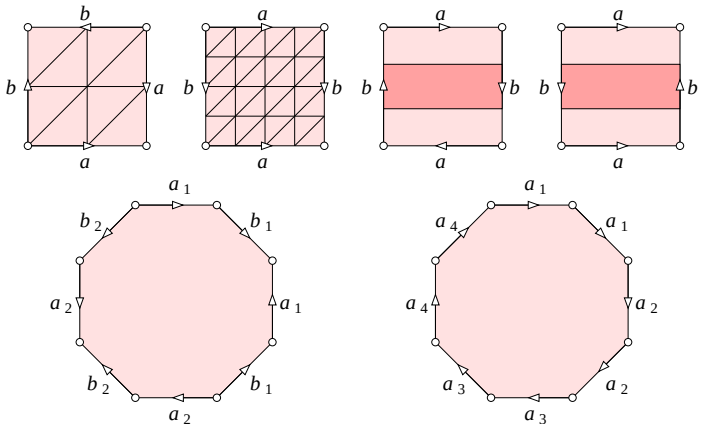

Figure II.3: Top from left to right: the sphere, the torus, the projective plane, and the Klein bottle. After removing the (darker) Möbius strip from the last two, we are left with a disk in the case of the projective plane and another Möbius strip in the case of the Klein bottle. Bottom: the polygonal schema in standard form for the double torus on the left and the double Klein bottle on the right.

polygonal schema. Figure II.3 shows that the sphere, the torus, the projective

plane, and the Klein bottle can all be constructed from the square. More generally, we have a $4g$-gon for a sphere with g tubes and a $2g$-gon for a sphere with g cross-caps attached to it. The gluing pattern is shown in the second row of Figure II.3. Note that the square of the torus is in standard form but that of the Klein bottle is not.

Classification Theorem for Compact 2-manifolds. The two infinite families $\mathbb{S}^2, \mathbb{T}^2, \mathbb{T}^2\#\mathbb{T}^2, \ldots$ and $\mathbb{P}^2, \mathbb{P}^2\#\mathbb{P}^2, \ldots$ exhaust the family of compact 2-manifolds without boundary.

The first family of orientable, compact 2-manifolds consists of the sphere, the torus, the double torus, and so on. The second family of non-orientable, compact 2-manifolds consists of the projective plane, the Klein bottle, the triple projective plane, and so on. To get a classification of the connected, compact 2-manifolds with boundary we can take one without boundary and make h holes by removing the same number of open disks. Each starting compact 2-manifold and each $h \geq 1$ give a different surface and they exhaust all possibilities.

Triangulations. To triangulate a 2-manifold we decompose it into triangular regions, each a disk whose boundary circle is cut at three points into three paths. We may think of the region and its boundary as the homeomorphic image of a triangle. By taking a geometric triangle for each region and arranging them so they share vertices and edges the same way as the regions we obtain a piecewise linear model which is a *triangulation* if it is homeomorphic to the 2-manifold. See Figure II.4 for a triangulation of the sphere. The condition

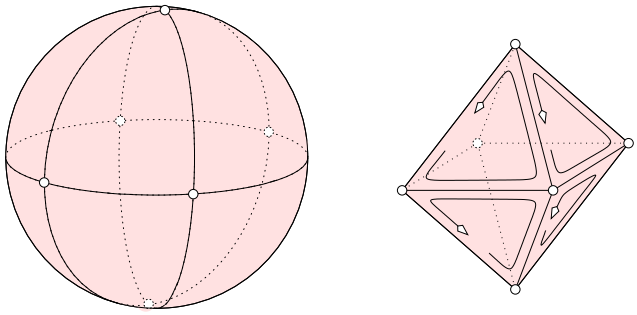

Figure II.4: The sphere is homeomorphic to the surface of an octahedron, which is a triangulation of the sphere.

of homeomorphism requires that any two triangles are either disjoint, share an

edge, or share a vertex. Sharing two edges is not permitted for then the two triangles would be the same. It is also not permitted that two vertices of a triangle are the same. To illustrate these conditions we note that the triangulation of the first square in Figure II.3 is not a valid triangulation of the sphere, but the triangulation of the second square is a valid triangulation of the torus.

Given a triangulation of a 2-manifold $\mathbb{M}$, we may orient each triangle. Two triangles sharing an edge are *consistently oriented* if they induce opposite orientations on the shared edge, as in Figure II.4. Then $\mathbb{M}$ is orientable iff the triangles can be oriented in such a way that every adjacent pair is consistently oriented.

Euler characteristic. Recall that a triangulation is a collection of triangles, edges, and vertices. We are only interested in finite triangulations. Letting n, m, and ℓ be the numbers of vertices, edges, and triangles, same as in the previous chapter, the *Euler characteristic* is their alternating sum, $\chi = n - m + \ell$. We have seen that the Euler characteristic of the sphere is $\chi = 2$, no matter how we triangulate. More generally, the Euler characteristic is independent of the triangulation for every 2-manifold.

EULER CHARACTERISTIC OF COMPACT 2-MANIFOLDS. A sphere with g tubes has $\chi = 2 - 2g$ and a sphere with g cross-caps has $\chi = 2 - g$.

The number g is the *genus* of $\mathbb{M}$; it is the maximum number of disjoint closed curves along which we can cut without disconnecting $\mathbb{M}$. To see this result we may triangulate the polygonal schema of $\mathbb{M}$. For a sphere with g tubes we have $\ell = 1$ region, $m = 2g$ edges, and $n = 1$ vertex. Further decomposing the edges and regions does not change the alternating sum, so we have $\chi = 2 - 2g$. For a sphere with g cross-caps we have $\ell = 1$ region, $m = g$ edges, and $n = 1$ vertex giving $\chi = 2 - g$.

Observe that adding a tube decreases the Euler characteristic by two while adding a cross-cap decreases it by only one. Indeed, we can substitute k handles for $2k$ cross-caps and obtain the g-fold projective plane from the k-fold torus by gluing $g - 2k$ cross-caps, provided $g > 2k$. Note that non-orientability cannot be cancelled by the connected sum. Hence, this operation can get us from the orientable to the non-orientable manifolds but not back.

Doubling. The compact, non-orientable 2-manifolds can be obtained from the orientable 2-manifolds by identifying points in pairs. For example, if we identify opposite (*antipodal*) points of the sphere we get the projective plane.

We can also go the other direction, constructing orientable manifolds from non-

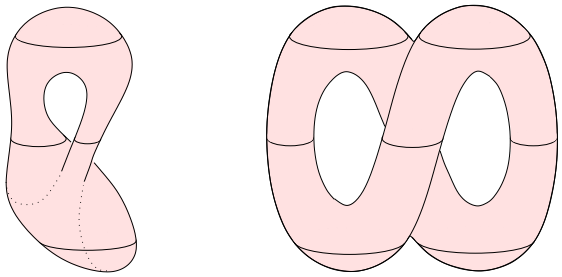

Figure II.5: Doubling the Klein bottle produces the torus.

orientable ones; see Figure II.5. Imagine a triangulation of a connected, compact, non-orientable 2-manifold $\mathbb{N}$ in $\mathbb{R}^3$, drawn with self-intersections, which we ignore. Make two copies of each triangle, edge, and vertex off-setting them slightly, one on either side of the manifold. Here sidedness is local and therefore well defined. The triangles fit together locally, and because $\mathbb{N}$ is connected, they form the triangulation of a connected 2-manifold, $\mathbb{M}$. It is orientable because one side is consistently facing $\mathbb{N}$. Since all triangles, edges, vertices are doubled, we have $\chi(\mathbb{M}) = 2\chi(\mathbb{N})$. Using the relation between genus and Euler characteristic we have $\chi(\mathbb{N}) = 2 - g(\mathbb{N})$ and therefore $\chi(\mathbb{M}) = 4 - 2g(\mathbb{N}) = 2 - 2g(\mathbb{M})$. It follows that $\mathbb{M}$ has $g(\mathbb{M}) = g(\mathbb{N}) - 1$ tubes. As listed in Table II.1, the doubling operation constructs the sphere from the projective plane, the torus from the Klein bottle, etc.. The double is sometimes called the *double cover*, since the reverse operation of re-identifying doubled regions maps $\mathbb{M}$ to $\mathbb{N}$ covering it twice.

$\chi(\mathbb{N})$	$g(\mathbb{N})$	$\mathbb{N}$	$\mathbb{M}$	$g(\mathbb{M})$	$\chi(\mathbb{M})$
1	1	$\mathbb{P}^2$	$\mathbb{S}^2$	0	2
0	2	$\mathbb{P}^2\#\mathbb{P}^2$	$\mathbb{T}^2$	1	0
-1	3	$\mathbb{P}^2\#\mathbb{P}^2\#\mathbb{P}$	$\mathbb{T}^2\#\mathbb{T}^2$	2	-2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Table II.1: Doubling turns the non-orientable 2-manifold on the left into the orientable 2-manifold on the right.

Bibliographic notes. The confusing aspects of non-orientable 2-manifolds have been captured in a delightful novel about the life within such a surface [1]. The classification of compact 2-manifolds is sometimes credited to Brahana [2]

and other times to Dehn and Heegard [3]. The classification of 3-manifolds, on the other hand, is an ongoing project within mathematics. With the proof of the Poincaré conjecture by Perelman, there is new hope that this can be soon accomplished. In contrast, recognizing whether two triangulated 4-manifolds are homeomorphic is undecidable [4]. The classification of manifolds beyond dimension three is therefore a hopeless undertaking.

[1] E. A. ABBOT. *Flatland.* Dover, New York, 1952.

[2] H. R. BRAHANA. Systems of circuits on two-dimensional manifolds. *Ann. Math.* **23** (1922), 144–168.

[3] M. DEHN AND P. HEEGARD. Analysis situ. *Enz. Math. Wiss. III A B 3*, Leipzig (1907).

[4] A. A. MARKOV. Insolubility of the problem of homeomorphy. In *Proc. Int. Congr. Math.*, 1958, 14–21.

II.2 Searching a Triangulation

Many algorithms benefit from a convenient data structure that represents a surface by storing its triangulation. In this section, we describe such a data structure and show how to use it to determine the topological type of a surface.

Ordered triangles. We begin with the description of the core piece of the data structure, which is a representation of the symmetry group of the standard triangle. Its main function will be to keep track of direction and orientation when we navigate the triangulation. This group is isomorphic to the group of permutations of three elements, the vertices of the triangle. We call each permutation an *ordered triangle* and use cyclic shifts and transpositions to move between them. As illustrated in Figure II.6, the cyclic shift from *abc* to

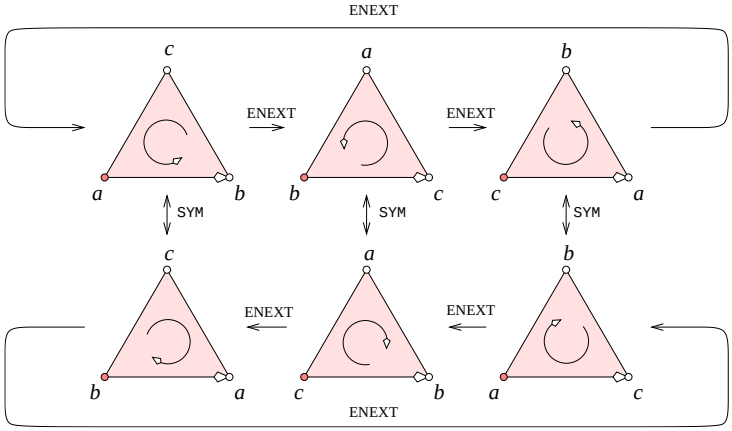

Figure II.6: The symmetry group of the standard triangle consists of six ordered versions. The cyclic shifts partition the group into two orientations, each consisting of three ordered triangles.

bca corresponds to advancing the leading directed edge to next position, from *ab* to *bc*. The transposition of the leading two vertices corresponds to reversing the direction of the lead edge while keeping the third vertex fixed.

We store each triangle in a single node of the data structure to be described shortly. A reference to the triangle consists of a pointer to this node, μ, together with a three-bit integer, ι, identifying the ordered version of the triangle. Using the first bit to identify the orientation, we represent $abc, bca, cab, bac, cba, acb$ by $\iota = 0, 1, 2, 4, 5, 6$, in this sequence. Moving between different ordered versions

of the same triangle can be done with simple arithmetic operations on ι. To advance the lead edge, we increment using modulo arithmetic.

```
ordTri ENEXT(μ, ι)
  if ι ≤ 2 then return (μ, (ι + 1) mod 3)
           else return (μ, (ι + 1) mod 3 + 4)
  endif.
```

To reverse the direction of the lead edge, we flip the first bit.

```
ordTri SYM(μ, ι)
  return (μ, (ι + 4) mod 8).
```

We see that encoding the symmetry group requires very little overhead, just a few bits whenever we point to a triangle.

Data structure. We are now ready to describe the data structure representing the triangulation K of a connected, compact, 2-manifold without boundary. We store the vertices of K in a linear array, $V[1..n]$, and the triangles in the nodes of a graph. The arcs connect nodes of neighboring triangles defined by shared edges. Since every triangle has exactly three neighbors, the degree of every node is three. Inside a node, we store pointers to the three neighbors as well as to the three vertices, which are indices into V.

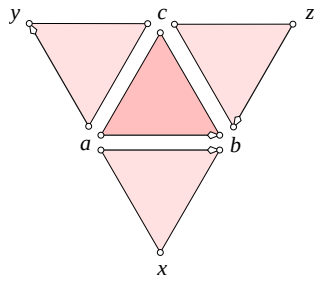

Figure II.7: The triangle abc with its three neighbors. The arrowheads identify the directed lead edges.

Let abc be a triangle and x, y, z the respective third vertices of the neighbor triangles. Each ordered version of the triangle points to its lead vertex and the ordered neighbor triangle that shares the directed lead edge. To describe this in an example, we assume the nodes μ, μ_x, μ_y, μ_z store the four triangles

with $\iota = 0$ corresponding to the ordered versions *abc*, *abx*, *ayc*, *zbc*, as drawn in Figure II.7. Assuming a is stored at positions i in V and observing that *ab* is the lead edge of *abx*, the ordered triangle *abc* stores pointers $(\mu, 0).org = i$ and $(\mu, 0).fnext = (\mu_x, 0)$. Assuming furthermore that b and c are stored at positions j and k of the vertex array, the other five ordered triangles in μ store pointers to the positions j, k, j, k, i and to the ordered triangles $(\mu_z, 1)$, $(\mu_y, 2)$, $(\mu_x, 4)$, $(\mu_z, 5)$, $(\mu_y, 6)$, in this sequence. To move around in the triangulation, we use simple functions to retrieve this information.

```
ordTri FNEXT(μ, ι)
  return (μ, ι).fnext.

int ORG(μ, ι)
  return (μ, ι).org.
```

There is clearly redundancy left in the proposed data structure, but we resist further optimizations to keep the implementation transparent.

Depth-first Search. A common operation is visiting all triangles of the triangulation. This corresponds to searching the entire representing graph. Two of the most popular strategies are Breadth-first Search and Depth-first Search. As suggested by the name, Bepth-first Search proceeds along an advancing front that expands around an initial node. In contrast, Depth-first Search ventures directly into the unknown and covers the neighborhood only after returning from the adventure. We implement the latter strategy using a recursive function. Assuming all nodes are initially unmarked, we start the search by calling that function for an arbitrary first node, μ_0.

```
void VISIT(μ)
  if μ is unmarked then mark μ; P1;
                        forall neighbors ν of μ do
                                VISIT(ν)
                        endfor; P2
                   else P3
  endif.
```

The search proceeds along a spanning tree of the graph defined by calling a neighboring node ν a *child* of μ if the first visit to ν originates from μ. The root of this tree is μ_0. To customize the function, we would add instructions at the three indicated places:

P1. steps to be executed the first time the node is visited;

P2. steps to be executed after all children have been processed;

P3. steps to be executed each time the node is revisited.

We will see examples of such customizations shortly. After searching the graph once, we will typically search it once more to remove all the marks and prepare the graph for further processing. Without accounting for the additional instructions, the running time of Depth-first Search is linear in $n + m$, the number of nodes and arcs in the graph. Indeed, each arc is traversed exactly twice, once in each direction.

Orientability. We use Depth-first Search to decide whether a connected, compact 2-manifold without boundary given by a triangulation K is orientable. We do this by orienting all triangles in a consistent manner and report non-orientability if the attempt fails. In other words, we choose one of two orientations for each triangle such that the shared edge between neighboring triangles are directed in opposite ways. Assuming none of the orientations are yet chosen, we start the process by calling the function for an arbitrary first ordered triangle, (μ_0, ι_0).

```
boolean isOrientable(μ, ι)
  if μ is unmarked then mark μ and choose orientation containing ι;
      b_x = isOrientable(fnext(sym(μ, ι)));
      b_y = isOrientable(fnext(enext(sym(μ, ι))));
      b_z = isOrientable(fnext(enext^2(sym(μ, ι))));
      return b_x and b_y and b_z
    else return [orientation of μ contains ι]
  endif.
```

Here we orient μ at P1, we unwind the for-loop, and we return a boolean value at P2 and another at P3. The latter value indicates whether or not we have consistent orientations in spite of the triangle μ having been oriented prior to the current visit. The boolean value returned at P2 indicates whether or not we have found a contradiction to orientability. A single value of FALSE anywhere during the computation is propagated to the root of the search tree telling us that the surface is non-orientable. Since each triangle has only three neighbors, the running time of the algorithm is linear in the number of triangles.

Classification. Recall from the preceding section that the type of a connected, compact 2-manifold without boundary is uniquely determined by its

genus and whether or not it is orientable. Since every triangle has three edges and every edge belongs to two triangles, we have $3\ell = 2m$ and therefore $2n - \ell = 4 - 4g$ in the orientable case and $2n - \ell = 4 - 2g$ in the non-orientable case. Assuming we know the number of vertices from the size of the array, we just need to count the triangles, which we do again by Depth-first Search.

```
int #Triangles(μ, ι)
  if μ is unmarked then mark μ;
      ℓ_x = #Triangles(fnext(μ, ι));
      ℓ_y = #Triangles(fnext(enext(μ, ι)));
      ℓ_z = #Triangles(fnext(enext^2(μ, ι)));
      return ℓ_x + ℓ_y + ℓ_z + 1
    else return 0
  endif.
```

Combining the information, it is now easy to determine the genus.

```
int Genus(μ, ι)
  ℓ = #Triangles(μ, ι);
  if isOrientable(μ, ι) then return (ℓ − 2n + 4)/4
                        else return (ℓ − 2n + 4)/2
  endif.
```

In summary, we can decide the topological type of a triangulated, compact 2-manifold without boundary in time linear in the number of triangles. We can clearly not do it faster since the entire triangulation must be searched, else we could alter the type by a small modification. By adding another search counting the boundaries, we can extend this result to compact 2-manifolds with boundary.

Bibliographic notes. Data structures for storing triangulated 2-manifolds have been described in the computer science literature since Baumgart [3]; see also the doubly-linked edge lists in [7] and the quad-edge structure in [6]. These data structures differ in their details from the graph representation described in this section but are functionally very similar. Extensions to storing 3- and higher-dimensional complexes can be found in [5] and in [2]. Searching graphs is a core topic in computer science and descriptions of Depth-first Search can be found in most algorithms texts, including [1] and [4].

[1] A. V. Aho, J. E. Hopcroft and J. D. Ullman. *The Design and Analysis of Computer Algorithms.* Addison-Wesley, Reading, Massachusetts, 1973.

[2] E. BRISSON. Representing geometric structures in d dimensions: topology and order. *Discrete Comput. Geom.* **9** (1993), 387–426.

[3] B. BAUMGART. A polyhedron representation for computer vision. In "Proc. Natl. Comput. Conf., 1975", 589–596.

[4] T. H. CORMEN, C. E. LEISERSON, R. L. RIVEST AND C. STEIN. *Introduction to Algorithms.* Second edition, McGraw Hill, Boston, 2001.

[5] D. P. DOBKIN AND M. J. LASZLO. Primitives for the manipulation of three-dimensional subdivisions. *Algorithmica* **4** (1989), 3–32.

[6] L. J. GUIBAS AND J. STOLFI. Primitives for the manipulation of general subdivisions and the computation of Voronoi diagrams. *ACM Trans. Graphics* **4** (1985), 74–123.

[7] F. P. PREPARATA AND M. I. SHAMOS. *Computational Geometry: an Introduction.* Springer-Verlag, New York, 1985.

II.3 Self-intersections

Since non-orientable, compact 2-manifolds without boundary cannot be embedded in three-dimensional Euclidean space, all their models in that space occur with self-intersections. In contrast, all orientable, compact 2-manifolds have embeddings, but their models may have accidental self-intersections. Removes those is a core topic in repairing surface models of solid shapes.

Mapping into space. Let $\mathbb{M}$ be a compact 2-manifold without boundary. We want to say what it means for M to be smooth and for a continuous map $f : \mathbb{M} \to \mathbb{R}^3$ to be a smooth mapping. We define a *coordinate chart* $\{(U, \phi)\}$ to be an open set $U \subset \mathbb{M}$ together with a continuous map $\phi : U \to R^2$ that is a homeomorphism onto its image. Two coordinate charts $\{(U, \phi)\}$ and $\{(V, \psi)\}$ are *compatible* if U and V are disjoint, or the map

$$\phi \circ \psi^{-1} : \psi(U \cap V) \to \phi(U \cap V)$$

extends to a smooth (C^∞) map $\mathbb{R}^2 \to \mathbb{R}^2$. We define $\mathbb{M}$ to be smooth if it is covered by a *maximal* collection $\{(U, \phi)\}$ of compatible coordinate charts. A continuous function $f : \mathbb{M} \to \mathbb{R}$ is smooth if for each coordinate chart $\{(U, \phi)\}$, $f \circ \phi^{-1}$ is smooth. A map to $f : \mathbb{M} \to \mathbb{R}^3$ is smooth if each of the component functions $f_i = \pi_i \circ f$ is smooth, where π_i denotes projection onto the i-th factor.

For the time being, we assume that $\mathbb{M}$ and f are smooth. If we choose a coordinate chart, we get a local parameterization of $\mathbb{M}$ with two variables. We then think of the coordinate functions f_i as mapping pairs (s_1, s_2) to x_i, for $i = 1, 2, 3$. Collecting the gradients of the coordinate functions in a matrix, we get the *Jacobian* of f,

$$J \quad = \quad \begin{bmatrix} \frac{\partial f_1}{\partial s_1} & \frac{\partial f_1}{\partial s_2} \\ \\ \frac{\partial f_2}{\partial s_1} & \frac{\partial f_2}{\partial s_2} \\ \\ \frac{\partial f_3}{\partial s_1} & \frac{\partial f_3}{\partial s_2} \end{bmatrix}.$$

While this Jacobian matrix depends on the choice of local coordinates, its rank does not. Notice that the rank of the Jacobian is at most two. The mapping f is an *immersion* if the Jacobian has full rank two at all points of $\mathbb{M}$. It is an *embedding* if f is a homeomorphism onto its image, an embedding is necessarily an immersion, but not vice-versa. For smooth mappings, there are three types of generic self-intersections, all illustrated in Figure II.8. The most interesting of the three is the branch point, which comes in several guises. We

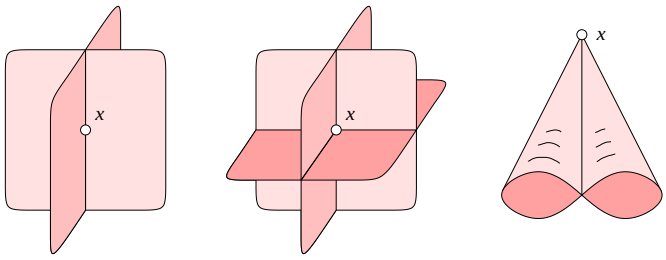

Figure II.8: From left to right: a double point, a triple point, a branch point.

can construct it by cutting a disk from two sides toward the center, folding it, and re-glueing the sides as shown in Figure II.9. Embeddings have no self-intersections at all and immersions have only the first two types and no branch points.

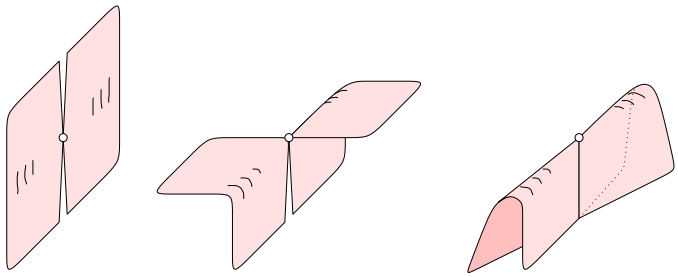

Figure II.9: Constructing the Whitney umbrella from a disk.

The piecewise linear case. The classification of generic self-intersections is similar in the piecewise linear case in which $\mathbb{M}$ is given by a finite triangulation, K. However, in contrast to the smooth case, the enumeration of the generic types is elementary. Since $\mathbb{M}$ is a 2-manifold, the triangles that contain a vertex form a disk. It is not difficult to see that imposing this condition on the vertices suffices to guarantee that K triangulates a 2-manifold without boundary. On the other hand, requiring that each edge belongs to exactly two triangles is not sufficient.

We put K into space by mapping each vertex to a point in $\mathbb{R}^3$. The edges and triangles are mapped to the convex hulls of the images of their vertices.

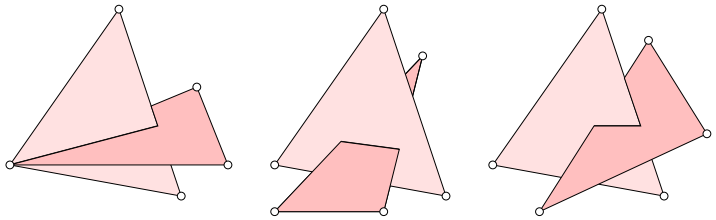

Figure II.10: The three ways two triangles whose vertices are in general position in $\mathbb{R}^3$ can cross each other.

This mapping is an embedding iff any two triangles are either disjoint or they share a vertex or they share an edge. Any other type of intersection is improper and referred to as a *crossing*. It is convenient to assume that the points are in general position, that is, no three are collinear and no four are coplanar. Under this assumption, there are only three types of crossings possible between two triangles, all shown in Figure II.10. Each crossing is a line segment common to two triangles. In the first case, one of the endpoints of the line segment coincides with the image of a vertex, which necessarily belongs to both crossing triangles. In the other two cases, each endpoint of the line segment lies on the images of an edge in the triangulation.

Recognizing crossings. We reduce the recognition problem from two triangles to an edge and a triangle and further to four points in space. Writing a_1, a_2, a_3 for the coordinates of the point a in space and similarly for the points x, y, and z, we say the sequence $axyz$ has *positive orientation* if the matrix

$$\Delta(a,x,y,z) \quad = \quad \begin{bmatrix} 1 & a_1 & a_2 & a_3 \\ 1 & x_1 & x_2 & x_3 \\ 1 & y_1 & y_2 & y_3 \\ 1 & z_1 & z_2 & z_3 \end{bmatrix}$$

has positive determinant. We observe that this corresponds to the case in which a sees xyz make a right-turn in space. The four points lie in a common plane iff the determinant vanishes. Finally, we say $axyz$ has *negative orientation* if $\det \Delta(a,x,y,z) < 0$.

Using the ability to decide the orientation of a sequence of four points, we now return to the next more complicated problem given by five points, a, b, x, y, z in $\mathbb{R}^3$. We say the edge ab *stabs* the triangle xyz if the two have an improper intersection. Assuming the five points are distinct and in general position, we

have only two cases, namely either the intersection is empty or a point in the common interior of the edge and the triangle. Thus, ab stabs xyz iff a and b lie on different sides of the plane spanned by xyz and ab forms the same orientation with the three directed edges xy, yz, and zx.

```
boolean doesStab(a, b, x, y, z)
  return sign det Δ(a, x, y, z) ≠ sign det Δ(b, x, y, z) and
    sign det Δ(a, b, x, y) = sign det Δ(a, b, y, z) = sign det Δ(a, b, z, x).
```

We finally return to the original recognition problem formulated for two triangles, abc and xyz. We first consider the case in which they share one of the points, $a = x$. Then we have a crossing iff one of the respective opposite edges stabs the other triangle. We second consider the case in which the six points are distinct. Then the triangles are disjoint iff none of the six edges stabs the other triangle, and the triangles cross iff exactly two edges stab the other triangle. Assuming general position, there are no other cases. If the two stabbing edges belong to the same triangle, we have the case in the middle in Figure II.10, and if they belong to different triangles, we have the case on the right.

Curves and preimages. Returning to the case on the left in Figure II.10, we see that one endpoint of the line segment lies on the image of an edge of the triangulation. There is a unique triangle on the other side of that edge that continues the intersections. Similarly, there are unique continuations of the intersection in the middle and the right case. Starting at a crossing, we can therefore trace the intersection triangle by triangle, adding a line segment at a time. Since we only have finitely many triangles, the curve must either end or close up by coming back to where it started. These are the only two cases:

- a path that starts at the image of a vertex and ends at the image of another vertex;
- a closed curve that avoids the images of all vertices in the triangulation.

Almost all points of such a path or closed curve are double points. Exceptions are triple points at which the curves intersects each other or themselves. The number of triple points is at most the number of ways we can choose three triangles, which is finite, and generically there are no points that belong to more than three triangles.

When we trace a path or a closed curve in space, we can, at the same time, trace its preimage under the mapping f. In the case of a path, we get two

arcs starting at a common vertex and ending at another common vertex of the triangulation. In the case of the closed curve, we get either two loops or one loop whose image covers the curve twice. The three cases are illustrated in Figure II.11. The most interesting case is the double-covering loop. Such a

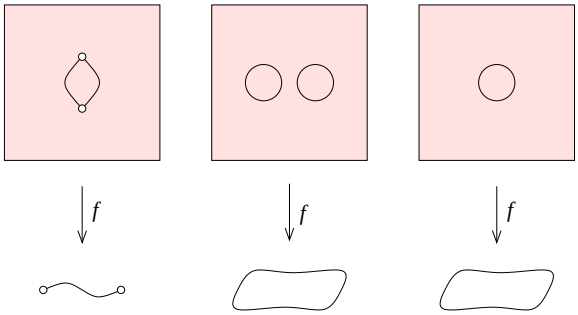

Figure II.11: The preimage of an intersection curve. From left to right: two arcs with common endpoints, two loops, one loop covering the closed curve twice.

loop is necessarily orientation-reversing. To see this, we may again trace the closed curve, its image in $\mathbb{R}^3$, and this time draw parallel curves to the left and the right on one of the two intersecting sheets. At the time we come back to where we started, the parallel curves have moved to the other sheet. There is either a clockwise or a counterclockwise rotation of the first sheet to the second that maps each curve locally to itself. If the rotation is clockwise, as seen by looking in the direction of the curve, then it is clockwise at all points of the curve. Same for counterclockwise. This implies that after another round we map the first sheet to itself but with reversed orientation. The double-covering loop can thus only happen if $\mathbb{M}$ is non-orientable. No conclusion can be drawn if the preimage consists of two loops.

To construct an example of a double-covering loop, we sweep the midpoint of a rod (a line segment) along a circle in space. The rod is normal to the circle at all times but it may rotate within the normal plane as we sweep along. If there is no rotation then the rod sweeps out a cylinder, and if the rotation is π after one time around then we get a Möbius strip. However, if the rotation is $\frac{\pi}{2}$, we need a second time around to complete the surface. We thus get a Möbius strip that crosses itself along the center circle, which is covered twice.

Immersions of the Klein bottle. We have seen a first picture of the Klein bottle in Figure II.2. The surface in that drawing intersects itself along a

path which ends at two branch points. In the smooth case, we get rank-deficient Jacobians at the branch points implying that this is not the image of an immersion. However, the Klein bottle can also be mapped without branch points and we conclude this section with the description of two such mappings.

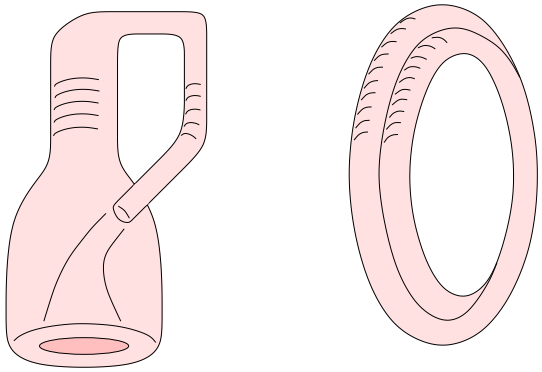

Figure II.12: Two immersions of the Klein bottle. Both models intersect themselves in a closed curve whose preimage are two loops. On the left, these loops are orientation-preserving and on the right, they are orientation-reversing.

In the first immersion, the neck of the bottle extends and turns back to the body, like a sleeping Flamingo, but then continues and passes through the surface, as sketched in Figure II.12 on the left. The closed intersection curve is the common image of two orientation-preserving loops. The second immersion is obtained by sweeping the cross point of a figure-8 along a circle in space. Similar to the rod example above, we keep the figure-8 normal to the circle at all times but we rotate within the normal plane. Turning the figure-8 upside down during one time around we exchange the lobes and form a surface that intersects itself along the circle, as sketched in Figure II.12 on the right. The preimage of the circle consists of two loops, both of which are orientation-reversing.

Bibliographic notes. The way surfaces in three-dimensional space intersect each other and themselves is discussed in length and with many illustrations by Carter [2]. In the generic case, a smooth mapping to $\mathbb{R}^3$ has only three types of singularities, double points, triple points, and branch points. Whitney proved that every d-manifold has an immersion in $\mathbb{R}^{2d-1}$ [4]. This implies that every 2-manifold can be immersed in $\mathbb{R}^3$, meaning there are smooth mappings without branch points. For the projective plane, we must have a branch point or a triple

point which implies that every immersion has a triple point [1]. Whitney also proved that every d-manifold can be embedded in $\mathbb{R}^{2d}$ [3], implying that every 2-manifold can be embedded in $\mathbb{R}^4$.

[1] T. F. Banchoff. Triple points and surgery of immersed surfaces. *Proc. Amer. Math. Soc.* **46** (1974), 403–413.

[2] J. S. Carter. *How Surfaces Intersect in Space. An Introduction to Topology.* Second edition, World Scientific, Singapore, 1995.

[3] H. Whitney. The self-intersections of a smooth n-manifold in $2n$-space. *Annals of Math.* **45** (1944), 220–246.

[4] H. Whitney. The singularities of a smooth n-manifold in $(2n-1)$-space. *Annals of Math.* **45** (1944), 247–293.

II.4 Surface Simplification

In applications, it is often necessary to simplify the data or its representation. One reason is measurement noise, which we would like to eliminate, another are features, which we look for at various levels of resolution. In this section, we study edge contractions used in simplifying triangulated surface models of solid shapes.

Edge contraction. Suppose K is a triangulation of a 2-manifold without boundary. We recall this means that edges are shared by pairs and vertices by rings of triangles, as depicted in Figure II.13. Let a and b be two vertices and ab the connecting edge in K. By the *contraction* of ab we mean the operation that identifies a with b and removes duplicates from the triangulation. Calling the new vertex c, we get the new triangulation L from K by

- removing ab, abx, and aby;
- substituting c for a and for b wherever they occur in the remaining set of vertices, edges, and triangles;
- removing resulting duplications making sure L is a set.

As a consequence of the operation, there are new incidences between edges and triangles that did not exist in K; see Figure II.13.

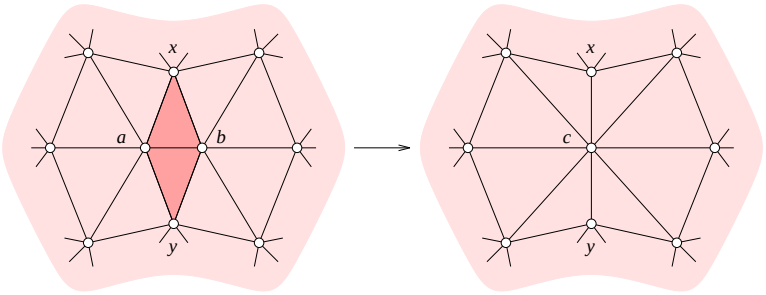

Figure II.13: To contract ab, we remove the two dark triangles and repair the hole by gluing their two left edges to their two right edges.

Algorithm. To simplify a triangulation, we iterate the edge contraction operation. In the abstract setting, any edge is as good as any other. In a practical

situation, we will want to prioritize the edges so that contractions that preserve the shape of the manifold are preferred. To give meaning to this statement, we will define shape to mean the topological type of the surface as well as the geometric form we get when we embed the triangulation in $\mathbb{R}^3$. We will discuss the latter meaning later and for now assume we have a function that assigns to each edge ab a non-negative real number $\textsc{Error}(ab)$ assessing the damage the contraction of ab causes to the geometric form. Small numbers will mean little damage. To write the algorithm, we assume a priority queue storing all edges ordered by the mentioned numerical error assessment. This is a data structure that supports the operations of returning the top priority edge as well as of inserting and deleting an edge, each in time at most logarithmic in the number of edges in the queue. Specifically, we assume a a function isEmpty that tests whether or not the priority queue still contains edges, and a function MinExtract that removes the edge with minimum error from the priority queue and returns it. Furthermore, we assume the availability of a boolean test isSafe that decides whether or not the contraction of an edge preserves the topological type of the surface.

```
while not isEmpty do ab = MinExtract;
                     if isSafe(ab) then contract ab endif
endwhile.
```

Some modifications are necessary to recognize edges that no longer belong to the triangulation and to put edges back into the priority queue when they become safe for contraction. Details are omitted. The running time of the algorithm depends on the size of local neighborhoods in the triangulation and on the data structure we maintain to represent it. Under reasonable assumptions, the most time-consuming step is the maintenance of the priority queue, which for each step is only logarithmic in the number of edges.

Topological type. We now consider the question whether or not the contraction of an edge preserves the topological type. Define the *link* of an edge ab as the set of vertices that span triangles with ab, and the link of a vertex a as the set of vertices that span edges with a and the set of edges that span triangles with a,

$$\begin{aligned} \operatorname{Lk} ab &= \{x \in K \mid abx \in K\}; \\ \operatorname{Lk} a &= \{x, xy \in K \mid ax, axy \in K\}. \end{aligned}$$

Since the topological type of K is that of a 2-manifold without boundary, each edge link is a pair of vertices and each vertex link is a closed curve made up of

edges and vertices in K. Let L be obtained from K by contracting the edge ab. We show that the contraction of the edge ab preserves the topological type of the surface iff the links of the endpoints, a and b, meet in exactly two points, namely in the vertices x and y in the link of ab, as in Figure II.13. We will simplify language by blurring the difference between a triangulation and the topological space it triangulates.

LINK CONDITION LEMMA. The triangulations K and L have the same topological type iff $\operatorname{Lk} ab = \operatorname{Lk} a \cap \operatorname{Lk} b$.

PROOF. We have $\operatorname{Lk} ab \subseteq \operatorname{Lk} a, \operatorname{Lk} b$, by definition. The only possible violation to the link condition is therefore an extra edge or vertex in the intersection of the two vertex links. If $\operatorname{Lk} a$ and $\operatorname{Lk} b$ share an edge then the contraction of ab creates a triangle sticking out of the surface, contradicting that L triangulates a 2-manifold. Similarly, if the two vertex links share a vertex $z \notin \operatorname{Lk} ab$ then the contraction of ab creates an edge cz that belongs to four triangles, again contradicting that L triangulates a 2-manifold.

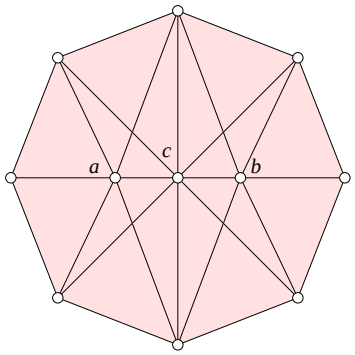

Figure II.14: Mapping the neighborhood of c in L to a triangulated polygon and overlaying it with a similar mapping of the neighborhoods of a and b in K.

To prove the other direction, we draw the link of c in L as a convex polygon in $\mathbb{R}^2$; see Figure II.14. Using Tutte's Theorem from the previous chapter, we can decompose the polygon by drawing the triangles incident to c in L. Similarly, we can decompose the polygon by drawing the triangles incident to a and b in K. We superimpose the two triangulations and refine to get a new triangulation, if necessary. The result is mapped back to K and to L, effectively refining the neighborhoods of a and b in K and that of c in L. The link of c and everything outside that link is untouched by the contraction. Hence, on

and outside the link K and L are the same and inside the link K and L are now isomorphic by refinement. It follows that K and L are isomorphic and therefore have the same topological type. □

Squared distance. To talk about the geometric meaning of shape, we now assume that K is embedded in $\mathbb{R}^3$, with straight edges and flat triangles. To develop an error measure, we use the planes spanned by the triangles. Letting $u \in \mathbb{S}^2$ be the unit normal of a plane h and $\delta \in \mathbb{R}$ its offset, we can write h as the set of points $y \in \mathbb{R}^3$ for which $\langle y, u \rangle = -\delta$. Using matrix notation for the scalar product, the *signed distance* of a point $x \in \mathbb{R}^3$ from h is

$$d(x,h) \quad = \quad (x-y)^T \cdot u \quad = \quad x^T \cdot u + \delta.$$

Defining $\mathbf{x}^T = (x^T, 1)$ and $\mathbf{u}^T = (u^T, \delta)$, we can write this as a four-dimensional scalar product, $\mathbf{x}^T \cdot \mathbf{u}$. We use this to express the sum of squared distances from a set of planes in matrix form. Letting H be a finite set of planes, this gives a function $E_H : \mathbb{R}^3 \to \mathbb{R}$ defined by

$$\begin{aligned} E_H(x) &= \sum_{h_i \in H} d^2(x, h_i) \\ &= \sum_{h_i \in H} (\mathbf{x}^T \cdot \mathbf{u}_i)(\mathbf{u}_i^T \cdot \mathbf{x}) \\ &= \mathbf{x}^T \cdot \left(\sum_{h_i \in H} \mathbf{u}_i \cdot \mathbf{u}_i^T \right) \cdot \mathbf{x}. \end{aligned}$$

Hence $E_H(x) = \mathbf{x}^T \cdot \mathbf{Q} \cdot \mathbf{x}$, where

$$\mathbf{Q} \quad = \quad \sum_{h_i \in H} (\mathbf{u}_i \cdot \mathbf{u}_i^T) \quad = \quad \begin{bmatrix} A & P & Q & U \\ P & B & R & V \\ Q & R & C & W \\ U & V & W & Z \end{bmatrix}$$

is a symmetric, four-by-four matrix we refer to as the *fundamental quadric* of the map E_H. Writing $x^T = (x_1, x_2, x_3)$ we get

$$\begin{aligned} E_H(x) \quad = \quad & Ax_1^2 + Bx_2^2 + Cx_3^2 + 2(Px_1x_2 + Qx_1x_3 + Rx_2x_3) \\ & + 2(Ux_1 + Vx_2 + Wx_3) + Z. \end{aligned}$$

We see that E_H is a quadratic map that is non-negative and unbounded.

Error assessment. In the application, we are interested in measuring the damage to the geometric form caused by contracting the edge ab to the new vertex c. We think of the operation as a map between vertices, $\varphi : \operatorname{Vert} K \to \operatorname{Vert} L$, defined by $\varphi(a) = \varphi(b) = c$ and $\varphi(x) = x$ for all $x \neq a, b$. Letting K_0 be the initial triangulation, we obtain L by a sequence of edge contractions giving rise to a composition of vertex maps, which is again a vertex map, $\varphi_0 : \operatorname{Vert} K_0 \to \operatorname{Vert} L$. The vertices in $V_c = \varphi_0^{-1}(c) \subseteq \operatorname{Vert} K_0$ all map to c and we let H be the set of planes spanned by triangles in K_0 incident to at least one vertex in V_c. Finally, we define the *error* of the contraction of ab as the minimum, over all possible placements of c as a point in $\mathbb{R}^3$, of the sum of squared distances from the planes,

$$\textsc{Error}(ab) \quad = \quad \min_{c \in \mathbb{R}^3} E_H(c).$$

For generic sets of planes, this minimum is unique and easy to compute. The gradient of $E = E_H$ at a point x is the vector of steepest increase, $\nabla E(x) = (\frac{\partial E}{\partial x_1}(x), \frac{\partial E}{\partial x_2}(x), \frac{\partial E}{\partial x_3}(x))$. It is zero iff x minimizes E. The derivative with respect to x_i can be computed using the multiplication rule,

$$\begin{aligned} \frac{\partial E}{\partial x_i} &= \frac{\partial \mathbf{x}^T}{\partial x_i} \cdot \mathbf{Q} \cdot \mathbf{x} + \mathbf{x}^T \cdot \mathbf{Q} \cdot \frac{\partial \mathbf{x}}{\partial x_i} \\ &= \mathbf{Q}[i]^T \cdot \mathbf{x} + \mathbf{x}^T \cdot \mathbf{Q}[i], \end{aligned}$$

where $\mathbf{Q}[i]$ is the i-th column and $\mathbf{Q}[i]^T$ is the i-th row of $\mathbf{Q}$. The point $c \in \mathbb{R}^3$ that minimizes E can thus be computed by setting $\frac{\partial E}{\partial x_i}$ to zero, for $i = 1, 2, 3$, and solving the resulting system of three linear equations.

Maintenance of the error measure. It can be expensive to compute the fundamental quadric from scratch but relatively inexpensive to maintain it throughout the algorithm. When we contract an edge ab we associate the new vertex with the union of the two plane sets, $H_c = H_a \cup H_b$. Unfortunately, this is not a disjoint union and we cannot just add the two quadrics. Instead, we use inclusion-exclusion and subtract the quadric of $H_{ab} = H_a \cap H_b$, which we store with the contracted edge. We describe how this works from the beginning.

Starting with the initial complex, K_0, we store a quadric with every vertex, every edge, and every triangle. For a triangle, abx, we store the quadric $\mathbf{Q}_{abx}$ defined by the one plane that contains the triangle. An edge, ab, is shared by two triangles, abx and aby, and we store the quadric defined by the two corresponding planes, $\mathbf{Q}_{ab} = \mathbf{Q}_{abx} + \mathbf{Q}_{aby}$. A vertex, a, is shared by the ring of triangles in its star and we initialize its quadric, $\mathbf{Q}_a$, to the sum of the quadrics of these triangles. Note that the triangles that share the edge ab are

precisely the ones that share both endpoints, a and b. This gives rise to a simple relationship between the sets of planes.

INVARIANT. Let abx be a triangle in the surface triangulation, with edges ab, ax, ay and vertices a, b, x. Then $H_{ab} = H_a \cap H_b$ and $H_{abx} = H_{ax} \cap H_{bx}$.

To maintain these two relations past an edge contraction, it is important that we limit ourselves to those that satisfy the Link Condition Lemma and therefore the topological type of the surface. The relations are therefore indeed invariants of the algorithm. Now consider the contraction of the edge ab. By the Invariant, the set of planes associated with the edge is the intersection of those of the endpoints. Hence we can compute the quadric of the new vertex as $\mathbf{Q}_c = \mathbf{Q}_a + \mathbf{Q}_b - \mathbf{Q}_{ab}$. We also get two new edges, cx and cy, and to maintain the Invariant, we associate each with the union of plane sets of the corresponding old edges. By the Invariant, these two sets overlap in the plane set of the shared triangle, which consists of a single plane. Hence, we get $\mathbf{Q}_{cx} = \mathbf{Q}_{ax} + \mathbf{Q}_{bx} - \mathbf{Q}_{abx}$ and $\mathbf{Q}_{cy} = \mathbf{Q}_{ay} + \mathbf{Q}_{by} - \mathbf{Q}_{aby}$.

Bibliographic notes. The algorithm described in this section is essentially the surface simplification algorithm by Garland and Heckbert [2]. They combine edge contractions with the error measure remembering the original form through accumulated quadrics. However, instead of maintaining the quadric through inclusion-exclusion, they take a short-cut and compute the quadric of the new vertex as the sum of quadrics of the endpoints of the contracted edge, without removing duplicates. In practice, this makes little difference because planes contribute at most in triplicates The test for maintaining the topological type has been added later and more general versions of the Link Condition Lemma can be found in [1]. Priority queues are standard tools in computer science and implementations are described in most texts on algorithms, including volume three of Knuth's pioneering series [3].

[1] T. DEY, H. EDELSBRUNNER, S. GUHA AND D. V. NEKHAYEV. Topology preserving edge contraction. *Publ. Inst. Math. (Beograd) (N.S.)* **66** (1999), 23-45.

[2] M. GARLAND AND P. S. HECKBERT. Surface simplification using quadric error metrics. *Computer Graphics*, Proc. SIGGRAPH, 1997, 209–216.

[3] D. E. KNUTH. *Sorting and Searching. The Art of Computer Programming, Vol. 3.* Addison-Wesley, Reading, Massachusetts, 1973.

Exercises

The credit assignment reflects a subjective assessment of difficulty. A typical question can be answered using knowledge of the material combined with some thought and analysis.

1. **Classifying 2-manifolds** (two credits). Characterize the two surfaces depicted in Figure II.15 in terms of genus, boundary, and orientability.

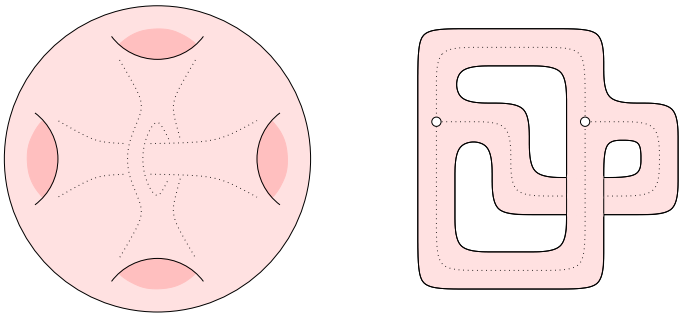

Figure II.15: Left: a 2-manifold without boundary obtained by adding tunnels inside the sphere. We see four tunnel openings and one tunnel passing though a fork of the other. Right: a 2-manifold with boundary obtained by thickening a graph.

2. **2-coloring** (two credits). Let K be a triangulation of an orientable 2-manifold without boundary. Construct L by decomposing each edge into two and each triangle into six. To do this, we add a new vertex in the interior of each edge. Similarly, we add a new vertex in the interior of each triangle, connecting it to the six vertices in the boundary of the triangle. The resulting structure is the same as the barycentric subdivision of K, which we will define in Chapter III.

   (i) Show that the vertices of L can be 3-colored such that no two neighboring vertices receive the same color.

   (ii) Prove that the triangles of L can be 2-colored such that no two triangles sharing an edge receive the same color.

3. **Klein bottle** (two credits). Cut and paste the standard polygonal schema for the Klein bottle (a, a, b, b) to obtain the polygonal schema in which opposite edges of a square are identified (a, b, a^{-1}, b); see Figure II.3.

4. **Triangulation of 2-manifold** (two credits). Let $V = \{1, 2, \ldots, n\}$ be a set of n vertices and $F \subseteq \binom{V}{3}$ a set of $\ell = \operatorname{card} F$ triangles. Give an algorithm that takes time at most proportional to $n + \ell$ for the following tasks:

   (i) decide whether or not every edge is shared by exactly two triangles;

   (ii) decide whether or not every vertex belongs to a set of triangles whose union is a disk.

5. **Intersection tests in $\mathbb{R}^3$** (two credits). Let $a, b, c \in \mathbb{R}^3$ and $u, v, w \in \mathbb{R}^3$ be the vertices of two triangles in space. Write numerical tests for the following questions:

   (i) does u see a, b, c form a left-turn or a right-turn?

   (ii) does the line segment with endpoints u and v cross the plane that passes through a, b, c?

   (iii) are the boundaries of the two triangles linked in $\mathbb{R}^3$?

6. **Irreducible triangulations** (three credits). An *irreducible* triangulation is one in which every edge contraction changes its topological type. Prove that the only irreducible triangulation of $\mathbb{S}^2$ is the boundary of the tetrahedron, which consists of four triangles sharing six edges and four vertices.

7. **Graphs on Möbius strip** (one credit). Is every graph that can be embedded on the Möbius strip planar?

8. **Squared distance minimization** (two credits). Let S be a finite set of points in $\mathbb{R}^3$ and $f : \mathbb{R}^3 \to \mathbb{R}$ be defined by $f(x) = \sum_{p \in S} \|x - p\|^2$.

   (i) Show that f is a quadratic function and has a unique minimum.

   (ii) At which point does f attain its minimum?

Chapter III

Complexes

There are many ways to represent a topological space, one being a decomposition into simple pieces. This decomposition qualifies to be called a complex if the pieces are topologically simple and their common intersections are lower-dimensional pieces of the same kind. Within these requirements, we still have a great deal of freedom. Particularly attractive are the extreme choices: few complicated or many simple pieces. The former choice lends itself to hand-calculations of topological invariants but also to the design of aesthetically pleasing shapes, such as car bodies and the like. The latter choice is preferred in computation and automation. Since we focus on computational aspects of topology, we favor the latter extreme choice of which the simplicial complex is the prime example.

III.1 Simplicial Complexes

In this book, we use simplicial complexes as the prime data structure to represent topological spaces. In this section, we introduce them in their geometric as well as abstract forms. The main technical result is the existence of simplicial maps that approximate continuous maps arbitrarily closely.

Simplices. Let $u_0, u_1, \ldots, u_k$ be points in $\mathbb{R}^d$. A point $x = \sum_{i=0}^{k} \lambda_i u_i$ is an *affine combination* of the u_i if the λ_i sum to 1. The *affine hull* is the set of affine combinations. It is a *k-plane* if the $k+1$ points are *affinely independent* by which we mean that any two affine combinations, $x = \sum \lambda_i u_i$ and $y = \sum \mu_i u_i$, are the same iff $\lambda_i = \mu_i$ for all i. The $k+1$ points are affinely independent iff the k vectors $u_i - u_0$, for $1 \leq i \leq k$, are linearly independent. In $\mathbb{R}^d$ we can have at most d linearly independent vectors and therefore at most $d+1$ affinely independent points.

An affine combination, $x = \sum \lambda_i u_i$, is a *convex combination* if all λ_i are non-negative. The *convex hull* is the set of convex combinations. A *k-simplex* is the convex hull of $k+1$ affinely independent points, $\sigma = \operatorname{conv}\{u_0, u_1, \ldots, u_k\}$. We sometimes say the u_i *span* σ. Its *dimension* is $\dim \sigma = k$. We use special names for the first few dimensions, *vertex* for 0-simplex, *edge* for 1-simplex, *triangle* for 2-simplex, and *tetrahedron* for 3-simplex; see Figure III.1. Any subset of

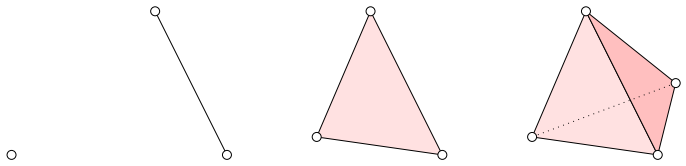

Figure III.1: From left to right: a vertex, an edge, a triangle, and a tetrahedron. We note that an edge has two vertices, a triangle has three edges, and a tetrahedron has four triangles as faces.

affinely independent points is again affinely independent and therefore also defines a simplex. A *face* of σ is the convex hull of a non-empty subset of the u_i and it is *proper* if the subset is not the entire set. We sometimes write $\tau \leq \sigma$ if τ is a face and $\tau < \sigma$ if it is a proper face of σ. If τ is a (proper) face of σ we call σ a *(proper) coface* of τ. Since a set of size $k+1$ has 2^{k+1} subsets, including the empty set, σ has $2^{k+1} - 1$ faces, all of which are proper except for σ itself. The *boundary* of σ, denoted as $\operatorname{bd}\sigma$, is the union of all proper faces, and the *interior* is everything else, $\operatorname{int}\sigma = \sigma - \operatorname{bd}\sigma$. A point $x \in \sigma$ belongs to $\operatorname{int}\sigma$ iff

all its coefficients λ_i are positive. It follows that every point $x \in \sigma$ belongs to the interior of exactly one face, namely the one spanned by the points u_i that correspond to positive coefficients λ_i.

Simplicial complexes. We are interested in sets of simplices that are closed under taking faces and that have no improper intersections.

DEFINITION. A *simplicial complex* is a finite collection of simplices K such that $\sigma \in K$ and $\tau \leq \sigma$ implies $\tau \in K$, and $\sigma, \sigma_0 \in K$ implies $\sigma \cap \sigma_0$ is either empty or a face of both.

The *dimension* of K is the maximum dimension of any of its simplices. The *underlying space*, denoted as $|K|$, is the union of its simplices together with the topology inherited from the ambient Euclidean space in which the simplices live. A *polyhedron* is the underlying space of a simplicial complex. A *triangulation* of a topological space $\mathbb{X}$ is a simplicial complex K together with a homeomorphism between $\mathbb{X}$ and $|K|$. The topological space is *triangulable* if it has a triangulation. A *subcomplex* of K is a simplicial complex $L \subseteq K$. It is *full* if it contains all simplices in K spanned by vertices in L. A particular subcomplex is the *j-skeleton* consisting of all simplices of dimension j or less, $K^{(j)} = \{\sigma \in K \mid \dim \sigma \leq j\}$. The 0-skeleton is also referred to as the *vertex set*, $\operatorname{Vert} K = K^{(0)}$. Skeleta are generally not full.

A subset of a simplicial complex useful in talking about local neighborhoods is the *star* of a simplex τ consisting of all cofaces of τ, $\operatorname{St} \tau = \{\sigma \in K \mid \tau \leq \sigma\}$. Generally, the star is not closed under taking faces. We can make it into a complex by adding all missing faces. The result is the *closed star*, $\overline{\operatorname{St}}\, \tau$, which is the smallest subcomplex that contains the star. The *link* consists of all simplices in the closed star that are disjoint from τ, $\operatorname{Lk} \tau = \{\upsilon \in \overline{\operatorname{St}}\, \tau \mid \upsilon \cap \tau = \emptyset\}$. It τ is a vertex then the link is just the difference between the closed star and the star. More generally, it is the closed star minus the stars of all faces of τ. For example if K triangulates a 2-manifold without boundary then the link of an edge is a pair of points, a 0-sphere, and the link of a vertex is a cycle of edges and vertices, a 1-sphere.

Abstract simplicial complex. It is often easier to construct a complex abstractly and to worry about how to put it into Euclidean space later, if at all.

DEFINITION. An *abstract simplicial complex* is a finite collection of sets A such that $\alpha \in A$ and $\beta \subseteq \alpha$ implies $\beta \in A$.

The sets in A are its *simplices*. The *dimension* of a simplex is $\dim \alpha = \operatorname{card} \alpha - 1$ and the dimension of the complex is the maximum dimension of any of its simplices. A *face* of α is a non-empty subset $\beta \subseteq \alpha$, which is *proper* if $\beta \neq \alpha$. The *vertex set* is the union of all simplices, $\operatorname{Vert} A = \bigcup A$, the collection of all α such that $\alpha \in A$ for some simplex A. A *subcomplex* is an abstract simplicial complex $B \subseteq A$. Two abstract simplicial complexes are *isomorphic* if there is a bijection $b : \operatorname{Vert} A \to \operatorname{Vert} B$ such that $\alpha \in A$ iff $b(\alpha) \in B$. The largest abstract simplicial complex with a vertex set of size $n+1$ is the n-dimensional simplex with a total number of $2^{n+1} - 1$ faces. Given a (geometric) simplicial complex K, we can construct an abstract simplicial complex A by throwing away all simplices and retaining only their sets of vertices. We call A a *vertex scheme* of K. Symmetrically, we call K a *geometric realization* of A. Constructing geometric realizations is surprisingly easy if the dimension of the ambient space is sufficiently high.

Geometric Realization Theorem. Every abstract simplicial complex of dimension d has a geometric realization in $\mathbb{R}^{2d+1}$.

Proof. Let $f : \operatorname{Vert} A \to \mathbb{R}^{2d+1}$ be an injection whose image is a set of points in general position. Specifically, any $2d+2$ or fewer of the points are affinely independent. Let α and α_0 be simplices in A with $k = \dim \alpha$ and $k_0 = \dim \alpha_0$. The union of the two has size $\operatorname{card}(\alpha \cup \alpha_0) = \operatorname{card} \alpha + \operatorname{card} \alpha_0 - \operatorname{card}(\alpha \cap \alpha_0) \leq k + k_0 + 2 \leq 2d + 2$. The points in $\alpha \cup \alpha_0$ are therefore affinely independent, which implies that every convex combination x of points in $\alpha \cup \alpha_0$ is unique. Hence, x belongs to $\sigma = \operatorname{conv} f(\alpha)$ as well as to $\sigma_0 = \operatorname{conv} f(\alpha_0)$ iff x is a convex combination of $\alpha \cap \alpha_0$. This implies that the intersection of σ and σ_0 is either empty of the simplex $\operatorname{conv} f(\alpha \cap \alpha_0)$, as required. ◻

Simplicial maps. The natural equivalent of continuous maps between topological spaces are simplicial maps between simplicial complexes, which we now introduce. Let K be a simplicial complex with vertices $u_0, u_1, \ldots, u_n$. Every point $x \in |K|$ belongs to the interior of exactly one simplex in K. Letting $\sigma = \operatorname{conv}\{u_0, u_1, \ldots, u_k\}$ be this simplex, we have $x = \sum_{i=0}^{k} \lambda_i u_i$ with $\sum_{i=0}^{k} \lambda_i = 1$ and $\lambda_i > 0$ for all i. Setting $b_i(x) = \lambda_i$ for $0 \leq i \leq k$ and $b_i(x) = 0$ for $k+1 \leq i \leq n$ we have $x = \sum_{i=0}^{n} b_i(x) u_i$ and we call the $b_i(x)$ the *barycentric coordinates* of x in K.

We use these coordinates to construct a piecewise linear, continuous map from a particular kind of map between the vertices of two simplicial complexes. A *vertex map* is a function $\varphi : \operatorname{Vert} K \to \operatorname{Vert} L$ with the property that the vertices of every simplex in K map to vertices of a simplex in L. Then φ can

be extended to a continuous map $f : |K| \to |L|$ defined by

$$f(x) \quad = \quad \sum_{i=0}^{n} b_i(x)\varphi(u_i),$$

the *simplicial map* induced by φ. There is an alternative way to think of this construction. Fix a vertex u_j and consider the map $b_j : |K| \to \mathbb{R}$ which maps each point x to its j-th barycentric coordinate. The graph of this map has the shape of a hat, increasing from zero on and outside the link to one at u_j. The map b_j is continuous and is sometimes referred to as a basis function. The simplicial map is thus the weighted sum of the $n + 1$ basis functions. To emphasize that the simplicial map is linear on every simplex, we usually drop the underlying space from the notation and write $f : K \to L$.

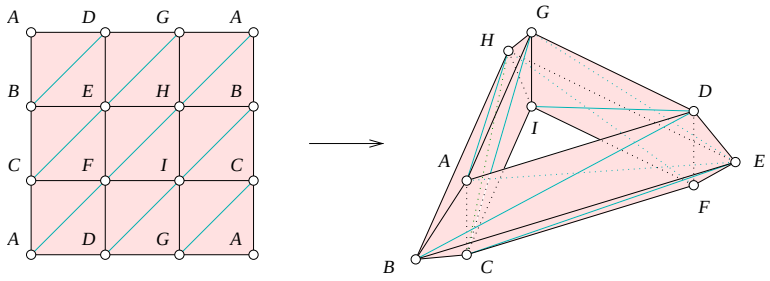

Figure III.2: A vertex map and its induced simplicial map from the square to the torus.

As an example, we consider the simplicial map $f : [0, 1]^2 \to \mathbb{T}^2$ illustrated in Figure III.2. Given the vertex map, the simplicial map is unique and glues the simplices of the triangulation of the square to obtain a triangulation of the torus. If the vertex map $\varphi : \operatorname{Vert} K \to \operatorname{Vert} L$ is bijective and $\varphi^{-1} : \operatorname{Vert} L \to \operatorname{Vert} K$ is also a vertex map then the induced simplicial map f is a homeomorphism. In this case we call f a *simplicial homeomorphism* or an *isomorphism* between K and L.

Subdivisions. A simplicial complex L is a *subdivision* of another simplicial complex K if $|L| = |K|$ and every simplex in L is contained in a simplex in K. There are many ways to construct subdivisions. A particular one is the *barycentric subdivision*, $L = \operatorname{Sd} K$, illustrated in Figure III.3. A crucial concept in its construction is the *barycenter* of a simplex, which is the average of its vertices. We proceed by induction over the dimension. To get started, the

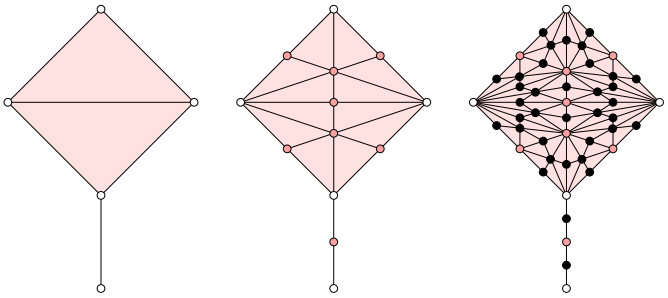

Figure III.3: Left: a simplicial complex consisting of two triangles, six edges, and five vertices. Middle and right: its first two barycentric subdivisions.

barycentric subdivision of the 0-skeleton is the same, $\mathrm{Sd}K^{(0)} = K^{(0)}$. Assuming we have the barycentric subdivision of $K^{(j-1)}$, we construct $\mathrm{Sd}K^{(j)}$ by adding the barycenter of every j-simplex as a new vertex and connecting it to the simplices that subdivide the boundary of the j-simplex.

The *diameter* of a set in Euclidean space is the supremum over the distances between its points. Since the simplices of K are point sets in Euclidean space, their diameters are well defined. The *mesh* of K is the maximum diameter of any simplex or, equivalently, the length of its longest edge.

Mesh Lemma. Letting δ be the mesh of the d-dimensional simplicial complex K, the mesh of $\mathrm{Sd}K$ is at most $\frac{d}{d+1}\delta$.

Proof. Let τ and υ be complementary faces of a simplex $\sigma \in K$, that is, $\tau \cap \upsilon = \emptyset$ and $\dim\tau + \dim\upsilon = \dim\sigma - 1$. The line segment connecting the barycenters of τ and υ has length at most δ, and it splits into two edges in $\mathrm{Sd}K$, in proportions $1 + \dim\upsilon$ to $1 + \dim\tau$. The fraction of length is therefore between $\frac{1}{k+1}$ and $\frac{k}{k+1}$, where $k = \dim\sigma$. Both edges have therefore length at most $\frac{k}{k+1} \leq \frac{d}{d+1}$ times δ. ◻

By the Mesh Lemma, we can make the diameters of the simplices as small as we like by iterating the subdivision operation. For $n \geq 1$, the *n-th barycentric subdivision* of K is $\mathrm{Sd}^n K = \mathrm{Sd}(\mathrm{Sd}^{n-1}K)$. As n goes to infinity, the mesh of $\mathrm{Sd}^n K$ goes to zero.

Simplicial approximations. It is sometimes convenient to think of a vertex star as an open set of points. Formally, we define $N(u) = \bigcup_{\sigma \in \mathrm{St}\, u} \mathrm{int}\,\sigma$. Let K

and L be simplicial complexes. A continuous map $g : |K| \to |L|$ satisfies the *star condition* if the image of every vertex star in K is contained in a vertex star in L, that is, for each vertex $u \in K$ there is a vertex $v \in L$ such that $g(N(u)) \subseteq N(v)$. Let $\varphi : \operatorname{Vert} K \to \operatorname{Vert} L$ map u to the vertex $\varphi(u) = v$ that exists by the star condition. To understand this new function, we take a point x in the interior of a simplex σ in K. Its image, $g(x)$, lies in the interior of a unique simplex τ in L. It follows that the star of every vertex u of σ maps into the star of a vertex v in L that contains the interior of τ. But this implies that v is a vertex of τ. We conclude that each vertex u of σ maps to a vertex $\varphi(u)$ of τ. Hence, φ is a vertex map and thus induces a simplicial map $f : K \to L$. This map satisfies the condition of an *simplicial approximation* of g, namely $g(N(u)) \subseteq N(f(u))$ for each vertex u of K.

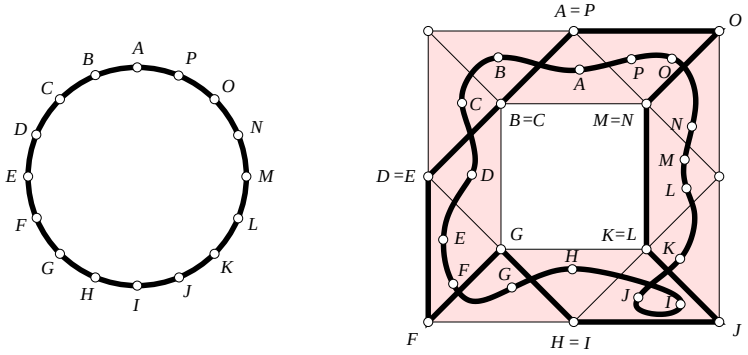

Figure III.4: The circle on the left is mapped into the closed annulus by a continuous map and a simplicial approximation of that map. Corresponding vertices are labeled by the same letter.

We illustrated the definitions in Figure III.4. The image we have in mind is that g and f are not too different. In particular, $g(x)$ and $f(x)$ belong to a common simplex in L for every $x \in |K|$. Given a continuous map $g : |K| \to |L|$, it is plausible that we can subdivide K sufficiently finely so that a simplicial approximation exists. To be sure we prove this fact.

Simplicial Approximation Theorem. If $g : |K| \to |L|$ is continuous then there is a sufficiently large integer n such that g has a simplicial approximation $f : \mathrm{Sd}^n K \to L$.

Proof. Cover $|K|$ with open sets of the form $g^{-1}(N(v))$, $v \in \operatorname{Vert} L$. Since $|K|$ is compact there is a positive real number λ such that any set of diameter less

than λ is contained in one of the sets in the open cover. Choose n such that each simplex in $\mathrm{Sd}^n K$ has diameter less than half of λ. Then each star in K has diameter less than λ implying it lies in one of the sets $g^{-1}(N(v))$. Hence g satisfies the star condition implying the existence of a simplicial approximation. □

Bibliographic notes. The terminology we use for abstract and geometric simplicial complexes follows the one in Munkres [3]. The geometric realization of a d-dimensional abstract simplicial complex in $\mathbb{R}^{2d+1}$ goes back to Karl Menger. We have seen that $2d+1$ dimensions suffice for the geometric realization of any d-dimensional abstract simplicial complex. Complexes that require that many dimensions have been described by Flores [1] and van Kampen [5]. An example of such a complex is the d-skeleton of the $(2d+2)$-simplex, which does not embed in $\mathbb{R}^{2d}$. For $d = 1$ this is the complete graph of five vertices, which does not embed in the plane as discussed in Chapter I.

A stronger version of the Simplicial Approximation Theorem played an important role in the development of combinatorial topology during the first half of the twentieth century. Known as the Hauptvermutung (German for "main conjecture"), it claimed that any two simplicial complexes that triangulate the same topological space have isomorphic subdivisions. This turned out to be correct for simplicial complexes of dimension 2 and 3 but not higher. The first counterexample found by Milnor was a simplicial complex of dimension 7 [2]. We refer to the book edited by Ranicki [4] for further information on the topic.

[1] A. Flores. Über n-dimensionale Komplexe die in $\mathbb{R}_{2n+1}$ selbstverschlungen sind. *Ergeb. Math. Koll.* **6** (1933/34), 4–7.

[2] J. Milnor. Two complexes which are homeomorphic but combinatorially distinct. *Ann. of Math.* **74** (1961), 575–590.

[3] J. R. Munkres. *Elements of Algebraic Topology.* Perseus, Cambridge, Massachusetts, 1984.

[4] A. A. Ranicki (editor). *The Hauptvermutung Book.* Kluwer, Dordrecht, the Netherlands, 1996.

[5] E. R. van Kampen. Komplexe in euklidischen Räumen. *Abh. Math. Sem. Univ. Hamburg* **9** (1933), 72-78.

III.2 Convex Set Systems

Simplicial complexes often arise as intersection patterns of collections of sets. We begin with two fundamental results for convex sets and then proceed to the special case in which the sets are geometric balls.

Sets with common points. Let F be a finite collection of convex sets in $\mathbb{R}^d$. The smaller the dimension of the ambient Euclidean space, d, the more restrictive are the intersection patterns we observe. For example, if $d = 1$ and we have three intervals that intersect in pairs then it is not possible that they do not intersect as a triplet. This result generalizes to higher dimensions.

Helly's Theorem. Let F be a finite collection of closed, convex sets in $\mathbb{R}^d$. Every $d+1$ of the sets have a non-empty common intersection iff they all have a non-empty common intersection.

Proof. We prove only the non-obvious direction, by induction over the dimension, d, and the number of sets, $n = \operatorname{card} F$. The implication is clearly true for $d = 1$ and all n as well as for $n = d+1$. Now suppose we have a minimal counterexample consisting of $n > d+1$ closed, convex sets in $\mathbb{R}^d$, which we denote as $X_1, X_2, \ldots, X_n$. By minimality of the counterexample, the set $Y_n = \bigcap_{i=1}^{n-1} X_i$ is non-empty and disjoint from X_n. Because Y_n and X_n are both closed and convex, we can find a $(d-1)$-dimensional plane h that separates and is disjoint from both sets, as in Figure III.5. Let F' be the collection of sets $Z_i = X_i \cap h$,

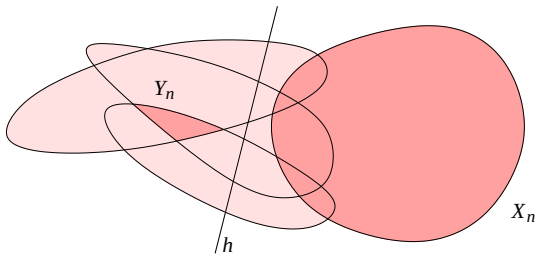

Figure III.5: The $(d-1)$-plane separates the n-th set from the common intersection of the first $n-1$ sets in F.

for $1 \leq i \leq n-1$, each a non-empty, closed, convex set in $\mathbb{R}^{d-1}$. By assumption, any d of the first $n-1$ sets X_i have a common intersection with X_n. It follows that the common intersection of the d sets contains points on both sides of h

implying that any d of the sets Z_i have a non-empty common intersection. By minimality of the counterexample, this implies $\bigcap F' \neq \emptyset$. This intersection is

$$\bigcap F' \quad = \quad \bigcap_{i=1}^{n-1} (X_i \cap h) \quad = \quad Y_n \cap h.$$

But this contradicts the choice of h as a $(d-1)$-plane disjoint from Y_n. ◻

Convexity is a convenient but unnecessarily strong requirement in Helly's Theorem. Indeed, the conclusion holds if the sets in F are closed and all their non-empty common intersections are contractible, a property we will define shortly.

Homotopy type. We prepare the next step by introducing a notion of equivalence between topological spaces that is weaker than topological equivalence. We begin by considering two continuous maps, $f, g : \mathbb{X} \to \mathbb{Y}$. A *homotopy* between f and g is another continuous map $H : \mathbb{X} \times [0,1] \to \mathbb{Y}$ that agrees with f for $t = 0$ and with g for $t = 1$, that is, $H(x,0) = f(x)$ and $H(x,1) = g(x)$ for all $x \in \mathbb{X}$. We may think of $t \in [0,1]$ as time and the homotopy as a time-series of functions $f_t : \mathbb{X} \to \mathbb{Y}$ defined by $f_t(x) = H(x,t)$. It starts at $f_0 = f$ and ends at $f_1 = g$. Noting that this defines an equivalence relation, we write $f \simeq g$ and call f and g if there is a homotopy between them.

This notion can be used to relate spaces. Beginning with a special case, we call $\mathbb{Y} \subseteq \mathbb{X}$ a *retract* of $\mathbb{X}$ if there is a continuous map $r : \mathbb{X} \to \mathbb{Y}$ with $r(y) = y$ for all $y \in \mathbb{Y}$. The map r is called a *retraction*. We call $\mathbb{Y}$ a *deformation retract* and r a *deformation retraction* if there is a homotopy between r and the identity on $\mathbb{X}$, $r \simeq \mathrm{id}_{\mathbb{X}}$. Clearly, every deformation retract is a retract but not the other way round. For example, a connected interval on the circle is a retract but not a deformation retract of $\mathbb{S}^1$. An insubstantial generalization of the notion of deformation retract is obtained by considering maps in both directions. Specifically, we call two not necessarily nested topological spaces, $\mathbb{X}$ and $\mathbb{Y}$, *homotopy equivalent* if there are continuous maps $f : \mathbb{X} \to \mathbb{Y}$ and $g : \mathbb{Y} \to \mathbb{X}$ such that $g \circ f \simeq \mathrm{id}_{\mathbb{X}}$ and $f \circ g \simeq \mathrm{id}_{\mathbb{Y}}$. This gives an equivalence relation and we write $\mathbb{X} \simeq \mathbb{Y}$ and say they have the same *homotopy type* if they are homotopy equivalent. The maps f and g are sometimes referred to as *homotopy equivalences* and as *homotopy inverses* of each other.

To see that having the same homotopy type indeed generalizes being a deformation retract we note if $r : \mathbb{X} \to \mathbb{Y}$ is a deformation retraction then $f = r$ and $g = \mathrm{id}_{\mathbb{Y}}$ are continuous maps that satisfy the conditions and thus establish $\mathbb{X} \simeq \mathbb{Y}$. If $\mathbb{Y}$ is a single point then $\mathbb{X}$ has the homotopy type of a point and we say $\mathbb{X}$ is *contractible*.

Nerves. We now return to our finite collection of sets, F. Without assuming the sets are convex, we define the *nerve* to consist of all non-empty subcollections whose sets have a non-empty common intersection,

$$\text{Nrv}\, F \quad = \quad \{X \subseteq F \mid \bigcap X \neq \emptyset\}.$$

It is always an abstract simplicial complex, no matter what sets we have in F. Indeed, if $\bigcap X \neq \emptyset$ and $Y \subseteq X$ then $\bigcap Y \neq \emptyset$. We can realize the nerve geometrically in some Euclidean space, so it makes sense to talk about its topology type and its homotopy type. We will sometimes do this without explicit construction of the geometric realization. As an example, consider the collection of four sets in Figure III.6 whose union is obviously not homotopy equivalent to the nerve. Nevertheless, taking the nerve preserves the homotopy type if the sets in the collection are convex. This is a fundamental result which we state formally but without proof.

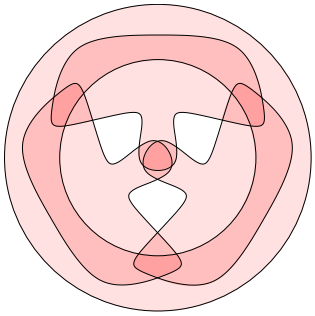

Figure III.6: A collection of four sets whose union is a disk with three holes in the plane. The nerve is the boundary complex of the tetrahedron which has the homotopy type of a sphere.

Nerve Theorem. Let F be a finite collection of closed, convex sets in Euclidean space. Then the nerve of F and the union of the sets in F have the same homotopy type.

Similar to Helly's Theorem, the requirement on the sets can be relaxed without sacrificing the conclusion. Specifically, if $\bigcup F$ is triangulable, all sets in F are closed, and all non-empty common intersections are contractible then $\text{Nrv}\, F \simeq \bigcup F$. We note that Helly's Theorem can be interpreted as a constraint on the structure of the nerve. Specifically, if the sets live in $\mathbb{R}^d$ then a subcollection of $k \geq d+1$ sets cannot have all $\binom{k}{d+1}$ d-simplices in the nerve without having the entire k-simplex in the nerve.

Čech complexes. We now consider the special case in which the convex sets are closed geometric balls, all of the same radius, r. Let S be a finite set of points in $\mathbb{R}^d$ and write $B_x(r) = x + r\mathbb{B}^d$ for the closed ball with center x and radius r. The *Čech complex* of S and r is isomorphic to the nerve of this collection of balls,

$$\check{\text{C}}\text{ech}(r) \quad = \quad \{\sigma \subseteq S \mid \bigcap_{x \in \sigma} B_x(r) \neq \emptyset\}.$$

Clearly, a set of balls has a non-empty intersection iff their centers lies inside a common ball of the same radius. Indeed, a point y belongs to all balls iff $\|x - y\| \leq r$ for all centers x. An easy consequence of Helly's Theorem is therefore that every $d+1$ points in S are contained in a common ball of radius r iff all points in S are. This is Jung's Theorem which predates the more general theorem by Helly. The Čech complex does not necessarily have a geometric realization in $\mathbb{R}^d$ but it is fine as an abstract simplicial complex; see Figure III.7. For larger radius, the disks are bigger and create more overlaps while

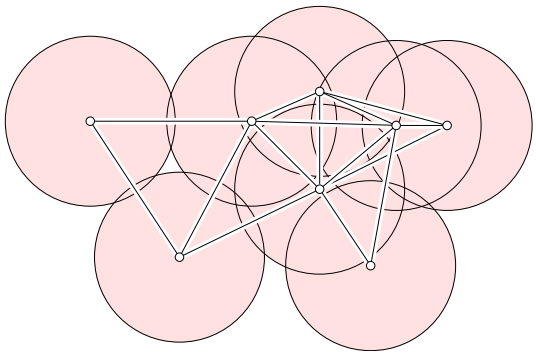

Figure III.7: Nine points with pairwise intersections among the disks indicated by straight edges connecting their centers. The Čech complex fills nine of the ten possible triangles as well as the two tetrahedra. The only difference between the Vietoris-Rips and the Čech complexes is the tenth triangle, which belongs only to the former.

retaining the ones for smaller radius. Hence $\check{\text{C}}\text{ech}(r_0) \subseteq \check{\text{C}}\text{ech}(r)$ whenever $r_0 \leq r$. If we continuously increase the radius, from 0 to ∞, we get a discrete family of nested Čech complexes. We will come back to this construction later.

Smallest enclosing balls. Beyonds sets of two points it seems cumbersome to recognize the ones that form simplices in the Čech complex. Nevertheless, there is a fast algorithm for the purpose.

Let $\sigma \subseteq S$ be a subset of the given points. We have seen that deciding whether or not σ belongs to Čech(r) is equivalent to deciding whether or not σ fits inside a ball of radius r. Let the *miniball* of σ be the smallest closed ball that contains σ, which we note is unique. The radius of the miniball is smaller than or equal to r iff $\sigma \in$ Čech(r), so finding it solves our problem. Observe that the miniball is already determined by a subset of $k+1 \leq d+1$ of the points, which all lie on its boundary. If we know this subset then we can verify the miniball by testing that it indeed contains all the other points. In a situation in which we have many more points than dimensions, the chance that a point belongs to this subset is small and discarding it is easy. This is the strategy of the Miniball Algorithm. It takes two disjoint subsets τ and υ of σ and returns the miniball that contains all points of τ in its interior and all points of υ on its boundary. To get the miniball of σ we call MiniBall$(\sigma, \emptyset)$.

```
ball MiniBall(τ, υ)
  if τ = ∅ then compute the miniball B of υ directly
           else choose a random point u ∈ τ;
                B = MiniBall(τ − {u}, υ);
                if u ∉ B then
                   B = MiniBall(τ − {u}, υ ∪ {u})
                endif
  endif; return B.
```

When τ is empty, we have a set υ of at most $d+1$ points, which we know all lie on the boundary. Assuming the dimension, d, is a constant, we can compute their miniball directly and in constant time. To analyze the running time, we ask how often we execute the test "$u \notin B$". Let $t_j(n)$ be the expected number of such tests for calling MiniBall with n points in τ and $j = d + 1 - \operatorname{card} \upsilon$ possibly open positions on the boundary of the miniball. Obviously, $t_j(0) = 0$, and it is reassuring that the constant amount of work needed to compute the ball for the at most $d+1$ points in υ is payed for by the test that initiated the call. Consider $n > 0$. We have one call with parameters $n-1$ and j, one test "$u \notin B$", and one call with parameters $n-1$ and $j-1$. The probability that the second call indeed happens is at most j out of n. Hence,

$$t_j(n) \quad \leq \quad t_j(n-1) + 1 + \frac{j}{n} \cdot t_{j-1}(n-1).$$

Setting $j = 0$ we get $t_0(n) \leq t_0(n-1) + 1$ and therefore $t_0(n) \leq n$. Similarly, $t_1(n) \leq t_1(n-1) + 2 \leq 2n$. More generally, we get $t_j(n) \leq (j+1)!n$, which is a constant times n since $j \leq d+1$ is a constant. In summary, for constant dimension the algorithm takes expected constant time per point.

Vietoris-Rips complexes. Instead of checking all subcollections, we may just check pairs and add 2- and higher-dimensional simplices whenever we can. This simplification leads to the *Vietoris-Rips complex* of S and r consisting of all subsets of diameter at most $2r$,

$$\text{Vietoris-Rips}(r) \quad = \quad \{\sigma \subseteq S \mid \operatorname{diam} \sigma \leq 2r\}.$$

Clearly, the edges in the Vietoris-Rips complex are the same as in the Čech complex. Furthermore, $\check{\text{C}}\text{ech}(r) \subseteq \text{Vietoris-Rips}(r)$ because the latter contains every simplex warranted by the given edges. We now prove that the containment relation can be reversed if we are willing to increase the radius of the Čech complex by a multiplicative constant.

Vietoris-Rips Lemma. Letting S be a finite set of points in some Euclidean space and $r \geq 0$, we have $\text{Vietoris-Rips}(r) \subseteq \check{\text{C}}\text{ech}(\sqrt{2}r)$.

Proof. A simplex is *regular* if all its edges have the same length. A convenient representation for dimension d is the *standard d-simplex*, Δ^d, spanned by the endpoints of the unit coordinate vectors in $\mathbb{R}^{d+1}$; see Figure III.8. Each edge of

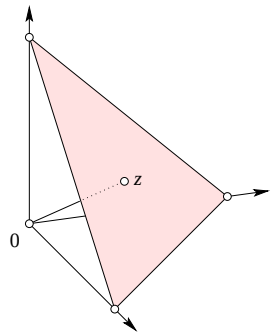

Figure III.8: The standard triangle connecting the unit coordinate vectors in $\mathbb{R}^3$.

Δ^d has length $\sqrt{2}$. By symmetry, the distance of the origin from the standard simplex is its distance from the barycenter, the point z whose $d+1$ coordinates are all equal to $\frac{1}{d+1}$. That distance is therefore $\|z\| = 1/\sqrt{d+1}$. The barycenter is also the center of the smallest d-sphere that passes through the vertices of Δ^d. Writing r_d for the radius of that sphere, we have $r_d^2 = 1 - \|z\|^2 = \frac{d}{d+1}$. For dimension 1, this is indeed half the length of the interval, and for dimension 2, it is the radius of the equilateral triangle. As the dimension goes to infinity, the radius grows and approaches 1 from below. Any set of $d+1$ or fewer points

for which the same d-ball of radius r_d is the miniball has a pair at distance $\sqrt{2}$ or larger. It follows that every simplex of diameter $\sqrt{2}$ or less belongs to Čech(r_d). Multiplying with $\sqrt{2}r$ we get Vietoris-Rips$(r) \subseteq$ Čech$(\sqrt{2}rr_d)$. Since $r_d \leq 1$ for all d, the latter is a subcomplex of Čech$(\sqrt{2}r)$, which implies the claimed subcomplex relationship. □

Bibliographic notes. Helly proved his theorem at the beginning of last century, first for convex sets and then for sets with contractible common intersections [4, 5]. The concept of nerve has been introduced at about the same time by Alexandrov [1]. The Nerve Theorem goes back to Borsuk [2], Leray [6], and others. It has a complicated literature, with version differing in the generality of the assumption and the strength of the conclusion. The Čech complexes are inspired by the theory of Čech homology, from which they borrow their name. The Vietoris-Rips complex appears in Vietoris [7] and in later work by Rips; see [3]. Algorithms for finding the smallest ball enclosing a finite set of points have been studied in computational geometry, culminating in the randomized minidisk algorithm of Welzl which has versions that are efficient even for large sets in high dimensions [8].

[1] P. S. Alexandrov. Über den allgemeinen Dimensionsbegriff und seine Beziehungen zur elementaren geometrischen Anschauung. *Math. Ann.* **98** (1928), 617–635.

[2] K. Borsuk. On the imbedding of systems of compacta in simplicial complexes. *Fund. Math.* **35** (1948), 217–234.

[3] M. Gromov. *Hyperbolic groups.* In *Essays in Group Theory*, MSRI Publ. **8**, Springer-Verlag, 1987.

[4] E. Helly. Über Mengen konvexer Körper mit gemeinschaftlichen Punkten. *Jahresber. Deutsch. Math.-Verein.* **32** (1923), 175–176.

[5] E. Helly. Über Systeme von abgeschlossenen Mengen mit gemeinschaftlichen Punkten. *Monatsh. Math. Physik* **37** (1930), 281–302.

[6] J. Leray. Sur la forme des espaces topologiques et sur les points fixes des représentations. *J. Math. Pures Appl.* **24** (1945), 95–167.

[7] L. Vietoris. Über den höheren Zusammenhang kompakter Räume and eine Klasse von zusammenhangstreuen Abbildungen. *Math. Ann.* **97** (1927), 454–472.

[8] E. Welzl. Smallest enclosing disks (balls and ellipsoids). In *New Results and New Trends in Computer Science*, H. A. Maurer (ed.), Springer-Verlag, Lecture Notes in Computer Science **555** (1991), 359–370.

III.3 Delaunay Complexes

In this section, we introduce a geometric constructions that limits the dimension of the simplices we get from a nerve. The main new structures are the Voronoi diagram and the Delaunay complex of a finite set of points. We begin by studying the inversion of space.

Inversion. Recall that $\mathbb{S}^d$ is the d-dimensional sphere with center at the origin and unit radius in $\mathbb{R}^{d+1}$. To invert $\mathbb{R}^{d+1}$, we map each point $x \neq 0$ to the point on the same half-line whose distance from the origin is one over the distance of x from 0. More formally, the *inversion* maps x to $\iota(x) = x/\|x\|^2$. It exchanges inside with outside and leaves points on $\mathbb{S}^d$ fixed. Clearly, $\iota(\iota(x)) = x$. We construct the image of a point x inside $\mathbb{S}^d$ by drawing right-angled triangles. First, we get a point $p \in \mathbb{S}^d$ such that $0xp$ has a right angle at x. Second, we choose x' on the half-line of x such that $0px'$ has a right angle at p. The angle at 0 is the same in both so the two triangles are similar. Hence, $\|x\| : \|p\| = \|p\| : \|x'\|$ which implies $\|x\|\|x'\| = \|p\|^2 = 1$ and thus $x' = \iota(x)$. We use this construction to show that the inversion maps spheres to spheres. We note, however, that it generally does not map centers to centers.

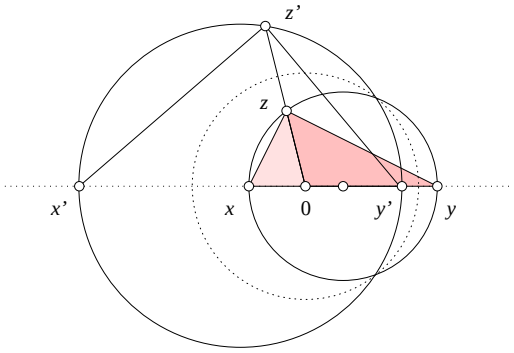

Figure III.9: As z sweeps out the circle passing through x and y, its image, $z' = \iota(z)$, sweeps out the circle passing through x' and y'.

Inversion Lemma. Let Σ be a d-sphere in $\mathbb{R}^{d+1}$. If $0 \notin \Sigma$ then $\iota(\Sigma)$ is a d-sphere and if $0 \in \Sigma$ then $\iota(\Sigma)$ is a d-plane.

Proof. Consider first the case in which Σ does not pass through the origin, as in Figure III.9. If 0 is the center of Σ then the result is obvious, so assume 0 is

not the center. Draw the line passing through 0 and the center; it intersects Σ in points x and y, which we invert to get points $x' = \iota(x)$ and $y' = \iota(y)$. Let z be another point on Σ and $z' = \iota(z)$ its inverse. Then $\|x\|\|x'\| = \|z\|\|z'\| = 1$ which implies that the triangles $0xz$ and $0z'x'$ are similar. By the same token, $0yz$ and $0z'y'$ are similar. But xyz has a right angle at z implying the angles at x' and y' inside $x'y'z'$ add up to a right angle. It follows that $x'y'z'$ has a right angle at z'. As z travels on Σ, the sphere with diameter xy, the image z' travels on $\iota(\Sigma)$, the sphere with diameter $x'y'$. What happens when Σ passes through the origin, say $0 = x$? Then the triangle $0y'z'$ has a right angle at y'. Equivalently, the image of Σ is the plane normal to the vector y and passing through the point y'. □

The Inversion Lemma suggests we think of a d-plane as a special kind of d-sphere, namely one that passes through the point at infinity.

Stereographic projection. The inversion can be defined relative to any center $z \in \mathbb{R}^{d+1}$ and any radius $r > 0$, that is, $\iota_{z,r}(x) = r \cdot \iota(\frac{x-z}{r}) + z$. It is not difficult to check that x and $x' = \iota_{z,r}(x)$ indeed lie on the same half-line emanating from z and the product of their distances is $\|x - z\|\|x' - z\| = r^2$, as desired. We consider the special case in which the center is the point $N =$

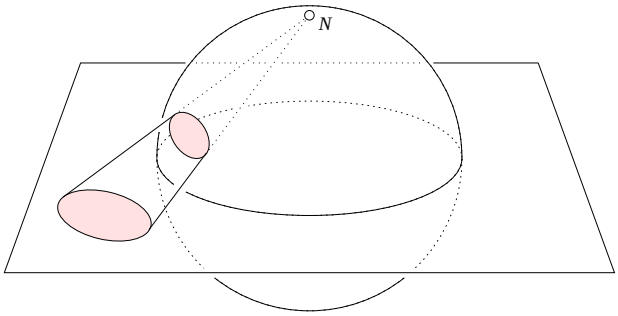

Figure III.10: The stereographic projection maps a circle on the unit sphere to a circle in the plane. If the circle on the sphere passes through the north-pole then its image is a line, that is, a circle that passes through the point at infinity.

$(0, \ldots, 0, 1)$, the north-pole of $\mathbb{S}^d$, and the radius is $r = \sqrt{2}$, the Euclidean distance between the north-pole and the equator. The image of $\mathbb{S}^d$ is the d-plane of points with vanishing $(d+1)$-st coordinates, which we denote as $\mathbb{R}^d$. The *stereographic projection* is the restriction of this particular inversion to the unit sphere, that is, $\varsigma : \mathbb{S}^d - \{N\} \to \mathbb{R}^d$ defined by $\varsigma(x) = \iota_{N,\sqrt{2}}(x)$, as

sketched in Figure III.10. Similar to the inversion, the stereographic projection preserves spheres.

Stereographic Projection Lemma. Let Σ' be a $(d-1)$-sphere on $\mathbb{S}^d$. If $N \notin \Sigma'$ then $\varsigma(\Sigma')$ is a $(d-1)$-sphere and if $N \in \Sigma'$ then $\varsigma(\Sigma')$ is a $(d-1)$-plane in $\mathbb{R}^d$.

Indeed, every $(d-1)$-sphere considered in the lemma is the intersection of $\mathbb{S}^d$ with another d-sphere. Its image is therefore the intersection of $\mathbb{R}^d$ with the image of the d-sphere, which is either a d-sphere or a d-plane. The intersection is thus either a $(d-1)$-sphere or a $(d-1)$-plane. As before, we consider a plane as a special sphere that passes through the point at infinity.

Voronoi diagram. We use the stereographic projection and the more general inversion to elucidate the construction of a particular simplicial complex from a finite set $S \subseteq \mathbb{R}^d$. The *Voronoi cell* of a point u in S is the set of points for which u is the closest, $V_u = \{x \in \mathbb{R}^d \mid \|x-u\| \leq \|x-v\|,\ v \in S\}$. It is the

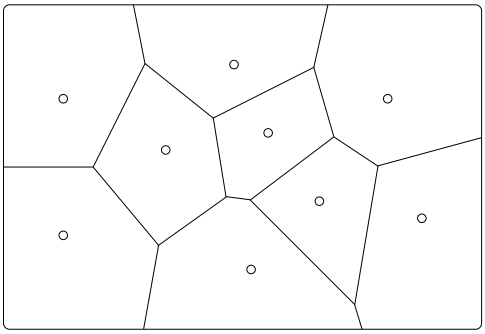

Figure III.11: The Voronoi diagram of nine points in the plane. By definition, each vertex of the diagram is equally far from the points that generate the incident Voronoi cells and further from all other points in S.

intersection of half-spaces of points at least as close to u as to v, over all points v in S. Hence, V_u is a convex polyhedron in $\mathbb{R}^d$. Any two Voronoi cells meet at most in a common piece of their boundary, and together the Voronoi cells cover the entire space, as illustrated in Figure III.11. The *Voronoi diagram* of S is the collection of Voronoi cells of its points.

We will shortly use a generalization of the concept to points u with real weights w_u. The *weighted squared distance*, or *power*, of a point $x \in \mathbb{R}^d$ from

u is then $\pi_u(x) = \|x - u\|^2 - w_u$. For positive weight, we can interpret the weighted point as the sphere with center u and square radius w_u. For a point x outside this sphere, the power is positive and equal to the square length of a tangent line segment from x to the sphere. For x on the sphere the power vanishes, and for x inside the sphere the power is negative. The *bisector* of two weighted points is the set of points with equal power from both. Just like in the unweighted case, the bisector is a plane normal to the line connecting the two points, except that is not necessarily halfway between them; see Figure III.12. Given a finite set of weighted points, we can thus define the *weighted*

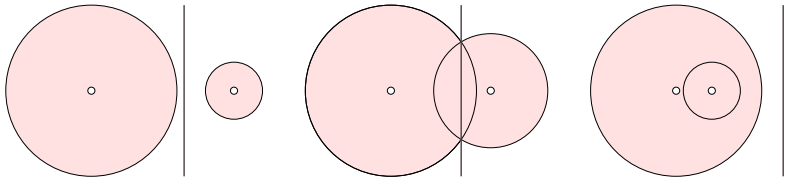

Figure III.12: The bisectors of pairs of weighted points. From left to right: two disjoint circles side by side, two intersecting circles, and two nested circles.

Voronoi cell, or *power cell*, of u as the set of points $x \in \mathbb{R}^d$ with $\pi_u(x) \leq \pi_v(x)$ for all weighted points v in the set. Finally, the *weighted Voronoi diagram*, or *power diagram*, is the set of power cells of the weighted points.

Lifting. We get a different and perhaps more illuminating view of the Voronoi diagram by lifting its cells to one higher dimension. Let S be a finite set of points in $\mathbb{R}^d$, as before, but draw them in $\mathbb{R}^{d+1}$, adding zeros as $(d+1)$-st coordinates. Map each point $u \in S$ to $\mathbb{S}^d$ using the inverse of the stereographic projection, and let Π_u be the d-plane tangent to $\mathbb{S}^d$ touching the sphere in the point $\varsigma^{-1}(u)$, as illustrated in Figure III.13. Using inversion, we now map each d-plane Π_u to the d-sphere $\Sigma_u = \iota(\Pi_u)$. It passes through the north-pole and is tangent to $\mathbb{R}^d$, the preimage of $\mathbb{S}^d$. The arrangements of planes and of spheres are closely related to the Voronoi diagram. We focus on the spheres first.

First Sphere Lemma. A point $x \in \mathbb{R}^d$ belongs to the Voronoi cell of $u \in S$ iff the first intersection of the directed line segment from x to N is with the d-sphere Σ_u.

Proof. Interpret the sphere Σ_u as a weighted point, namely its center with weight equal to the square of its radius. The power of a point x is the squared length of a tangent line segment, which is equal to $\|x - u\|^2$ if $x \in \mathbb{R}^d$. It

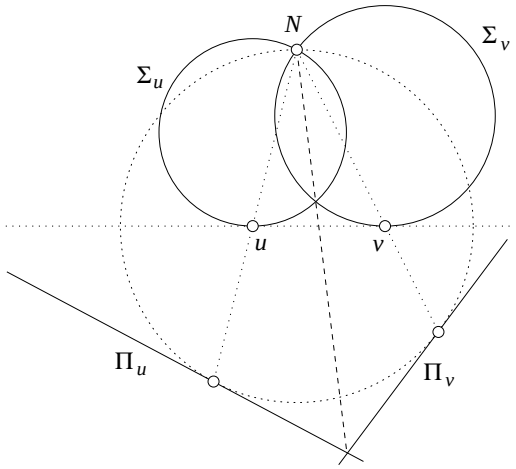

Figure III.13: We map the points u and v in $\mathbb{R}^1$ to the lines Π_u and Π_v tangent to $\mathbb{S}^1$ and further to the circles Σ_u and Σ_v passing through N and tangent to $\mathbb{R}^1$. The dashed line connecting N and the midpoint between u and v passes through the intersection of the two circles and the intersection of the two lines.

follows that the weighted Voronoi cell of the weighted center intersect $\mathbb{R}^d$ in the Voronoi cell of u. The claim follows because all bisectors of the weighted points pass through N. ◻

Switching from spheres to planes we get a similar characterization of the Voronoi diagram in terms of tangent planes.

First Plane Lemma. A point $x \in \mathbb{R}^d$ belongs to the Voronoi cell of $u \in S$ iff the first intersection of the directed line segment from N to x is with the d-plane Π_u.

Delaunay triangulation. The *Delaunay complex* of a finite set $S \subseteq \mathbb{R}^d$ is isomorphic to the nerve of the Voronoi diagram,

$$\text{Delaunay} \quad = \quad \{\sigma \subseteq S \mid \bigcap_{u \in \sigma} V_u \neq \emptyset\}.$$

We say the set S is in general position if no $d+2$ of the points lie on a common $(d-1)$-sphere. This assumption implies that no $d+2$ Voronoi cells have a non-empty common intersection. Equivalently, the dimension of any simplex

in the Delaunay complex is at most d. Assuming general position, we get a geometric realization by taking convex hulls of abstract simplices, as in Figure III.14. The result is often referred to as the *Delaunay triangulation* of S. To

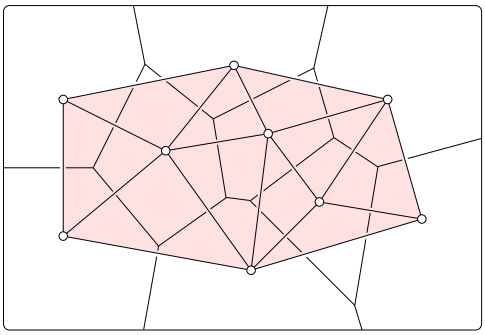

Figure III.14: The Delaunay triangulation superimposed on the Voronoi diagram. No four of the given points are cocircular implying the Delaunay complex has simplices of dimension at most two and a canonical geometric realization in $\mathbb{R}^2$.

see that this construction gives indeed a geometric realization of the Delaunay complex, we lift the points to the set $\varsigma^{-1}(S)$ on $\mathbb{S}^d$. Similarly, we lift a general point $x \in \mathbb{R}^d$ to the d-plane Π_x tangent to $\mathbb{S}^d$ at the point $\varsigma^{-1}(x)$. Keeping the same normal direction, we move this plane toward N. This corresponds to growing a $(d-1)$-sphere around x. The first point encountered by the plane corresponds to the first point encountered by the sphere, which is therefore the nearest to x. This suggests we add N to the set of lifted points and we take the convex hull in $\mathbb{R}^{d+1}$. The boundary of the resulting convex polytope consists of faces up to dimension d, some of which share N as a vertex. We are interested in the other faces, since they are spanned by points that correspond to Voronoi cells with a non-empty common intersection. Using central projection from N, we map these faces to $\mathbb{R}^d$. By convexity of the polytope, the images of the faces have no improper intersections. Indeed, we get the geometric realization of the Delaunay complex, as promised.

Similar to the Voronoi diagram, we can generalize the Delaunay complex to a finite set of points with real weights. Specifically, the *weighted Delaunay complex* is the abstract simplicial complex that contains a subset of the weighted points iff their weighted Voronoi cells have a non-empty common intersection. In contrast to the unweighted case, the cell of a weighted point can be empty, a difference that is sometimes overlooked. As a consequence, the vertex set of the weighted Delaunay triangulation is a subset and not necessarily the entire set

of given weighted points. Assuming general position, this complex can again be geometrically realized by taking convex hulls of the abstract simplices. The appropriate notion of general position is that no point of $\mathbb{R}^d$ has the same power from more than $d+1$ of the weighted points. This property is satisfied with probability one, a necessary requirement for a general position assumption.

Bibliographic notes. Voronoi diagrams are named after Georgy Voronoi [4] and Delaunay triangulations after Boris Delaunay (also Delone) [2]. Both structures have been studied centuries earlier by others, including Dirichlet, Gauß, and Descartes. Weighted Voronoi diagrams are perhaps as old as the unweighted ones and are known under a plethora of different names, including Thiessen polygons, Dirichlet tessellations, and power diagrams; see the survey article by Aurenhammer [1]. Their dual weighted Delaunay triangulations are also known under a variety of names, including regular triangulations and coherent triangulations; see e.g. [3].

[1] F. Aurenhammer. Voronoi diagrams — a study of a fundamental geometric data structure. *ACM Comput. Surveys* **23** (1991), 345–405.

[2] B. Delaunay. Sur la sphère vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk* **7** (1934), 793–800.

[3] I. M. Gelfand, M. M. Kapranov and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants.* Birkhäuser, Boston, Massachusetts, 1994.

[4] G. Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *J. Reine Angew. Math.* **133** (1907), 97–178, and **134** (1908), 198–287.

III.4 Alpha Complexes

In this section, we use a radius constraint to introduce a family of subcomplexes of the Delaunay complex. These complexes are similar to the Čech complexes but differ from them by having canonical geometric realizations.

Union of balls. Let S be a finite set of points in $\mathbb{R}^d$ and r a non-negative real number. For each $u \in S$, we let $B_u(r) = u + r\mathbb{B}^d$ be the closed ball with center u and radius r. The union of these balls is the set of points at distance at most r from at least one of the points in S. To decompose the union, we intersect each ball with the corresponding Voronoi cell, $R_u(r) = B_u(r) \cap V_u$. Since balls and Voronoi cells are convex, the $R_u(r)$ are also convex. Any two of them are disjoint or overlap along a common piece of their boundaries, and together the $R_u(r)$ cover the entire union, as in Figure III.15. The *alpha complex* is isomorphic to the nerve of this cover,

$$\text{Alpha}(r) \quad = \quad \{\sigma \subseteq S \mid \bigcap_{u \in \sigma} R_u(r) \neq \emptyset\}.$$

Since $R_u(r) \subseteq V_u$, the alpha complex is a subcomplex of the Delaunay complex.

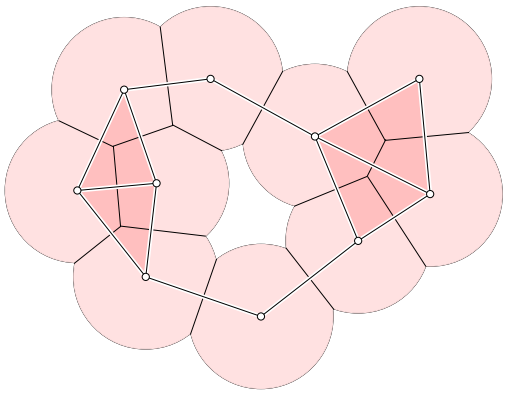

Figure III.15: The union of disks is decomposed into convex regions by the Voronoi diagram. The corresponding alpha complex is superimposed.

It follows that for a set S in general position, we get a geometric realization by taking convex hulls of abstract simplices, same as in the previous section. Furthermore, $R_u(r) \subseteq B_u(r)$ which implies $\text{Alpha}(r) \subseteq \check{\text{C}}\text{ech}(r)$. Since the $R_u(r)$ are closed and convex and together they cover the union, the Nerve

Theorem implies that the union of balls and Alpha(r) have the same homotopy type, $\bigcup_{u \in S} B_u(r) \simeq |\text{Alpha}(r)|$.

Weighted alpha complexes. For many applications, it is useful to permit balls with different sizes. An example of significant importance is the modeling of biomolecules, such as proteins, RNA, and DNA. Each atom is represented by a ball whose radius reflects the range of its van der Waals interactions and thus depends on the atom type. Let therefore S be a finite set of points u with real weights w_u. Same as in the previous section, we think of u as a ball B_u with center u and squared radius $r_u^2 = w_u$. We again consider the union of the balls, which we decompose into convex regions, now using weighted Voronoi cells, $R_u = B_u \cap V_u$. This is illustrated in Figure III.16. In analogy to

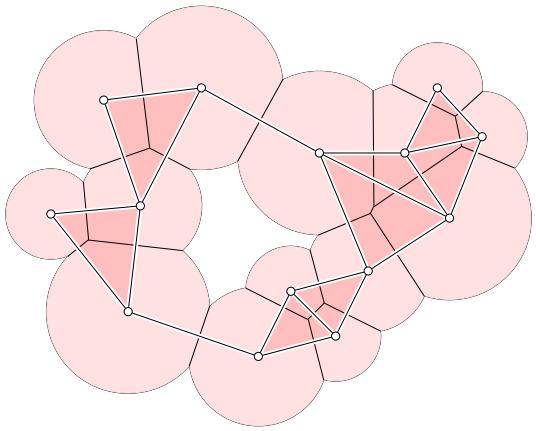

Figure III.16: Convex decomposition of a union of disks and the weighted alpha complex superimposed.

the unweighted case, the *weighted alpha complex* of S is isomorphic to the nerve of the regions R_u, that is, the set of abstract simplices $\sigma \subseteq S$ such that $\bigcap_{u \in \sigma} R_u \neq \emptyset$. The weighted alpha complex is a subcomplex of the weighted Delaunay complex. Assuming the weighted points are in general position, we get again a geometric realization by taking convex hulls of abstract simplices. It will be convenient to blur the difference, which we do by using the exact same notation and dropping the term weighted unless it is essential.

Filtration. There is a free parameter, r, which we may vary to get smaller and larger unions and smaller and larger alpha complexes. Sometimes, there is

a best choice of r but more often it does not exist. Indeed, the more interesting object is the family of alpha complexes, since it represents the data at different scale levels, if you will, and it allows us to draw conclusions from comparisons between different complexes in the same family.

We first explain the construction in the relatively straightforward unweighted case. Given a finite set $S \subseteq \mathbb{R}^d$, we continuously increase the radius and thus get a 1-parameter family of nested unions. Correspondingly, we get a 1-parameter family of nested alpha complexes, but because they are all subcomplexes of the same Delaunay complex, which is finite, only finitely many them are distinct. Writing K_i for the i-th alpha complex in this sequence, we get

$$\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_m,$$

which we call a *filtration* of $K_m = \text{Delaunay}$. What we have here is a stepwise assembly of the final complex in such a way that every set along the way is a simplicial complex.

There is more than one way to generalize this construction to the weighted case. For example, we could grow the corresponding balls uniformly. Starting with B_u, which has radius $\sqrt{w_u}$, we would increase the radius to $\sqrt{w_u} + r$ for $r > 0$. This makes sense in many applications, including the modeling of biomolecules, but has the complicating side effect that the Voronoi diagram of the set of balls for different values of r are not necessarily the same. Hence, the resulting alpha complexes are not necessarily nested. Instead, we let $B_u(r)$ be the ball with center u and squared radius $w_u + r^2$. The points x with equal power from $B_u(r)$ and $B_v(r)$ satisfy $\|x-u\|^2 - (w_u + r^2) = \|x-v\|^2 - (w_v + r^2)$. The squared radius cancels, implying that the same points x form the bisector for all choices of r. Hence, the union of balls are decomposed into convex sets by the same weighted Voronoi diagram for any r. Similarly, the weighted alpha complexes are all subcomplexes of the weighted Delaunay triangulation of the given points. More specifically, the alpha complex for r_0 is a subcomplex of that for r whenever $r_0 \leq r$ and we get again a filtration that starts with the empty complex and ends with the entire weighted Delaunay complex, same as in the unweighted case.

The structure of a simplex. We are interested in the difference between two contiguous complexes in the filtration, $K_{i+1} - K_i$. For this purpose, we study the structure of an abstract simplex, and not just because it arises as element of the alpha complex. Recall that an abstract d-simplex, α, is a set of $d+1$ points. It has 2^{d+1} subsets, including the empty set and α itself. In the *Hasse diagram* of this set system, we draw a node for each subset of α and

an arc for each subset relation, avoiding arcs that are implied by transitivity. Drawing the containing sets above the contained ones and keeping subsets of same cardinality in common rows, we get a picture like in Figure III.17. It looks

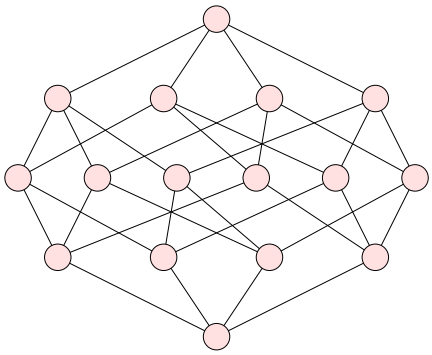

Figure III.17: The Hasse diagram of an abstract 3-simplex. From top to bottom, the rows of nodes contain the 3-simplex, the four 2-simplices, the six 1-simplices, the four 0-simplices, and the empty set.

like the edge-skeleton of the $(d+1)$-dimensional cube, and not by coincidence. Indeed, we can construct the Hasse diagram inductively, first drawing the Hasse diagram of a $(d-1)$-face. By inductive assumption, this is the edge-skeleton of a d-cube. When we add the last point, u_d, to the simplex, we get a new set $\beta \cup \{u_d\}$ for each old set β. To update the Hasse diagram, we add a second copy of the d-cube and connect corresponding sets. This is precisely the recipe for drawing the $(d+1)$-cube.

Another useful method constructs the Hasse diagram one pair of adjacent nodes at a time. We describe this in the other direction, disassembling the diagram one pair at a time. Specifically, we allowed ourselves to remove a pair $\beta_0 \subset \beta$ if β is the only remaining set that properly contains β_0. Note that β is necessarily maximal and we have $\dim \beta_0 = \dim \beta - 1$ because the operation maintains the system as an abstract simplicial complex. It is easy to see that disassembling the Hasse diagram of the d-simplex this way is possible, for example by removing the pairs $\beta_0 \subset \beta_0 \cup \{u_d\}$ in the order of decreasing dimension of β_0.

Collapses. Now suppose we have a geometric d-simplex, σ, and we consider the Hasse diagram of its system of faces, to which we add the empty set to be consistent with before. The operation of removing a pair $\beta_0 \subset \beta$ corresponds to

removing a pair of faces $\tau_0 < \tau$. The condition is that τ is the only remaining proper coface of τ_0. The operation of removing the pair $\tau_0 < \tau$ is referred to as an *elementary collapse*, or a $(k, k+1)$-*collapse* when $k = \dim \tau_0$. As illustrated in Figure III.18, a d-simplex can be reduced to a single simplex by a sequence of $2^d - 1$ elementary collapses. Since the elementary collapses maintain the set as a simplicial complex, the remaining simplex is necessarily a vertex. We can apply elementary collapses more generally to any simplicial complex, K, and not just that consisting of all faces of a simplex. Letting L be the result of the collapse, we note that there is a deformation retraction from $|K|$ to $|L|$. This implies that K and L have the same homotopy type. We call K *collapsible* if there is a sequence of elementary collapses that reduces K to a single vertex. Since collapses preserve the homotopy type, this is only possible if $|K|$ is contractible. As it turns out, not every simplicial complex with contractible underlying space is also collapsible.

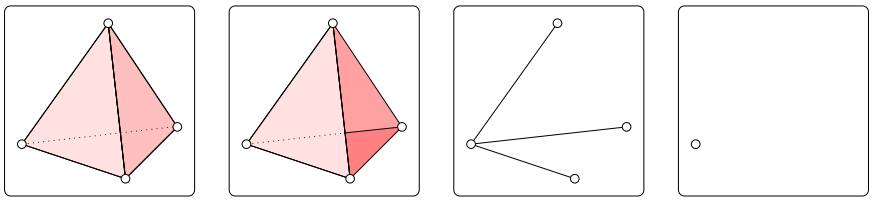

Figure III.18: From left to right: a tetrahedron, the three triangles left after a $(2, 3)$-collapse, the three edges left after three additional $(1, 2)$-collapses, and the vertex left after three additional $(0, 1)$-collapses.

It is convenient to extend the notion of collapse and consider pairs of simplices $\tau < \upsilon$ whose dimensions differ by one or more. Instead of requiring that υ is the only proper coface of τ, we now require that all cofaces of τ are faces of υ. Letting $k = \dim \tau$ and $\ell = \dim \upsilon$, we get $\binom{\ell-k}{i}$ simplices of dimension $i + k$ and therefore a total of $2^{\ell-k} = \sum_{i=0}^{\ell-k} \binom{\ell-k}{i}$ simplices between τ and υ, including the two. The Hasse diagram of this set of faces has the structure of an $(\ell - k - 1)$-simplex, which we have seen can be collapsed down to a vertex by a sequence of $2^{\ell-k-1} - 1$ elementary collapses. Each $(i, i+1)$-collapse in this sequence corresponds to an $(i + k + 1, i + k + 2)$-collapse in the sequence that removes the cofaces of τ. We append a $(k, k+1)$-collapse which finally removes τ together with the last remaining proper coface. We refer to this sequence of $2^{\ell-k-1}$ elementary collapses as a (k, ℓ)-*collapse*. Since elementary collapses preserve the homotopy type so do the more general collapses.

Critical and regular events. Let r_i be the smallest radius such that $K_i = \text{Alpha}(r_i)$. A simplex τ belongs to K_{i+1} but not to K_i if the balls with radius r_{i+1} have a non-empty common intersection with the corresponding intersection of Voronoi cells but the balls with radius r_i do not; see Figure III.19. Assuming general position and $\dim \tau = k$, the intersection of Voronoi cells, $V_\tau = \bigcap_{u \in \tau} V_u$, is a convex polyhedron of dimension $d - k$. By definition of r_{i+1}, the balls $B_u(r_{i+1})$ intersect V_τ in a single point, x.

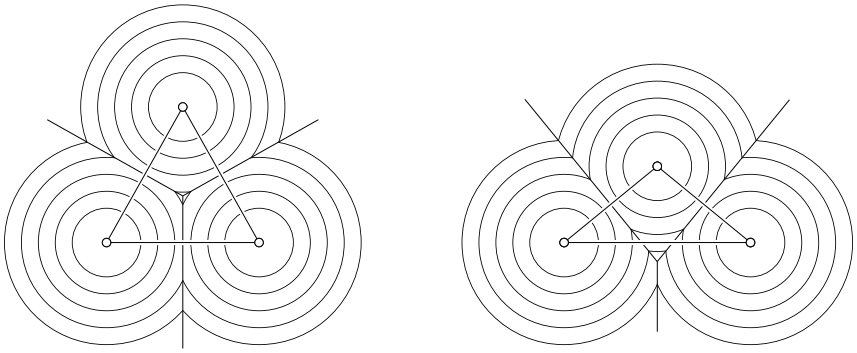

Figure III.19: Left: three points spanning an acute triangle. In the alpha complex evolution, the three edges appear before a critical event adds the triangle. Right: three points spanning an obtuse triangle. Two edges appear before a regular event adds the triangle together with the third edge.

Consider first the case that x lies on the boundary of V_τ. Then there are other Voronoi polyhedra for which x is the first contact with the union of balls. Assume V_τ is the polyhedron with highest dimension in this collection and let V_υ be the polyhedron with lowest dimension. Correspondingly, τ is the simplex with lowest dimension in $K_{i+1} - K_i$ and υ is the simplex with highest dimension. The other simplices in $K_{i+1} - K_i$ are the faces of υ that are cofaces of τ. In other words, we obtain K_i from K_{i+1} by a (k, ℓ)-collapse, where $k = \dim \tau$ and $\ell = \dim \upsilon$. We call this collapse a *regular event* in the evolution of the alpha complex.

Consider second the case that x lies in the interior of V_τ and it is not the first contact for any higher-dimensional Voronoi polyhedron. In other words, τ is the only simplex in $K_{i+1} - K_i$. We call the addition of τ a *critical!event* because it changes the homotopy type of the complex. Since the union of balls has the homotopy type of the complex, we know that also the union changes its type when the radius reaches r_{i+1}.

Bibliographic notes. Alpha complexes have been introduced for points in $\mathbb{R}^2$ by Edelsbrunner, Kirkpatrick, and Seidel [2]. They have been extended to $\mathbb{R}^3$ in [3] and to weighted points in general, fixed dimension in [1]. The three-dimensional software written by Ernst Mücke has been popular in many areas of science and engineering, including structural molecular biology where they serve as an efficient representation of proteins and other biomolecules. Alpha complexes have been the starting point of the work on persistent homology to be discussed in Chapter VII. The difference between critical and regular events in the evolution of the alpha complex reminds us of the difference between critical and regular points of a Morse function, which will be studied in Chapter VI. The connection is direct but made technically difficult because Morse theory has been developed principally for smooth functions [4]. A lesser known development of the same ideas for non-smooth functions is based on the concept of a topological Morse function [5] of which the Euclidean distance and power functions for a finite points set are examples.

[1] H. Edelsbrunner. The union of balls and its dual shape. *Discrete Comput. Geom.* **13** (1995), 415–440.

[2] H. Edelsbrunner, D. G. Kirkpatrick and R. Seidel. On the shape of a set of points in the plane. *IEEE Trans. Inform. Theory* **IT-29** (1983), 551–559.

[3] H. Edelsbrunner and E. P. Mücke. Three-dimensional alpha shapes. *ACM Trans. Graphics* **13** (1994), 43–72.

[4] J. Milnor. *Morse Theory.* Princeton Univ. Press, New Jersey, 1963.

[5] M. Morse. Topologically non-degenerate functions on a compact n-manifold. *J. Analyse Math.* **7** (1959), 189–208.

Exercises

The credit assignment reflects a subjective assessment of difficulty. A typical question can be answered using knowledge of the material combined with some thought and analysis.

1. **Deciding isomorphism** (three credits). What is the computational complexity of recognizing isomorphic abstract simplicial complexes?
2. **Order complex** (two credits). A *flag* in a simplicial complex K in $\mathbb{R}^d$ is a nested sequence of proper faces, $\sigma_0 < \sigma_1 < \ldots < \sigma_k$. The collection of flags form an abstract simplicial complex A sometimes referred to as the *order complex* of K. Prove that A has a geometric realization in $\mathbb{R}^d$.
3. **Barycentric subdivision** (one credit). Let K consist of a d-simplex σ and its faces.
   (i) How many d-simplexes belong to the barycentric subdivision, $\mathrm{Sd}K$?
   (ii) What is the d-dimensional volume of the individual d-simplices in $\mathrm{Sd}K$?
4. **Covering a tree** (one credit). Let P be a finite collection of closed paths that cover a tree, that is, each node and each edge of the tree belongs to at least one path.
   (i) Prove that the nerve of P is contractible.
   (ii) Is the nerve still contractible if we allow subtrees in the collection? What about sub-forests?
5. **Nerve of stars** (one credit). Let K be a simplicial complex. Prove that K is a geometric realization of the nerve of the collection of vertex stars in K.
6. **Helly for boxes** (two credits). The *box* defined by two points $a = (a_1, a_2, \ldots, a_d)$ and $b = (b_1, b_2, \ldots, b_d)$ in $\mathbb{R}^d$ consists of all points x whose coordinates satisfy $a_i \leq x_i \leq b_i$ for all i. Let F be a finite collection of boxes in $\mathbb{R}^d$. Prove that if every pair of boxes has a non-empty intersection then the entire collection has a non-empty intersection.
7. **Alpha complexes** (two credits). Let $S \subseteq \mathbb{R}^d$ be a finite set of points in general position. Recall that $\check{\mathrm{C}}\mathrm{ech}(r)$ and $\mathrm{Alpha}(r)$ are the Čech and alpha complexes for radius $r \geq 0$. Is it true that $\mathrm{Alpha}(r) = \check{\mathrm{C}}\mathrm{ech}(r) \cap \mathrm{Delaunay}$? If yes, prove the following two subcomplex relations. If no, give examples to show which subcomplex relations are not valid.

(i) $\text{Alpha}(r) \subseteq \check{\text{C}}\text{ech}(r) \cap \text{Delaunay}$.
(ii) $\check{\text{C}}\text{ech}(r) \cap \text{Delaunay} \subseteq \text{Alpha}(r)$.

8. **Collapsibility** (three credits). Call a simplicial complex *collapsible* if there is a sequence of collapses that reduce the complex to a single vertex. The existence of such a sequence implies that the underlying space of the complex is contractible. Describe a finite 2-dimensional simplicial complex that is not collapsible although its underlying space is contractible.

Chapter IV

Homology

Homology is a mathematical formalism for talking in a quantitative and unambiguous manner about how a space is connected. Compared to most other, competing formalisms, homology has faster algorithms but captures less of the topological information. We should keep in mind, however, that detailed classifications are not within our computational reach in any case. Specifically, the question whether or not two triangulated 4-manifolds are homeomeomorphic or homotopy equivalent are both undecidable. In practice, having fast algorithms is a definitive advantage and being insensitive to some topological information is not necessarily a disadvantage. More useful than knowing everything is being able to assess the importance of information and to rank it accordingly, a topic we will address directly in Chapter VII. Before we get there, we need to learn the ropes, which we do in this chapter.

IV.1 Homology Groups

Homology groups provide a mathematical language for the holes in a topological space. Perhaps surprisingly, they capture holes indirectly, by focusing on what surrounds them. Their main ingredients are group operations and maps that relate topologically meaningful subsets of a space with each other. In this section, we introduce the various groups involved in the setup.

Chain complexes. Let K be a simplicial complex and p a dimension. A *p-chain* is a formal sum of p-simplices in K. The standard notation for this is $c = \sum a_i \sigma_i$, where the σ_i are the p-simplices and the a_i are the *coefficients*. In computational topology, we mostly work with coefficients a_i that are either 0 or 1, called *modulo 2 coefficients*. Coefficients can, however, be more complicated numbers like integers, rational numbers, real numbers, elements of a field, or elements of a ring. Since we are working modulo 2, we can think of a chain as a set of p-simplices, namely those σ_i with $a_i = 1$. But when we do consider chains with other coefficient groups, this way of thinking is more cumbersome, so we will use it sparingly.

Two p-chains are added componentwise, like polynomials. Specifically, if $c = \sum a_i \sigma_i$ and $c' = \sum b_i \sigma_i$ then $c + c' = \sum (a_i + b_i)\sigma_i$, where the coefficients satisfy $1 + 1 = 0$. In set notation, the sum of two p-chains is their symmetric difference. The p-chains together with the addition operation form the *group of p-chains* denoted as $(\mathsf{C}_p, +)$, or simply $\mathsf{C}_p = \mathsf{C}_p(K)$ if the operation is understood. Associativity follows from associativity of addition modulo 2. The neutral element is $0 = \sum 0\sigma_i$. The inverse of c is $-c = c$ since $c + c = 0$. Finally, C_p is abelian because addition modulo 2 is abelian. We have a group of p-chains for each integer p. For p less than zero and greater than the dimension of K this group is trivial, consisting only of the neutral element. To relate these groups, we define the *boundary* of a p-simplex as the sum of its $(p-1)$-dimensional faces. Writing $\sigma = [u_0, u_1, \ldots, u_p]$ for the simplex spanned by the listed vertices, its boundary is

$$\partial_p \sigma \quad = \quad \sum_{j=0}^{p} [u_0, \ldots, \hat{u}_j, \ldots, u_p],$$

where the hat indicates that u_j is omitted. For a p-chain, $c = \sum a_i \sigma_i$, the boundary is the sum of the boundaries of its simplices, $\partial_p c = \sum a_i \partial_p \sigma_i$. Hence, taking the boundary maps a p-chain to a $(p-1)$-chain, and we write $\partial_p : \mathsf{C}_p \to \mathsf{C}_{p-1}$. Notice also that taking the boundary commutes with addition, that is, $\partial_p(c + c') = \partial_p c + \partial_p c'$. This is the defining property of a *homomorphism*,

a map between groups that commutes with the group operation. We will therefore refer to ∂_p as the *boundary homomorphism* or, shorter, the *boundary map* for chains. The *chain complex* is the sequence of chain groups connected by boundary homomorphisms,

$$\ldots \xrightarrow{\partial_{p+2}} \mathsf{C}_{p+1} \xrightarrow{\partial_{p+1}} \mathsf{C}_p \xrightarrow{\partial_p} \mathsf{C}_{p-1} \xrightarrow{\partial_{p-1}} \ldots$$

It will often be convenient to drop the index from the boundary homomorphism since it is implied by the dimension of the chain it applies to.

Cycles and boundaries. We distinguish two particular types of chains and use them to define homology groups. A *p-cycle* is a p-chain with empty boundary, $\partial c = 0$. Since ∂ commutes with addition, we have a *group of p-cycles*, denoted as $\mathsf{Z}_p = \mathsf{Z}_p(K)$, which is a subgroup of the group of p-chains. In other words, the group of p-cycles is the kernel of the p-th boundary homomorphism, $\mathsf{Z}_p = \ker \partial_p$. Since the chain groups are abelian so are their cycle subgroups. Consider $p = 0$ as an example. The boundary of every vertex is zero ($\mathsf{C}_{-1} = 0$), hence, $\mathsf{Z}_0 = \ker \partial_0 = \mathsf{C}_0$. For $p > 0$, however, Z_p is usually not all of C_p.

A *p-boundary* is a p-chain that is the boundary of a $(p+1)$-chain, $c = \partial d$ with $d \in \mathsf{C}_{p+1}$. Since ∂ commutes with addition, we have a *group of p-boundaries*, denoted by $\mathsf{B}_p = \mathsf{B}_p(K)$, which is again a subgroup of the p-chains. In other words, the group of p-boundaries is the image of the $(p+1)$-st boundary homomorphism, $\mathsf{B}_p = \operatorname{im} \partial_{p+1}$. Since the chain groups are abelian so are their boundary subgroups. Consider $p = 0$ as an example. Every 1-chain consists of some number of edges, each with two endpoints. Taking the boundary cancels duplicate endpoints in pairs, leaving an even number of distinct vertices. Now suppose the complex is connected. Then for any even number of vertices, we can find paths that connect them in pairs and we can add the paths to get a 1-chain whose boundary consists of the given vertices. Hence, every even set of vertices is a 0-boundary and every odd set of vertices is not. If K is connected this implies that exactly half the 0-cycles are 0-boundaries. The fundamental property that makes homology work is that the boundary of a boundary is necessarily zero.

Fundamental Lemma of Homology. $\partial_p \partial_{p+1} d = 0$ for every integer p and every $(p+1)$-chain d.

Proof. We just need to show that $\partial_p \partial_{p+1} \tau = 0$ for a $(p+1)$-simplex τ. The boundary, $\partial_{p+1}\tau$, consists of all p-faces of τ. Every $(p-1)$-face of τ belongs to exactly two p-faces, so $\partial_p(\partial_{p+1}\tau) = 0$. ∎

It follows that every p-boundary is also a p-cycle or, equivalently, that B_p is a subgroup of Z_p. Figure IV.1 illustrates the subgroup relations among the three types of groups and their connection across dimensions established by the boundary homomorphisms.

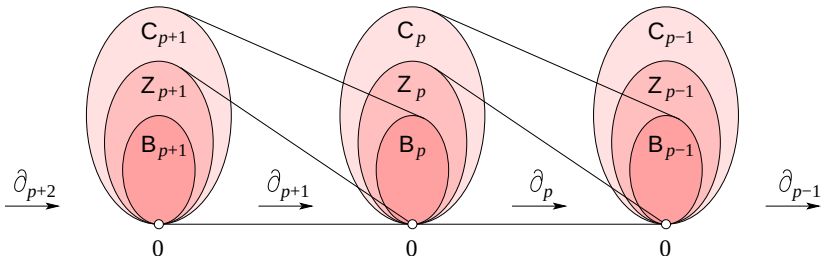

Figure IV.1: The chain complex consisting of a linear sequence of chain, cycle, and boundary groups connected by homomorphisms.

Homology groups. Since the boundaries form subgroups of the cycle groups, we can take quotients. In other words, we can partition each cycle group into classes of cycles that differ from each other by boundaries. This leads to the notion of homology groups and their ranks, which we now define and discuss.

DEFINITION. The *p-th homology group* is the p-th cycle group modulo the p-th boundary group, $\mathsf{H}_p = \mathsf{Z}_p/\mathsf{B}_p$. The *$p$-th Betti number* is the rank of this group, $\beta_p = \operatorname{rank} \mathsf{H}_p$.

Each element of $\mathsf{H}_p = \mathsf{H}_p(K)$ is obtained by adding all p-boundaries to a given p-cycle, $c + \mathsf{B}_p$ with $c \in \mathsf{Z}_p$. If we take any other cycle $c' = c + c''$ in this class, we get the same class, $c' + \mathsf{B}_p = c + \mathsf{B}_p$, since $c'' + \mathsf{B}_p = \mathsf{B}_p$ for every $c'' \in \mathsf{B}_p$. This class is thus a coset of H_p and is referred to as a *homology class*. Any two cycles in the same homology class are said to be *homologous*, which is denoted as $c \sim c'$. We may take c as the representative of this class but any other cycle in the class does as well. Similarly, addition of two classes, $(c + \mathsf{B}_p) + (c_0 + \mathsf{B}_p) = (c + c_0) + \mathsf{B}_p$, is independent of the representatives and is therefore well defined. We thus see that H_p is indeed a group, and because Z_p is abelian so is H_p.

The cardinality of a group is called its *order*. Since we use modulo 2 coefficients, a group with n generators has order 2^n. For example, the base 2 loga-

rithm of the order of C_p is the number of p-dimensional simplices in the complex. Furthermore, the group is isomorphic to $\mathbb{Z}_2^n$, the group of bit-vectors of length n together with the exclusive-or operation. This is an n-dimensional vector space generated by n bit-vectors, for example the n unit vectors. The dimension is referred to as the *rank* of the vector space, $n = \operatorname{rank} \mathbb{Z}_2^n = \log_2 \operatorname{ord} \mathbb{Z}_2^n$. The cycles and boundaries exhibit the same vector space structure, except that their dimension is often less than that of the chains. The number of cycles in each homology class is the order of B_p, hence the number of classes in the homology group is $\operatorname{ord} \mathsf{H}_p = \operatorname{ord} \mathsf{Z}_p / \operatorname{ord} \mathsf{B}_p$. Equivalently, the rank is the difference, $\beta_p = \operatorname{rank} \mathsf{H}_p = \operatorname{rank} \mathsf{Z}_p - \operatorname{rank} \mathsf{B}_p$. This suggests two alternative methods for

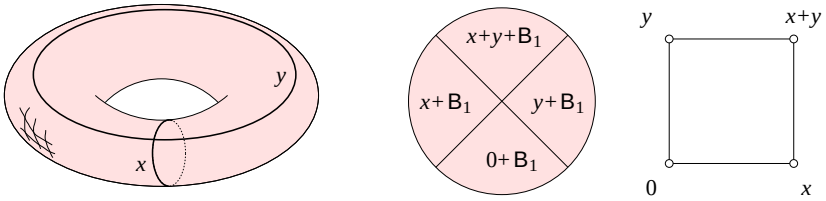

Figure IV.2: The first homology group of the torus has order 4 and rank 2. In the middle, the four elements are drawn as the cosets in the group of 1-cycles. On the right, the four elements are the vertices of a square.

illustrating a homology group, as a partition of the set of cycles or the hypercube of dimension β_p. As an example consider the torus in Figure IV.2. There are only four homology classes in H_1, namely B_1, $x + \mathsf{B}_1$, $y + \mathsf{B}_1$, and $(x+y) + \mathsf{B}_1$, where x and y are the non-bounding 1-cycles that go once around the arm and the hole of the torus. The two corresponding cosets, $x + \mathsf{B}_1$ and $y + \mathsf{B}_1$, generate the first homology group.

The homology of a ball. We define a ball to be any triangulated topological space that is homeomorphic to $\mathbb{B}^k$, the subset of points at distance at most one from the origin in $\mathbb{R}^k$. What is the homology of a ball? Given our intuition that homology should measure holes, it should be trivial. This turns out to almost be true, actually if K is a ball, then $\mathsf{H}_p(K) = 0$ except when $p = 0$ where it has rank 1. This is surprisingly hard to prove, however! It is usually done with a lot of machinery like simplicial approximations and homotopy equivalences. For now, let's at least see this directly when K is the set of faces of a single simplex of dimension k. In this case, the p-chains of K have rank equal to the number of p-faces, which is $\binom{k+1}{p+1}$. Let the vertices be $u_0, u_1, \ldots, u_k$ and consider a p-chain c with simplices of the form $[u_{i_0}, u_{i_1}, \ldots, u_{i_p}]$. The condition

$\partial c = 0$ is equivalent to every $(p-1)$-simplex occurring an even number of times as a face of a p-simplex in c. Assuming $p > 0$, we can construct a $(p+1)$-chain d with boundary $\partial d = c$. This will imply that every p-cycle is also a p-boundary and, equivalently, that H_p is trivial. Specifically, we let d be the set of $(p+1)$-simplices of the form $[u_0, u_{i_0}, u_{i_1}, \ldots, u_{i_p}]$. In words, d picks up a $(p+1)$-simplex for each p-simplex in c that does not contain u_0 as a vertex. To see that c is the boundary of d, we distinguish three types of p-faces of simplices in d. A p-simplex in c that does not contain u_0 occurs exactly once as a face of a $(p+1)$-simplex in d and therefore belongs to ∂d. A p-simplex τ not in c occurs an even number of times, namely once for each time the $(p-1)$-face σ obtained by dropping u_0 occurs in the boundary of a simplex in c. By the same argument, we get a p-simplex τ in c that contains u_0 an odd number of times because one of the p-simplices in c that contains the $(p-1)$-face σ does not give rise to a $(p+1)$-simplex in d, namely τ itself.

This covers all positive dimensions. For $p = 0$, we have already observed that exactly half the cycles are boundaries. Hence, $\mathsf{H}_0 = \mathsf{Z}_0/\mathsf{B}_0$ is isomorphic to $\mathbb{Z}_2$ and $\beta_0 = 1$, as claimed.

Reduced homology. There is something dissatisfying about the 0-th homology group behaving differently for the ball than the others. The reason for the difference is that we have set up things so that β_0 counts the components, but if there is one component there is no hole. More satisfying would be to count one for two components, namely for the one gap between them. This is achieved by a small but often useful modification of homology, namely adding the *augmentation map* $\epsilon : \mathsf{C}_0 \to \mathbb{Z}_2$ defined by $\epsilon(u) = 1$ for each vertex u to the chain complex. We thus get

$$\ldots \xrightarrow{\partial_2} \mathsf{C}_1 \xrightarrow{\partial_1} \mathsf{C}_0 \xrightarrow{\epsilon} \mathbb{Z}_2 \xrightarrow{0} 0 \longrightarrow \ldots$$

Cycles and boundaries are defined as before and the only difference we notice is for Z_0 which now requires that each 0-cycle has an even number of vertices. This results in the *reduced homology groups*, $\tilde{\mathsf{H}}_p$, and the *reduced Betti numbers*, $\tilde{\beta}_p = \operatorname{rank} \tilde{\mathsf{H}}_p$. Assuming K is non-empty, we have $\tilde{\beta}_p = \beta_p$ for all $p \geq 1$ and $\tilde{\beta}_0 = \beta_0 - 1$. For $K = \emptyset$ we have $\tilde{\beta}_{-1} = 1$ since both elements of $\mathbb{Z}_2$ are (-1)-cycles, they belong to the kernel, but only one is a (-1)-boundary, it belongs to the image of the augmentation map.

Induced maps. A continuous map from one topological space to another maps cycles to cycles and boundaries to boundaries. We can therefore use the images to construct new homology groups. The are not necessarily the same as

the ones of the original space since cycles can become boundaries, for example trivial cycles. We describe this more formally for two simplicial complexes and a simplicial map, $f : K \to L$, between them. Recall that f takes each simplex of K linearly to a simplex of L. It induces a map from the chains of K to the chains of the same dimension of L. Specifically, if $c = \sum a_i \sigma_i$ is a p-chain in K, then $f_\#(c) = \sum a_i \tau_i$, where $\tau_i = f(\sigma_i)$ if it has dimension p and $\tau_i = 0$ if $f(\sigma)$ has dimension less than p. Writing ∂_K and ∂_L for the boundary maps in the two complexes, we note that $f_\# \circ \partial_K = \partial_L \circ f_\#$, that is, the induced map commutes with the boundary map. This is obvious when $f(\sigma_i)$ has dimension p, since then all $(p-1)$-faces of σ_i map to the corresponding $(p-1)$-faces of τ_i. If, on the other hand, $f(\sigma_i)$ has dimension less than p, then the $(p-1)$-faces of σ_i map to simplices of dimension less than $p-1$, with the possible exception of exactly two $(p-1)$-faces whose images coincide and cancel each other. So both $f_\#(\partial_K \sigma_i)$ and $\partial_L f_\#(\sigma)$ are zero. Note that in the case when $f : K \to L$ is the inclusion of one simplicial complex into another, simplices always keep their dimension, so the induced map, $f_\#$, is a little easier to understand.

The fact that the induced map commutes with the boundary map implies that $f_\#$ takes cycles to cycles, $f_\#(\mathsf{Z}_p(K)) \subseteq f_\#(\mathsf{Z}_p(L))$, and boundaries to boundaries, $f_\#(\mathsf{B}_p(K)) \subseteq f_\#(\mathsf{B}_p(L))$. Therefore, it defines a map on the quotients, which we call the *induced map on homology*, written $f_* : \mathsf{H}_p(K) \to \mathsf{H}_p(L)$. The rank of the image is bounded from above by both Betti numbers, $\operatorname{rank} f_*(\mathsf{H}_p(K)) \leq \min\{\beta_p(K), \beta_p(L)\}$.

Degree of a map. We now present a first application of the concept of induced maps. We describe it for general continuous maps, appealing to the Simplicial Approximation Theorem proved in Section III.1 when we need triangulations and an approximating simplicial map. Let $g : \mathbb{S}^p \to \mathbb{S}^p$ be a continuous map and let c be the unique generator of the k-th homology group of the p-sphere. Then $g(c)$ is either homologous to c or to 0. In other words, $g(c) \sim \alpha c$ and $\alpha \in \{0, 1\}$ is called the *modulo* 2 *degree* of g. If g is the identity then $\alpha = 1$. However, if g extends a continuous map $g_0 : \mathbb{B}^{p+1} \to \mathbb{S}^p$ then the induced map on homology, $g_* : \mathsf{H}_p(\mathbb{S}^p) \to \mathsf{H}_p(\mathbb{S}^p)$ is the composite of two induced maps, $\mathsf{H}_p(\mathbb{S}^p) \to \mathsf{H}_p(\mathbb{B}^{p+1}) \to \mathsf{H}_p(\mathbb{S}^p)$, where the first is induced by inclusion. The middle group is trivial, hence $\alpha = 0$. We are now ready to prove a classic result on fixed points of continuous maps.

Brouwer's Fixed Point Theorem. A continuous map $f : \mathbb{B}^{p+1} \to \mathbb{B}^{p+1}$ has at least one fixed point $x = f(x)$.

PROOF. Let $A, B : \mathbb{S}^p \to \mathbb{S}^p$ be maps defined by $A(x) = (x - f(x))/\|x - f(x)\|$ and $B(x) = x$. Since B is the identity its modulo 2 degree is 1. If f has no fixed point then A is well defined and has modulo 2 degree 0 because it extends a map from the $(p+1)$-ball to the p-sphere. We now construct $H : \mathbb{S}^p \times [0,1] \to \mathbb{S}^p$ defined by $H(x,t) = (x - tf(x))/\|x - tf(x)\|$. For $t = 1$ we have $x \neq f(x)$ because there is no fixed point and for $t < 1$ we have $x \neq tf(x)$ because $\|x\| = 1 > \|tf(x)\|$. We conclude that H is a homotopy between A and B which implies that the modulo 2 degree of the two are the same, a contradiction. □

Bibliographic notes. Like many other concepts in topology, homology groups were introduced by Henri Poincaré in one of a series of papers on "analysis situ" [5]. He named the ranks of the homology groups after another mathematician, Betti, who introduced a slightly different version years earlier. The field experienced a rapid development during the twentieth century. There were many competing theories, simplicial and singular homology just being two examples, which have been consolidated by axiomizing the assumptions under which homology groups exist [1]. Today we have a number of well established textbooks in the field. We refer to Giblin [2] for an intuitive introduction and to Hatcher [3], Munkres [4], and Spanier [6] for more comprehensive sources.

[1] S. EILENBERG AND N. STEENROD. *Foundations of Algebraic Topology.* Princeton Univ. Press, New Jersey, 1952.

[2] P. J. GIBLIN. *Graphs, Surfaces and Homology.* Chapman and Hall, London, 1981.

[3] A. HATCHER. *Algebraic Topology.* Cambridge Univ. Press, England, 2002.

[4] J. R. MUNKRES. *Elements of Algebraic Topology.* Addison-Wesley, Redwood City, California, 1984.

[5] H. POINCARÉ. Complément à l'analysis situs. *Rendiconti del Circolo Matematico di Palermo* **13** (1899), 285–343. .

[6] E. H. SPANIER. *Algebraic Topology.* Springer-Verlag, New York, 1966.

IV.2 Matrix Reduction

The homology groups of a triangulated space can be computed from the matrices representing the boundary homomorphisms. Their reduced versions readily provide the ranks of the cycle and boundary groups, and their differences give the Betti numbers. Summing these same differences leads to a proof of the Euler-Poincaré formula which generalizes the Euler relation for planar graphs.

Euler-Poincaré formula. Recall that the Euler characteristic of a simplicial complex is the alternating sum of the number of simplices in each dimension. Similarly, recall that the rank of the p-th homology group is the rank of the p-th cycle group minus the rank of the p-th boundary group. Writing $n_p = \operatorname{rank} \mathsf{C}_p$ for the number of p-simplices in K, $z_p = \operatorname{rank} \mathsf{Z}_p$ and $b_p = \operatorname{rank} \mathsf{B}_p$ for the ranks of the cycle and boundary groups, we have $n_p = z_p + b_{p-1}$. This is the general fact that for any linear transformation between vector spaces $f : U \to V$, the dimension of U equals the sum of the dimension of the kernel of f and the dimension of the image of f. The Euler characteristic is the alternating sum of the n_p, which is therefore

$$\begin{aligned} \chi &= \sum_{p \geq 0} (-1)^p (z_p + b_{p-1}) \\ &= \sum_{p \geq 0} (-1)^p (z_p - b_p), \end{aligned}$$

which is the same as the alternating sum of Betti numbers. To appreciate the beauty of this result, we need to know that homology groups do not depend on the triangulation chosen for a topological space. The technical proof of this claim is difficult and we refer the reader to more advanced texts, but even the more general result that homotopy equivalent spaces have isomorphic homology groups is plausible. For example, we can free ourselves from the triangulation entirely and define chains in terms of continuous maps from the standard simplex into the space $\mathbb{X}$. This gives rise to singular homology, which can be shown to give groups isomorphic to the ones we get by simplicial homology, the theory we describe in this chapter. If we now have a continuous map $f : \mathbb{X} \to \mathbb{Y}$ we can map the chains from $\mathbb{X}$ to those of $\mathbb{Y}$ by simply composing. If f is a homotopy equivalence then it turns out that $\mathbb{X}$ and $\mathbb{Y}$ have isomorphic homology groups. This also implies that the Euler characteristic is an invariant of the space, that is, it does not depend on the simplicial complex we use to triangulate it.

Euler-Poincaré Theorem. The Euler characteristic of a topological space is the alternating sum of its Betti numbers, $\chi = \sum_{p \geq 0} (-1)^p \beta_p$.

Boundary matrices. To compute homology, we combine information from two sources, one representing the cycles and the other the boundaries, just as in the proof of the Euler-Poincaré Theorem. Let K be a simplicial complex. Its *p-th boundary matrix* represents the $(p-1)$-simplices as rows and the p-simplices as columns. Assuming an arbitrary but fixed ordering of the simplices, for each dimension, this matrix is $\partial_p = [a_i^j]$, where i ranges from 1 to n_{p-1}, j ranges from 1 to n_p, and $a_i^j = 1$ if the i-th $(p-1)$-simplex is a face of the j-th p-simplex and $a_i^j = 0$, otherwise. Given a p-chain, $c = \sum a_i \sigma_i$, the boundary can be computed by matrix multiplication,

$$\partial_p c = \begin{bmatrix} a_1^1 & a_1^2 & \dots & a_1^{n_p} \\ a_2^1 & a_2^2 & \dots & a_2^{n_p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n_{p-1}}^1 & a_{n_{p-1}}^2 & \dots & a_{n_{p-1}}^{n_p} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n_p} \end{bmatrix}.$$

In words, a collection of columns represents a p-chain and the sum of these columns gives its boundary. Similarly, a collection of rows represents a $(p-1)$-chain and the sum of these rows gives its coboundary, a concept that will be defined in the next chapter.

Row and column operations. The rows of the matrix ∂_p form a basis of the $(p-1)$-st chain group, C_{p-1}, and the columns form a basis of the p-th chain group, C_p. We use two types of column operations to modify the matrix without changing its rank: exchanging columns k and l and adding column k to column l. Both can be expressed by multiplying with a matrix $V = [v_i^j]$ from the right. To exchange two columns, we have $v_k^l = v_l^k = 1$ and $v_i^i = 1$ for

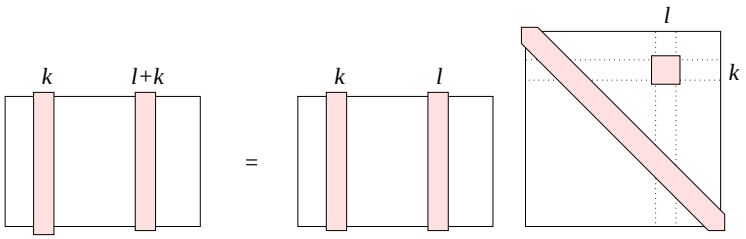

Figure IV.3: The effect of a single off-diagonal one in the matrix V is the addition of one column in the boundary matrix to another. The effect on the basis of C_p is similar.

all $i \neq k, l$. All other entries are zero. To add column k to column l, we have

$v_k^l = 1$ and $v_i^i = 1$ for all i. All other entries are zero. As indicated in Figure IV.3, the effect of the operation is that the l-th column now represents the sum of the k-th and the l-th p-simplices, or the sum of whatever the two columns represented before the operation. Similarly, we have two row operations, one exchanging two rows and the other adding one row to another. This translates to multiplication with a matrix $U = [u_i^j]$ from the left. To exchange two rows, we again have $u_k^l = u_l^k = 1$, $u_i^i = 1$ for $i \neq k, l$, and all other entries zero. To add the k-th to the l-th row we have $u_l^k = 1$, $u_i^i = 1$ for all i, and all other entries zero, as in Figure IV.4. The effect of this operation is that the k-th row

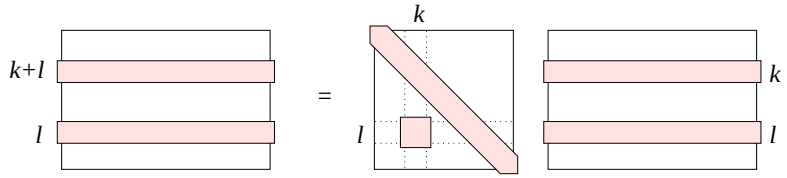

Figure IV.4: The effect of a single off-diagonal one in the matrix U is the addition of one row in the boundary matrix to another. The effect on the basis of C_{p-1} is that the row that was added now represents the sum of $(p-1)$-chains, the opposite of a column operation.

now represents the sum of the k-th and the l-th $(p-1)$-simplices, or the sum of whatever the two rows represented before the operation. Although the $(p-1)$ and p-chains represented by the rows and columns change as we perform row and column operations, they always represent bases of the two chain groups.

Smith normal form. Using row and column operations, we can reduce the p-th boundary matrix to *Smith normal form*. For modulo 2 arithmetic, this means an initial segment of the diagonal is 1 and everything else is 0, as in Figure IV.5. Recall that $n_p = \operatorname{rank} \mathsf{C}_p$ is the number of columns of the p-th boundary matrix. Let $n_p = b_{p-1} + z_p$ so that the leftmost b_{p-1} columns have ones in the diagonal and the rightmost z_p columns are zero. The former represent p-chains whose non-zero boundaries generate the group of $(p-1)$-boundaries. The latter represent p-cycles that generate Z_p. Once we have all boundary matrices in normal form, we can extract the Betti numbers as differences between ranks, $\beta_p = \operatorname{rank} \mathsf{Z}_p - \operatorname{rank} \mathsf{B}_p$ for $p \geq 0$. To get the bases of the boundary and cycle groups, we keep track of the matrix products that represent the row and column operations. Writing U_{p-1} and V_p for the left and right products, we get the normal form as $N_p = U_{p-1} \partial_p V_p$. The new basis for the cycle group is given in the last z_p columns of V_p. Similarly, the new basis

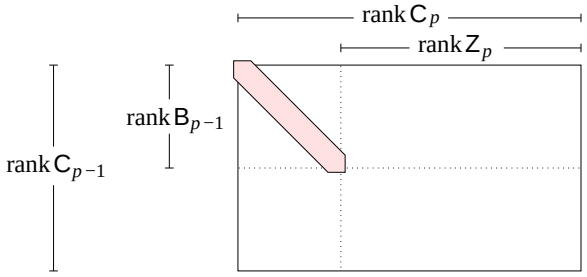

Figure IV.5: The entries in the shaded, initial portion of the diagonal are 1 and all other entries are 0. The ranks of the boundary and cycle groups are readily available as the numbers of non-zero and zero columns.

for the boundary group is encoded in U_{p-1} and we get the basis vectors from the first b_{p-1} columns of the inverse.

Reduction. To reduce ∂_p, we proceed similar to Gaussian elimination for solving a system of linear equations. In at most two exchange operations, we move a 1 to the upper left corner, and with at most $n_{p-1} - 1$ row and $n_p - 1$ column additions, we zero out the rest of the first column and first row. We then recurse for the submatrix obtained by removing the first row and first column. We start the reduction by initializing the matrix to $N_p[i, j] = a_i^j$ for all i and j, and by calling the function for $x = 1$, the position of the considered diagonal element.

```
void REDUCE(x)
  if there exist k ≥ x, l ≥ x with N_p[k, l] = 1 then
    exchange rows x and k;  exchange columns x and l;
    for i = x + 1 to n_{p-1} do
      if N_p[i, x] = 1 then add row x to row i endif
    endfor;
    for j = x + 1 to n_p do
      if N_p[x, j] = 1 then add column x to column j endif
    endfor;
    REDUCE(x + 1)
  endif.
```

We have at most n_{p-1} row and n_p column operations per recursive call and hence at most $(n_{p-1} + n_p) \min\{n_{p-1}, n_p\}$ row and column operations in total.

Multiplying with their lengths, we thus get a running time of a constant times $2n_{p-1}n_p \min\{n_{p-1}, n_p\}$. The amount of memory is at most some constant times $(n_{p-1} + n_p)^2$ needed to store the matrices. In summary, we reduce the boundary matrices in time at most cubic and in memory at most quadratic in the number of simplices in K.

Example. To get a feeling for the algorithm, we use it to compute the reduced homology group of the 3-ball triangulated by the faces of a single tetrahedron. We do the computations one dimension at a time and this way get the reduced Betti numbers of all skeleta of the complex as we go. The 0-skeleton consists

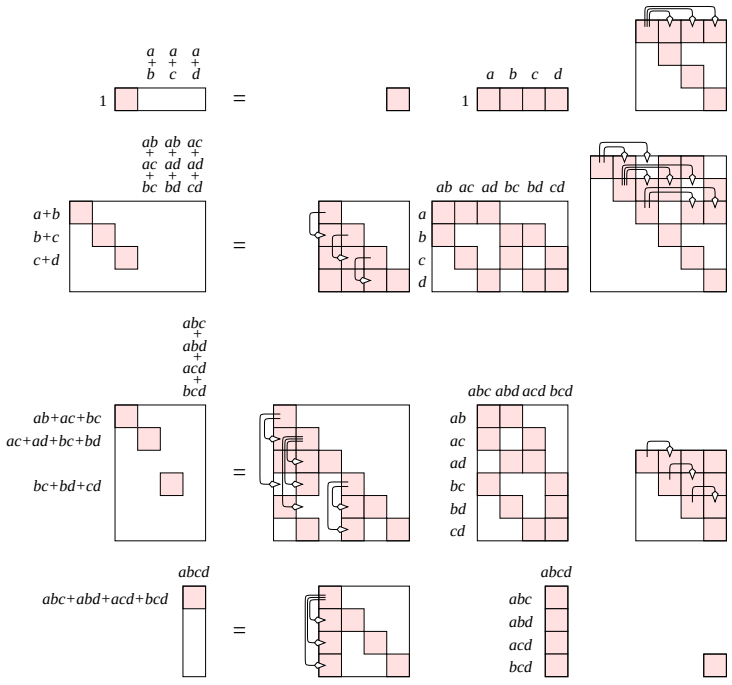

Figure IV.6: From top to bottom: the matrix equations $N_p = U_{p-1}\partial_p V_p$ for reducing the zeroth, first, second, and third boundary matrices of the tetrahedron. The ones are shaded and the zeros are white. The bases are indicated both for the boundary and the normal form matrices. For clarity, no exchanges are performed.

of four vertices and its sole non-trivial boundary matrix is ∂_0 consisting of a row of ones, shown as part of the first equation in Figure IV.6. Three column

operations remove three of the four ones and we get $\tilde{\beta}_0 = 3$, the number or zero columns in N_0. Proceeding to the 1-skeleton, we add the six edges and consider the first boundary matrix. After reduction, it has three ones in the diagonal, shown as part of the second equation in Figure IV.6. Combining the information from N_0 and N_1 we get $\tilde{\beta}_0 = 3 - 3 = 0$, and counting the zero columns in N_1 we get $\tilde{\beta}_1 = 3$. Proceeding to the 2-skeleton, we add the four triangles and thus get a triangulation of the 2-sphere. After reduction, the boundary matrix has again three ones in the diagonal, shown as part of the third equation in Figure IV.6. The triangles do not affect the zeroth homology group and we have $\tilde{\beta}_0 = 0$, same as before. Combining the information from N_1 and N_2 we get $\tilde{\beta}_1 = 3 - 3 = 0$, and counting the zero columns in N_2 we get $\tilde{\beta}_2 = 1$. We finally get the triangulation of the 3-ball by adding the one tetrahedron. After reduction, the boundary matrix has a single one in the diagonal, shown as part of the fourth equation in Figure IV.6. The first two reduced Betti numbers remain unaffected and the other two also vanish, so we get $\tilde{\beta}_0 = \tilde{\beta}_1 = \tilde{\beta}_2 = \tilde{\beta}_3 = 0$, as expected.

Bibliographic notes. The generalization of the Euler relation for planar graphs to the Euler-Poincaré Theorem has an interesting history analyzed from a philosophical viewpoint by Lakatos [2]. The result of reducing the boundary matrix is sometimes referred to as Smith normal form [4]. We describe the algorithm for modulo 2 arithmetic, but other, more elaborate coefficient groups can also be used. Already for integers, this complicates matters significantly and it is no longer straightforward to guarantee a running time that is polynomial in the number of simplices [3]. However, improvements to polynomial time are possible [1, 5].

[1] R. KANNAN AND A. BACHEM. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Comput.* **8** (1979), 499–507.

[2] I. LAKATOS. *Proofs and Refutations: the Logic of Mathematical Discovery.* Cambridge Univ. Press, Cambridge, England, 1976.

[3] J. R. MUNKRES. *Elements of Algebraic Topology.* Addison-Wesley, Redwood City, California, 1984.

[4] H. J. SMITH. On systems of indeterminate equations and congruences. *Philos. Trans.* **151** (1861), 293–326.

[5] A. STORJOHANN. Near optimal algorithm for computing Smith normal forms of integer matrices. *In* "Proc. Internat. Sympos. Symbol. Algebraic Comput., 1997", 267–274.

IV.3 Relative Homology

We extend homology beyond closed spaces by considering nested pairs of closed spaces and studying their difference. We need two new concepts to relate the homology of such a pair to the homology of the individual closed spaces, induced maps and exact sequences.

Relative homology groups. Homology groups have been defined for triangulated spaces, which are therefore necessarily closed. To extend them to other spaces, we introduce homology groups for pairs of closed spaces. Let K be a simplicial complex and K_0 a subcomplex of K. The *relative chain groups* are quotients of the chain groups of K and of K_0, $\mathsf{C}_p(K, K_0) = \mathsf{C}_p(K)/\mathsf{C}_p(K_0)$. Taking this quotient partitions $\mathsf{C}_p(K)$ into cosets, $c + \mathsf{C}_p(K_0)$, whose p-chains possibly differ in the p-simplices in K_0 but not in the ones in $K - K_0$. The *boundary map* is induced by the one for K, that is, $\partial_p(c + \mathsf{C}_p(K_0)) = \partial_p c + \mathsf{C}_{p-1}(K_0)$. As before, ∂ commutes with addition and taking the boundary twice gives zero. We thus define *relative cycle groups*, *relative boundary groups*, and *relative homology groups* as kernels, images, and quotients,

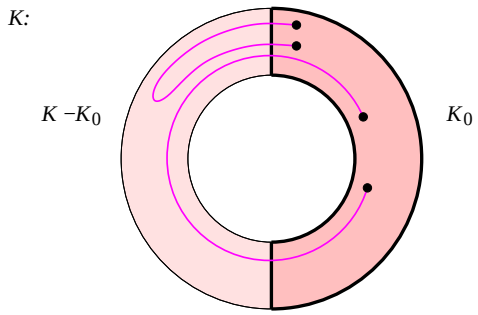

Figure IV.7: The pair (K, K_0), where K triangulates the annulus and K_0 the right half of the annulus. The displayed paths are neither boundaries nor cycles in K but are both relative cycles and one is a relative boundary in (K, K_0). Which one?

$$\begin{aligned} \mathsf{Z}_p(K, K_0) &= \ker \partial_p; \\ \mathsf{B}_p(K, K_0) &= \operatorname{im} \partial_{p+1}; \\ \mathsf{H}_p(K, K_0) &= \ker \partial_p / \operatorname{im} \partial_{p+1}, \end{aligned}$$

just as before. Let $c + \mathsf{C}_p(K_0)$ be a relative p-chain. It is a relative p-cycle iff ∂c is carried by K_0, which includes the possibility that ∂c is zero. Furthermore,

it is a relative p-boundary is there is a $(p+1)$-chain d of K such that $c - \partial d$ is carried by K_0; see Figure IV.7.

Excision. By construction, relative homology depends only on the part of K outside K_0 and ignores the part inside K_0. Hence, we can remove simplices from both complexes without changing the homology.

EXCISION THEOREM. Let $K_0 \subseteq K$ and $L_0 \subseteq L$ be pairs of simplicial complexes that satisfy $L \subseteq K$ and $L - L_0 = K - K_0$. Then they have isomorphic relative homology groups, that is, $\mathsf{H}_p(K, K_0) \simeq \mathsf{H}_p(L, L_0)$ for all dimensions p.

Instead of giving an algebraic proof of this fairly obvious fact, we take a look at the Smith Normal Form Reduction for Relative Homology. Ordering the simplices in K_0 before the ones in $K - K_0$, all the relevant information is contained in the lower right submatrices that belong to rows and columns of simplices in $K - K_0$. We reduce these submatrices, ignoring the rows and columns of simplices in K_0. As illustrated in Figure IV.8, we get the ranks of the relative boundary and cycle groups by counting the non-zero and zero columns in the submatrices. Using the same ordering of simplices, we get the boundary matrices of L by removing rows and columns that correspond to simplices in $K - L$. By definition of L and L_0, these rows and columns correspond to simplices in K_0. The lower right submatrices defined by $L - L_0$ are therefore the same as before. This implies $\mathsf{H}_p(K, K_0) \simeq \mathsf{H}_p(L, L_0)$ for all dimensions p, as claimed in the Excision Theorem.

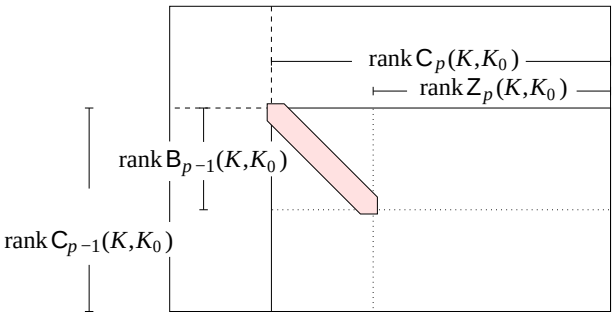

Figure IV.8: By ordering the simplices of K_0 before the others, we get the incidences between simplices in $K - K_0$ in the lower right submatrix, which we reduce to compute the homology of the pair (K, K_0).

We could have deleted the rows and columns of simplices in K_0 but chose to keep them because they contain the information that relates the relative homology groups of (K, K_0) with the (absolute) homology groups of K and K_0. We need new concepts to describe this connection.

Induced maps. A continuous map from one topological space to another maps cycles to cycles and boundaries to boundaries. We can therefore use the images to construct new homology groups. They are not necessarily the same as the ones of the original space since cycles can become boundaries, for example trivial cycles. We describe this more formally for two simplicial complexes and a simplicial map, $f : K \to L$, between them. Recall that f takes each simplex of K linearly to a simplex of L. It induces a map from the chains of K to the chains of the same dimension of L. Specifically, if $c = \sum a_i \sigma_i$ is a p-chain in K, then $f_\#(c) = \sum a_i \tau_i$, where $\tau_i = f(\sigma_i)$ if it has dimension p and $\tau_i = 0$ if $f(\sigma_i)$ has dimension less than p. Writing ∂_K and ∂_L for the boundary maps in the two complexes, we note that $f_\# \circ \partial_K = \partial_L \circ f_\#$, that is, the induced map commutes with the boundary map. This is obvious when $f(\sigma_i)$ has dimension p, since then all $(p-1)$-faces of σ_i map to the corresponding $(p-1)$-faces of τ_i. If, on the other hand, $f(\sigma_i)$ has dimension less than p, then the $(p-1)$-faces of σ_i map to simplices of dimension less than $p-1$, with the possible exception of exactly two $(p-1)$-faces whose images coincide and cancel each other. So both $f_\# \circ \partial_K(\sigma_i)$ and $\partial_L \circ f_\#(\sigma)$ are zero. Note that in the common case when $f : K \to L$ includes one simplicial complex into the other, simplices always keep their dimension, so the induced map, $f_\#$, is a little easier to understand.

The fact that the induced map commutes with the boundary map implies that $f_\#$ takes cycles to cycles, $f_\#(\mathsf{Z}_p(K)) \subseteq \mathsf{Z}_p(L)$, and boundaries to boundaries, $f_\#(\mathsf{B}_p(K)) \subseteq \mathsf{B}_p(L)$. Therefore, it defines a map on the quotients, which we call the *induced map on homology*, written $f_* : \mathsf{H}_p(K) \to \mathsf{H}_p(L)$. Similarly, we have an induced map on reduced homology. The order of the image is of course bounded from above by the order of the domain and of the range, and hence, $\operatorname{rank} f_*(\mathsf{H}_p(K)) \leq \min\{\beta_p(K), \beta_p(L)\}$. It is easy to construct examples for which the rank of the image is strictly smaller than both Betti numbers.

Degree of a map. We present a first application of the concept of induced maps. Describing it for general continuous maps, we appeal to the Simplicial Approximation Theorem proved in Section III.1 when we need triangulations and an approximating simplicial map. Let $g : \mathbb{S}^d \to \mathbb{S}^d$ be a continuous map and let c be the unique generator of the d-th reduced homology group of the d-sphere. Then $g(c)$ is either homologous to c or to 0. In other words, $g(c) \sim \alpha c$ and $\alpha \in \{0, 1\}$ is called the *modulo* 2 *degree* or just the *degree* of g. If g is the

identity then $\alpha = 1$. However, if g extends a continuous map $g_0 : \mathbb{B}^{d+1} \to \mathbb{S}^d$ then the induced map on reduced homology, $g_* : \tilde{\mathsf{H}}_d(\mathbb{S}^d) \to \tilde{\mathsf{H}}_d(\mathbb{S}^d)$ is the composite of two induced maps, $\tilde{\mathsf{H}}_d(\mathbb{S}^d) \to \tilde{\mathsf{H}}_d(\mathbb{B}^{d+1}) \to \tilde{\mathsf{H}}_d(\mathbb{S}^d)$, where the first is induced by inclusion. The middle group is trivial, hence $\alpha = 0$. We note that a homotopy cannot change the degree of a map. With this, we are ready to prove a classic result on fixed points of continuous maps.

BROUWER'S FIXED POINT THEOREM. A continuous map $f : \mathbb{B}^{d+1} \to \mathbb{B}^{d+1}$ has at least one fixed point $x = f(x)$.

PROOF. Let $A, B : \mathbb{S}^d \to \mathbb{S}^d$ be maps defined by $A(x) = (x - f(x))/\|x - f(x)\|$ and $B(x) = x$. Since B is the identity, its degree is one. If f has no fixed point then A is well defined and has degree zero because it extends to a map from the $(d+1)$-ball to the d-sphere. We now construct $H : \mathbb{S}^d \times [0,1] \to \mathbb{S}^d$ defined by $H(x,t) = (x - tf(x))/\|x - tf(x)\|$. For $t = 1$, we have $x \neq f(x)$ because there is no fixed point, and for $t < 1$, we have $x \neq tf(x)$ because $\|x\| = 1 > \|tf(x)\|$. We conclude that H is a homotopy between A and B which implies that the degree of the two are the same, a contradiction. □

Maps between vector spaces. Since we use modulo 2 arithmetic, the induced map on homology is a linear transformation between vector spaces. We discuss such maps in generality, without burdening ourselves with the interpretation that these vector spaces are obtained by taking quotients, or what have you. Letting $f : \mathsf{U} \to \mathsf{V}$ be a linear transformation between vector spaces, the *kernel*, *image*, and *cokernel* are defined as usual,

$$\begin{aligned} \ker f &= \{u \in \mathsf{U} \mid f(u) = 0 \in \mathsf{V}\}; \\ \operatorname{im} f &= \{v \in \mathsf{V} \mid \text{there exists } u \in \mathsf{U} \text{ with } f(u) = v\}; \\ \operatorname{cok} f &= \mathsf{V}/\operatorname{im} f. \end{aligned}$$

For example, if f is represented by a matrix, like ∂, we can reduce and get the kernel spanned by the zero columns, the image by the non-zero rows, and the cokernel by the zero rows. All three are vector spaces in their own right, so we can take direct sums, recalling that this is like taking Cartesian products and using the group operations componentwise. A fundamental result from linear algebra states that U and V are completely described by the three. Specifically, U is isomorphic to the direct sum of the kernel and the image and V is isomorphic to the direct sum of the image and the cokernel,

$$\begin{aligned} \mathsf{U} &\simeq \ker f \oplus \operatorname{im} f; \\ \mathsf{V} &\simeq \operatorname{im} f \oplus \operatorname{cok} f. \end{aligned}$$

Again, this has obvious interpretations in terms of the reduced matrix representing f. If we have three vector spaces and two linear transformations, $f : \mathsf{U} \to \mathsf{V}$ and $g : \mathsf{V} \to \mathsf{W}$, we say the sequence $\mathsf{U} \to \mathsf{V} \to \mathsf{W}$ is *exact* at V if $\operatorname{im} f = \ker g$; see Figure IV.9. Note that this implies $g \circ f = 0$, thus the sequence might be three terms in a chain complex, and exactness would mean that the homology group at V was 0. But we will use this concept in more general ways than that. If $0 \to \mathsf{U} \to \mathsf{V}$ is a sequence, then exactness at U is

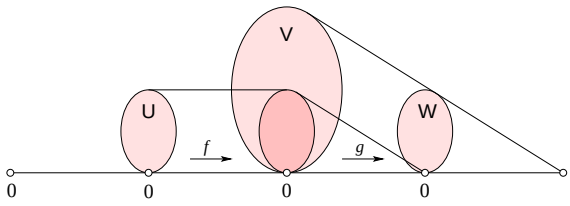

Figure IV.9: A short exact sequence of vector spaces. It starts and ends with zero and is exact at each of the three vector spaces between the two ends.

equivalent to injectivity of $\mathsf{U} \to \mathsf{V}$. Similarly, if $\mathsf{V} \to \mathsf{W} \to 0$, then exactness at W is equivalent to surjectivity of $\mathsf{V} \to \mathsf{W}$. A *short exact sequence* is a sequence of length five,

$$0 \to \mathsf{U} \xrightarrow{f} \mathsf{V} \xrightarrow{g} \mathsf{W} \to 0,$$

that starts and ends with the trivial vector space and is exact at U, V, and W. By what we said above, $f : \mathsf{U} \to \mathsf{V}$ is injective and $g : \mathsf{V} \to \mathsf{W}$ is surjective. In this situation, it is always true that the middle vector space is isomorphic to the direct sum of the adjacent vector spaces, $\mathsf{V} \simeq \mathsf{U} \oplus \mathsf{W}$. Thus if we somehow already know U and W then we have calculated V.

Exact sequence of a pair. Sequences that are exact are convenient means to express otherwise cumbersome relationships between homology groups. Exceptionally powerful are *long exact sequences* which are infinite sequences of vector spaces that are exact at all of them. A long exact sequence is like a chain complex, but with trivial homology throughout. A particular example relates the relative homology groups of a pair with the absolute homology groups of the spaces forming the pair.

Exact Sequence of a Pair Theorem. Let K be a simplicial complex and $K_0 \subseteq K$ a subcomplex. Then there is a long exact sequence

$$\ldots \to \mathsf{H}_p(K_0) \to \mathsf{H}_p(K) \to \mathsf{H}_p(K, K_0) \to \mathsf{H}_{p-1}(K_0) \to \ldots$$

The same statement holds if we substitute the reduced homology groups of K and K_0 for their non-reduced homology groups.

The next section will give a general method for constructing long exact sequences, including that of a pair. Therefore we will content ourselves here with a brief description of the maps between the groups. The map $\mathsf{H}_p(K_0) \to \mathsf{H}_p(K)$ is just the map on homology induced by the inclusion $K_0 \subseteq K$. The map $\mathsf{H}_p(K) \to \mathsf{H}_p(K, K_0)$ is also induced by inclusion, $K \subseteq K$, that is, a class generated by a cycle c in K is mapped to the relative class generated by $c + \mathsf{C}_p(K_0)$. The third map, $\mathsf{H}_p(K, K_0) \to \mathsf{H}_{p-1}(K_0)$, is called the *connecting homomorphism* and is the crucial piece of the construction. To describe it, let $c = \sum a_i \sigma_i$ generate a relative p-cycle, that is, $\partial c \in \mathsf{C}_{p-1}(K_0)$. In K_0, the boundary of c is clearly a cycle and therefore represents a class in $\mathsf{H}_{p-1}(K_0)$. This defines the connecting homomorphism, mapping the relative homology class generated by $c + \mathsf{C}_p(K_0)$ to the absolute homology class generated by ∂c. Indeed, any cycle in the same relative class with c can be written as $c + c' + c_0$, where $c' \in \mathsf{B}_p(K)$ and $c_0 \in \mathsf{C}_p(K_0)$. But then $\partial c' = 0$ and ∂c_0 is a boundary in K_0. Hence, $\partial(c + c' + c_0) = \partial c + \partial c_0$ is homologous to ∂c as a cycle in K_0.

As an example, consider the pair $(\mathbb{B}^3, \mathbb{S}^1)$, the 3-ball modulo its equator, triangulated by (K, K_0). We can use the exact homology sequence to figure the relative homology of this pair. Using reduced homology, all groups of $\mathbb{B}^3$ are zero. Similarly, the only non-zero reduced homology group of $\mathbb{S}^1$ is the first one, which has rank one. Except for dimension $p = 2$, we therefore have

$$\ldots \to 0 \to \mathsf{H}_p(\mathbb{B}^3, \mathbb{S}^1) \to 0 \to \ldots,$$

implying that $\mathsf{H}_p(\mathbb{B}^3, \mathbb{S}^1)$ itself is zero. For $p = 2$ we have the only non-trivial portion of the long exact sequence,

$$\ldots \to 0 \to \mathsf{H}_2(\mathbb{B}^3, \mathbb{S}^1) \to \tilde{\mathsf{H}}_1(\mathbb{S}^1) \to 0 \to \ldots$$

The map between the middle two groups is thus injective as well as surjective, which implies that $\mathsf{H}_2(\mathbb{B}^3, \mathbb{S}^1)$ has rank one, same as $\tilde{\mathsf{H}}_1(\mathbb{S}^1)$. Indeed, we have a single non-trivial relative homology class of dimension 2, namely the one generated by the disk spanned by the equator circle. The connecting homomorphism maps this class to the absolute homology class of dimension 1 generated by the circle itself.

Bibliographic notes. Relative homology groups were introduced in the 1920s by Solomon Lefschetz for application to his fixed point theorem. They seem barely more than an afterthought to absolute homology groups. Nevertheless, they have many applications, including the study of the local homology

of a space, see e.g. [3, 4], and the computation of absolute homology groups via exact sequences. Brouwer's Fixed Point Theorem impresses by its generality and is popular also outside mathematics. He proved the 3-dimensional case in 1910 [1] and the general case in 1912 [2].

[1] L. E. J. Brouwer. Über eineindeutige, stetige Transformationen von Flächen in sich. *Math. Ann.* **69** (1910), 176–180.

[2] L. E. J. Brouwer. Über Abbildungen von Mannigfaltigkeiten. *Math. Ann.* **71** (1912), 97–115.

[3] A. Hatcher. *Algebraic Topology.* Cambridge Univ. Press, England, 2002.

[4] J. R. Munkres. *Elements of Algebraic Topology.* Addison-Wesley, Redwood City, California, 1984.

IV.4 Exact Sequences

As we have seen above, long exact sequences are handy for deriving homology groups from others. In this section, we introduce a general method for constructing such sequences and use it to get a divide-and-conquer formulation of homology, known as the Mayer-Vietoris sequence.

Chain complexes and chain maps. Freeing ourselves from the simplicial complex background, we consider a sequence of vector spaces with homomorphisms between them, $\mathcal{U} = (\mathsf{U}_p, u_p)$ with $u_p : \mathsf{U}_p \to \mathsf{U}_{p-1}$. If $u_p u_{p+1} = 0$ for every p then we call $\mathcal{U}$ a *chain complex* and the u_p its *boundary maps*. The vanishing of the pairwise compositions of maps is all we need to define cycle groups, $\mathsf{Z}_p(\mathcal{U}) = \ker u_p$, boundary groups, $\mathsf{B}_p(\mathcal{U}) = \operatorname{im} u_{p+1}$, and homology groups, $\mathsf{H}_p(\mathcal{U}) = \ker u_p / \operatorname{im} u_{p+1}$, in the usual way. Of course, the best example is the chain complex of a simplicial complex, $\mathcal{C}(K) = (\mathsf{C}_p(K), \partial_d)$.

Letting $\mathcal{V} = (\mathsf{V}_p, v_p)$ be another chain complex, a *chain map* is a sequence of homomorphisms $\phi_p : \mathsf{U}_p \to \mathsf{V}_p$, one for each dimension p, that commute with the boundary maps. Specifically, $v_p \phi_p = \phi_{p-1} u_p$, for every p, but we will often drop the indices and just write $v\phi = \phi u$ to express this property. Commutativity guarantees that cycles go to cycles, $\phi_p(\mathsf{Z}_p(\mathcal{U})) \subseteq \mathsf{Z}_p(\mathcal{V})$, and boundaries go to boundaries, $\phi_p(\mathsf{B}_p(\mathcal{U})) \subseteq \mathsf{B}_p(\mathcal{V})$. Just as in the case of the induced map defined in the previous section, this implies that the chain map induces a map on homology, $(\phi_p)_* : \mathsf{H}_p(\mathcal{U}) \to \mathsf{H}_p(\mathcal{V})$, for every dimension p.

Letting $\mathcal{W} = (\mathsf{W}_p, w_p)$ be a third chain complex and the sequence of $\psi_p : \mathsf{V}_p \to \mathsf{W}_p$ a second chain map, we call $\mathcal{U} \to \mathcal{V} \to \mathcal{W}$ *exact* at $\mathcal{V}$ if $\ker \psi_p = \operatorname{im} \phi_p$ for every p. A *short exact sequence* of chain complexes is a sequence of length five,

$$0 \to \mathcal{U} \xrightarrow{\phi} \mathcal{V} \xrightarrow{\psi} \mathcal{W} \to 0,$$

that begins and ends with the trivial chain complex and is exact at $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$. Equivalently, we have a short exact sequence of vector spaces, $0 \to \mathsf{U}_p \to \mathsf{V}_p \to \mathsf{W}_p \to 0$, for each dimension p. Recall that this implies that each ϕ_p is injective, each ψ_p is surjective, and each V_p is isomorphic to the direct sum of U_p and W_p, although there is no natural choice for this isomorphism.

The snake or zig-zag. We are now ready to explain the general method for constructing long exact sequences of homology groups from short exact sequences of chain complexes.

SNAKE LEMMA. Let $0 \to \mathcal{U} \xrightarrow{\phi} \mathcal{V} \xrightarrow{\psi} \mathcal{W} \to 0$ be a short exact sequence of chain complexes. Then there is a well-defined map $D : \mathsf{H}_p(\mathcal{W}) \to \mathsf{H}_{p-1}(\mathcal{U})$, called the *connecting homomorphism*, such that

$$\ldots \to \mathsf{H}_p(\mathcal{U}) \to \mathsf{H}_p(\mathcal{V}) \to \mathsf{H}_p(\mathcal{W}) \xrightarrow{D} \mathsf{H}_{p-1}(\mathcal{U}) \to \ldots$$

is a long exact sequence.

Other than the connecting homomorphism, the maps in the long exact sequence are induced by the chain maps. Before looking at the algebraic details of the construction, let us see how the Snake Lemma gives rise to the Exact Homology Sequence of a Pair described in the previous section. We have a simplicial complex, K, and a subcomplex, $K_0 \subseteq K$. Inclusion of K_0 in K and K in K induces a short exact sequence of chain complexes,

$$0 \to \mathcal{C}(K_0) \to \mathcal{C}(K) \to \mathcal{C}(K, K_0) \to 0,$$

where $\mathcal{C}(K) = (\mathsf{C}_p(K), \partial_p)$ and similar for K_0 and (K, K_0). Indeed, $\mathcal{C}(K_0) \to \mathcal{C}(K)$ is injective and $\mathcal{C}(K) \to \mathcal{C}(K, K_0)$ is surjective. Finally, we have exactness in the middle because a chain of K is carried by K_0 iff it is zero in (K, K_0). The implied long exact sequence is the exact homology sequence of (K, K_0),

$$\ldots \to \mathsf{H}_p(K_0) \to \mathsf{H}_p(K) \to \mathsf{H}_p(K, K_0) \xrightarrow{D} \mathsf{H}_{p-1}(K_0) \to \ldots$$

As always, the crucial piece of the sequence is the connecting homomorphism. We now give a detailed description of its construction, in the general setting of the Snake Lemma. We omit the proof that the long exact sequence is in fact exact, leaving that as an exercise to the interested reader.

Connecting homomorphism. We construct D in four steps using the portion of the short exact sequence of chain complexes shown below. The vertical arrows are boundary maps and the horizontal arrows are chain maps. To simplify the discussion, we will frequently suppress the subscripts that indicate the dimensions on the maps ϕ, ψ, u, v, w as they can be determined from the domain and the clutter they introduce is more confusing that it is helpful.

$$\begin{array}{ccccccccccc}
 & & & & & & \mathsf{V}_{p+1} & \to & \mathsf{W}_{p+1} & \to & 0 \\
 & & & & & & \downarrow & \square_3 & \downarrow & & \\
0 & \to & \mathsf{U}_p & \to & \mathsf{V}_p & \to & \mathsf{W}_p & \to & 0 & & \\
 & & \downarrow & \square_2 & \downarrow & \square_0 & \downarrow & & & & \\
0 & \to & \mathsf{U}_{p-1} & \to & \mathsf{V}_{p-1} & \to & \mathsf{W}_{p-1} & \to & 0 & & \\
 & & \downarrow & \square_1 & \downarrow & & & & & & \\
0 & \to & \mathsf{U}_{p-2} & \to & \mathsf{V}_{p-2} & & & & & &
\end{array}$$

Notice the labeled commutative squares in the diagram. We will refer to them when we want to emphasize that the maps around their boundaries commute. For example, the fact that $\square_0$ is a commutative square means that $w\psi = \psi v$ as a map from V_p to W_{p-1}. With this in mind, here are the steps in establishing the connecting homomorphism.

Step 1: define γ. We begin with a cycle $\alpha \in \mathsf{W}_p$ representing a class in $\mathsf{H}_p(\mathcal{W})$. Since ψ is surjective, there exists a chain $\beta \in \mathsf{V}_p$ with $\psi(\beta) = \alpha$. Since α has zero boundary and $\square_0$ is commutative, the boundary of β lies in the kernel of the second chain map, $v(\beta) \in \ker \psi$. Exactness at V_{p-1} then implies that there exists a chain $\gamma \in \mathsf{U}_{p-1}$ whose image under the first chain map is the boundary of β, $\phi(\gamma) = v(\beta)$. We summarize the situation by extracting the relevant piece of the above diagram, and a little more:

$$\begin{array}{ccccc} & & \beta & \xrightarrow{\psi} & \alpha \\ & & \downarrow & \square_0 & \downarrow \\ \gamma & \xrightarrow{\phi} & v(\beta) & \xrightarrow{\psi} & 0 \\ \downarrow & \square_1 & \downarrow & & \\ 0 & \xrightarrow{\phi} & 0 & & \end{array}$$

Step 2: γ is a cycle. By commutativity of $\square_1$ and the fact that $vv = 0$, we have $\phi u(\gamma) = 0$. But ϕ is injective, so this implies that $u(\gamma) = 0$, which means that γ is a cycle; see the drawing above. Hence, γ represents a class in $\mathsf{H}_{p-1}(\mathcal{U})$ and this class is the image of the class represented by α under the connecting homomorphism. The map goes left, from α to β, then down to $v(\beta)$, and then left again, to γ. We may draw this as a snake or a zig-zag cutting through the diagram, thus the name. Notice, however, that we have made choices for α and β and we need to show that our answer does not depend on them.

Step 3: choice of β. Suppose first that we make another choice for β, call it β_0, and let γ_0 to be the unique element of U_{p-1} such that $\phi(\gamma_0) = v(\beta_0)$. We again summarize the situation by extracting a piece of the diagram:

$$\begin{array}{ccccc} \mu & \xrightarrow{\phi} & \beta, \beta_0 & \xrightarrow{\psi} & \alpha \\ \downarrow & \square_2 & \downarrow & \square_0 & \downarrow \\ \gamma, \gamma_0 & \xrightarrow{\phi} & v(\beta), v(\beta_0) & \xrightarrow{\psi} & 0 \end{array}$$

We have $\psi(\beta) = \psi(\beta_0) = \alpha$ and therefore $\beta + \beta_0 \in \ker \psi = \operatorname{im} \phi$, so there exists a chain $\mu \in \mathsf{U}_p$ with $\phi(\mu) = \beta + \beta_0$. Since $\square_2$ commutes, $\phi u(\mu) = \phi(\gamma) + \phi(\gamma_0)$.

But ϕ is injective, so $u(\mu) = \gamma + \gamma_0$. In words, γ and γ_0 differ by a boundary, namely $u(\mu)$, and therefore represent the same homology class.

Step 4: choice of α. Finally, we consider what happens with a different choice of α, say α_0. Let β_0 and γ_0 be defined from α_0, the same way β and γ are defined from α. Since α and α_0 are two choices of representative for the same homology class in $\mathsf{H}_p(\mathcal{W})$, there exists a chain ν in W_{p+1} such that $w(\nu) = \alpha + \alpha_0$. Since ψ is surjective, there exists a chain $\varrho \in \mathsf{V}_{p+1}$ with $\psi(\varrho) = \nu$. The situation is again summarized in a portion of the diagram:

$$\begin{array}{ccccc}
 & & \varrho & \xrightarrow{\psi} & \nu \\
 & & \downarrow & \square_3 & \downarrow \\
\mu' & \xrightarrow{\phi} & v(\varrho), \beta, \beta_0 & \xrightarrow{\psi} & \alpha, \alpha_0 \\
\downarrow & \square_2 & \downarrow & \square_0 & \downarrow \\
\gamma, \gamma_0 & \xrightarrow{\phi} & 0, v(\beta), v(\beta_0) & \xrightarrow{\psi} & 0.
\end{array}$$

By commutativity of $\square_3$, $v(\varrho)$ and $\beta+\beta_0$ both map to $\alpha+\alpha_0$. This implies that their sum lies in $\ker\psi = \operatorname{im}\phi$ and there is a chain μ' in U_p with $\phi(\mu') = v(\varrho) + \beta+\beta_0$. Using commutativity of $\square_2$ and $vv = 0$, we see that $\phi u(\mu') = v(\beta+\beta_0)$. But injectivity of ϕ implies that the preimage of $v(\beta+\beta_0)$ is $\gamma+\gamma_0$ and hence $u(\mu') = \gamma + \gamma_0$. We see that γ and γ_0 differ by a boundary and thus represent the same homology class, as required. This finishes the construction of the connecting homomorphism, D.

Mayer-Vietoris sequence. We use the Snake Lemma to derive the divide-and-conquer formulation of homology known as the Mayer-Vietoris sequence. Given two spaces, it relates their homology to the homology of the union and the intersection.

Mayer-Vietoris Sequence Theorem. Let K be a simplicial complex and K', K'' subcomplexes such that $K = K' \cup K''$. Let $A = K' \cap K''$. Then there exists a long exact sequence

$$\ldots \to \mathsf{H}_p(A) \to \mathsf{H}_p(K') \oplus \mathsf{H}_p(K'') \to \mathsf{H}_p(K) \to \mathsf{H}_{p-1}(A) \to \ldots$$

and similarly for the reduced homology groups.

Proof. On the level of chains, $\mathsf{C}_p(A)$ is a subgroup of both $\mathsf{C}_p(K')$ and $\mathsf{C}_p(K'')$. Forming the direct sums, $\mathsf{C}_p(K') \oplus \mathsf{C}_p(K'')$, for all dimensions p, we get a chain complex $\mathcal{C}(K') \oplus \mathcal{C}(K'')$ with boundary map defined componentwise. We

have two copies of $\mathsf{C}_p(A)$, and can kill one off with the image of $\mathsf{C}_p(A)$ via the diagonal, and the quotient is easily identified with $\mathsf{C}_p(K)$. Stated more formally, let $i' : A \to K'$ and $i'' : A \to K''$ be the inclusions of A, and let $j' : K' \to K$ and $j'' : K'' \to K$ be the inclusions into K. Set $i(a) = (i'(a), i''(a))$ and $j(x, y) = j'(x) + j''(y)$. Then it is not difficult to see that we have a short exact sequence of chain complexes, namely

$$0 \to \mathcal{C}(A) \xrightarrow{i} \mathcal{C}(K') \oplus \mathcal{C}(K'') \xrightarrow{j} \mathcal{C}(K) \to 0.$$

The long exact sequence implied by the Snake Lemma is the Mayer-Vietoris sequence. The above is easily adapted to the reduced sequence as well. ◻

Exactness of the Mayer-Vietoris sequence at $\mathsf{H}_p(K)$ tells us that this group is isomorphic to the image of $j_* : \mathsf{H}_p(K') \oplus \mathsf{H}_p(K'') \to \mathsf{H}_p(K)$ direct sum with the kernel of $i_* : \mathsf{H}_{p-1}(A) \to \mathsf{H}_{p-1}(K') \oplus \mathsf{H}_{p-1}(K'')$. This distinguishes two types of homology classes in K. A class in $\operatorname{im} j_*$ lives in K', in K'', or in both. A class in $\ker i_*$ corresponds to a $(p-1)$-dimensional cycle $\gamma \in A$ that bounds both in K' and K''. If we write $\gamma = \partial\alpha' = \partial\alpha''$, with α' a p-chain in K' and α'' a p-chain in K'', then $\alpha = \alpha' + \alpha''$ is a cycle in K that represents this second type of class; see Figure IV.10. It is useful to check through the four steps

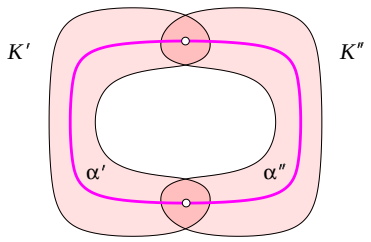

Figure IV.10: The 1-cycle α decomposes into 1-chains α' in K' and α'' in K''. The common boundary of the two 1-chains is a pair of points, a reduced 0-cycle in A.

constructing the connecting homomorphism, D. They take a class in $\mathsf{H}_p(K)$ and define one in $\mathsf{H}_{p-1}(A)$ as follows. Represent the class by α, a p-cycle of K. As before, there exists β, a p-chain in $\mathsf{C}_p(K') \oplus \mathsf{C}_p(K'')$, such that $j(\beta) = \alpha$. In fact, there are several and we get them by writing $\alpha = \alpha' + \alpha''$, with α' in K', α'' in K'', and setting $\beta = (\alpha', \alpha'')$. Different ways of decomposing α give different β, but note that any two differ by something in A. Now take $\partial\beta = (\partial\alpha', \partial\alpha'')$. The fact that α is a cycle tells us that $\partial\alpha' = \partial\alpha''$ and lies in A. Thus, the cycle γ in the construction of D is $\partial\alpha'$.

The sphere, $\mathbb{S}^d$. To illustrate the utility of the Mayer-Vietoris sequence, we use it to compute the homology of the d-dimensional sphere, $\mathbb{S}^d$. Specifically, we show that

$$\tilde{\beta}_p(\mathbb{S}^d) = \begin{cases} 1 & \text{if } p = d; \\ 0 & \text{if } p \neq d. \end{cases}$$

Writing the sphere as the union of its upper and lower hemisphere, $\mathbb{S}^d = U \cup L$, we get the equator as the intersection, $A = U \cap L$. Each hemisphere is a ball and the equator is a sphere of dimension $d-1$. This allows us to compute the homology of $\mathbb{S}^d$ inductively, using the reduced Mayer-Vietoris sequence,

$$\ldots \to \tilde{\mathsf{H}}_p(A) \to \tilde{\mathsf{H}}_p(U) \oplus \tilde{\mathsf{H}}_p(L) \to \tilde{\mathsf{H}}_p(\mathbb{S}^d) \to \tilde{\mathsf{H}}_{p-1}(E) \to \ldots$$

For $d = 0$, the sphere is two points, so its reduced homology has rank one in dimension 0, and rank zero otherwise. This established the induction basis. For general d, the sequence decomposes into pieces of the form

$$0 \oplus 0 \to \tilde{\mathsf{H}}_p(\mathbb{S}^d) \to \tilde{\mathsf{H}}_{p-1}(\mathbb{S}^{d-1}) \to 0 \oplus 0,$$

where $0 \oplus 0$ is of course the zero element in the direct sum of the homology groups of the two hemispheres. This implies that the rank of the p-th reduced homology group of $\mathbb{S}^d$ is the same as the rank of the $(p-1)$-st reduced homology group of $\mathbb{S}^{d-1}$, namely one for $p = d$ and zero otherwise, as claimed. Note that the generator of $\tilde{\mathsf{H}}_d(\mathbb{S}^d)$ is the second type of class, consisting of two chains, one from each hemisphere, whose boundary is the generating cycle of $\tilde{\mathsf{H}}_{d-1}(\mathbb{S}^{d-1})$. In particular, it is represented by the sum of all its d-dimensional simplices.

The real projective space, $\mathbb{P}^d$. As another example, we consider the real projective d-dimensional space which is the quotient space of the antipodal map, $f(x) = -x$. In other words, $\mathbb{P}^d$ is obtained by gluing $\mathbb{S}^d$ to itself by identifying antipodal points in pairs. Specifically, we show that the reduced Betti numbers are

$$\tilde{\beta}_p(\mathbb{P}^d) = \begin{cases} 1 & \text{for } 1 \leq p \leq d; \\ 0 & \text{otherwise.} \end{cases}$$

For dimensions $d = 0, 1$ we have familiar spaces, namely the point, $\mathbb{P}^0$, and the circle, $\mathbb{P}^1$. We already know their homology and their reduced Betti numbers agree with the claimed formula. This establishes the induction basis. For general d, we decompose $\mathbb{S}^d$ into three subspaces by limiting the d-th coordinate to $x_d \leq -1/2$, $-1/2 \leq x_d \leq 1/2$, and $1/2 \leq x_d$. The first and the last are identified by f and give a single subspace $B \subseteq \mathbb{P}^d$, which is a ball. The middle

subspace becomes a space M that is homotopy equivalent to the quotient space of the equator, where $x_d = 0$, which is in turn homeomorphic to $\mathbb{P}^{d-1}$. The middle subspace intersects the union of the other two in two spheres of dimension $d-1$. Taking the quotient identifies the two spheres, implying that B and M intersect in a single sphere of dimension $d-1$. Since the reduced homology of B vanishes in all dimensions p, the Mayer-Vietoris sequence decomposes into pieces of the form

$$0 \to 0 \oplus \tilde{\mathsf{H}}_p(\mathbb{P}^{d-1}) \to \tilde{\mathsf{H}}_p(\mathbb{P}^d) \to 0,$$

for $p < d-1$. By induction, this establishes $\tilde{\beta}_0(\mathbb{P}^d) = 0$ and $\tilde{\beta}_p(\mathbb{P}^d) = 1$ for $1 \leq p \leq d-1$. We still need to show that the d-th reduced Betti number is equal to one. The piece of the Mayer-Vietoris sequence we use for this is

$$0 \oplus 0 \to \tilde{\mathsf{H}}_d(\mathbb{P}^d) \xrightarrow{D} \tilde{\mathsf{H}}_{d-1}(\mathbb{S}^{d-1}) \xrightarrow{g_*} 0 \oplus \tilde{\mathsf{H}}_{d-1}(\mathbb{P}^{d-1}) \to \tilde{\mathsf{H}}_{d-1}(\mathbb{P}^d) \to 0.$$

We claim that the map g_* is 0. This will imply that D is injective as well as surjective and hence $\tilde{\beta}_d(\mathbb{P}^d) = 1$, as required. To see that g_* is zero, we use the inductive assumption, namely that $\tilde{\beta}_{d-1}(\mathbb{P}^{d-1}) = 1$. The corresponding homology group has a unique generator, namely the sum of all $(d-1)$-simplices triangulating $\mathbb{P}_{d-1}$. The map g takes each simplex in the triangulation of $\mathbb{S}^{d-1}$ to its quotient, which means each simplex in $\mathbb{P}^{d-1}$ is counted twice. The top-dimensional simplices cancel in pairs, which completes the calculation.

Bibliographic notes. The introduction of exact sequences is often contributed to Eilenberg and sometimes to Lyndon, but see also [1]. The Snake Lemma is a major achievement of algebraic topology and the construction of the connecting homomorphism is its critical piece. A complete proof can be found in many algebraic topology texts, including [3]. The Mayer-Vietoris sequences are older than the Snake Lemma and go back to work by Mayer [2] and by Vietoris [4].

[1] J. L. KELLEY AND E. PITCHER. Exact homomorphism sequences in homology theory. *Ann. of Math.* **48** (1947), 682–709.

[2] W. MAYER. Über abstrakte Topologie. *Monatschr. Math. Phys.* **36** (1929), 1–42 and 219–258.

[3] J. R. MUNKRES. *Elements of Algebraic Topology.* Addison-Wesley, Redwood City, California, 1984.

[4] L. VIETORIS. Über die Homologiegruppen der Vereinigung zweier Komplexe. *Monatschr. Math. Phys.* **37** (1930), 159–162.

Exercises

The credit assignment reflects a subjective assessment of difficulty. A typical question can be answered using knowledge of the material combined with some thought and analysis.

1. **Sperner Lemma** (three credits). Let K be a triangulated triangular region as in Figure IV.11. We 3-color the vertices such that

   - the three corners receive three different colors;
   - the vertices on each side of the region are 2-colored.

   Prove that there is a triangle in K whose vertices receive three different colors.

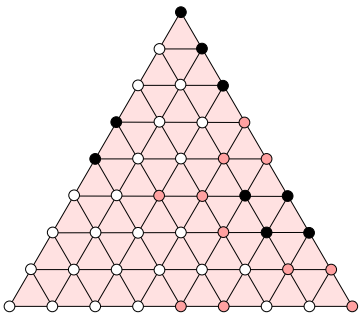

Figure IV.11: Each vertex receives one of three colors, white, shaded, or black.

2. **Isomorphic homology** (one credit). Construct two topological spaces that have isomorphic homology groups but are not homotopy equivalent.
3. **Fixed point** (two credits). Let $f : \mathbb{B}^d \to \mathbb{B}^d$ be a continuous map with the property that there is a $\delta < 1$ such that $\|f(x) - f(y)\| \leq \delta \|x - y\|$ for all points $x, y \in \mathbb{B}^d$. In words, the distance between any two points diminishes by at least a constant factor $\delta < 1$ each time we apply f. Prove that such a map f has a unique fixed point $x = f(x)$. [On orientation maps, this point is usually marked as "you are here".]
4. **Klein bottle** (one credit). Show that the Betti numbers of the 2-dimensional Klein bottle are $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$. Which other 2-manifold has the same Betti numbers?

5. **Dunce cap** (three credits). The *dunce cap* is constructed from a piece of cloth in the shape of an equilateral triangle as follows. Orienting two edges away from a common origin we glue them to each other as prescribed by their orientation. This gives a piece of a cone with a rim (the third edge) and a seam (the glued first two edges). Now we orient the rim and glue it along the seam, again such that orientations match. The result reminds us of the shell of a snail, perhaps.

   (i) Give a triangulation of the dunce cap.

   (i) Show that the reduced Betti numbers of the dunce cap vanish in all dimensions.

   (ii) Show that the dunce cap is contractible but any triangulation of it is not collapsible.

6. **3-torus** (three credits). Consider the 3-*dimensional torus* obtained from the unit cube by gluing opposite faces in pairs, without twisting. That is, each point $(x, y, 0)$ is identified with $(x, y, 1)$, $(x, 0, z)$ with $(x, 1, z)$, and $(0, y, z)$ with $(1, y, z)$. Show that the Betti numbers of this space are $\beta_0 = \beta_3 = 1$ and $\beta_1 = \beta_2 = 3$.

7. **The Steenrod Five Lemma** (two credits). Suppose we have a commutative diagram of vector spaces and homomorphisms,

$$\begin{array}{ccccccccc} \mathsf{U}_1 & \to & \mathsf{U}_2 & \to & \mathsf{U}_3 & \to & \mathsf{U}_4 & \to & \mathsf{U}_5 \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ \mathsf{V}_1 & \to & \mathsf{V}_2 & \to & \mathsf{V}_3 & \to & \mathsf{V}_4 & \to & \mathsf{V}_5, \end{array}$$

   where the horizontal sequences are exact at the middle three vector spaces and the first two and last two vertical arrows are isomorphisms. Prove that then the middle vertical arrow is also an isomorphism.

8. **Exact sequence of a triple** (one credit). Let C be a simplicial complex with subcomplexes $A \subseteq B \subseteq C$. Prove the existence of the following *exact homology sequence of the triple*:

$$\ldots \to \mathsf{H}_p(B, C) \to \mathsf{H}_p(A, C) \to \mathsf{H}_p(A, B) \to \mathsf{H}_{p-1}(B, C) \to \ldots$$

Chapter V

Duality

Instead of computing homology from a triangulation, we can also work with different decompositions and get isomorphic groups. The alpha complex and the dual Voronoi decomposition of a union of balls come to mind. Generalizing this geometric idea beyond Euclidean space, and in particular beyond manifolds, runs into difficulties. This is the motivation for taking the issue to the algebraic level, where it leads to the concept of cohomology groups. For modulo 2 arithmetic, these are isomorphic to the corresponding homology groups, but the isomorphisms are not natural. For nice topological spaces, such as manifolds and manifolds with boundary, there are relations between the homology and the cohomology groups that go beyond the general relations. In this chapter, we will see three of them, Poincaré duality, Lefschetz duality, and Alexander duality. The last of the three has algorithmic ramifications for subsets of three-dimensional Euclidean space.

V.1 Cohomology

In this section, we introduce cohomology groups. They are similar to homology groups but less geometric and motivated primarily by algebraic considerations. They belong to the standard tool-set of an algebraic topologist and appear in modern statements of the duality results discussed in the subsequent three sections.

Groups of maps. Let $\mathsf{G} = \mathbb{Z}_2$, the group of two elements, 0 and 1, together with addition modulo 2. All abelian groups we have encountered so far are vector spaces isomorphic to G^n for some integer n. Let U be such a vector space and $\varphi : \mathsf{U} \to \mathsf{G}$ a homomorphism. To define φ, it suffices to specify its values on the generators of U. If φ_0 is a second such homomorphism, their sum is defined by $(\varphi + \varphi_0)(u) = \varphi(u) + \varphi_0(u)$. This is again a homomorphism because

$$\begin{aligned}(\varphi + \varphi_0)(u + v) &= \varphi(u + v) + \varphi_0(u + v)\\ &= \varphi(u) + \varphi(v) + \varphi_0(u) + \varphi_0(v)\\ &= (\varphi + \varphi_0)(u) + (\varphi + \varphi_0)(v).\end{aligned}$$

It is easy to see that addition of homomorphisms is associative. We also have a neutral element, the zero homomorphism that sends every $u \in \mathsf{U}$ to $0 \in \mathsf{G}$, and an inverse, which for modulo 2 arithmetic is the identity, $-\varphi = \varphi$. We thus have a *group of homomorphisms* from U to G, denoted as $\mathrm{Hom}(\mathsf{U}, \mathsf{G})$. Think for example of U as the group of p-chains of a simplicial complex and $\mathrm{Hom}(\mathsf{U}, \mathsf{G})$ as the group of labelings of the p-simplices by 0 and 1. The vector spaces U and $\mathrm{Hom}(\mathsf{U}, \mathsf{G})$ are isomorphic, although the isomorphism requires us to pick a basis of U. Specifically, if U has the basis $e_1, e_2, \ldots, e_n$, then $\mathrm{Hom}(\mathsf{U}, \mathsf{G})$ has the basis $f_1, f_2, \ldots, f_n$, where $f_i(e_j) = \delta_{i,j}$, the Kronecker delta that is 1 if $i = j$ and 0 otherwise, and the isomorphism is defined by mapping e_i to f_i for all i. If we choose a different basis, the isomorphism changes. As a specific example, take $\mathsf{U} = \mathsf{G}^2$, with basis $e_1 = (1, 0)$ and $e_2 = (0, 1)$. Then the vector from the origin to (a, b) is written as $ae_1 + be_2$. The isomorphism from U to $\mathrm{Hom}(\mathsf{U}, \mathsf{G})$ takes $w = (a, b)$ to the map $f_w = af_1 + bf_2$ whose value on another vector $(x, y) = xe_1 + ye_2$ is $ax + by$. Suppose instead that we take the basis $e'_1 = e_1 + e_2$ and $e'_2 = e_2$. Then $(a, b) = ae'_1 + (b - a)e'_2$, and the new isomorphism takes $w = (a, b)$ to the map $f'_w = af'_1 + (b - a)f'_2$, whose value on the vector $(x, y) = xe'_1 + (y - x)e'_2$ is $ax + (b - a)(y - x)$. We see that f_w and f'_w assign generally different values which shows that the two isomorphisms are indeed different. This is the reason cohomology is worth defining at all, because if

there was a natural isomorphism between U and $\mathrm{Hom}(\mathsf{U}, \mathsf{G})$, the theories of homology and cohomology would be the same.

Given another vector space V and a homomorphism $f : \mathsf{U} \to \mathsf{V}$, there is an induced *dual homomorphism*, $f^* : \mathrm{Hom}(\mathsf{V}, \mathsf{G}) \to \mathrm{Hom}(\mathsf{U}, \mathsf{G})$, that maps $\psi : \mathsf{V} \to \mathsf{G}$ to the composite $f^*(\psi) = \psi \circ f : \mathsf{U} \to \mathsf{G}$. The map f^* is indeed a homomorphism since

$$\begin{aligned} f^*(\psi + \psi_0)(u) &= (\psi + \psi_0) \circ f(u) \\ &= \psi \circ f(u) + \psi_0 \circ f(u) \\ &= f^*(\psi)(u) + f^*(\psi_0)(u) \end{aligned}$$

for every $u \in \mathsf{U}$. The group of homomorphisms and the dual homomorphism can be defined for more general abelian groups U, V, and G but this will not be necessary for our purposes.

Simplicial cohomology. Let K be a simplicial complex. We construct cohomology groups by turning chain groups into groups of homomorphisms and boundary maps into their dual homomorphisms. To begin, we define a *p-cochain* as a homomorphism $\varphi : \mathsf{C}_p \to \mathsf{G}$, where $\mathsf{G} = \mathbb{Z}_2$ as before. Given a p-chain $c \in \mathsf{C}_p$, the cochain evaluates c by mapping it to 0 or 1. It is common to write this evaluation like a scalar product, $\varphi(c) = \langle \varphi, c \rangle$. Letting ℓ be the number of p-simplices σ in c with $\varphi(\sigma) = 1$, we have $\langle \varphi, c \rangle = 1$ iff ℓ is odd. Considering chains and cochains as sets, the evaluation thus distinguishes odd from even intersections.

The p-dimensional cochains form the *group of p-cochains*, $\mathsf{C}^p = \mathrm{Hom}(\mathsf{C}_p, \mathsf{G})$. Recall that the boundary map is a homomorphism $\partial_p : \mathsf{C}_p \to \mathsf{C}_{p-1}$. It thus defines a dual homomorphism, the *coboundary map*

$$\delta^{p-1} : \mathrm{Hom}(\mathsf{C}_{p-1}, \mathsf{G}) \to \mathrm{Hom}(\mathsf{C}_p, \mathsf{G}),$$

or simply $\delta : \mathsf{C}^{p-1} \to \mathsf{C}^p$. It is worth looking at this construction in more detail. Let φ be a $(p-1)$-cochain and ∂c a $(p-1)$-chain. By definition of dual homomorphism, φ applied to ∂c is the same as $\delta\varphi$ applied to c, that is, $\langle \varphi, \partial c \rangle = \langle \delta\varphi, c \rangle$. Suppose for example that φ evaluates a single $(p-1)$-simplex to one and all others to zero. Then $\delta\varphi$ evaluates all p-dimensional cofaces of this simplex to one and all others to zero. This gives a concrete interpretation of the coboundary map which will allow us to construct more elaborate examples shortly. Since the coboundary map runs in a direction opposite to the boundary map, it raises the dimension. Its kernel is the *group of cocycles* and its image

is the *group of coboundaries*,

$$\begin{aligned} \mathsf{Z}^p &= \ker \delta^p : \mathsf{C}^p \to \mathsf{C}^{p+1}, \\ \mathsf{B}^p &= \operatorname{im} \delta^{p-1} : \mathsf{C}^{p-1} \to \mathsf{C}^p. \end{aligned}$$

Recall the Fundamental Lemma of Homology according to which $\partial \circ \partial : \mathsf{C}_{p+1} \to \mathsf{C}_{p-1}$ is the zero homomorphism. We therefore have $\langle \delta\delta\varphi, c\rangle = \langle \delta\varphi, \partial c\rangle = \langle \varphi, \partial\partial c\rangle = 0$. In other words, $\delta \circ \delta : \mathsf{C}^{p-1} \to \mathsf{C}^{p+1}$ is also the zero homomorphism. Hence, the coboundary groups are subgroups of the cocycle groups and we have the familiar picture, except that the maps now go from right to left, as in Figure V.1.

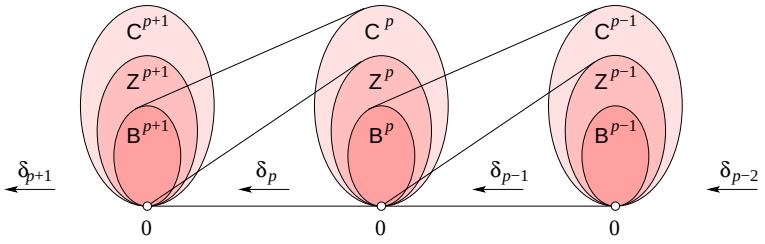

Figure V.1: The cochain complex consisting of a linear sequence of cochain, cocycle, and coboundary groups connected by coboundary homomorphisms.

DEFINITION. The *p-th cohomology group* is the quotient of the p-th cocycle group modulo the p-th coboundary group, $\mathsf{H}^p = \mathsf{Z}^p/\mathsf{B}^p$, for all p.

Reduced cohomology. Similar to homology, it is often useful to modify the definition slightly and to define the *reduced cohomology groups*, denoted as $\tilde{\mathsf{H}}^p$. Recall that for homology, this is done by introducing the augmentation map $\epsilon : \mathsf{C}_0 \to \mathbb{Z}_2$ defined by $\epsilon(u) = 1$ for each vertex u. The (-1)-st cochain group, $\mathsf{C}^{-1} = \operatorname{Hom}(\mathbb{Z}_2, \mathsf{G})$, has two elements, the map ϕ_0 mapping 1 to 0 and the map ϕ_1 mapping 1 to 1. The dual homomorphism of the augmentation map, $\epsilon^* : \operatorname{Hom}(\mathbb{Z}_2, \mathsf{G}) \to \mathsf{C}^0$, maps ϕ_0 to ψ_0, which evaluates every vertex to zero, and ϕ_1 to ψ_1, which evaluates every vertex to one. With this we have

$$\ldots \xleftarrow{\delta^1} \mathsf{C}^1 \xleftarrow{\delta^0} \mathsf{C}^0 \xleftarrow{\epsilon^*} \operatorname{Hom}(\mathbb{Z}_2, \mathsf{G}) \xleftarrow{0} 0 \xleftarrow{0} \ldots$$

Before the modification, the only 0-coboundary was the trivial 0-cochain, ψ_0. Now we have two 0-coboundaries, ψ_0 and ψ_1. The net effect of this modification

is that the rank of the zeroth cohomology group drops by one, same as the rank of the zeroth homology group when we add the augmentation map. As an exception to this rule, the ranks of H^0 and $\tilde{\mathsf{H}}^0$ are the same if C^0 is trivial, in which case $\operatorname{rank} \tilde{\mathsf{H}}^{-1} = 1$, again same as in reduced homology.

An example. To get a better feeling for cohomology, let us consider the triangulation of the annulus in Figure V.2. The 0-cochain that evaluates every single vertex to one is a 0-cocycle because every edge has exactly two vertices, which implies that the coboundary of this particular 0-cochain is the zero homomorphism. This is the only non-trivial 0-cocycle, and since for dimensional reasons there are no non-trivial 0-coboundaries, this implies that the zeroth cohomomology group, H^0, has rank one. Correspondingly, the zeroth reduced cohomology group has rank zero.

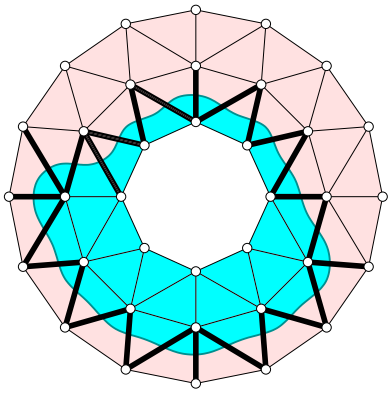

Figure V.2: The 1-cocycle is drawn by highlighting the edges it evaluates to one. They all cross the "dual" closed curve. The 1-cocycle is a 1-coboundary because it is the coboundary of the 0-cochain that evaluates a vertex to one iff it lies in the shaded region inside the closed curve.

One dimension up, we consider a 1-cochain $\varphi : \mathsf{C}_1 \to \mathsf{G}$. Its coboundary is the 2-chain $\delta\varphi : \mathsf{C}_2 \to \mathsf{G}$ that evaluates a triangle to one iff it is the coface of an odd number of edges evaluated to one by φ. Hence, φ is a 1-cocycle iff every triangle is incident to an even number of edges evaluating to one. A 1-cocycle thus looks like a picket fence; see Figure V.2. In this example, we can draw a closed curve such that an edge evaluates to one iff it crosses the curve, and a 1-chain is evaluated to the parity of the number of times it crosses that curve. (We can actually do this in general, but the curve may have more than one

component.) If the 1-chain is a 1-cycle then this number is necessarily even and the evaluation is zero. The 1-cocycle in Figure V.2 is also a 1-coboundary. To get a 1-cocycle that is not the image of a 0-cochain, we construct a picket fence that starts with an outer boundary edge of the annulus and ends with an inner boundary edge. All such picket fences are cohomologous and any one of them can be used as representative of the cohomology class that generates the first cohomology group. It follows that the rank of H^1 is one.

Another dimension up, we have $\mathsf{Z}^2 = \mathsf{C}^2$ simply because every 2-cochain maps to zero, the sole element of C^3, and is therefore also a 2-cocycle. We also have $\mathsf{B}^2 = \mathsf{C}^2$. To see this, note that the 2-cochain that evaluates a single triangle to one and all others to zero is a 2-coboundary. Indeed, we can draw three curves from a point in the interior of the triangle to the boundary of the annulus and get a "dual" 1-cochain as the sum of three picket fences, one for each curve, whose coboundary is the 2-cochain. Other 2-cochains are obtained as coboundaries of sums of such triplets of picket fences. It follows that the second cohomology group, H^2, has rank 0. Observe that the ranks of the cohomology groups are the same as the ranks of the corresponding homology groups. This is not a coincidence.

Coboundary matrix. Recall that we can get the rank of the p-th homology group from two boundary matrices transformed into normal form by row and column operations. Recall also that $\operatorname{rank} \mathsf{H}_p = \operatorname{rank} \mathsf{Z}_p - \operatorname{rank} \mathsf{B}_p$. As illustrated in Figure V.3, the right hand side of this equation is the number of zero columns in the p-th matrix minus the number of non-zero rows in the $(p+1)$-st matrix; see Figure IV.5. As we have seen earlier, a cochain evaluates a single p-simplex

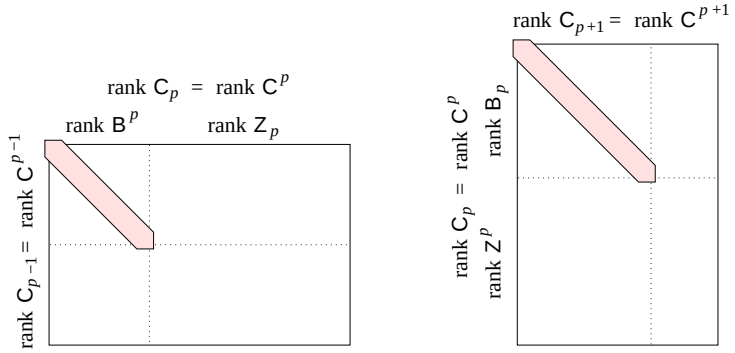

Figure V.3: The p-th and $(p+1)$-st boundary matrices in normal form. They are also the $(p-1)$-st and p-th coboundary matrices in normal form transposed.

to one and all others to zero iff its coboundary evaluates each $(p+1)$-coface of this p-simplex to one and all other $(p+1)$-simplices to zero. It follows that the coboundary matrices are the boundary matrices transposed. The normal form of the boundary matrices thus already contains the information we need to get at the ranks of the cohomology groups. Specifically, $\operatorname{rank} \mathsf{H}^p = \operatorname{rank} \mathsf{Z}^p - \operatorname{rank} \mathsf{B}^p$, the rank of the cocycle group is the number of zero rows in the $(p+1)$-st boundary matrix, and the rank of the coboundary group is the number of non-zero columns in the p-th boundary matrix, both in normal form. The number of columns of the p-th matrix is the number of rows of the $(p+1)$-st matrix, hence $\operatorname{rank} \mathsf{B}^p + \operatorname{rank} \mathsf{Z}_p = \operatorname{rank} \mathsf{Z}^p + \operatorname{rank} \mathsf{B}_p$; see Figure V.3. This implies

$$\begin{aligned} \operatorname{rank} \mathsf{H}^p &= \operatorname{rank} \mathsf{Z}^p - \operatorname{rank} \mathsf{B}^p \\ &= \operatorname{rank} \mathsf{Z}_p - \operatorname{rank} \mathsf{B}_p \quad = \quad \operatorname{rank} \mathsf{H}_p. \end{aligned}$$

Since homology and cohomology groups have the same rank, there is no concept of co-Betti number. For modulo 2 arithmetic, the rank determines the group, hence homology and cohomology groups are isomorphic, $\mathsf{H}_p \simeq \mathsf{H}^p$ for all p. This is the $\mathbb{Z}_2$-version of a standard result in algebraic topology. For more general coefficient groups, it relates the free parts and torsion parts of the homology groups with those of the cohomology groups. A more complete statement of the result for $\mathbb{Z}_2$-coefficients is the following.

Universal Coefficient Theorem. Given a topological space, $\mathbb{X}$, there are maps $\mathsf{H}^p(\mathbb{X}) \to \operatorname{Hom}(\mathsf{H}_p(\mathbb{X}), \mathsf{G}) \to \mathsf{H}_p(\mathbb{X})$ in which the first map is a natural isomorphism and the second is an isomorphism that is not natural.

We saw at the beginning of this section that the second isomorphism depends on a choice of basis and is therefore not natural. The first isomorphism does not depend on such a choice. It is natural in the sense that if $\mathbb{Y}$ is another topological space and $f : \mathbb{X} \to \mathbb{Y}$ is a continuous map, then the diagram

$$\begin{array}{ccc} \mathsf{H}^p(\mathbb{X}) & \to & \operatorname{Hom}(\mathsf{H}_p(\mathbb{X}), \mathsf{G}) \\ \uparrow & & \uparrow \\ \mathsf{H}^p(\mathbb{Y}) & \to & \operatorname{Hom}(\mathsf{H}_p(\mathbb{Y}), \mathsf{G}) \end{array}$$

of induced maps commutes. The fact that the isomorphism between H^p and $\operatorname{Hom}(\mathsf{H}_p, \mathsf{G})$ is natural is the reason there is no need to introduce a theory of co-cohomology.

Bibliographic notes. Similar to homology, cohomology is an established topic within algebraic topology today, but it took some time to become clearly

established. Cohomology has a long and complicated history with a variety of precursors that go back to Poincaré, Alexander, Lefschetz, De Rham, Pontryagin, Kolmogorov, Whitney, Čech, Eilenberg, Steenrod, Spanier and others. All these approaches were unified with the clear statement of a set of axioms that characterize homology and cohomology theories [1]. The Universal Coefficient Theorem and the duality theorems in the coming three sections were originally proven in more elementary forms before being reformulated in terms of homology and cohomology as we describe them here [2].

[1] S. EILENBERG AND N. STEENROD. *Foundations of Algebraic Topology*. Princeton Univ. Press, New Jersey, 1952.

[2] J. R. MUNKRES. *Elements of Algebraic Topology*. Addison-Wesley, Redwood City, California, 1984.

V.2 Poincaré Duality

For sufficiently nice topological spaces, there are relations between the homology and cohomology groups that go beyond the ones we have already seen. These relationships go under the name of duality. The first and most important of these is Poincaré duality, which we describe in this and the next section.

Combinatorial manifolds. In the rest of this chapter, we work only with triangulations of manifolds that satisfy a condition on the topology of the links. Specifically, a *combinatorial d-manifold* is a manifold of dimension d together with a triangulation such that the link of every i-simplex triangulates the sphere of dimension $d-i-1$. The condition implies that the closed star of every simplex has the topology of the d-dimensional ball, $\mathbb{B}^d$. To describe this in greater detail, we introduce the *join* of two topological spaces, $\mathbb{X}$ and $\mathbb{Y}$, which we denote as $\mathbb{X}*\mathbb{Y}$. Begin with the product $\mathbb{X}\times[0,1]\times\mathbb{Y}$. For each $x_0 \in \mathbb{X}$ identify all points $(x_0, 0, y)$ together, and for each $y_0 \in \mathbb{Y}$ identify all points $(x, 1, y_0)$ together. The quotient space of these identifications is $\mathbb{X} * \mathbb{Y}$. Also, $\mathbb{X}*\emptyset = \mathbb{X}$. Figure V.4 illustrates the construction by showing the *suspension* of $\mathbb{X}$, that is, the join with the 0-sphere, denoted as $\Sigma\mathbb{X} = \mathbb{X} * \mathbb{S}^0$. Geometrically, we can think of the join as a union of line segments connecting $\mathbb{X}$ to $\mathbb{Y}$ that are disjoint except possibly sharing the endpoint in $\mathbb{X}$ or the endpoint in $\mathbb{Y}$.

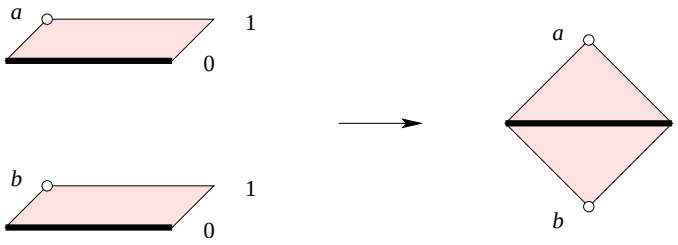

Figure V.4: Constructing the join of a line segment and a pair of points. Left: the product of the two with the unit interval. Right: the suspension obtained from the product by identification.

Returning to the definition of a combinatorial manifold, we recall that the star of a simplex, σ, consists of all simplices τ that contain σ as a face. Besides σ, each simplex in the star is the join of σ with a simplex in the link of σ. If σ is an i-simplex then $\mathrm{Lk}\,\sigma$ is a $(d-i-1)$-sphere. Taking the join, we get a d-ball, as mentioned earlier.

Exotic manifolds. Not every triangulation of a manifold satisfies the conditions on the links given above. We describe the construction of a triangulation of the 5-sphere that has a vertex whose link is not a 4-sphere. We begin with a triangulation, P, of the Poincaré 3-sphere. This space is homologically the same as but topologically different from the 3-sphere, $\mathbb{S}^3$. There are many ways to describe it. A particularly convenient way uses three complex numbers to write a point in $\mathbb{R}^6$. Letting x_1 to x_6 be the coordinates, we set $x = x_1 + ix_2$, $y = x_3 + ix_4$, $z = x_5 + ix_6$ and recall that their conjugates are $\bar{x} = x_1 - ix_2$, $\bar{y} = x_3 - ix_4$, $\bar{z} = x_5 - ix_6$. Consider the following two equations:

$$\begin{aligned} x\bar{x} + y\bar{y} + z\bar{z} &= 1; \\ x^2 + y^3 + z^5 &= 0. \end{aligned}$$

The first equation describes the 5-sphere. The second equation is really two equations, one for the real and the other for the imaginary parts, and it define a 4-dimensional space whose points have neighborhoods homeomorphic to $\mathbb{R}^4$ except at the origin, where the space is singular. The intersection of the two spaces is the Poincaré 3-sphere. It is triangulable and we let P be a triangulation of the Poincaré 3-sphere. Next, we take two suspension steps to construct a triangulation of the 5-sphere. Writing this in terms of triangulations, wet get

$$\begin{aligned} \Sigma P &= \{a, b\} \cup \{\sigma, a * \sigma, b * \sigma \mid \sigma \in P\}; \\ \Sigma^2 P &= \{u, v\} \cup \{\tau, u * \tau, v * \tau \mid \tau \in \Sigma P\}. \end{aligned}$$

The shared link of the vertices a and b in ΣP is P, which is not a triangulation of $\mathbb{S}^3$. It follows that a and b do not have neighborhoods homeomorphic to $\mathbb{R}^3$. Hence, the underlying space of ΣP is not even a manifold. Taking the suspension twice is the same as forming the join with a circle. Hence, $\Sigma^2 P$ triangulates the join of the Poincaré 3-sphere with $\mathbb{S}^1$. As it turns out, this join is homeomorphic to $\mathbb{S}^5$. The proof of this fact is not easy and omitted. But now we have a triangulation of a 5-manifold, namely $\Sigma^2 P$, that violates the condition on the links. Specifically, the shared link of the vertices u and v in $\Sigma^2 P$ is ΣP, which is not even a 4-manifold.

Dual blocks. Let now $\mathbb{M}$ be a compact, combinatorial d-manifold triangulated by K. Recall that the barycentric subdivision, $\mathrm{Sd}K$, is obtained by connecting the barycenters of the simplices in K; see Section III.1. It is not difficult to show that if K has the link property required for a combinatorial manifold, then so does $\mathrm{Sd}K$. Label each vertex in $\mathrm{Sd}K$ by the dimension of the corresponding simplex in K and note that each simplex in $\mathrm{Sd}K$ has distinct labels on its vertices. The vertex with smallest label is therefore unique. Letting u be the barycenter of σ in K, the *dual block*, denoted by $\hat{\sigma}$, is the union

of the simplices in the barycentric subdivision for which u is the vertex with minimum label; see Figure V.5. We let B be the set of dual blocks and call it the *dual block decomposition* of $\mathbb{M}$. For example, in the case of a combina-

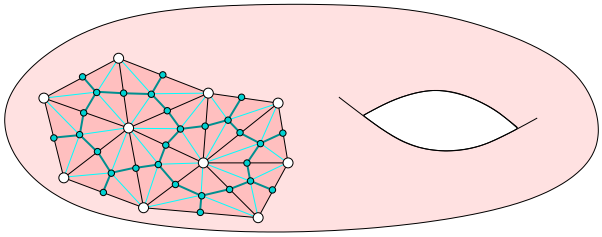

Figure V.5: A small piece of a triangulation of the torus, the barycentric subdivision, and the dual block decomposition.

torial 3-manifold, the dual blocks to a vertex, edge, triangle, and tetrahedron are, respectively, a ball, a disk, an interval, and a point. The relationship between K and B is much like that between the Delaunay triangulation and its dual Voronoi diagram. In particular, if the p-simplex σ is a face of the $(p+1)$-simplex τ, then the dual block $\hat{\sigma}$ contains $\hat{\tau}$ in its boundary. In fact, the boundary of $\hat{\sigma}$ is the union of dual blocks $\hat{\tau}$ over all proper cofaces τ of σ. We denote this boundary by $\operatorname{bd}\hat{\sigma}$, noting that $\hat{\sigma}$ is the join of $\operatorname{bd}\hat{\sigma}$ with the barycenter of σ. Since we have a combinatorial manifold, $\operatorname{bd}\hat{\sigma}$ has the topology of the $(q-1)$-sphere, where $p+q=d$.

We construct a new chain complex from the dual block decomposition as follows. Choosing complementary dimensions $p+q=d$, a *block chain* of dimension q is a formal sum $\sum a_i\hat{\sigma}_i$, where the σ_i are the p-simplices of K and the $\hat{\sigma}_i$ are the dual blocks of dimension q, with modulo 2 coefficients as usual. The collection of block chains of dimension q form an abelian group, D_q. The boundary homomorphism connecting the q-th group to the $(q-1)$-st group is defined by mapping $\hat{\sigma}_i$ to $\partial_q\hat{\sigma}_i = \sum\hat{\tau}_j$, where the sum is over all proper cofaces τ_j of σ_i whose dimension is $p+1$. The full boundary homomorphism, $\partial_q : \mathsf{D}_q \to \mathsf{D}_{q-1}$, is the linear extension to block chains. It is easy to see that $\partial_{q-1}\circ\partial_q = 0$ so that $(\mathsf{D}_q, \partial_q)$ is indeed a chain complex.

Blocks or simplices. We now have three ways to compute the homology of $\mathbb{M}$, using the simplices in K, using the simplices in $\operatorname{Sd}K$, or using the dual blocks in B. We formally prove what is to be expected, namely that $\operatorname{Sd}K$ and B give the same homology. Write $\mathcal{C} = (\mathsf{C}_p, \partial_p)$ for the chain complex defined by $\operatorname{Sd}K$ and $\mathcal{D} = (\mathsf{D}_q, \partial_q)$ for the chain complex defined by B. Mapping

each q-dimensional dual block to the sum of q-simplices it contains, we get a homomorphism $b_q : \mathsf{D}_q \to \mathsf{C}_q$. The maps b_q commute with the boundary maps and thus form a chain map between the two chain complexes, which we denote as $b : \mathcal{D} \to \mathcal{C}$.

BLOCK COMPLEX LEMMA. The chain map $b : \mathcal{D} \to \mathcal{C}$ induces $b_* : \mathsf{H}_p(\mathcal{D}) \to \mathsf{H}_p(\mathcal{C})$ which is an isomorphism for each dimension p.

PROOF. Let X_p be the subcomplex of $\mathrm{Sd}K$ consisting of all simplices that lie in blocks of dimension at most p. Clearly, $X_0 \subseteq X_1 \subseteq \ldots \subseteq X_d = \mathrm{Sd}K$. The p-th relative homology group of the pair (X_p, X_{p-1}) is isomorphic to D_p. More generally,

$$\mathsf{H}_p(X_q, X_{q-1}) \quad \simeq \quad \begin{cases} \mathsf{D}_p & \text{if } p = q; \\ 0 & \text{if } p \neq q. \end{cases}$$

Indeed, each pair $(\hat{\sigma}, \mathrm{bd}\,\hat{\sigma})$ has the homology of a q-ball relative to its boundary. Next, consider the long exact sequence of the pair (X_q, X_{q-1}),

$$\ldots \to \mathsf{H}_{p+1}(X_q, X_{q-1}) \to \mathsf{H}_p(X_{q-1}) \to \mathsf{H}_p(X_q) \to \mathsf{H}_p(X_q, X_{q-1}) \to \ldots$$

The relative groups are all zero, except possibly $\mathsf{H}_q(X_q, X_{q-1})$. Hence, the maps from $\mathsf{H}_p(X_{q-1})$ to $\mathsf{H}_p(X_q)$ are isomorphism for $p + 1 < q$. Composing these isomorphism for q from $p+2$ to d implies that $\mathsf{H}_p(X_{p+1})$ is isomorphic to $\mathsf{H}_p(\mathrm{Sd}K)$. The main tool in this proof is a two-dimensional diagram connecting pieces of the long exact sequences of the pairs (X_q, X_{q-1}) for $q = p-1, p, p+1$. We write this diagram identifying $\mathsf{H}_q(X_q, X_{q-1})$ with D_q.

$$\begin{array}{ccccccccc}
 & & \mathsf{D}_{p+1} & & & & 0 = \mathsf{H}_{p-1}(X_{p-2}) & & \\
 & & \downarrow e & \searrow & & & \downarrow & & \\
0 = \mathsf{H}_p(X_{p-1}) & \longrightarrow & \mathsf{H}_p(X_p) & \xrightarrow{f} & \mathsf{D}_p & \xrightarrow{g} & \mathsf{H}_{p-1}(X_{p-1}) & & \\
 & & \downarrow l & & & \searrow & \downarrow h & & \\
 & & \mathsf{H}_p(X_{p+1}) & & & & \mathsf{D}_{p-1} & & \\
 & & \downarrow & & & & & & \\
 & & 0 = \mathsf{H}_p(X_{p+1}, X_p) & & & & & &
\end{array}$$

We see the block chain complex run diagonally, from the upper left to the lower right in the diagram. The two triangles in the diagram commute. As mentioned above, the relative homology groups off the main diagonal are zero, which explains the trivial group at the bottom of the diagram. We also note that $\mathsf{H}_p(X_q) = 0$ for all $q < p$ simply because the dimension of X_q is less than p. This gives two additional trivial groups in the diagram.

We are now ready for some diagram chasing. The subgroup of p-cycles in D_p is the kernel of $\partial_p = h \circ g$. Since h is injective, this group is also the kernel of g. By exactness of the horizontal sequence, we have $\ker g = \operatorname{im} f$, and since f is injective, this implies that $\mathsf{H}_p(X_p)$ is isomorphic to the group of p-cycles. The subgroup of p-boundaries in D_p is the image of $\partial_{p+1} = f \circ e$. Since f is injective, this group is isomorphic to the image of e. The p-th homology group is the quotient of the two, $\mathsf{H}_p(\mathcal{D}) = \mathsf{H}_p(X_p)/\operatorname{im} e$. By exactness of the first vertical sequence, this is equal to $\mathsf{H}_p(X_p)/\ker l$. But l is surjective, so this quotient is isomorphic to $\mathsf{H}_p(X_{p+1})$ and therefore to $\mathsf{H}_p(\mathrm{Sd}K)$, as required. □

First form of Poincaré duality. There is a fairly direct translation between chains formed by dual blocks and cochains formed by the corresponding simplices. We have all results lined up to prove the main result of this section.

Poincaré Duality Theorem (first form). Let $\mathbb{M}$ be a compact, combinatorial d-manifold. Then there is an isomorphism between $\mathsf{H}_p(\mathbb{M})$ and $\mathsf{H}^q(\mathbb{M})$ for every pair of complementary dimensions $p + q = d$.

Proof. Let K be a triangulation of $\mathbb{M}$ and define q such that $p + q = d$. If σ is a q-simplex of K, let σ^* be the dual q-cochain defined by $\langle \sigma^*, \sigma \rangle = 1$ and $\langle \sigma^*, \tau \rangle = 0$ if $\tau \neq \sigma$. The map $PD_q : \mathsf{D}_p \to \mathsf{C}^q$ is defined on the chain level by setting $PD_q(\hat{\sigma}) = \sigma^*$ and extending linearly. It is, of course, an isomorphism. To prove Poincaré duality, we only need to show that PD_q commutes with boundary and coboundary:

$$PD_{q-1} \circ \partial_p \quad = \quad \delta^q \circ PD_q.$$

But this is easy since $\langle \delta^q(\sigma^*), \tau \rangle = \langle \sigma^*, \partial(\tau) \rangle = 1$ iff σ is a face of τ, which is exactly the definition of ∂_q. □

Recall that the Universal Coefficient Theorem states that $\mathsf{H}_p(\mathbb{M})$ is isomorphic to $\mathsf{H}^p(\mathbb{M})$. Together with the Poincaré Duality Theorem, we thus have $\mathsf{H}_p(\mathbb{M}) \simeq \mathsf{H}_q(\mathbb{M})$ for all $p + q = d$.

Bibliographic notes. Poincaré mentioned a form of his duality in a paper in 1893, without giving a proof. He tried a proof in his 1895 Analysis situ paper [4] based on intersection theory (see the next section), which he invented. Criticism of his work by Poul Heegard led him to realize that his proof was flawed, and he gave a new proof in two complements of the Analysis situ paper, [5, 6], now based on dual triangulations. Poincaré duality took on its modern form in the 1930s when Eduard Čech and Hassler Whitney invented the cup and cap products of cohomology.

In this book, we have assumed that we are working with combinatorial manifolds. The construction of a triangulation of the 5-sphere described in this section is due to Edwards [1]. See [7] for further exotic manifolds, including some for which all triangulations violate the condition on the links. While the restriction to combinatorial manifolds is a loss of generality, the Poincaré Duality Theorem nevertheless holds for arbitrary triangulated manifolds [3]. In fact, if we use singular homology, Poincaré duality holds for arbitrary topological manifolds and even for non-compact manifolds if we use what is called *cohomology with compact support*. A nice proof of this can be found in [2, Chapter 20].

[1] R. D. EDWARDS. Approximating certain cell-like maps by homeomorphisms. *Notices Amer. Math. Soc.* **24** (1977), A647.

[2] J. P. MAY. *A Concise Course in Algebraic Topology.* Chicago Lectures in Mathematics, Univ. Chicago Press, Chicago, Illinois, 1999.

[3] J. R. MUNKRES. *Elements of Algebraic Topology.* Addison-Wesley, Redwood City, California, 1984.

[4] H. POINCARÉ. Analysis situs. *J. Ecole Polytechn.* **1** (1895), 1–121.

[5] H. POINCARÉ. Complément à l'analysis situs. *Rend. Circ. Mat. Palermo* **13** (1899), 285–343.

[6] H. POINCARÉ. Cinquième complément a l'analysis situs. *Rend. Circ. Mat. Palermo* **18** (1904), 45–110.

[7] A. A. RANICKI (EDITOR). *The Hauptvermutung Book.* Kluwer, Dordrecht, the Netherlands, 1996.

V.3 Intersection Theory

There is a second version of Poincarè duality which can be stated purely in terms of homology. It is based on an intersection pairing between homology classes of complementary dimensions introduced in this section.

Counting intersections modulo 2. Let $\mathbb{M}$ be a combinatorial manifold of dimension d and K a triangulation of $\mathbb{M}$. Furthermore, let p and q be integers such that $p + q = d$. As explained in Section V.2, if σ is a p-simplex in K then its dual block, $\hat{\sigma}$, is q-dimensional. The two meet in a single point, the barycenter of σ. If τ is another p-simplex then $\sigma \neq \tau$ implies that σ and $\hat{\tau}$ are disjoint. We therefore define

$$\sigma \cdot \hat{\tau} \quad = \quad \begin{cases} 1 & \text{if } \sigma = \tau; \\ 0 & \text{if } \sigma \neq \tau. \end{cases}$$

We are mainly interested in intersections of cycles. Suppose that $c = \sum_i a_i \sigma_i$ is a p-cycle in K and $d = \sum_j b_j \hat{\tau}_j$ is a q-cycle in the dual block decomposition. Then the *intersection number* of the two cycles is

$$c \cdot d \quad = \quad \sum_{i,j} a_i b_j (\sigma_i \cdot \hat{\tau}_j),$$

counting the intersections modulo 2. In other words, $c \cdot d = 0$ if the two cycles are disjoint or meet in an even number of points, and $c \cdot d = 1$ if they meet in an odd number of points. As an example, consider the center circle of the Möbius strip and a pulled off copy, that is, a nearby closed curve that meets the center circle in a finite number of points, as sketched in Figure V.6. The

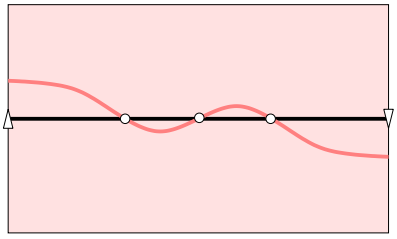

Figure V.6: The black center circle of the Möbius strip intersects the gray pulled off copy in three points.

topology of the Möbius strip forces an odd number of intersections. This is

unlike the orientable case in which a pulled off closed curve meets the original in an even number of points.

It is not difficult to show that if we replace c or d by a homologous cycle, then the intersection number does not change. For example, if $c \sim c_0$, we consider the intersection of d with a $(p+1)$-chain γ in K for which $\partial\gamma = c + c_0$. Let τ be a $(p+1)$-simplex of γ and $\hat{\sigma}$ a block of dimension $q = d-p$. The key observation is that τ and $\hat{\sigma}$ are disjoint unless σ is a face of τ in which case they intersect in the edge connecting the barycenter of τ to the barycenter of σ. Completing the intersection between γ and d, the edge extends to either a closed curve or a path with two endpoints. These points either lie both on c, or both on c_0, or one on c and the other on c_0. The total number of endpoints is even, which implies that the intersection numbers are the same, that is, $c \cdot d = c_0 \cdot d$.

Pairings. Since the intersection number is invariant under choosing different representatives of a homology class, we have a map $\# : \mathsf{H}_p(\mathbb{M}) \times \mathsf{H}_q(\mathbb{M}) \to \mathsf{G}$ defined by $\#(\gamma, \delta) = c \cdot d$, where c and d are representative cycles of γ and δ. We call this map the *intersection pairing* of the homology groups, where $p + q = d$, as before. Using the same notation as for simplices and cycles, we write $\gamma \cdot \delta = \#(\gamma, \delta)$ and call it the *intersection number* of $\gamma \in \mathsf{H}_p(\mathbb{M})$ and $\delta \in \mathsf{H}_q(\mathbb{M})$. The pairing is bilinear and symmetric, that is,

$$\begin{aligned} (a\gamma + a_0\gamma_0) \cdot \delta &= a(\gamma \cdot \delta) + a_0(\gamma_0 \cdot \delta); \\ \gamma \cdot (b\delta + b_0\delta_0) &= b(\gamma \cdot \delta) + b_0(\gamma \cdot \delta_0); \\ \gamma \cdot \delta &= \delta \cdot \gamma. \end{aligned}$$

Since we work modulo 2, we do not have to worry about orientations of simplices and manifolds. To define intersection theory over an arbitrary field, we would need to deal with this issue, and the intersection number would be an element of the field. In this case, bilinearity still holds but symmetry does not. Indeed, if γ is p-dimensional and δ is q-dimensional then $\gamma \cdot \delta = (-1)^{pq}(\delta \cdot \gamma)$.

Pairings can be defined more generally. For example, let U and V be vector spaces over $\mathsf{G} = \mathbb{Z}_2$. A bilinear pairing $\# : \mathsf{U} \times \mathsf{V} \to \mathsf{G}$ gives a natural homomorphism $\phi_\# : \mathsf{V} \to \mathrm{Hom}(\mathsf{U}, \mathsf{G})$ defined by $\phi_\#(v) = f_v$, where $f_v(u) = u \cdot v$. The pairing is *perfect* if for every non-zero $u \in \mathsf{U}$ there exists at least one $v_0 \in \mathsf{V}$ with $\#(u, v_0) = 1$ and, symmetrically, for every non-zero $v \in \mathsf{V}$ there exists at least one $u_0 \in \mathsf{U}$ with $\#(u_0, v) = 1$.

PERFECT PAIRING LEMMA. The pairing $\# : \mathsf{U} \times \mathsf{V} \to \mathsf{G}$ is perfect iff the implied natural homomorphism $\phi_\# : \mathsf{V} \to \mathrm{Hom}(\mathsf{U}, \mathsf{G})$ is an isomorphism.

PROOF. Suppose first that $\phi_\#$ is an isomorphism. If we take $v \neq 0$, then since $\phi_\#$ is injective, $f_v \neq 0$, which means there is at least one u_0 with $\#(u_0, v) = 1$. Furthermore, if $u \neq 0$, since $\phi_\#$ is surjective, there is a $v_0 \in V$ with $\phi_\#(v_0) = u^*$, and this means that $\#(u, v_0) = 1$.

Conversely, suppose that the paring is perfect. The map $\phi_\#$ is injective because if $f_v = 0$, then $\#(u, v) = 0$ for every u, so $\#$ perfect gives $v = 0$. Note that this implies $\operatorname{rank} \mathsf{V} \leq \operatorname{rank} \operatorname{Hom}(\mathsf{U}, \mathsf{G}) = \operatorname{rank} \mathsf{U}$. The similarly defined map from U to $\operatorname{Hom}(\mathsf{V}, \mathsf{G})$ is injective by the analogous argument, which implies $\operatorname{rank} \mathsf{U} \leq \operatorname{rank} \operatorname{Hom}(\mathsf{V}, \mathsf{G}) = \operatorname{rank} \mathsf{V}$. Thus $\phi_\#$ is an injective map between vector spaces of the same dimension, which implies it is an isomorphism. □

Since V and $\operatorname{Hom}(\mathsf{U}, \mathsf{G})$ are isomorphic, this implies that U and V are isomorphic. However, this isomorphism depends on a choice of basis.

Intersection and cohomology. We can define the Poincaré duality map using intersection numbers. Indeed, if σ is a p-simplex of K and $\hat{\sigma}$ is its dual block of dimension q, then $PD_q(\hat{\sigma}) = \sigma^*$. That is, $PD_q(\hat{\sigma})$ is the p-dimensional cochain for which

$$\langle \sigma^*, \tau \rangle \quad = \quad \begin{cases} 1 & \text{if } \sigma = \tau; \\ 0 & \text{if } \sigma \neq \tau. \end{cases}$$

Since the same holds for intersection numbers, we have $\langle PD_q(\hat{\sigma}), \tau \rangle = \hat{\sigma} \cdot \tau$. By linear extension, this formula holds for chains, and since it is the same for different representatives of the same class, the formula also holds for the induced map on homology, that is,

$$\langle PD_*(\gamma), \delta \rangle \quad = \quad \gamma \cdot \delta.$$

Using this formula, there is a second version of Poincaré duality.

POINCARÉ DUALITY THEOREM (SECOND FORM). Let $\mathbb{M}$ be a compact, combinatorial d-manifold. Then the pairing $\# : \mathsf{H}_p(\mathbb{M}) \times \mathsf{H}_q(\mathbb{M}) \to \mathsf{G}$ defined by $\#(\gamma, \delta) = \gamma \cdot \delta$ is perfect for all integers $p + q = d$.

The proof follows from the first form and is omitted.

The torus and the Klein bottle. To illustrate Poincaré duality formulated in terms of intersection numbers, we now consider the two examples sketched in Figure V.7. For the 2-dimensional torus, $\mathbb{S}^1 \times \mathbb{S}^1$, the most interesting case is in dimension 1 for which the second form of Poincaré duality gives a perfect pairing

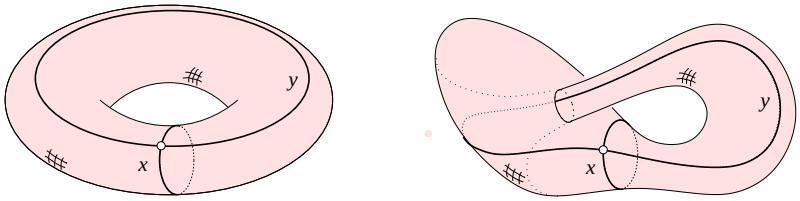

Figure V.7: The meridian and longitudinal curves of the torus on the left and of the Klein bottle on the right.

$\# : \mathsf{H}_1 \times \mathsf{H}_1 \to \mathsf{G}$. Natural generators of H_1 are the *meridian curve*, x, which bounds a disk in the solid region enclosed by the torus, and the *longitudinal curve*, y, which meets x in a single point and does not bound. The intersection numbers are easy to compute. Pushing off x and y give homologous closed curves that are disjoint from the originals or meet them in an even number of points. Hence, the intersection numbers between x and x and between y and y vanish and the intersection number between x and y is one, see Table V.1 on the left. Note that the determinant of the matrix of intersection numbers is one.

The modulo 2 homology of the Klein bottle is the same as that of the torus. However, the intersection pairing on H_1 is different. Like for the torus, we can take two curves x and y that generate H_1 with $x \cdot y = y \cdot x = 1$ and $x \cdot x = 0$. However, a neighborhood of the curve y is a Möbius strip, so pushing off y gives a closed curve that intersects y and odd number of times, that is, $y \cdot y = 1$. If we change the basis, we still do not get the same matrix as that of the torus. Once again, the matrix of intersection numbers, given in Table V.1 on the right, has determinant one.

	x	y	x	y
x	0	1	0	1
y	1	0	1	1

Table V.1: The intersection numbers of the meridian and the longitudinal curves for the torus on the left and the Klein bottle on the right.

Euler characteristic. By the Euler-Poincaré Theorem, the Euler characteristic of any space is the alternating sum of its Betti numbers. Letting $\mathbb{M}$ be a compact, combinatorial d-manifold, the Poincaré Duality and the Universal

Coefficient Theorems imply $\beta_i = \beta_{d-i}$ for all i. For odd d, this gives

$$\chi(\mathbb{M}) \quad = \quad \beta_0 - \beta_1 + \ldots + \beta_{d-1} - \beta_d,$$

which vanishes. For even d, this tells us that the terms above and below half the dimension contribute equal amounts to the Euler characteristic. Writing $d = 2k$, this gives

$$\chi(\mathbb{M}) \quad = \quad 2[\beta_0 - \beta_1 + \ldots \pm \beta_{k-1}] \mp \beta_k.$$

It follows the Euler characteristic is even iff β_k is even. However, the group $\mathsf{H}_k(\mathbb{M})$ is paired with itself and is therefore self-dual. If $\mathbb{M}$ is orientable, this can be used to show that β_k is indeed even, and so is the Euler characteristic. In contrast, homology and cohomology modulo 2 does not capture this subtlety.

Manifolds with boundary. If $\mathbb{M}$ is a manifold with boundary, Poincaré duality does not hold. For example, if we take the ball, $\mathbb{B}^d$, its 0-dimensional homology has rank one while its d-dimensional homology vanishes. There is a form for manifolds with boundary, however, called Lefschetz duality, which reduces to Poincaré duality when the boundary is empty. It relates an absolute homology or cohomology group to a relative one. Returning to our example, note that $\mathsf{H}_0(\mathbb{B}^d)$ and $\mathsf{H}_d(\mathbb{B}^d, \mathbb{S}_{d-1})$ both have rank one.

Lefschetz Duality Theorem (first form). Let $\mathbb{M}$ be a compact, combinatorial d-manifold with boundary $\partial\mathbb{M}$. Then for every pair of complementary dimensions $p + q = d$, there are isomorphisms $\mathsf{H}_p(\mathbb{M}, \partial\mathbb{M}) \simeq \mathsf{H}^q(\mathbb{M})$ and $\mathsf{H}_p(\mathbb{M}) \simeq \mathsf{H}^q(\mathbb{M}, \partial\mathbb{M})$.

Again this can be combined with the Universal Coefficient Theorem, $\mathsf{H}_p(\mathbb{M}) \simeq \mathsf{H}^p(\mathbb{M})$, to see that $\mathsf{H}_p(\mathbb{M}, \partial\mathbb{M}) \simeq \mathsf{H}_q(\mathbb{M})$ for all $p + q = d$. The proof of the Lefschetz Duality Theorem follows that of the Poincaré Duality Theorem exactly, inserting relative chains and cochains where needed. We omit the proof. There is also a second version of Lefschetz duality based on the extension of the intersection pairing to a pairing between absolute and relative classes. Again we omit the details and the proof.

Lefschetz Duality Theorem (second form). Let $\mathbb{M}$ be a compact, combinatorial d-manifold with boundary $\partial\mathbb{M}$. Then the intersection pairing $\# : \mathsf{H}_p(\mathbb{M}) \times \mathsf{H}_q(\mathbb{M}, \partial\mathbb{M}) \to \mathsf{G}$ is perfect for all $p + q = d$.

We illustrate Lefschetz duality formulated in terms of intersection numbers for the capped torus sketched in Figure V.8. Being homeomorphic to the

cylinder, the first homology group of the capped torus has a single generator, the meridian curve of the full torus. Similarly, the first relative homology group has a single generator, namely the portion of the longitudinal curve connecting points on the two boundary circles; see Figure V.8.

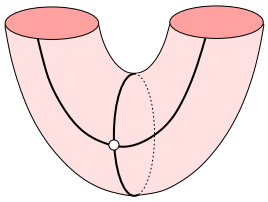

Figure V.8: The displayed generators of the first absolute and first relative homology groups of the capped torus meet in a single point.

Bibliographic Notes. Henry Poincaré invented intersection theory to prove his duality theorem in 1895 [3], but this attempt failed. It is also said that Alexander and Lefschetz founded the intersection theory of cycles on manifolds in the 1920s. Their theory was one of the precursors of cohomology. The Lefschetz Duality Theorem dates back to the 1920s when Solomon Lefschetz introduced it along with the concept of relative homology [1]. A good modern account can be found in [2].

[1] S. LEFSCHETZ. Manifolds with a boundary and their transformations. *Trans. Amer. Math. Soc.* **29** (1927), 429–462.

[2] C. R. F. MAUNDER. *Algebraic Topology.* Cambridge Univ. Press, England, 1980.

[3] H. POINCARÉ. Analysis situs. *J. Ecole Polytechn.* **1** (1895), 1–121.

V.4 Alexander Duality

Prisons in d-dimensional space are made of $(d-1)$-dimensional walls. This is because a wall of dimension $d-2$ or less cannot separate any portion of space from the rest. The topic of this section is a formal expression of a generalization of this statement and its use in the design of a fast algorithm for homology.

The theorem. The complement of a 2-sphere in the 3 sphere consists of two balls and thus has homology only in dimension 0. The complement of the torus, however, is two solid torii, and each of these has homology in both dimensions 0 and 1. This suggests a relationship between the homology of a subspace and its complement. In the general case of one submanifold in another such a statement exists, but the most famous and prettiest case is the case where the manifold is the sphere. We state that here.

Alexander Duality Theorem. Let K be a triangulation of $\mathbb{S}^d$ and $\mathbb{X} \subseteq \mathbb{S}^d$ be triangulated by a non-empty subcomplex $L \subseteq K$. Then $\tilde{\mathsf{H}}_p(\mathbb{X}) \simeq \tilde{\mathsf{H}}^{d-p-1}(\mathbb{S}^d - \mathbb{X})$.

Proof. We prove the claim using Lefschetz duality, excision, and the exact sequence of a pair. We begin by constructing a regular neighborhood N of the space $\mathbb{X}$. Consider the second barycentric subdivision of K, $\mathrm{Sd}^2 K = \mathrm{Sd}(\mathrm{Sd} K)$. Define N to be the closed star of L in that subdivision and E to be the closure of the complement of N in the same subdivision. It can be shown that K is a deformation retract of N, and E is a deformation retract of $\mathbb{S}^d - \mathbb{X}$. The complexes E and N share the same boundary which is the link of L in $\mathrm{Sd}^2 K$.

We now prove Alexander duality for $0 \leq p < d-1$ by showing the following chain of isomorphisms:

$$\tilde{\mathsf{H}}^{d-p-1}(\mathbb{S}^d - \mathbb{X}) \simeq \tilde{\mathsf{H}}^{d-p-1}(E) \simeq \mathsf{H}^{d-p-1}(E) \simeq \mathsf{H}_{p+1}(E, \partial E)$$

$$\simeq \tilde{\mathsf{H}}_{p+1}(E, \partial E) \simeq \tilde{\mathsf{H}}_{p+1}(\mathbb{S}^d, N) \simeq \tilde{\mathsf{H}}_p(N) \simeq \tilde{\mathsf{H}}_p(\mathbb{X}).$$

The first isomorphism follows from the fact that $\mathbb{S}^d - \mathbb{X}$ deformation retracts onto E. The second follows because cohomology and reduced cohomology are the same in dimensions greater than zero. The third isomorphism is Lefschetz duality for E, which is a d-manifold with boundary. The fourth follows like the second, this time for homology. The fifth isomorphism is excision, where we excise the interior of N to see that the inclusion of pairs $(E, \partial E) \to (\mathbb{S}^d, N)$ induces an isomorphism on homology. For the sixth we notice that the map $\tilde{\mathsf{H}}_{p+1}(\mathbb{S}^d, N) \to \tilde{\mathsf{H}}_p(N)$ is the connecting map in the reduced exact sequence

of the pair $(\mathbb{S}^d, N)$. Since $p+1 < d$, $\tilde{\mathsf{H}}_{p+1}(\mathbb{S}^d) = \tilde{\mathsf{H}}_p(\mathbb{S}^d) = 0$, we get the isomorphism. The final one follows from the fact that N deformation retracts onto L.

When $p = d-1$, the difference is in Lefschetz duality for E and in the fact that $\tilde{\mathsf{H}}^d(S^d)$ has rank one. We have

$$\mathsf{H}^0(\mathbb{S}^d - X) \simeq \mathsf{H}^0(E) \simeq \mathsf{H}_d(E, \partial E)$$

$$\simeq \mathsf{H}_d(\mathbb{S}^d, N) \simeq \mathsf{H}_{d-1}(N) \oplus \mathsf{G} \simeq \mathsf{H}_{d-1}(X) \oplus \mathsf{G} \simeq \tilde{\mathsf{H}}_{d-1}(X) \oplus \mathsf{G},$$

and $\mathsf{H}^0(\mathbb{S}^d - X) \simeq \tilde{\mathsf{H}}^0(\mathbb{S}^d - X) \oplus \mathsf{G}$ The extra copy of G is easily seen to match up, as it is the generator of $H^0(\mathbb{S}^d) \simeq H_n(\mathbb{S}^d)$. This implies the result. □

Knots in S^3. Let $N \subset S^3$ be a submanifold homeomorphic to a circle. We think of N as obtained by gluing the ends of a piece of string that we have tied in a knot, and so call N a knot itself. Define the *exterior* X of N to be the closure of the complement $S^3 - N$. In studying knots, we look for topological invariants of the X of N. By Alexander duality, $\tilde{\mathsf{H}}_1(X) \simeq \tilde{\mathsf{H}}^1(N) \simeq \mathsf{G}$, and the other reduced groups of X are 0. Thus homology doesn't distinguish knots in S^3! What is needed is the fundamental group of X, and we refer the reader to other texts on topology to learn about this.

Incremental algorithm. [[Explain the computation of homology by adding one simplex at a time.]]

[[When adding the simplex σ_i to K_{i-1} we define $K = K_{i-1} \cup \{\sigma\}$, $K_0 = K_{i-1}$, and consider the exact homology sequence of the pair (K, K_0),

$$\ldots \to \mathsf{H}_p(K_0) \to \mathsf{H}_p(K) \xrightarrow{g_*} \mathsf{H}_{p+1}(K, K_0) \to \mathsf{H}_{p-1}(K_0) \to \ldots$$

All the relative homology groups are zero except for $\mathsf{H}_p(K, K_0)$, where $p = \dim \sigma_i$. Then there are two cases, namely that g_* is the zero homomorphism or it is injective. In the first case, the rank of the $(p-1)$-st homology group drops by one. In the second case, the rank of the p-th homology group increases by one.]]

Union-find for first homology. [[Explain briefly how the union-find data structure can be used to maintain the rank of the first homology group in $\alpha(m)$ time per vertex and edge.]]

Alexander duality for second to the last homology. [[Reduce the $(d-1)$-st homology to the first homology formulated in terms of dual blocks.]]

[[In summary, we have an algorithm that computes the Betti numbers of a simplicial complex in $\mathbb{S}^3$ in time proportional to $m\alpha(m)$, where m is the number of simplices.]]

Bibliographic note. [[We may think of the Alexander Duality Theorem as a generalization of the Jordan Curve Theorem proved by C. Jordan in 1892.]]

[[Alexander duality was presaged by work of J. W. Alexander in 1915. This was later further developed, in particular by P. S. Alexandrov and Lev Pontryagin.]]

[[Reference Delfinado and Edelsbrunner [2] for incremental algortihm.]]

[1] J. W. Alexander. A proof of the invariance of certain constants of analysis situ. *Trans. Amer. Math. Soc.* **16** (1915), 148–154.

[2] C. J. A. Delfinado and H. Edelsbrunner. An incremental algorithm for Betti numbers of simplicial complexes on the 3-sphere. *Comput. Aided Geom. Design* **12** (1995), 771–784.

Exercises

The credit assignment reflects a subjective assessment of difficulty. A typical question can be answered using knowledge of the material combined with some thought and analysis.

1. **Coboundary** (one credit). Prove that the coboundary map can be thought of as taking each simplex to its cofaces of one dimension higher. Formally, $\langle \delta\varphi, \tau \rangle = 1$ iff $\langle \varphi, \sigma \rangle = 1$ for an odd number of faces σ of τ with dimension $\dim \sigma = \dim \tau - 1$.
2. **Universal Coefficient Theorem** (two credits). Let $\varphi \in \mathsf{Z}^p$ be a cocycle representing a cohomology class $\gamma \in \mathsf{H}^p$ and let $c \in \mathsf{Z}_p$ be a cycle representing a homology class $\alpha \in \mathsf{H}_p$. Let $j : \mathsf{H}^p \to \mathrm{Hom}(\mathsf{H}_p, \mathbb{Z}_2)$ be defined so that $j(\gamma)$ applied to α is equal to $\langle \varphi, c \rangle$.
   (i) Show that j is well defined, that is, it does not depend on the representatives chosen for γ and α.
   (ii) Show that j is an isomorphism.
3. **Dual vector spaces** (two credits), Let U be a vector space over $\mathsf{G} = \mathbb{Z}_2$ and $\mathsf{U}^* = \mathrm{Hom}(\mathsf{U}, \mathsf{G})$ be its dual.
   (i) Show that U^* is also a vector space and U and U^* are isomorphic. However, note that the isomorphism between U and U^* depends on a choice of basis and is thus not natural.

   Let $(\mathsf{U}^*)^* = \mathrm{Hom}(\mathsf{U}^*, \mathsf{G})$ be the dual of the dual of U. Let $j : \mathsf{U} \to (\mathsf{U}^*)^*$ be defined by mapping $u \in \mathsf{U}$ to $j(u) = \phi \in (\mathsf{U}^*)^*$ such that $\phi(f) = f(u)$ for every $f \in \mathsf{U}^*$.
   (ii) Prove that j is an isomorphism.
4. **Poincaré Duality** (two credits). Use the Perfect Pairing Lemma to prove the first form from the second form of the Poincaré Duality Theorem.
5. **Duality on Torus** (how many credits?).
6. **Poincaré Duality** (how many credits?). Show Poincaré Duality without assuming a PL triangulation.

However, there is a weaker property that is true and suffices for the correctness of the above argument on Euler characteristics. Letting D be a $(d-j)$-dimensional block, we write $\bar{D}$ for its closure and $\dot{D} = \bar{D} - D$ for its boundary. Then the relative homology of the pair $(\bar{D}, \dot{D})$ is that of the $(d-j)$-dimensional

ball relative its boundary, namely $\mathsf{H}_p(\bar{D}, \dot{D}) \simeq \mathbb{Z}_2$ if $p = d - j$ and it vanishes if $p \neq d - j$. This property of blocks can be used to prove the following striking symmetry of manifolds.

Intersection theory extended. We don't need $p + q = d$ in order to define intersections. In fact, if σ_1 and ${\sigma^*}_2$ are a p simplex and a dual q block, where σ_2 is a $d - p$ simplex, then the intersection $\sigma_1 \cap {\sigma^*}_2$ is the union of all $p + q - d$ in $\mathrm{Sd}\sigma_1 \subset \mathrm{Sd}K$ whose

Intersection theory for even-dimensional manifolds. For $2d$-dimensional manifolds, the intersection pairing in dimension d, $\mathsf{H}_d(\mathbb{M}) \times \mathsf{H}_d(\mathbb{M}) \to G$ is an important invariant. Choosing a basis of $\mathsf{H}_d(\mathbb{M})$, this bilinear map can be represented by a $\beta_d \times \beta_d$ matrix I, via the formula $v \cdot w = v^t I w$. Poincaré duality implies that this matrix is non-degenerate, which in turns tells us that it is non-singular (has determinant one).

Chapter VI

Morse Functions

The class of real-valued functions on a manifold is an unwieldy animal, and restricting it to continuous functions does not do a whole lot to tame it. Even smooth functions can be rather complicated in their behavior and it is best to add another requirement, namely genericity. What we get then is the class of Morse functions, which distinguishes itself by having only simple critical points. Most of the theory is concerned with the study of these critical points, their structure, and what they say about the manifold and the function. In spite of the fact that we rarely find Morse functions in actual applications, or smooth functions for that matter, knowing about their structure significantly benefits our understanding of general, smooth functions and even piecewise linear functions, as we will see.

VI.1 Generic Smooth Functions

Many questions in the sciences and engineering are posed in terms of real-valued functions. General such functions are a nightmare and continuous functions are not much better. Even smooth functions can be exceedingly complicated but when they are restricted to being generic they become intelligible.

The upright torus. We start with an example that foreshadows many of the results on generic smooth functions in an intuitive manner. Let $\mathbb{M}$ be the two-dimensional torus and $f(x)$ the height of the point $x \in \mathbb{M}$ above a horizontal plane on which the torus rests, as in Figure VI.1. We call $f : \mathbb{M} \to \mathbb{R}$ a *height*

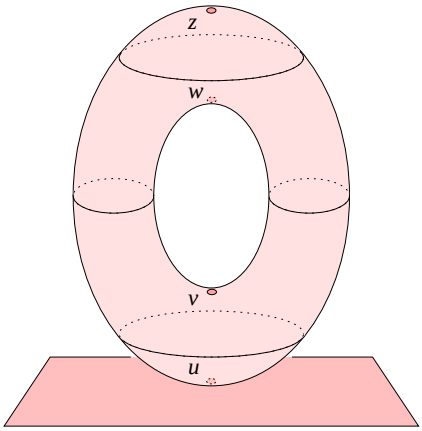

Figure VI.1: The vertical height function on the torus with critical points u, v, w, z and level sets between their height values.

function. Each real number a has a preimage, $f^{-1}(a)$, which we refer to as a *level set*. It consists of all points $x \in \mathbb{M}$ at height a. Accordingly, the *sublevel set* consists of all points at height at most a,

$$\mathbb{M}_a \quad = \quad f^{-1}(-\infty, a] \quad = \quad \{x \in \mathbb{M} \mid f(x) \leq a\}.$$

We are interested in the evolution of the sublevel set as we increase the threshold. Critical events occur when a passes the height values of the points u, v, w, z in Figure VI.1. For $a < f(u)$, the sublevel set is empty. For $f(u) < a < f(v)$, it is a disk, which has the homotopy type of a point. For $f(v) < a < f(w)$, the sublevel set is a cylinder. It has the homotopy type of a circle. We imagine it

obtained by gluing the two ends of an interval to the disk which is then shrunk to a point. For $f(w) < a < f(z)$, the sublevel set is a capped torus. It has the homotopy type of a figure-8 obtained by gluing the two ends of another interval to the cylinder which is then shrunk to a circle. Finally, for $f(z) < a$, we have the complete torus. It is obtained by gluing a disk to the capped torus. Figure VI.2 illustrates the three intermediate stages of the evolution. We need background in differential topology to explain in what sense this evolution of the sublevel set is representative of the general situation.

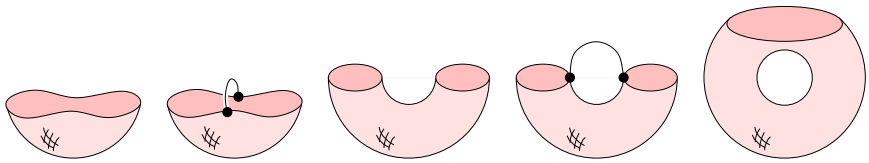

Figure VI.2: Going from a disk to a cylinder is homotopically the same as attaching a 1-handle. Similarly, going from the cylinder to the capped torus is homotopically the same as attaching another 1-handle.

Smooth functions. Let $\mathbb{M}$ be a smooth d-manifold, that is, $\mathbb{M}$ has an atlas of coordinate charts each diffeomorphic to an open ball in $\mathbb{R}^d$. We recall that a diffeomorphism is a homeomorphism that is smooth in both directions. Technically, being smooth means that derivatives of all orders exist. Practically, we just need derivatives of first and second order for most of the things we do, but it is easier to assume than to keep books. Denote the tangent space at a point $x \in \mathbb{M}$ by $\mathrm{T}\mathbb{M}_x$. It is the d-dimensional vector space consisting of all tangent vectors of $\mathbb{M}$ at x. A smooth mapping to another smooth manifold, $f : \mathbb{M} \to \mathbb{N}$, induces a linear mapping between the tangent spaces, the derivative $\mathrm{D}f_x : \mathrm{T}\mathbb{M}_x \to \mathrm{T}\mathbb{N}_{f(x)}$. We are primarily interested in real-valued functions for which $\mathbb{N} = \mathbb{R}$. Accordingly, we have linear maps $\mathrm{D}f_x : \mathrm{T}\mathbb{M}_x \to \mathrm{T}\mathbb{R}_{f(x)}$. The tangent space at a point of the real line is again a real line, so this is just a fancy way of saying that the derivatives are real-valued linear maps on the tangent spaces. Being linear, the image of such a map is either the entire line or just zero. We call $x \in \mathbb{M}$ a *regular point* of f if $\mathrm{D}f_x$ is surjective and we call x a *critical point* of f if $\mathrm{D}f_x$ is the zero map. If we have a local coordinate system $(x_1, x_2, \ldots, x_d)$ in a neighborhood of x then x is critical iff all its partial derivatives vanish,

$$\frac{\partial f}{\partial x_1}(x) \;=\; \frac{\partial f}{\partial x_2}(x) \;=\; \ldots \;=\; \frac{\partial f}{\partial x_d}(x) \;=\; 0.$$

The image of a critical point, $f(x)$, is called a *critical value* of f. All others are *regular values* of f. We use second derivatives to further distinguish between different types of critical points. The *Hessian* of f at the point x is the matrix of second derivatives,

$$H(x) \quad = \quad \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_d}(x) \\ \vdots & \ddots & \vdots & \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(x) & \frac{\partial^2 f}{\partial x_d \partial x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_d \partial x_d}(x) \end{bmatrix}.$$

A critical point x is *non-degenerate* if the Hessian is non-singular, that is, $\det H(x) \neq 0$. The points u, v, w, z in Figure VI.1 are examples of non-degenerate critical points. Examples of degenerate critical points are $x_1 = 0$ of the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x_1) = x_1^3$ and $(x_1, x_2) = (0, 0)$ of $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x_1, x_2) = x_1^3 - 3x_1x_2^2$. The degenerate critical point in that latter example is often referred to as a monkey saddle. Indeed, the graph of the function in a neighborhood goes up and down three times, providing convenient resting place for the two legs as well as the tail of the monkey.

Morse functions. At a critical point, all partial derivatives vanish. A local Taylor expansion has therefore no linear terms. If the critical point is non-degenerate then the behavior of the function in a small neighborhood is dominated by the quadratic terms. Even more, we can find local coordinates such that there are no higher-order terms.

MORSE LEMMA. Let u be a non-degenerate critical point of $f : \mathbb{M} \to \mathbb{R}$. There are local coordinates with $u = (0, 0, \dots, 0)$ such that

$$f(x) \quad = \quad f(u) - x_1^2 - \dots - x_q^2 + x_{q+1}^2 + \dots + x_d^2$$

for every point $x = (x_1, x_2, \dots, x_d)$ in a small neighborhood of u.

The number of minus signs in the quadratic polynomial is the *index* of the critical point, $\text{index}(u) = q$. The index classifies the non-degenerate critical points into $d + 1$ types. For a 2-manifold, we have three types, *minima* with index 0, *saddles* with index 1, and *maxima* with index 2. Examples of all three types can be seen in Figure VI.1. In Figure VI.3, we display them by showing the local evolution of the sublevel set. A consequence of the Morse Lemma is that non-degenerate critical points are isolated. In other words, each critical point has a local neighborhood that separates it from the others. This implies

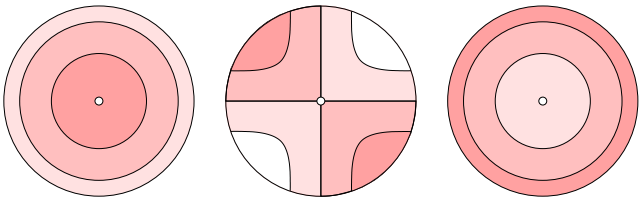

Figure VI.3: From left to right: the local pictures of a minimum, a saddle, a maximum. Imagine looking from above with the shading getting darker as the function shrinks away from the viewpoint.

that a Morse function on a compact manifold has at most a finite number of critical points. To contrast this with a function that is not Morse, take the height function of a torus, similar to Figure VI.1 but placing the torus sideways, the way it would naturally rest under the influence of gravity. This height function has an entire circle of minima and another circle of maxima. All these critical points are degenerate and their index is not defined.

DEFINITION. A *Morse function* is a smooth function on a manifold, $f : \mathbb{M} \to \mathbb{R}$, such that (i) all critical points are non-degenerate, and (ii) the critical points have distinct function values.

Sometimes the second condition is dropped but in this book we will always require both. For a geometrically perfect torus, the height function satisfies condition (i) for all but two directions, the ones parallel to the symmetry axis of the torus. Condition (ii) is violated for another two circles of directions along which the two saddles have the same height. The height function of $\mathbb{S}^2$ is a Morse function for all directions. The distance from a point is a Morse function for almost all points. Exceptions for the torus are points on the symmetry axis and on the center circle, but there are others. The only exception for the 2-sphere is the center.

Gradient vector field. A *vector field* on a manifold is a function $X : \mathbb{M} \to \mathrm{T}\mathbb{M}$ that maps every point $x \in \mathbb{M}$ to a vector $X(x)$ in the tangent space of $\mathbb{M}$ at x. Given $f : \mathbb{M} \to \mathbb{R}$ and X, we denote the directional derivative of f along the vector field by $X[f]$. It maps every point $x \in \mathbb{M}$ to the derivative of f at x in the direction $X(x)$. A particularly useful vector field is the one that points in the direction of steepest increase. To define it, we need to measure length, which we do by introducing a Riemannian metric, that is, a smoothly varying

inner product defined on the tangent spaces. For example, if $\mathbb{M}$ is smoothly embedded in some Euclidean space then the tangent spaces are linear subspaces of the same Euclidean space and we can borrow the metric. Given a smooth manifold $\mathbb{M}$, a Riemannian metric on $\mathbb{M}$, and a smooth function $f : \mathbb{M} \to \mathbb{R}$, we define the *gradient* of f as the vector field $\nabla f : \mathbb{M} \to \mathrm{T}\mathbb{M}$ characterized by $\langle X(x), \nabla f(x) \rangle = X[f]$ for every vector field X. Assuming local coordinates with orthonormal unit vectors x_i, the gradient at the point x is

$$\nabla f(x) \quad = \quad \left[\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \ldots, \frac{\partial f}{\partial x_d}(x) \right]^T .$$

We use the gradient to introduce a 1-*parameter group of diffeomorphisms* $\varphi : \mathbb{R} \times \mathbb{M} \to \mathbb{M}$. There are two characteristic properties of this group. First, the map $\varphi_t : \mathbb{M} \to \mathbb{M}$ defined by $\varphi_t(x) = \varphi(t, x)$ is a diffeomorphism of $\mathbb{M}$ to itself for each $t \in \mathbb{R}$, and second, $\varphi_{t+t_0} = \varphi_t \circ \varphi_{t_0}$ for all $t, t_0 \in \mathbb{R}$. Such a group defines a vector field by differentiation and we require that this vector field be the gradient vector field, modified by taking one over the original length:

$$\lim_{\varepsilon \to 0} \frac{f(\varphi_\varepsilon(x)) - f(x)}{\varepsilon} \quad = \quad \frac{\nabla f(x)}{\|\nabla f(x)\|^2}[f].$$

This group of diffeomorphisms follows the evolution of the sublevel set and can be used to prove that there are no topological changes that happen between contiguous critical values. Specifically, let $f : \mathbb{M} \to \mathbb{R}$ be smooth and $a < b$ such that $f^{-1}[a, b]$ is compact and contains no critical points of f. Then $\mathbb{M}_a$ is diffeomorphic to $\mathbb{M}_b$.

Attaching handles. The situation is different when we consider regular values $a < b$ such that $f^{-1}[a, b]$ is compact but contains one critical point of f. Let this critical point be u and its index be q. In this case, $\mathbb{M}_b$ has the homotopy type of $\mathbb{M}_a$ with a q-handle attached. To explain what this means, we recall that $\mathbb{B}^q$ is the q-dimensional unit ball and $\mathbb{S}^{q-1}$ is its boundary. Let $g : \mathbb{S}^{q-1} \to \mathrm{bd}\, \mathbb{M}_a$ be a continuous map. To *attach* the handle to $\mathbb{M}_a$, we identify each point $x \in \mathbb{S}^{q-1}$ with its image $g(x) \in \mathrm{bd}\, \mathbb{M}_a$. The only case that is a bit different is $q = 0$. Then $\mathbb{S}^{-1}$ is empty and attaching the 0-handle just means adding a disjoint point.

We illustrate this construction for a 3-manifold $\mathbb{M}$. There are four types of critical points, namely minima with index 0, saddles with index 1 or 2, and maxima with index 3. The two types of saddles deserve some attention. To illustrate the local evolution of the sublevel set, we draw spheres around them and shade the portion that belongs to the sublevel set, as in Figure VI.4. The

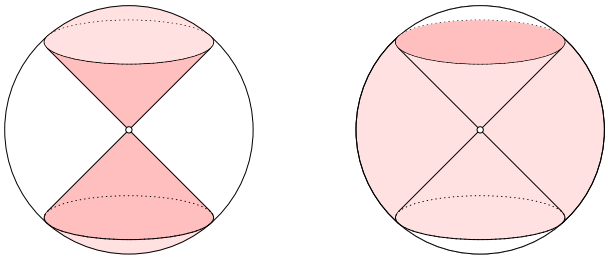

Figure VI.4: The double-cone neighborhood of the index 1 saddle on the left and of the index 2 saddle on the right. The volume occupied by the sublevel set is shaded.

level set that passes through the saddle forms locally a double-cone with the apex at the saddle. This is the same for both types, the only difference being the side on which the sublevel set resides. For the index 1 saddle, we imagine a two sheet hyperboloid approaching from two sides until the two sheets meet at the saddle. Thereafter, the sublevel set thickens around the saddle as its boundary moves out as a one sheet hyperboloid (an hour glass). Homotopically, this evolution is the same as attaching a 1-handle (an interval) connecting the two sheets. For the index 2 saddle, the sequence of events is reversed. Specifically, a one sheet hyperboloid approaches along a circle of directions until it reaches the saddle. Thereafter, the sublevel set thickens around the saddle as its boundary moves out as two sheets of a hyperboloid. Homotopically, this evolution is the same as attaching a 2-handle (a disk) closing the tunnel formed by the one sheet hyperboloid.

Bibliographic notes. Morse theory developed first in infinite dimensions, as part of the calculus of variations, see Morse [4]. The classic source on the subject for finite-dimensional manifolds is the text by Milnor [3], but see also Matsumoto [2] and Banyaga and Hurtubis [1].

[1] A. Banyaga and D. Hurtubis. *Lectures on Morse Homology.* Kluwer, Dordrecht, the Netherlands, 2004.

[2] Y. Matsumoto. *An Introduction to Morse Theory.* Translated from Japanese by K. Hudson and M. Saito, Amer. Math. Soc., 2002.

[3] J. Milnor. *Morse Theory.* Princeton Univ. Press, New Jersey, 1963.

[4] M. Morse. *The Calculus of Variations in the Large.* Amer. Math. Soc., New York, 1934.

VI.2 Transversality

Given a Morse function, we can follow the gradient flow and decompose the manifold depending on where the flow originates and where it ends. For this decomposition to form a complex, we require that the function satisfies an additional genericity assumption.

Integral lines. Recall the 1-parameter group of diffeomorphisms $\varphi : \mathbb{R} \times \mathbb{M} \to \mathbb{M}$ defined by a Morse function f on a manifold $\mathbb{M}$ with a Riemannian metric. The *integral line* that passes through a regular point $x \in \mathbb{M}$ is $\gamma = \gamma_x : \mathbb{R} \to \mathbb{M}$ defined by $\gamma(t) = \varphi(t, x)$; see Figure VI.5. It is the solution to

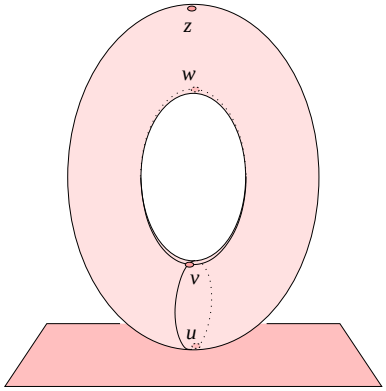

Figure VI.5: The upright torus with the four integral lines that end at the two saddles.

the ordinary differential equation defined by $\dot{\gamma}(t) = \nabla f(\gamma(t))$ and the initial condition $\gamma(0) = x$. Because φ and therefore γ are defined for all $t \in \mathbb{R}$, the integral line necessarily approaches a critical point, both for t going to plus and to minus infinity. We call these critical points the *origin* and the *destination* of the integral line,

$$\begin{aligned} \mathrm{org}(\gamma) &= \lim_{t \to -\infty} \gamma(t); \\ \mathrm{dest}(\gamma) &= \lim_{t \to \infty} \gamma(t). \end{aligned}$$

The function increases along the integral line which implies that $\mathrm{org}(\gamma) \neq \mathrm{dest}(\gamma)$. The Existence and Uniqueness Theorems of ordinary differential equations imply that the integral line that passes through another regular point y is either disjoint from or the same as the one passing through x, $\mathrm{im}\,\gamma_x = \mathrm{im}\,\gamma_y$

or $\operatorname{im}\gamma_x \cap \operatorname{im}\gamma_y = \emptyset$. This property suggests we decompose the manifold into integral lines or unions of integral lines with shared characteristics.

Stable and unstable manifolds. The *stable manifold* of a critical point u of f is the point itself together with all regular points whose integral lines end at u. Symmetrically, the *unstable manifold* of u is the point itself together with all regular points whose integral lines originate at u. More formally,

$$\begin{aligned} S(u) &= \{u\} \cup \{x \in \mathbb{M} \mid \operatorname{dest}(\gamma_x) = u\}; \\ U(u) &= \{u\} \cup \{y \in \mathbb{M} \mid \operatorname{org}(\gamma_y) = u\}. \end{aligned}$$

The function increases along integral lines. It follows that $f(u) \geq f(x)$ for all points x in the stable manifold of u. This is the reason why $S(u)$ is sometimes referred to as the *descending manifold* of u. Symmetrically, $f(u) \leq f(y)$ for all points y in the unstable manifold of u and $U(u)$ is sometimes referred to as the *ascending manifold* of u.

Suppose the dimension of $\mathbb{M}$ is d and the index of the critical point u is q. Then there is a $(q-1)$-sphere of directions along which integral lines approach u. It can be proved that together with u, these integral lines form an open ball of dimension q and that $S(u)$ is a submanifold homeomorphic to $\mathbb{R}^q$ that is immersed in $\mathbb{M}$. It is not embedded because distant points in $\mathbb{R}^q$ may map to arbitrarily close points in $\mathbb{M}$, as we can see in Figure VI.5. For example, the saddle v has a stable 1-manifold consisting of two integral lines that merge at v to form one open, connected interval. The two ends of the interval approach the minimum, u, which does not belong to the 1-manifold. While the map from $\mathbb{R}^1$ to $\mathbb{M}$ is continuous its inverse is not.

Morse-Smale functions. The stable manifolds do not necessarily form a complex. Specifically, it is possible that the boundary of a stable manifold is not the union of other stable manifolds of lower dimension. Take for example the upright torus in Figure VI.5. The stable 1-manifold of the upper saddle, w, reaches down to the lower saddle, v, but the latter is not a stable 0-manifold. The reason for this deficiency is a degeneracy in the gradient flow. In particular, we have an integral line that originates at a saddle and ends at another saddle. Equivalently, the integral line belongs to the stable 1-manifold of w and to the unstable 1-manifold of v. Generically, such integral lines do not exist.

DEFINITION. A *Morse-Smale function* is a Morse function, $f : \mathbb{M} \to \mathbb{R}$, whose stable and unstable manifolds intersect transversally.

Roughly, this requires that the stable and unstable manifolds cross when they intersect. More formally, let $\sigma : \mathbb{R}^q \to \mathbb{M}$ and $\upsilon : \mathbb{R}^p \to \mathbb{M}$ be two immersions. Letting $z \in \mathbb{M}$ be a point in their common image, we say that σ and υ intersect *transversally* at z if the derived images of the tangent spaces at preimages $x \in \sigma^{-1}(z)$ and $y \in \upsilon^{-1}(z)$ span the entire tangent space of $\mathbb{M}$ at z,

$$\mathrm{D}\sigma_x(\mathrm{T}\mathbb{R}^q_x) + \mathrm{D}\upsilon_y(\mathrm{T}\mathbb{R}^p_y) \quad = \quad \mathrm{T}\mathbb{M}_z.$$

We say that σ and υ are *transversal* to each other if they intersect transversally at every point z in their common image.

Complexes. Assuming transversality, the intersection of a stable q-manifold and an unstable p-manifold has dimension $q+p-d$. Furthermore, the boundary of every stable manifold is a union of stable manifolds of lower dimension. The set of stable manifolds thus forms a complex which we construct one dimension at a time.

0**-skeleton:** add all minima as stable 0-manifolds to initialize the complex;

1**-skeleton:** add all stable 1-manifolds, each an open interval glued at its endpoints to two points in the 0-skeleton;

2**-skeleton:** add all stable 2-manifolds, each an open disk glued along its boundary circle to a cycle in the 1-skeleton;

etc. It is possible that the two minima are the same so that the interval whose ends are both glued to it forms a loop. Similarly, the cycle in the 1-skeleton can

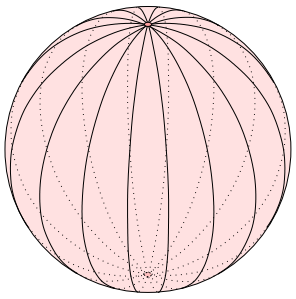

Figure VI.6: All integral lines of the height function of $\mathbb{S}^2$ originate at the minimum and end at the maximum. We therefore have two stable manifolds, a vertex for the minimum and an open disk for the maximum.

be degenerate, such as pinched or even just a single point. Similar situations

are possible for higher-dimensional stable manifolds. An example is the height function of the d-sphere. It has a single minimum, a single maximum, and no other critical points. The minimum has index 0 and forms a vertex in the complex. The maximum has index d and defines a stable d-manifold. It wraps around the sphere and its boundary is glued to a single point, the minimum, as illustrated for $d = 2$ in Figure VI.6.

Morse inequalities. If we take the alternating sum of the stable manifolds in the above example, we get $1 + (-1)^d$, which is the Euler characteristic of the d-sphere. This is not a coincidence. More generally, the alternating sum of stable manifolds gives the Euler characteristic, and this equation is one of the strong Morse inequalities. We state both, the weak and the strong Morse inequalities, writing c_q for the number of critical points of index q.

MORSE INEQUALITIES. Let $\mathbb{M}$ be a manifold of dimension d and $f : \mathbb{M} \to \mathbb{R}$ a Morse function. Then

(i) WEAK: $c_q \geq \beta_q(\mathbb{M})$ for all q;
(ii) STRONG: $\sum_{q=0}^{j}(-1)^{j-q}c_q \geq \sum_{q=0}^{j}(-1)^{j-q}\beta_q(\mathbb{M})$ for all j.

As mentioned above, the strong Morse inequality for $j = d$ is an equality. We can recover the weak inequalities from the strong ones. Indeed

$$\begin{aligned}\sum_{q=0}^{j}(-1)^{j-q}c_q &\geq \beta_j(\mathbb{M}) - \sum_{q=0}^{j-1}(-1)^{j-q-1}\beta_q(\mathbb{M}) \\ &\geq \beta_j(\mathbb{M}) - \sum_{q=0}^{j-1}(-1)^{j-q-1}c_q.\end{aligned}$$

Removing the common terms on both sides leaves $c_j \geq \beta_j(\mathbb{M})$, the j-th weak inequality. We omit the proof of the strong inequalities and instead refer to the proof of their PL versions in the next section.

Floer homology. Assuming a Morse-Smale function, we can intersect the stable and unstable manifolds and get a refinement of the two complexes which we refer to as the *Morse-Smale complex* of f. Its vertices are the critical points and its cells are the components of the unions of integral lines with common origin and common destination. It is quite possible that the stable manifold of a critical point intersects the unstable manifold of another critical point in

more than one component. By definition of transversality, the index difference between the origin and the destination equals the dimension of the cell. In particular, the edges are isolated integral lines connecting index $q-1$ with index q critical points.

To recover the homology of the manifold, we set up a chain complex. The q-chains are the formal sums of index q critical points. The boundary of an index q critical point, u, is the sum of index $q-1$ critical points connected to u by an edge in the Morse-Smale complex. If there are multiple edges, we add the index $q-1$ point multiple times. We illustrate this construction with the

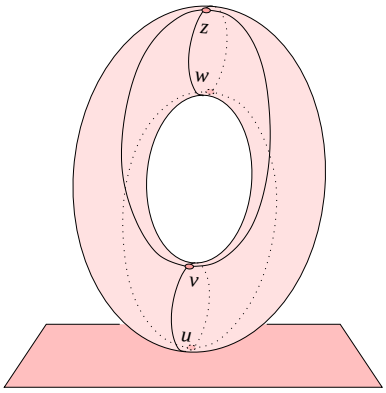

Figure VI.7: The Morse-Smale complex of the height function for the almost but not entirely upright torus.

example depicted in Figure VI.7. We have a slightly tilted torus whose height function is a Morse-Smale function. There are one minimum, two saddles, and one maximum. The non-trivial chain groups are therefore $\mathsf{C}_0 \simeq \mathsf{G}$, $\mathsf{C}_1 \simeq \mathsf{G}^2$, $\mathsf{C}_2 \simeq \mathsf{G}$, with $\mathsf{G} = \mathbb{Z}_2$, as usual. In this example, each critical point appears twice in the boundary of every other critical point, or not at all. Hence, the boundary of each one of the four critical points is zero. It follows that the boundary groups are trivial and the cycle groups as well as the homology groups are isomorphic to the chain groups. The Betti numbers are therefore $\beta_0 = 1$, $\beta_2 = 2$, $\beta_2 = 1$, which is consistent with what we already know about the torus.

Bibliographic notes. The concepts of integral lines and stable as well as unstable manifolds rely on fundamental properties of solutions to ordinary differential equations, in particular the Theorems of Existence and Uniqueness,

see e.g. Arnold [1]. The extra requirement of transversality between stable and unstable manifolds that distinguishes Morse from Morse-Smale complexes has been proven to be generic by Kupka [3] and Smale [4]. The chain complex whose groups are formal sums of critical points is sometimes referred to as Morse-Smale-Witten complex and the resulting homology theory is referred to as Floer homology [2].

[1] V. I. Arnold. *Ordinary Differential Equations.* Translated from Russian, MIT Press, Cambridge, Massachusetts, 1973.

[2] A. Floer. Witten's complex and infinite dimensional Morse theory. *J. Diff. Geom.* **30** (1989), 207–221.

[3] I. Kupka. Contribution à la théorie des champs génériques. *Contributions to Differential Equations* **2** (1963), 457–484.

[4] S. Smale. Stable manifolds for differential equations and diffeomorphisms. *Ann. Scuola Norm. Sup. Pisa* **17** (1963), 97–116.

VI.3 Piecewise Linear Functions

We rarely find smooth functions in practical situations. Instead, we often find non-smooth functions that approximate smooth ones or series of non-smooth functions that approach a smooth limit. In this section, we turn things around and use insights gained into the smooth case as a guide in our attempt to understand the piecewise linear case.

Lower star filtration. Let K be a simplicial complex with real values specified at all vertices. Using linear extension over the simplices, we obtain a *piecewise linear (PL) function* $f : |K| \to \mathbb{R}$. It is defined by $f(x) = \sum_i b_i(x) f(u_i)$, where the u_i are the vertices of K and the $b_i(x)$ are the barycentric coordinates of x; see Section III.1. It is convenient to assume that f is *generic*, by which we mean that the vertices have distinct function values. We can then order the vertices by increasing function value as $f(u_1) < f(u_2) < \ldots < f(u_n)$. For each $0 \leq i \leq n$, we let K_i be the full subcomplex defined by the first i vertices. In other words, a simplex $\sigma \in K$ belongs to K_i iff each vertex u_j of σ satisfies $j \leq i$. Recall that the star of a vertex u_i is the set of cofaces of u_i in K. The *lower star* is the subset of simplices for which u_i is the vertex with maximum function value,

$$\mathrm{St}_- u_i \quad = \quad \{\sigma \in \mathrm{St}\, u_i \mid x \in \sigma \Rightarrow f(x) \leq f(u_i)\}.$$

By assumption of genericity, each simplex has a unique maximum vertex and thus belongs to a unique lower star. It follows that the lower stars partition K. Furthermore, K_i is the union of the first i lower stars. This motivates us to call the nested sequence of complexes $\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K$ the *lower star filtration* of f. It will be useful to notice that the K_i are representative of the continuous family of sublevel sets. Specifically, for $f(u_i) \leq a < f(u_{i+1})$

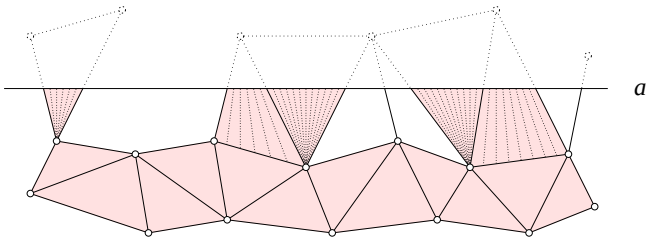

Figure VI.8: We retract $|K|_a$ to $|K_i|$ by shrinking the line segments decomposing the partial simplices from the top downward.

the sublevel set $|K|_a = f^{-1}(-\infty, a]$ has the same homotopy type as K_i. To prove this, consider each simplex with at least one vertex in K_i and at least one vertex in $K - K_i$. Write this simplex as a union of line segments connecting points on the maximal face in K_i with points on the maximal face in $K - K_i$. In other words, express the simplex as the join of these two faces; see Figure VI.8. The sublevel set contains only a fraction of each line segment, namely the portion from the lower endpoint x in $|K_i|$ to the upper endpoint y with $f(y) = a$. To get a deformation retraction, we let $(1-t)y + tx$ be the upper endpoint at time t. Going from time $t = 0$ to $t = 1$ proves that $|K|_a$ and $|K_i|$ have the same homotopy type.

PL critical points. We study the change from one complex to the next in the lower star filtration in more detail. Recall that the link of a vertex is the set of simplices in the closed star that do not belong to the star. Similarly, the *lower link* is the collection of simplices in the closed lower star that do not belong to the lower star. Equivalently, it is the collection of simplices in the link whose vertices have smaller function value than u_i,

$$\operatorname{Lk}_- u_i \quad = \quad \{\sigma \in \operatorname{Lk} u_i \mid x \in \sigma \Rightarrow f(x) < f(u_i)\}.$$

When we go from K_{i-1} to K_i, we attach the closed lower star of u_i, gluing it along the lower link to the complex K_{i-1}. Assume now that K triangulates a d-manifold. This restricts the possibilities dramatically since every vertex star is an open d-ball and every vertex link is a $(d-1)$-sphere. A few examples of lower stars and lower links in a 2-manifold are shown in Figure VI.9. We

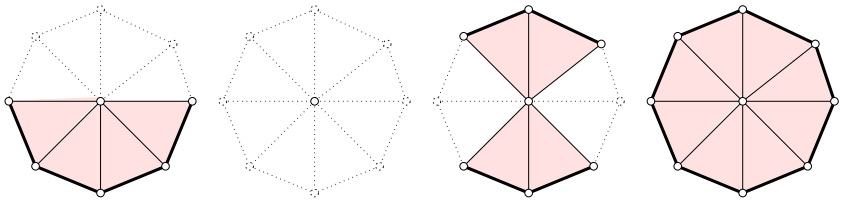

Figure VI.9: From left to right: the lower star and lower link of a regular vertex, a minimum, a saddle, and a maximum.

classify the vertices using the reduced Betti numbers of their lower links. Recall that $\tilde{\beta}_0$ is one less than β_0, the number of components. The only exception to this rule is the empty lower link for which we have $\tilde{\beta}_0 = \beta_0 = 0$ and $\tilde{\beta}_{-1} = 1$. Table VI.1 gives the reduced Betti numbers of the lower links in Figure VI.9. We call u_i a *PL regular vertex* if its lower link is non-empty but homologically

	$\tilde{\beta}_{-1}$	$\tilde{\beta}_0$	$\tilde{\beta}_1$
regular	0	0	0
minimum	1	0	0
saddle	0	1	0
maximum	0	0	1

Table VI.1: Classification of the vertices in a PL function on a 2-manifold.

trivial, and we call u_i a *simple PL critical vertex* of *index* q if its lower link has the reduced homology of the $(q-1)$-sphere. In other words, the only non-zero reduced Betti number of a simple PL critical vertex of index q is $\tilde{\beta}_{q-1} = 1$. We call a piecewise linear function $f : |K| \to \mathbb{R}$ on a manifold a *PL Morse function* if (i) each vertex is either PL regular or simple PL critical and (ii) the function values of the vertices are distinct.

Unfolding. In contrast to the smooth case, PL Morse functions are not dense among the class of all PL functions. Equivalently, a PL function on a manifold may require a substantial perturbation before it becomes PL Morse. As an example, consider the piecewise linear version of a monkey saddle displayed in Figure VI.10. It is therefore not reasonable to assume a PL Morse function

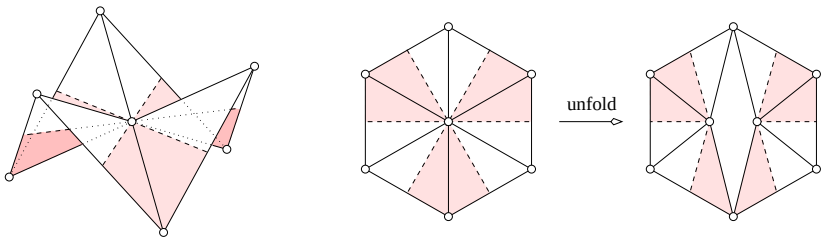

Figure VI.10: Left: a PL monkey saddle of a height function. The areas of points lower than the center vertex are shaded. Right: the unfolding of the monkey saddle into two simple saddles.

as input, but we can sometimes alter the triangulation locally to make it into a PL Morse function. In the 2-manifold case, a k-fold saddle is defined by $\tilde{\beta}_0 = k$. We can split it into k simple saddles by introducing $k-1$ new vertices and assigning appropriate function values close to that of the original, k-fold saddle; see Figure VI.10 for the case $k = 2$. It is less clear how to unfold possibly complicated PL critical points for higher-dimensional manifolds; see also Section X.8.

Alternating sum of indices. Let K be a triangulation of a d-manifold and $f : |K| \to \mathbb{R}$ a PL Morse function. It is not difficult to prove that the alternating sum of the simple PL critical points gives the Euler characteristic,

$$\chi(K) \quad = \quad \sum_u (-1)^{\text{index}(u)}.$$

Since it is easy and instructive, we give an inductive proof of this equation. To go from K_{i-1} to K_i, we add the lower star of u_i. By the Euler-Poincaré Theorem, the Euler characteristic of the lower link, $A = \text{Lk}_- u_i$, is

$$\begin{aligned} \chi(A) &= \sum_{q \geq 1} (-1)^{q-1} \beta_{q-1}(A) \\ &= 1 + \sum_{q \geq 0} (-1)^{q-1} \tilde{\beta}_{q-1}(A). \end{aligned}$$

By definition, this is 1 if u_i is PL regular and $1 + (-1)^{\text{index}(u_i)-1}$ if u_i is PL critical. Each j-simplex in the lower star corresponds to a $(j-1)$-simplex in the lower link, except for the vertex u_i itself. Adding the lower star to the complex thus increases the Euler characteristic by $1 - \chi(A)$, which is zero for a PL regular point and $(-1)^{\text{index}(u_i)}$ for a simple PL critical point. The claimed equation follows.

Mayer-Vietoris sequences. We prepare the proof of the complete set of Morse inequalities for PL Morse functions by recalling the Mayer-Vietoris sequence of a covering of a simplicial complex by two subcomplexes. Let $K = K' \cup K''$ be the covering and note that the intersection of the two subcomplexes, $A = K' \cap K''$, is also a subcomplex of K. As discussed in Section IV.4, the reduced version of the corresponding Mayer-Vietoris sequence is

$$\ldots \to \tilde{\mathsf{H}}_{p+1}(K) \xrightarrow{\varphi_p} \tilde{\mathsf{H}}_p(A) \xrightarrow{\psi_p} \tilde{\mathsf{H}}_p(K') \oplus \tilde{\mathsf{H}}_p(K'') \to \tilde{\mathsf{H}}_p(K) \to \tilde{\mathsf{H}}_{p-1}(A) \to \ldots$$

It is exact which means that the image of every homomorphism is equal to the kernel of the next homomorphism in the sequence. We are interested in the reduced p-th homology group of A, and write φ_p and ψ_p for the maps that connect it to its predecessor and successor groups in the sequence. Let k_p be the rank of the kernel of ψ_p. Similarly, let k^p the rank of the cokernel of φ_p, that is, of $\text{cok}\, \varphi_p = \tilde{\mathsf{H}}_p(A)/\text{im}\, \varphi_p$. By exactness at $\tilde{\mathsf{H}}_p(A)$, we have $\tilde{\beta}_p(A) = k_p + k^p$. As illustrated in Figure VI.11, exactness also implies that the rank of the image of ψ_p is k^p and the rank of $\tilde{\mathsf{H}}_{p+1}(K)/\ker \varphi_p$ is k_p.

We note that $\ker \psi_p$ and $\text{cok}\, \varphi_p$ distinguish two kinds of cycles in A. A cycle in the kernel bounds both in K' and in K'' and thus corresponds to a cycle of

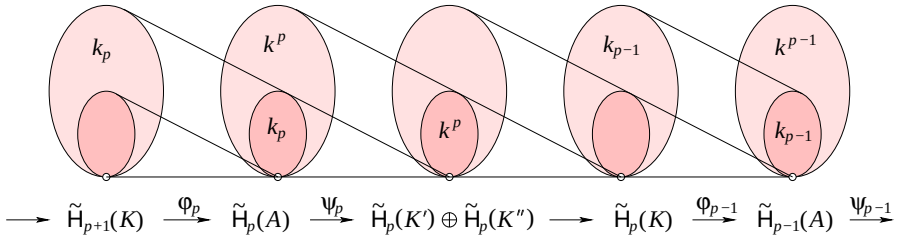

Figure VI.11: A portion of the Mayer-Vietoris sequence. By exactness, the rank of the kernel of every map complements the rank of the cokernel of the preceding map.

dimension $p+1$ in K. In contrast, a cycle in the cokernel is not in the image of the connecting homomorphism and thus represents a non-trivial homology class in K' or in K'' or in both.

PL Morse inequalities. We are now ready to state and prove the PL versions of the weak and strong Morse inequalities.

PL MORSE INEQUALITIES. Let K be a triangulation of a manifold of dimension d and $f : |K| \to \mathbb{R}$ a PL Morse function. Writing c_q for the number of index q PL critical points of f, we have

(i) WEAK: $c_q \geq \beta_q(K)$ for all q;

(ii) STRONG: $\sum_{q=0}^{j}(-1)^{j-q}c_q \geq \sum_{q=0}^{j}(-1)^{j-q}\beta_q(K)$ for all j.

PROOF. We prove the inequalities inductively, for each K_i. They hold initially, when K_0 is empty. For the inductive step, we note that K_i is the union of K_{i-1} and the closed lower star of u_i. To study the situation, we use the Mayer-Vietoris sequence obtained by setting $K = K_i$, $K' = K_{i-1}$, $K'' = \mathrm{St}_- u_i \cup \mathrm{Lk}_- u_i$, and $A = \mathrm{Lk}_- u_i$. Since K'' is the cone over a complex, it is homologically trivial. Referring to Figure VI.11, we let $\varphi_p : \tilde{\mathsf{H}}_{p+1}(K) \to \tilde{\mathsf{H}}_p(A)$ be the connecting homomorphism and $\psi_p : \tilde{\mathsf{H}}_p(A) \to \tilde{\mathsf{H}}_p(K') \oplus \tilde{\mathsf{H}}_p(K'')$ be induced by inclusion. Furthermore, $k_p = \operatorname{rank}\ker\psi_p$ and $k^p = \operatorname{rank}\operatorname{cok}\varphi_p$, as before. Since K'' is homologically trivial, the rank of $\tilde{\mathsf{H}}_p(K)$ is the rank of $\tilde{\mathsf{H}}_p(K')$ minus the rank of the image of ψ_p plus the rank of the kernel of ψ_{p-1}. Translating this back to the lower star filtration, we have

$$\operatorname{rank}\tilde{\mathsf{H}}_p(K_i) \quad = \quad \operatorname{rank}\tilde{\mathsf{H}}_p(K_{i-1}) - k^p + k_{p-1}.$$

By exactness of the sequence, $k_{p-1} + k^{p-1}$ is the rank of the reduced $(p-1)$-st Betti number of A. This number is 1 if u_i is a simple PL critical point of index p and 0 otherwise. Specifically, if u_i is PL regular then $k_{p-1} = k^{p-1} = 0$ for all p and the ranks of the homology groups do not change. Similarly, none of the counters of critical points change so all Morse inequalities remain valid. If $\text{index}(u_i) = p$ and $k_{p-1} = 1$ then both c_p and $\tilde{\beta}_p$ go up by one which maintains the validity of all Morse inequalities. On the other hand, if $\text{index}(u_i) = p$ and $k^{p-1} = 1$ then c_p goes up and $\tilde{\beta}_{p-1}$ goes down. Since the two have opposite signs, this maintains the validity of all Morse inequalities that contain both. The only strong Morse inequality that contains one but not both terms is the one for $j = p - 1$. It contains the relevant Betti number with a plus sign so this inequality is also preserved. □

We note that the strong Morse inequality for $j = d$ is actually an equality, namely the one we have proved above, before recalling the Mayer-Vietoris sequence. It contains both changing terms, in all cases, so there is never a chance that the two sides become different. We also note that the proof of the Morse inequalities in the smooth case is the same. Indeed, passing a non-degenerate critical point has the same effect as adding the lower star of a simple PL critical vertex of the same index.

Bibliographic notes. Piecewise linear functions on polyhedral manifolds have already been studied by Banchoff [1]. He defines the index of a vertex as the Euler characteristic of its lower link. This is coarser than our definition but leads to similar results, in particular a short and elementary proof that the Euler characteristic is equal to the alternating sum of critical points. However, it does not lend itself to a natural generalization of the other Morse inequalities to non-Morse PL functions. Our classification of PL critical points in terms of reduced Betti numbers can be found in [3], where it is used to compute the PL analog of the Morse-Smale complex for 2-manifolds. There are industrial applications of these ideas to surface design and segmentation based on curvature approximating and other shape-sensitive functions in $\mathbb{R}^3$ [2].

[1] T. F. Banchoff. Critical points and curvature for embedded polyhedra. *J. Differential Geometry* **1** (1967), 245–256.

[2] H. Edelsbrunner. Surface tiling with differential topology (extended abstract of invited talk). *In* "Proc. 3rd Eurographics Sympos. Geom. Process., 2005", 9–11.

[3] H. Edelsbrunner, J. Harer and A. Zomorodian. Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete Comput. Geom.* **30** (2003), 87–107.

VI.4 Reeb Graphs

The structure of a continuous function can sometimes be made explicit by visualizing the evolution of the components of the level set. This leads to the concept of the Reeb graph of the function. It has applications in medical imaging and other areas of science and engineering.

Iso-surface extraction. The practical motivation for studying Reeb graphs is the extraction of iso-surfaces for three-dimensional density data. In topological lingo, the density data is a continuous function, $f : [0, 1]^3 \to \mathbb{R}$, and an iso-surface is a level set, $f^{-1}(a)$. If f is smooth and a is a regular value then the level set is a 2-manifold, possibly with boundary. Similarly, if f is generic PL and a is not the value of a PL critical point then the level set is a 2-manifold, again possibly with boundary. Figure VI.12 illustrates this fact for a PL function on the unit square. Assuming we enter a triangle at a boundary

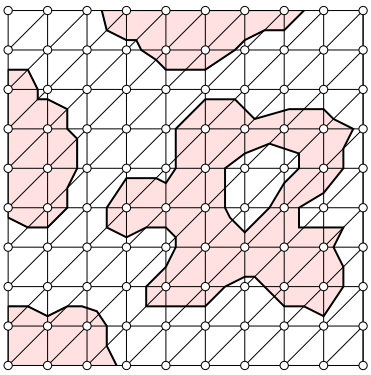

Figure VI.12: The level set of a generic PL function on a triangulation of the unit square. The superlevel set is white and the sublevel set is shaded.

point x with $f(x) = a$, there is a unique other boundary point y with $f(y) = a$ where we exit the triangle. We draw the line segment from x to y as part of the level set and repeat the construction by entering the next triangle at y. There is never any choice as we trace the curve until we arrive at its other end. The procedure is similar for a PL function on the unit cube, except that we use a graph search algorithm to collect the triangular and quadrangular surface pieces we get by slicing the tetrahedra with planes. The most popular choices are Breadth-first Search and Depth-first Search as described in Section II.2.

Given a first point on the level set, it easy to trace out the component that contains it. But to be sure we did not not miss any of the other components it seems we need to check every edge of the triangulation. The desire to avoid this costly computation leads to the introduction of the contour tree, which is a data structure that can be queried for initial points on components of the level set without checking the entire triangulation. It is based on the concept of a Reeb graph, which we discuss next.

Space of contours. Given a continuous map, $f : \mathbb{X} \to \mathbb{R}$, we note that the level sets form a partition of the topological space $\mathbb{X}$. We are interested in a possibly finer partition defined by calling two points $x, y \in \mathbb{X}$ *equivalent* if they belong to a common component of a level set of f. The thus defined equivalence classes are the *contours* of f. The *Reeb graph* of f is the set of contours, $R(f)$, together with its standard quotient topology. We recall that it is defined by taking all subsets whose preimages under $\psi : \mathbb{X} \to R(f)$ are open in $\mathbb{X}$, where $\psi(x)$ is of course the contour that contains x. Let $\pi : R(f) \to \mathbb{R}$ be the unique map whose composition with ψ is f. In other words, it is the map such that

$$\begin{array}{ccc} \mathbb{X} & \xrightarrow{f} & \mathbb{R} \\ & {\scriptstyle\psi}\searrow \quad \nearrow{\scriptstyle\pi} & \\ & R(f) & \end{array}$$

is a commutative diagram. We use it to explain how the Reeb graph speeds up the construction of a level set, $f^{-1}(a)$. Instead of going directly from $\mathbb{R}$ to $\mathbb{X}$, we first compute the preimage of a under π, a set of points in the Reeb graph. The level set consists of a number of contours, one for each point r in $\pi^{-1}(a)$. In a medical imaging application, $\mathbb{X}$ would be represented by a triangulation of the unit cube and the step back from a point r in $R(f)$ to $\mathbb{X}$ would be provided by a pointer to an edge in the triangulation that intersects the contour, $\psi^{-1}(r)$.

Besides using the Reeb graph as a data structure to accelerate the extraction of level sets, we may hope to learn something about the function or the topological space on which the function is defined. Even though the Reeb graph loses aspects of the original topological structure, there are some things that can be said. First of all, $\psi : \mathbb{X} \to R(f)$ maps components to components. Furthermore, the Reeb graph reflects the 1-dimensional connectivity of the space in some cases. To describe this, we refer to a 1-cycle in $R(f)$ as a *loop* and write #loops for the size of the basis. The preimage of a loop in $R(f)$ is necessarily non-contractible in $\mathbb{X}$, and two different loops correspond to non-homologous

1-cycles. Expressing the two properties in terms of Betti numbers, we get

$$\begin{aligned} \beta_0(R(f)) &= \beta_0(\mathbb{X}); \\ \beta_1(R(f)) &\leq \beta_1(\mathbb{X}). \end{aligned}$$

Hence, if $\mathbb{X}$ is contractible then the Reeb graph is a tree, independent of the function f. In medical imaging, the space is a cube and thus contractible, which justifies the practice of calling $R(f)$ a contour tree.

Reeb graphs of Morse functions. More can be said if $\mathbb{X} = \mathbb{M}$ is a manifold of dimension $d \geq 2$ and $f : \mathbb{M} \to \mathbb{R}$ is a Morse function, like in Figure VI.13. Recall that each point $u \in R(f)$ is the image of a contour in $\mathbb{M}$. We call u a

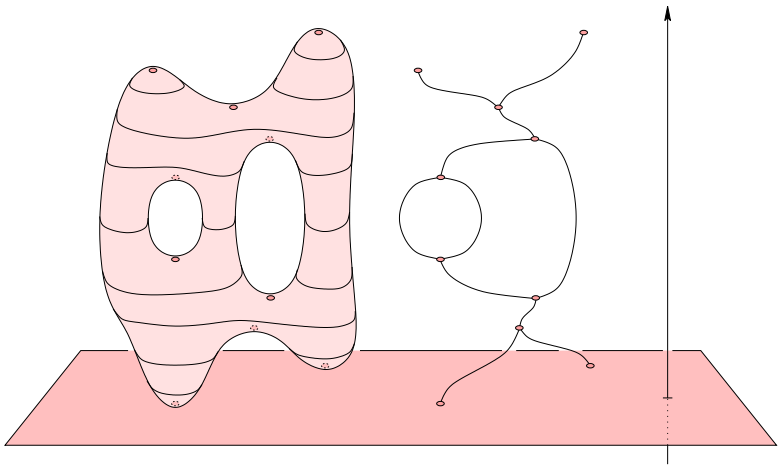

Figure VI.13: Level sets of the 2-manifold map to points on the real line and components of the level sets map to points of the Reeb graph.

node of the Reeb graph if $\psi^{-1}(u)$ contains a critical point or, equivalently, if u is the image of a critical point under ψ. By definition of Morse function, the critical points have distinct function values, which implies a bijection between the critical points of f and the nodes of $R(f)$. The rest of the Reeb graph is partitioned into *arcs* connecting the nodes. A minimum starts a contour and therefore corresponds to a degree one node. An index 1 saddle that merges two contours into one corresponds to a degree three node. Symmetrically, a maximum corresponds to a degree one node and an index $d-1$ saddle that splits a contour into two corresponds to a degree three node. All other critical points

correspond to nodes of degree two. Indeed, the only quadratic polynomials of the form $f(x) = -x_1^2 - \ldots - x_q^2 + x_{q+1}^2 + \ldots + x_d^2$ that have level sets with two components are the ones for $q = 1, d-1$.

We note that the Reeb graph is a one-dimensional topological space with points on arcs being individually meaningful objects. However, there is no preferred way to draw the graph in the plane or in space.

Loops in Reeb graphs. If $\mathbb{M}$ is an orientable 2-manifold then every saddle either merges two contours into one or it splits a contour into two. Either way, the saddle corresponds to a degree three node in the Reeb graph. We use this fact to show that the number of loops depends only on $\mathbb{M}$ and not on the function as long it is Morse. In the non-orientable case, we also have degree two nodes and therefore a number of loops that is no longer independent of the function.

Loop Lemma for 2-manifolds. The Reeb graph of a Morse function on a connected 2-manifold of genus g has g loops if the manifold is orientable and at most $\frac{g}{2}$ loops if it is non-orientable.

Proof. Let c_q be the number of critical points of index q and n_i the number of nodes with degree i in the Reeb graph. We first consider the orientable case for which the number of nodes is $n = n_1 + n_3$. We note that $n_1 = c_0 + c_2$ and $n_3 = c_1$. The number of arcs in the Reeb graph is $m = \frac{1}{2}(n_1 + 3n_3)$. The number of loops exceeds the surplus of arcs by one, that is, #loops $= 1 + m - n = 1 - \frac{1}{2}(c_0 - c_1 + c_2)$. By the last strong Morse inequality, the expression in parenthesis is the Euler characteristic, which for orientable 2-manifolds is $\chi = 2 - 2g$. It follows that #loops $= 1 - \frac{1}{2}(2 - 2g) = g$, as claimed.

In the non-orientable case, the number of nodes is $n = n_1 + n_2 + n_3$, where $n_1 = c_0 + c_2$ and $n_2 + n_3 = c_1$. The number of arcs is $m = \frac{1}{2}(n_1 + 2n_2 + 3n_3)$. The number of loops is again one more than the surplus of arcs, that is, #loops $= 1 + \frac{1}{2}(-n_1 + n_3) = 1 - \frac{1}{2}(c_0 - c_1 + c_2 + n_2)$. Substituting the Euler characteristic for the alternating sum of critical points, wet get #loops $= 1 - \frac{1}{2}(\chi + n_2)$. For a non-orientable 2-manifold, we have $\chi = 2 - g$ and therefore #loops $= \frac{1}{2}(g - n_2)$. Since the number of degree two nodes is non-negative, this is at most half the genus, as claimed. ◻

Coincidentally, the proof implies that the number of degree two nodes has the same parity as the genus. Subject to this constraint, it can be anywhere between zero and g which implies that the upper bound is tight and any integer number of loops between zero and half the genus can be achieved.

Constructing a Reeb graph. We finally consider the algorithmic problem of constructing the Reeb graph of a function on a 2-manifold. We assume the manifold is triangulated and the function, $f : \mathbb{M} \to \mathbb{R}$, is PL Morse. The algorithm sweeps the manifold in the order of increasing function values. We thus begin by sorting the vertices such that $f(u_i) < f(u_{i+1})$ for $1 \leq i < n$. Consider a corresponding sequence of interleaved values, $s_1 < f(u_1) < s_2 < \ldots < s_n < f(u_n) < s_{n+1}$. Since s_i is not the value of any vertex, its preimage is a 1-manifold, consisting of finitely many contours. Each contour is represented by a cyclic list of triangles in the triangulation. Every triangle contributes a line segment and any two contiguous triangles meet in an edge that contributes a shared endpoint of two line segments to the contour. The representation is the same for all values strictly between $f(u_{i-1})$ and $f(u_i)$. Adjustments need to be made when we move into the next open interval, between $f(u_i)$ and $f(u_{i+1})$.

CASE 1. u_i is a minimum. Add a degree one node to the Reeb graph. It starts a new arc associated with a new cyclic list initialized to the triangles in the star of u_i.

CASE 2. u_i is a regular vertex. Then two or more triangles in its star form a contiguous sequence in one of the cyclic lists. Except for the first and the last, all these triangles belong to the lower star. We remove the lower star triangles and replace them by the symmetrically defined upper star triangles of u_i.

CASE 3. u_i is a saddle. Then the triangles in its star form two contiguous sequences in the representation of the current level set. They may be part of the same cyclic list or of two different lists. Similar to Case 2, we keep the first and last triangle of each sequence and replace the lower star triangles in between by the corresponding upper star triangles of u_i. Either list can be empty. We do this by cutting the lists and regluing them when

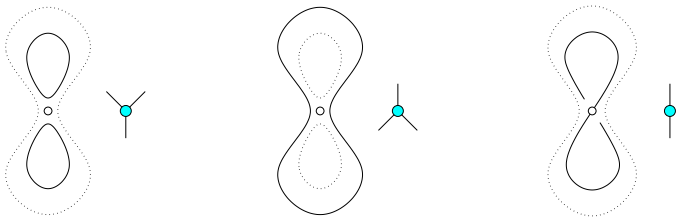

Figure VI.14: From left to right: merging two cyclic lists into one, splitting one list into two, reconnecting one list. Correspondingly, we add a down-fork, an up-fork, a degree two node to the Reeb graph.

we add the upper star triangles. The global effect of the operation depends on whether the cutting is done on one or two cyclic lists and which ends are glued together. There are three different cases, as illustrated in Figure VI.14. In each ase, we add a new node to the Reeb graph and represent the modified lists by arcs that end and start at that node.

CASE 4. u_i is a maximum. Remove the cyclic list of triangles in its star and end the corresponding arc by adding a new degree one node to the Reeb graph.

To implement the algorithm, we need a data structure that supports the following operations:

- CUT a cyclic list open by removing the links between two adjacent triangles;
- DROP a triangle from the end of an open list;
- APPEND a new triangle to the end of an open list;
- GLUE two ends of the same or of two different open lists;
- FIND the cyclic list that contains a specified triangle.

The cutting and gluing can be done without knowing whether the ends belong to the same or to different cyclic lists. However, to update the Reeb graph, we need to know and we use the FIND operation to determine the necessary information. All five operations are supported in time logarithmic in the length of the list if we store it in a so-called balanced search tree. Letting m be the number of edges in the triangulation, we thus get an algorithm that constructs the Reeb graph in time proportional to $m \log_2 m$. This is a significant improvement over the more straightforward algorithm that constructs the Reeb graph in time proportional to m^2. No such improvement is currently known for functions on manifolds of dimension three or higher.

Bibliographic notes. The most common method for extracting iso-surfaces from density data is the Marching Cube Algorithm due to Lorensen and Cline [3]. As the name suggests, it works with a cube complex rather than a triangulation. The portion of the iso-surface within a single cube can be complicated and the implementation of the algorithm requires some care. The idea of speeding up the iso-surface extraction with a contour tree is more recent [6]. This tree is really the Reeb graph of a PL function on a cube, which has no loops. The concept of the Reeb graph of a smooth function is much older [4]. The analysis of the number of loops and the Reeb Graph Algorithm for triangulated 2-manifolds are taken from a relatively recent source [2]. From a practical point

of view, the most demanding operations are CUT and GLUE as they require the splitting and melding of search trees. Particularly easy implementations of these operations are provided by the splay tree implementation of balanced search trees [5]. For contractible domains, the construction of the Reeb graph can be improved to time $m\alpha(m)$, where α is the extremely slow growing inverse of the Ackermann function [1]; see also Section II.2.

[1] H. CARR, J. SNOEYINK AND U. AXEN. Computing contour trees in all dimensions. *Comput. Geom. Theory Appl.* **24** (2002), 75–94.

[2] K. COLE-MCLAUGHLIN, H. EDELSBRUNNER, J. HARER, V. NATARAJAN AND V. PASCUCCI. Loops in Reeb graphs of 2-manifolds. *Discrete Comput. Geom.* **32** (2004), 231–244.

[3] W. E. LORENSEN AND H. E. CLINE. Marching cubes: a high resolution 3D surface construction algorithm. *Comput. Graphics* **21**, Proc. SIGGRAPH, 1987, 163–169.

[4] G. REEB. Sur les points singuliers d'une forme de Pfaff complèment intégrable ou d'une fonction numérique. *Comptes Rendus de L'Académie ses Séances, Paris* **222** (1946), 847–849.

[5] D. D. SLEATOR AND R. E. TARJAN. Self-adjusting binary search trees. *J. Assoc. Comput. Mach.* **32** (1985), 652–686.

[6] M. VAN KREVELD, R. VAN OOSTRUM, C. L. BAJAJ, V. PASCUCCI AND D. R. SCHIKORE. Contour trees and small seed sets for isosurface traversal. *In* "Proc. 13th Ann. Sympos. Comput. Geom., 1997", 212–220.

Exercises

The credit assignment reflects a subjective assessment of difficulty. A typical question can be answered using knowledge of the material combined with some thought and analysis.

1. **Hessian** (two credits). Compute the Hessian and, if defined, the index of the origin, which is critical for each function in the list below.

   (i) $f(x_1, x_2) = x_1^2 + x_2^2$.
   (ii) $f(x_1, x_2) = x_1 x_2$.
   (iii) $f(x_1, x_2) = (x_1 + x_2)^2$.
   (iv) $f(x_1, x_2, x_3) = x_1 x_2 x_3$.
   (v) $f(x_1, x_2, x_3) = x_1 x_2 + x_1 x_3 + x_2 x_3$.
   (vi) $f(x_1, x_2, x_3) = (x_1 + x_2 + x_3)^2$.

2. **Approximate Morse function** (two credits). Let $\mathbb{M}$ be a geometrically perfect torus in $\mathbb{R}^3$, that is, $\mathbb{M}$ is swept out by a circle rotating about a line that lies in the same plane but does not intersect the circle. Let $f : \mathbb{M} \to \mathbb{R}$ measure height parallel to the symmetry axis and note that f is not Morse.

   (i) Describe a Morse function $g : \mathbb{M} \to \mathbb{R}$ that differs from f by an arbitrarily small amount, $\|f - g\|_\infty < \varepsilon$.
   (ii) Draw the Reeb graphs of both functions.

3. **Morse-Smale complex** (two credits). Let $\mathbb{M}$ be the torus in Question 2 and let $f : \mathbb{M} \to \mathbb{R}$ measure height along a direction that is almost but not quite parallel to the symmetry axis of the torus.

   (i) Draw the Morse-Smale complex of the height function.
   (ii) Give the chain, cycle, boundary groups defined by Floer homology.

4. **Quadrangles** (three credits). Let $\mathbb{M}$ be a 2-manifold and $f : \mathbb{M} \to \mathbb{R}$ a Morse-Smale function.

   (i) Prove that each 2-dimensional cell of the Morse-Smale complex of f is a quadrangle. In other words, each 2-dimensional cell is an open disk whose boundary can be decomposed into four arcs each glued to an edge in the complex.
   (ii) Draw a case in which one edge is repeated so that the disk is glued to only three edges but twice to one of the three.

5. **Distance from a point** (three credits). Let $\mathbb{M}$ be the torus swept out by a unit circle rotating at unit distance from the x_3-axis. More formally, $\mathbb{M}$ consists of all solutions to $x_1^2 + x_2^2 = (2 \pm \sqrt{1 - x_3^2})^2$ in $\mathbb{R}^3$. For a point $z \in \mathbb{R}^3$ consider the function $f_z : \mathbb{M} \to \mathbb{R}$ defined by $f_z(x) = \|x - z\|$.

   (i) Describe the set of points z for which f_z violates property (i) of a Morse function.

   (ii) Describe the set of points z for which f_z is not a Morse function.

6. **Morse inequalities** (two credits). Recall that the unstable manifolds of a Morse function $f : \mathbb{M} \to \mathbb{R}$ are the stable manifolds of $-f$. Furthermore, if $\mathbb{M}$ is a d-manifold then an index p critical point of f is an index $d - p$ critical point of $-f$.

   (i) Use this symmetry to formulate collections of inequalities symmetric to the weak and strong Morse inequalities of f.

   (ii) Use these inequalities to prove that the Euler characteristic of $\mathbb{M}$ vanishes if d is odd.

7. **Reeb graph** (one credit). Consider the up-right torus at time $t = 0$ and imagine it falling down in slow motion until it rests on its side at time $t = 1$.

   (i) What is the corresponding time series of Reeb graphs of the height function of the torus?

   (ii) At which position (moment in time) does the Reeb graph of the height function not have a loop?

8. **BCC lattice** (two credits). Instead of the cubic lattice, we may consider constructing iso-surfaces from the body centered cubic lattice obtained by adding the centers of all integer unit cubes. More formally, this is the set of points $\mathbb{Z}^3 \cup \mathbb{Z}^3 + (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})^T$.

   (i) Show there is an (infinite) simplicial complex whose vertex set is the BCC lattice and whose tetrahedra are pairwise congruent, that is, one can be obtained from any other by a rigid transformation.

   (ii) Give a geometric description of the tetrahedron in (i), complete with all face, dihedral, and solid angles.

Chapter VII

Persistence

The central concept of this chapter is motivated by the practical need to cope with noise in data. This includes defining, recognizing, and possibly eliminating noise. These are lofty goals and the challenge can be overwhelming. Indeed, the distinction between noise and feature is not well-defined but rather a subjective notion in the eye of the beholder. In any particular case, the focus is on a range of scales and it is desired to ignore everything that is smaller or larger. In other words, we make ourselves the measure of all things and this way derive a unit, a point of view, an opinion. Motivated by this thought, we take an agnostic approach and withhold any judgement. Instead, we offer a means to measure scale, a tool that can be used to make a judgement based on quantitative information, if one so desires.

VII.1 Persistent Homology

Persistent homology can be used to measure the scale or resolution of a topological feature. There are two ingredients, one geometric, defining a function on a topological space, and the other algebraic, turning the function into measurements. The measurements make sense only if the function does.

The elder rule. We begin with a simplified scenario in which we develop our intuition. Let $\mathbb{X}$ be a connected topological space and $f : \mathbb{X} \to \mathbb{R}$ a continuous function. The thus defined sublevel sets form a 1-parameter family of nested subspaces, $\mathbb{X}_a \subseteq \mathbb{X}_b$ whenever $a \leq b$. It is convenient to write about this family as if it were one sublevel set that evolves as the threshold increases. We visualize this evolution by drawing each component of $\mathbb{X}_a$ as a point. The result is a 1-dimensional graph, $G(f)$, not unlike the Reeb graph discussed in the previous chapter. Thinking of f as a height function, we draw the graph from bottom to top. Since components never shrink, the arcs of the graph may merge but they never split. At the end, for large enough threshold a, we have a single component. It follows that $G(f)$ is a tree, and we refer to it as the *merge tree* of the function; see Figure VII.1. We decompose this tree into disjoint

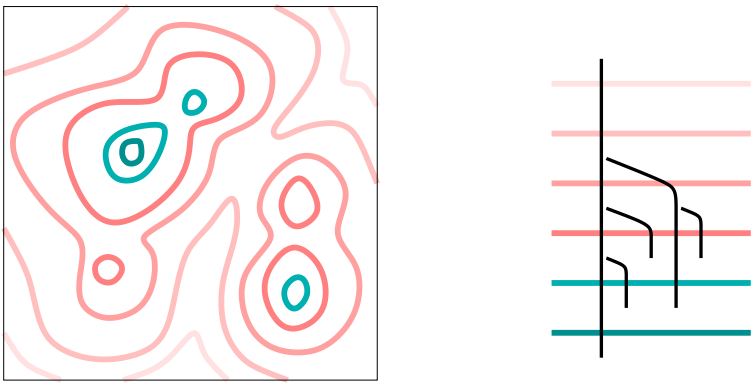

Figure VII.1: Left: a function on the unit square visualized by drawing six level sets with lighter colors indicating larger values. Right: the merge tree of the function.

paths that increase monotonically with f. To obtain the paths, we draw them from bottom to top, simultaneously while keeping their upper endpoints at the same height, a. Paths extend but before they merge, we end the one that started later. Thinking of the difference between two function values as age, we give precedence to the older path.

Elder Rule. At a juncture, the older of the two merging paths continues and the younger path ends.

Letting $a \leq b$ be two thresholds, we let $\beta(a, b)$ be the number of components in $\mathbb{X}_b$ that have a non-empty intersection with $\mathbb{X}_a$. In terms of the merge tree, this is the number of subtrees with topmost points at value b that reach down to level a or below. Each such subtree has a unique path, its longest, that spans the entire interval between a and b. It follows that $\beta(a, b)$ is also the number of paths in the path decomposition of $G(f)$ that span $[a, b]$. We note that any path decomposition that is not generated using the Elder Rule does not have this property. In particular, if f is Morse then the Elder Rule generates a unique path decomposition, which is therefore the only one for which the number of paths spanning $[a, b]$ is equal to $\beta(a, b)$ for all values of $a \leq b$.

Filtrations. We obtain persistence by formulating the Elder Rule for the homology groups of all dimensions. Consider a simplicial complex, K, and a function $f : K \to \mathbb{R}$. We require that f be *monotonic* by which we mean it is non-decreasing along increasing chains of faces, that is, $f(\sigma) \leq f(\tau)$ whenever σ is a face of τ. Monotonicity implies that the sublevel set, $K(a) = f^{-1}(-\infty, a]$, is a subcomplex of K for every $a \in \mathbb{R}$. Letting m be the number of simplices in K, we get $n + 1 \leq m + 1$ different subcomplexes, which we arrange as an increasing sequence,

$$\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K.$$

In other words, if $a_1 < a_2 < \ldots < a_n$ are the function values of the simplices in K and $a_0 = -\infty$ then $K_i = K(a_i)$ for each i. We call this sequence of complexes the *filtration* of f and think of it as a construction by adding chunks of simplices at a time. We have seen examples before, namely the Čech and the alpha complexes in Chapter III and the lower star filtration of a piecewise linear function in Section VI.3. More than in the sequence of complexes, we are interested in the topological evolution, as expressed by the corresponding sequence of homology groups. For every $i \leq j$ we have an inclusion map from the underlying space of K_i to that of K_j and therefore an induced homomorphism, $f_p^{i,j} : \mathsf{H}_p(K_i) \to \mathsf{H}_p(K_j)$, for each dimension p. The filtration thus corresponds to a sequence of homology groups connected by homomorphisms,

$$0 = \mathsf{H}_p(K_0) \to \mathsf{H}_p(K_1) \to \ldots \to \mathsf{H}_p(K_n) = \mathsf{H}_p(K),$$

again one for each dimension p. As we go from K_{i-1} to K_i, we gain new homology classes and we lose some when they become trivial or merge with

each other. We collect the classes that are born at or before a given threshold and die after another threshold in groups.

DEFINITION. The *p-th persistent homology groups* are the images of the homomorphisms induced by inclusion, $\mathsf{H}_p^{i,j} = \operatorname{im} f_p^{i,j}$, for $0 \le i \le j \le n$. The corresponding *p-th persistent Betti numbers* are the ranks of these groups, $\beta_p^{i,j} = \operatorname{rank} \mathsf{H}_p^{i,j}$.

Similarly, we define reduced persistent homology groups and reduced persistent Betti numbers. Note that $\mathsf{H}_p^{i,i} = \mathsf{H}_p(K_i)$. The persistent homology groups consist of the homology classes of K_i that are still alive at K_j or, more formally, $\mathsf{H}_p^{i,j} = \mathsf{Z}_p(K_i)/(\mathsf{B}_p(K_j) \cap \mathsf{Z}_p(K_i))$. We have such a group for each dimension p and each index pair $i \le j$. We can be more concrete about the classes counted by the persistent homology groups. Letting γ be a class in $\mathsf{H}_p(K_i)$, we say it is *born at* K_i if $\gamma \notin \mathsf{H}_p^{i-1,i}$. Furthermore, if γ is born at K_i then it *dies entering* K_j if it merges with an older class as we go from K_{j-1} to K_j, that is, $f_p^{i,j-1}(\gamma) \notin \mathsf{H}_p^{i-1,j-1}$ but $f_p^{i,j}(\gamma) \in \mathsf{H}_p^{i-1,j}$; see Figure VII.2. This is again the

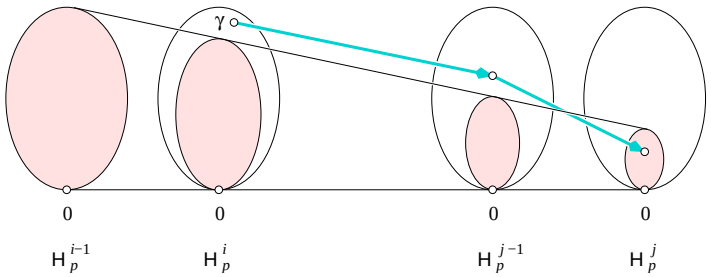

Figure VII.2: The class γ is born at K_i since it does not lie in the (shaded) image of H_p^{i-1}. Furthermore, γ dies entering K_j since this is the first time its image merges into the image of H_p^{i-1}.

Elder Rule. If γ is born at K_i and dies entering K_j then we call the difference in function value the *persistence*, $\operatorname{pers}(\gamma) = a_j - a_i$. Sometimes we prefer to ignore the actual function values and consider the difference in index, $j - i$, which we call the *index persistence* of the class. If γ is born at K_i but never dies then we set its persistence as well as its index persistence to infinity.

We note that births and deaths can also be defined for a sequence of vector spaces that are not necessarily homology groups. All we need is a finite sequence and homomorphisms from left to right which, for vector spaces, are usually referred to as linear maps.

Persistence diagrams. We visualize the collection of persistent Betti numbers by drawing points in two dimensions. Some of these points may have infinite coordinates and some might be the same, so we really talk about a multiset of points in the extended real plane, $\bar{\mathbb{R}}^2$. Letting $\mu_p^{i,j}$ be the number of p-dimensional classes born at K_i and dying entering K_j, we have

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j}),$$

for all $i < j$ and all p. Indeed, the first difference on the right hand side counts the classes that are born at or before K_i and die entering K_j, while the Similarly, the second difference counts the classes that are born at or before K_{i-1} and die entering K_j. Drawing each point (a_i, a_j) with multiplicity $\mu_p^{i,j}$, we get the *p-th persistence diagram* of the filtration, denoted as $\mathrm{Dgm}_p(f)$. It represents a class by a point whose vertical distance to the diagonal is the persistence. Since the multiplicities are defined only for $i < j$, all points lie above the diagonal. For technical reasons which will become clear in the next chapter, we add the points on the diagonal to the diagram, each with infinite multiplicity. It is easy to read off the persistent Betti numbers. Specifically, $\beta_p^{k,l}$ is the number of points in the upper, left quadrant with corner point (a_k, a_l). A class that is born at K_i and dies entering K_j is counted iff $a_i \leq a_k$ and $a_j > a_l$. The quadrant is therefore closed along its vertical right side and open along its horizontal lower side.

Fundamental Lemma of Persistent Homology. Let $\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K$ be a filtration. For every pair of indices $0 \leq k \leq l \leq n$ and every dimension p, the p-th persistent Betti number is $\beta_p^{k,l} = \sum_{i \leq k} \sum_{j > l} \mu_p^{i,j}$.

This is an important property. It says the diagram encodes the entire information about persistent homology groups.

Matrix reduction. Besides having a compact description in terms of diagrams, persistence can also be computed efficiently. The particular algorithm we use is a version of matrix reduction. Perhaps surprisingly, we can get all the information with a single reduction. To describe this, we use a *compatible ordering* of the simplices, that is, a sequence $\sigma_1, \sigma_2, \ldots, \sigma_m$ such that $f(\sigma_i) < f(\sigma_j)$ implies $i < j$ and so does σ_i being a face of σ_j. Such an ordering exists because f is monotonic. Note that every initial subsequence of simplices forms a subcomplex of K. We use this sequence when we set up the boundary matrix, ∂, which stores the simplices of all dimension in one place, that is,

$$\partial[i,j] = \begin{cases} 1 & \text{if } \sigma_i \text{ is a co-dimension one face of } \sigma_j; \\ 0 & \text{otherwise.} \end{cases}$$

In words, the rows and columns are ordered like the simplices in the total ordering and the boundary of a simplex is recorded in its column. The algorithm uses column operations to reduce ∂ to another 0-1 matrix R. Let $low(j)$ be the row index of the lowest one in column j. If the entire column is zero then $low(j)$ is undefined. We call R *reduced* if $low(j) \neq low(j_0)$ whenever $j \neq j_0$ specify two non-zero columns. The algorithm reduces ∂ by adding columns from left to right.

```
R = ∂;
for j = 1 to m do
  while there exists j0 < j with low(j0) = low(j) do
    add column j0 to column j
  endwhile
endfor.
```

The running time is at most cubic in the number of simplices. In matrix notation, the algorithm computes the reduced matrix as $R = \partial \cdot V$; see Figure VII.3. Since each simplex is preceded by its proper faces, ∂ is upper triangular. The j-th column of V encodes the columns in ∂ that add up to give the j-th column in R. Since we only add from left to right, V is also upper triangular and so is R. To get the ranks of the homology groups of K, we count the zero

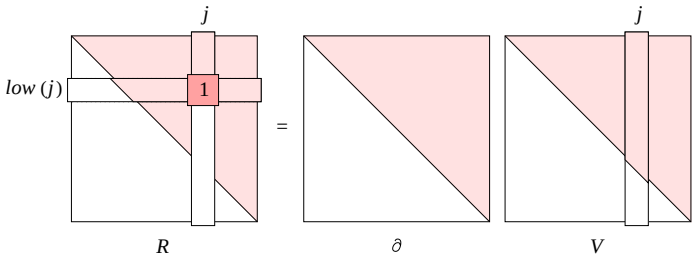

Figure VII.3: Reducing ∂ expressed as matrix multiplication. White areas are necessarily zero while entries in shaded areas can be either zero or one.

columns that correspond to p-simplices with $\#Zero_p(R)$ and the lowest ones in rows that correspond to p-simplices with $\#Low_p(R)$. Comparing the reduced matrix with the normal form matrices, we notice that $\#Zero_p(R) = \operatorname{rank} \mathsf{Z}_p$ and $\#Low_p(R) = \operatorname{rank} \mathsf{B}_p$. It follows that $\beta_p = \#Zero_p - \#Low_p$ for all p.

Pairing. However, there is significantly more information that we can harvest. To see this, we need to understand how the lowest ones relate to the

persistent homology groups. We begin by showing that they are unique, and this in spite of the fact that the reduced matrix, R, is not. Indeed, R is characterized by being reduced and obtained by left-to-right column operations. But we may or may not continue the operations once we reached a reduced matrix. To see that the lowest ones are unique, we consider the lower, left submatrix R_i^j of R whose corner element is $R[i,j]$. In other words, R_i^j is obtained from R by removing the first $i-1$ rows and the last $n-j$ columns. Since left-to-right column operations preserve the rank of every such submatrix, we have $\operatorname{rank} R_i^j = \operatorname{rank} \partial_i^j$ for all i and j. We consider the expression

$$r_R(i,j) \quad = \quad \operatorname{rank} R_i^j - \operatorname{rank} R_{i+1}^j + \operatorname{rank} R_{i+1}^{j-1} - \operatorname{rank} R_i^{j-1}$$

and note that $r_R(i,j) = r_\partial(i,j)$ for all i and j. To evaluate this expression, we observe that the linear combination of any collection of non-zero columns in R_i^j is again non-zero. It follows that the rank of R_i^j is equal to its number of non-zero columns. Now, if $R[i,j]$ is a lowest one then R_i^j has one more non-zero column than the other three submatrices, which implies $r_R(i,j) = 1$. If $R[i,j]$ is not a lowest one then we consider two subcases. If none of the columns from 1 to $j-1$ has its lowest one in row i then R_i^j and R_{i+1}^j have the same number of non-zero columns and so do R_i^{j-1} and R_{i+1}^{j-1}. Second, if one of these columns has its lowest one in row i then R_i^j has one more non-zero column than R_{i+1}^j and R_i^{j-1} has one more non-zero column than R_{i+1}^{j-1}. In either case, $r_R(i,j) = 0$. Since the ranks of the submatrices of R are the same as those of ∂, we have a characterization of the lowest ones that does not depend on the reduction process.

Pairing Lemma. We have $i = low(j)$ iff $r_\partial(i,j) = 1$. In particular, the pairing between rows and columns defined by the lowest ones in the reduced matrix does not depend on R.

Now that we know for sure that the lowest ones are not an artifact of the particular strategy used for reduction, we ask what exactly they mean. Note that column j reaches its final form at the end of the j-th iteration of the outer loop. At this moment, we have the reduced matrix for the complex consisting of the first j simplices in the total ordering. We distinguish the case in which column j ends up zero from the other in which it has a lowest one.

Case 1. column j of R is zero. We call σ_j *positive* since its addition creates a new cycle and thus gives birth to a new homology class.

Case 2. column j of R is non-zero. It stores the boundary of the chain accumulated in column j of matrix V and is thus a cycle. We call σ_j *negative* because its addition gives death to a homology class.

The class that dies in Case 2 is represented by column j. We still need to verify that it is born at the time the simplex of its lowest one, σ_i with $i = low(j)$, is added. But this is clear because the cycle in column j of R just died and all other cycles that die with it have ones below row i, else we could further reduced the matrix and obtain $low(j) < i$, which contradicts the algorithm. It follows that the lowest ones indeed correspond to the points in the persistence diagrams. More precisely, (a_i, a_j) is a finite point in $\mathrm{Dgm}_p(f)$ iff $i = low(j)$ and σ_i is a simplex of dimension p. In this case, σ_j is a simplex of dimension $p + 1$. We have (a_i, ∞) in $\mathrm{Dgm}_p(f)$ iff column i is zero but row i does not contain a lowest one. In other words, σ_i is positive but it does not get paired with a negative simplex.

An example. We illustrate the definitions with a small example. Let K consist of a triangle and its faces. To get a filtration, we first add the vertices, then the edges, and finally the triangle, numbering them in this order from 1 to 7. To make the exercise more interesting, we add the non-zero element of the (-1)-st reduced chain group as a dummy simplex of index 0 to compute reduced rather than ordinary homology. We recall that the augmentation map defines the boundary of each vertex as this dummy simplex. The resulting boundary matrix is shown as part of the matrix equation in Figure VII.4. We reduce it as described and get four non-zero columns in R. The first lowest

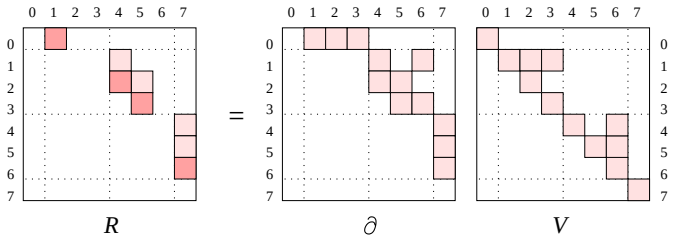

Figure VII.4: Reducing the boundary matrix of the complex consisting of a triangle and its faces. The shaded squares mark ones in the matrices. The dark shaded squares mark lowest ones in the reduced matrix.

one in R is in row 0 and column 1 and corresponds to the (-1)-dimensional reduced homology class that dies when we add vertex 1. The second lowest one is in row 2 and column 4. In words, the vertex 2 gives birth to the 0-cycle that the edge 4 kills. Similarly, the vertex 3 gives birth to the 0-cycle that the edge 5 kills. Adding the edge 6 does not kill anything, which we see in the matrix since column 6 is zero. It corresponds to a 1-cycle obtained by adding the prior

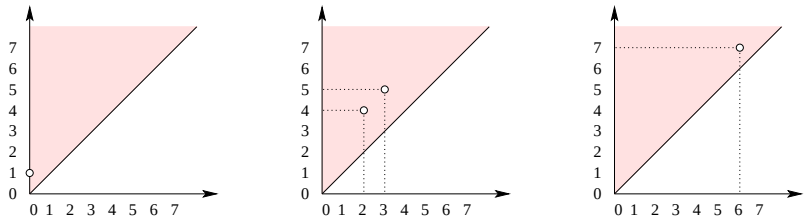

Figure VII.5: From left to right: the minus first, the zeroth, and the first persistence diagram of the filtration that constructs a complex by first adding the three vertices, then the three edges, and finally the triangle.

columns 4, 5, and 6, as indicated in V. The edge 6 thus gives birth to a 1-cycle that is then killed by the triangle 7. Figure VII.5 shows the corresponding three persistence diagrams which are drawn assuming the function value of a simplex is the same as its index. This particular function is injective so all points in the diagrams have multiplicity one.

Bibliographic notes. The concept of persistent homology has been introduced for components by Frosini and Landi [3] and for general homology groups by Robins [4] and independently by Edelsbrunner, Letscher, and Zomorodian [2]. The latter paper gives the first fast algorithm for persistence, the same as described in this section but with the sparse matrix implementation discussed in the next section. A generalization of the notion of persistence to coefficient groups that are fields can be found in [5]. A recent survey on persistent homology is [1].

[1] H. Edelsbrunner and J. Harer. Persistent homology — a survey. *Surveys on Discrete and Computational Geometry. Twenty Years Later*, eds. J. E. Goodman, J. Pach and R. Pollack, Contemporary Mathematics **453**, 257–282, Amer. Math. Soc., Providence, Rhode Island, 2008.

[2] H. Edelsbrunner, D. Letscher and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.* **28** (2002), 511–533.

[3] P. Frosini and C. Landi. Size theory as a topological tool for computer vision. *Pattern Recognition and Image Analysis* **9** (1999), 596–603.

[4] V. Robins. Toward computing homology from finite approximations. *Topology Proceedings* **24** (1999), 503–532.

[5] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.* **33** (2005), 249–274.

VII.2 Efficient Implementations

For practical applications, the number of simplices can be large so that storing the entire boundary matrix becomes prohibitive. As an alternative, we present a sparse matrix implementation of the Persistence Algorithm and give bounds on its running time that are better than cubic in the input size for some cases.

Sparse matrix representation. Same as in the previous section, we assume a monotonic function on a simplicial complex, $f : K \to \mathbb{R}$, and a compatible ordering of the simplices, $\sigma_1, \sigma_2, \ldots, \sigma_m$. We store the data using a linear array, $\partial[1..m]$, and a linked list of simplices per entry. The list in $\partial[j]$ corresponds to the j-th column of the boundary matrix, storing the co-dimension one faces of σ_j. By the end of the algorithm, the list in the j-th array entry corresponds to the column of the reduced matrix whose lowest one is in the j-th row. If there is no such column then the list will be empty. To emphasize the transition, we change the name for the array from ∂ at the beginning to R at the end of the algorithm. All lists are sorted in the order of decreasing index so that the most recently added simplex is readily available at the top; see Figure VII.6. We see

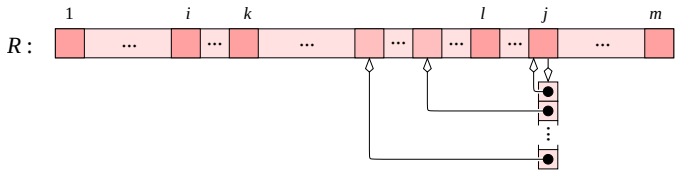

Figure VII.6: The sparse matrix representation of the reduced matrix with only one linked list shown.

a general migration of the lists from right to left. To describe the algorithm that governs this migration, we write L for the linked list of the j-th array entry, and $i = \text{TOP}(L)$ for the index of its top simplex. We call the i-th array entry *occupied*, if it stores a non-empty list, and *unoccupied*, otherwise.

$R = \partial$;
for $j = 1$ **to** m **do**
  $L = \partial[j].cycle$; $R[j].cycle = \text{NULL}$;
  while $L \neq \text{NULL}$ **and** $R[i]$ with $i = \text{TOP}(L)$ is occupied **do**
    $L = L + R[i].cycle$
  endwhile;
  if $L \neq \text{NULL}$ **then** $R[i].cycle = L$ **endif**
endfor.

Adding two lists means merging them while deleting both copies of every duplicate simplex. Since we store the lists in consistent sorted order, each addition can be done in parallel scans. It is instructive to compare this sparse matrix version of the Persistence Algorithm with its standard matrix implementation.

Analysis. The main structure of the sparse matrix implementation is that of two nested loops, the outer and the inner loop. The addition of two lists is another loop in disguise, so the running time is at most cubic in the input size. To improve on this first estimate, we define a *collision* as an attempt to deposit the list L that fails because the entry is occupied. Each collision requires the merging of two lists, which takes time proportional to the sum of their lengths. The loop ends when L runs empty or when the non-empty list L is successfully deposited. The first case identifies σ_j as giving birth to a homology class. The second case identifies σ_j as giving death and the simplex, σ_i, where the deposit happens as triggering the corresponding birth. Each list $R[k].cycle$ contains σ_k as its topmost simplex. Similarly, σ_k is the topmost simplex in L when it collides with the list in $R[k]$. Using modulo 2 arithmetic, σ_k gets deleted which implies that the topmost simplex in the merged list has index less than k. The inner loop thus proceeds monotonically from right to left. It follows that collisions for a simplex σ_j happen only at entries between i and j, where $i = 1$ if σ_j gives birth and i is the index of the corresponding birth if σ_j gives death. Note that in the latter case, $j - i$ is what we call the index persistence of σ_j. Consider now the inner loop for σ_j. A collision at entry k can happen only if σ_k gave birth to a class that died at σ_l before σ_j is reached. We have $i < k < l < j$, as in Figure VII.6. Similarly, the collisions during the inner loop for σ_l correspond to birth-death pairs nested within $[k, l]$. Inductively, this implies that the lists added at collisions contain only faces of simplices with index in $[i, j]$. Letting p be the dimension of σ_j, the number of such faces is at most $p + 1$ times the number of indices in the interval. The time to merge two lists is therefore at most proportional to this number. In summary, the running time of the inner loop for a p-simplex σ_j is at most $(p + 1)(j - i)^2$.

There are situations in which we know ahead of time which simplices give birth and which give death. For example, if the complex is geometrically realized in $\mathbb{R}^3$, the Incremental Betti Number Algorithm described in Section V.4 gives such a classification. With this information, we can then save the effort for the simplices that give birth so that the total running time of the algorithm becomes output-sensitive, and in particular bounded by the dimension times the sum of squares of the index persistences. Assuming constant dimension, this is at most proportional to m^3 but for most practical data it is significantly smaller than that.

Zeroth diagram. The structure of the lists used to compute the 0-th persistence diagram is simpler than for dimensions beyond zero. This diagram depends solely on the vertices and edges of K and on their sequence in the compatible ordering. A vertex has no boundary and always gives birth to a component, so no choice there. An edge σ_j has two vertices as its boundary, $\partial\sigma_j = u + w$. Suppose u comes first, that is, $u = \sigma_i$, $w = \sigma_k$, and $i < k$. The first step of the algorithm is then its attempt to deposit the list L consisting of u and w in $R[k]$. If $L_k = R[k].cycle$ is empty then the deposit is successful, σ_k, σ_j is a pair, and the inner loop ends. Otherwise, L_k is itself a list of two vertices, v and w in which v comes first. Adding the two lists gives $L + L_k$, which consists of u and v. Indeed, all lists have length two so that each addition takes only constant time. This implies that the total effort for dimension 0 is at most the sum of indices, for edges that give birth, and at most the sum of index persistences, for edges that give death. In any case, this is bounded from above by m^2.

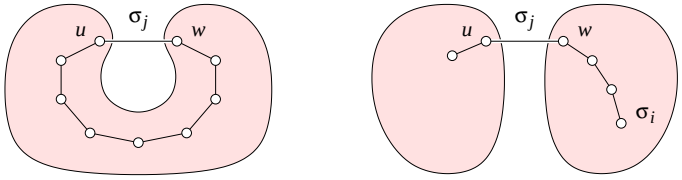

Figure VII.7: Adding the edge σ_j on the left gives birth to a 1-cycle while on the right it gives death to a component.

But we can do even better. Consider again the two cases for the edge with boundary $\partial\sigma_j = u+w$. It gives birth iff u and w belong to the same component of K_{j-1}, the complex right before we add σ_j; see Figure VII.7 on the left. Starting with σ_j, the algorithm adds an edge to the growing path at each collision, and L keeps track of its boundary, the two endpoints. Eventually, the two ends meet, L becomes empty, and the path becomes a 1-cycle. The edge σ_j gives death iff u and w belong to two different components of K_{j-1}; see Figure VII.7 on the right. The inner loop ends when one of the ends of the growing path reaches the first (oldest) vertex, σ_i, of one component. Since the inner loop works monotonically from right to left, this implies that the oldest vertex of the other component is even older. Following the Elder Rule, L gets deposited in $R[i]$ and σ_i, σ_j form a pair. Note that the outcome is predictable. All we need to know is whether or not u and w belong to different components in K_{j-1}, and if they do, which are the oldest vertices of these components. This is exactly the kind of information we can extract from the union-find data

structure, as explained in Chapter I. Recall that this data structure stores each component as a tree of vertices. Given a vertex, we traverse the path up to the root to determine the name of the component. Using the index of the oldest vertex as the name gives the information we need at negligible cost. In summary, we compute the 0-th persistence diagram in time at most proportional to $m\alpha(m)$, where α is the inverse of the Ackermann function which, for all practical purposes, is bounded from above by a constant.

Surfaces. We now consider a simplicial complex, K, that triangulates a 2-manifold. This case is of some practical importance and it allows for a fast implementation of the Persistence Algorithm. Let $f : |K| \to \mathbb{R}$ be obtained by piecewise linear interpolation of its values at the vertices, as explained in Section III.1. There is possibly non-trivial information in the 0-th and the 1-st persistence diagrams of f but not in any of the others. To compute these two diagrams fast, we need to answer two questions.

1. How can we turn the 1-parameter family of sublevel sets into a filtration that we can feed to our algorithm?
2. How can we improve the slower running time for the 1-st persistence diagram to roughly the time needed for the 0-th diagram.

We deal with the first question now and defer the second question to later. Assume for simplicity that the restriction of f to the vertices of K is injective. As defined in Chapter VI, the lower star filtration is then the sequence $\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K$, where K_i is the union of the lower stars of the first i vertices in the ordering by f. It is also the filtration generated by the monotonic function $g : K \to \mathbb{R}$ defined by mapping each simplex to $g(\sigma) = \max_{x \in \sigma} f(x)$. The diagrams of f are defined by the homology groups of the sublevel sets of f, $|K|_a = f^{-1}(-\infty, a]$, while those of g are defined by the homology groups of the sublevel sets of g, $K_a = g^{-1}(-\infty, a]$. By definition of lower star filtration, we have $|K_a| \subseteq |K|_a$ and the inclusion is a homotopy equivalence; see Figure VI.8 and the discussion around it. It follows that the vertical maps in the following diagram are isomorphisms:

$$\begin{array}{ccc} \mathsf{H}_p(|K|_a) & \longrightarrow & \mathsf{H}_p(|K|_b) \\ \uparrow & & \uparrow \\ \mathsf{H}_p(K_a) & \longrightarrow & \mathsf{H}_p(K_b), \end{array}$$

where p is any dimension and $a \leq b$ are any two real numbers. All four maps are induced by inclusion, implying the square commutes. Indeed, these two conditions suffice for the persistence diagrams defined by the two sequences to be the same.

PERSISTENCE EQUIVALENCE THEOREM. Consider two sequences of vector spaces connected by homomorphisms $\phi_i : \mathsf{U}_i \to \mathsf{V}_i$,

$$\begin{array}{ccccccccc} \mathsf{V}_0 & \to & \mathsf{V}_1 & \to & \ldots & \to & \mathsf{V}_{n-1} & \to & \mathsf{V}_n \\ \uparrow & & \uparrow & & & & \uparrow & & \uparrow \\ \mathsf{U}_0 & \to & \mathsf{U}_1 & \to & \ldots & \to & \mathsf{U}_{n-1} & \to & \mathsf{U}_n. \end{array}$$

If the ϕ_i are isomorphisms and all squares commute then the persistence diagram defined by the U_i is the same as that defined by the V_i.

The proof is not difficult but tedious and therefore omitted. As explained above, the 0-th persistence diagram of g can be computed in time at most proportional to $m\alpha(m)$. The equivalence with the 0-th persistence diagram of f thus implies that the latter can be computed in the same amount of time.

First diagram. Instead of computing the 1-st persistence diagram of f directly, we construct the 0-th persistence diagram of $-f$ and derive the diagram of f from it. We begin by describing the relation between $\mathrm{Dgm}_1(f)$ and $\mathrm{Dgm}_0(-f)$, omitting proofs since the relations are consequences of the more general theorems given in the next section.

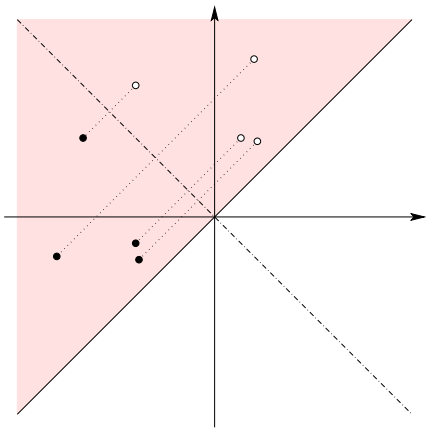

Figure VII.8: The white points of $\mathrm{Dgm}_1(f)$ are reflections of the black points of $\mathrm{Dgm}_0(-f)$ across the minor diagonal.

The 1-st persistence diagram of f consists of the diagonal, a finite portion of off-diagonal points (a, b), and an infinite portion of off-diagonal points (c, ∞). We construct the finite portion from the 0-th persistence diagram of $-f$. Specifically, the point (a, b) marks the birth of a 1-dimensional homology class at a

and its death at b. Looking at $-f$ is like taking the complement and going backward. We thus have the birth of a 0-dimensional homology class at $-b$ and its death at $-a$. It follows a point (a, b) belongs to $\mathrm{Dgm}_1(f)$ iff the point $(-b, -a)$ belongs to $\mathrm{Dgm}_0(-f)$. In other words, the finite portion of $\mathrm{Dgm}_1(f)$ can be obtained by reflecting the finite portion of $\mathrm{Dgm}_0(-f)$ across the minor diagonal, as illustrated in Figure VII.8. We get the points at infinity by partitioning the set of edges in the complex into three subsets: edges that give death in the lower star filtration of f, edges that give death in the lower star filtration of $-f$, and the rest. The first two contribute coordinates to the finite portions of the 0-th and the 1-st diagrams of f. For each edge in the third set, we have a point at infinity in the 1-st diagram, namely a class born when the edge is added and living on even when the complex K is complete. In summary, we have a three pass algorithm for computing the persistence diagrams of a piecewise linear function f on a triangulated 2-manifold in time at most proportional to $m\alpha(m)$.

Bibliographic notes. The original paper on persistent homology by Edelsbrunner, Letscher, and Zomorodian [3] describes the sparse matrix version of the Persistence Algorithm explained in this section. Furthermore, the paper focuses on cases in which birth and death information is available using the Incremental Betti Number Algorithm by Delfinado and Edelsbrunner [1]. The standard matrix reduction version of the Persistence Algorithm came historically later and brought with it a more general appeal at the expense of increased computational resources. The Persistence Equivalence Theorem relating diagrams of different functions has first appeared in [4]. The algorithm for triangulated surfaces is useful in combination with Morse theoretic ideas and has already lead to industrial applications [2].

[1] C. J. A. Delfinado and H. Edelsbrunner. An incremental algorithm for Betti numbers of simplicial complexes on the 3-sphere. *Comput. Aided Geom. Design* **12** (1995), 771–784.

[2] H. Edelsbrunner. Surface tiling with differential topology (extended abstract of invited talk). *In* "Proc. 3rd Eurographics Sympos. Geom. Process., 2005", 9–11.

[3] H. Edelsbrunner, D. Letscher and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.* **28** (2002), 511–533.

[4] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.* **33** (2005), 249–274.

VII.3 Extended Persistence

In this section, we discuss an extension of persistence that is motivated by the problem of fitting shapes to each other. This arises when we solve a puzzle but also in the assembly of mechanical shapes, in the reconstruction of broken artifacts, and in protein docking.

Elevation. Let $\mathbb{M}$ be a smoothly embedded 2-manifold in $\mathbb{R}^3$. Given a direction $u \in \mathbb{S}^2$, the *height function* in this direction, $f_u : \mathbb{M} \to \mathbb{R}$, is defined by mapping each point x to $f_u(x) = \langle x, u \rangle$. We usually draw u vertically going up and think of the height as the signed distance from a horizontal base plane, as in Figure VII.9. Given a threshold $a \in \mathbb{R}$, we recall that the sublevel set

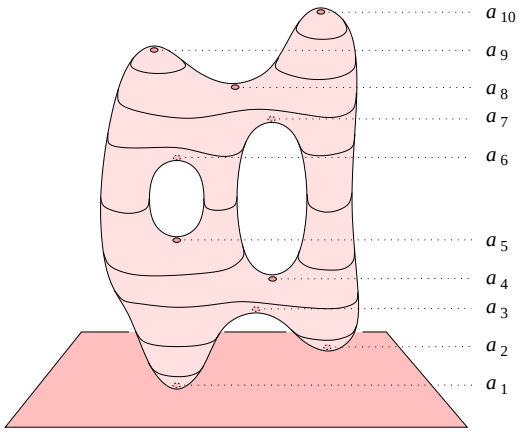

Figure VII.9: A smoothly embedded 2-manifold with level sets shown and critical points of the vertical height function marked.

consists of all points with height a or less, $\mathbb{M}_a = f_u^{-1}(-\infty, a]$. As mentioned in the previous sections, the sublevel sets are nested and define persistence through the corresponding sequence of homology groups. For a generic smooth surface, the homological critical values of a height function are the height values of isolated critical points. If furthermore the direction is generic then there are only three different types, minima which start components, saddles which merge components or complete loops, and maxima which fill holes. Assuming the critical points have distinct heights, the points in the persistence diagrams of f correspond to pairs of critical points. We have minimum-saddle pairs in the 0-th diagram and saddle-maximum pairs in the 1-st diagram.

Based on the family of height functions, we introduce the *elevation function* as a measure of the local protrusion or cavity. Of course, the challenges are the choice of the local neighborhood and of the direction in which the measurement is taken. We finesse both difficulties by exploiting the entire 2-parameter family of height functions. Recall that a point $x \in \mathbb{M}$ is critical for the height function in direction $u = \pm\mathbf{n}_x$, where $\mathbf{n}_x$ is the unit normal at x. If x is paired with another critical point y, we define the elevation of x and y as their absolute height difference, $|f_u(x) - f_u(y)|$, where $u = \pm\mathbf{n}_x = \pm\mathbf{n}_y$. Since x is critical twice, for $u = \pm\mathbf{n}_x$, we need to make sure that the pairing is the same in both directions, else we get contradictory assignments of elevation. We also need all critical points to be paired, else we get white areas in which elevation remains undefined. The latter is the reason we extend persistence and the former is a constraint we need to observe in this extension.

Extended filtration. Let $a_1 < a_2 < \ldots < a_n$ be the homological critical values of the height function $f_u : \mathbb{M} \to \mathbb{R}$. At interleaved values $b_0 < a_1 < b_1 < a_2 < \ldots < a_n < b_n$ we get sublevel sets $\mathbb{M}_{b_i} = f^{-1}(-\infty, b_i]$ which are 2-manifolds with boundary. Symmetrically, we define *superlevel sets* $\mathbb{M}^{b_i} = [b_i, \infty)$, which are complementary 2-manifolds with the same boundary. Finally, we use both to construct a sequence of homology groups going up and a sequence of relative homology groups coming back down,

$$\begin{array}{rcccccl} 0 & = & \mathsf{H}_p(\mathbb{M}_{b_0}) & \to \ldots \to & \mathsf{H}_p(\mathbb{M}_{b_n}) & & \\ & \to & \mathsf{H}_p(\mathbb{M}, \mathbb{M}^{b_n}) & \to \ldots \to & \mathsf{H}_p(\mathbb{M}, \mathbb{M}^{b_0}) & = & 0 \end{array}$$

for each dimension p. The homomorphisms are induced by inclusion. We recall that for modulo 2 arithmetic, the homology groups are isomorphic to the cohomology groups. Furthermore, Lefschetz duality implies $\mathsf{H}^p(\mathbb{M}_b) \simeq \mathsf{H}_{d-p}(\mathbb{M}, \mathbb{M}^b)$. This shows that the construction is intrinsically symmetric although not necessarily within the same dimension. Since we go from the trivial group to the trivial group, everything that gets born eventually dies. As a consequence, all births will be paired with corresponding deaths, as desired.

Tracing what gets born and dies in the relative homology groups is a bit less intuitive than for the absolute homology groups going up. However, we can translate the events between the absolute homology of $\mathbb{M}^b$ and the relative homology of the pair $(\mathbb{M}, \mathbb{M}^b)$. Coming down, the threshold decreases so the superlevel set grows. We call a homology class in the superlevel set *essential*, if it lives all the way down to b_0, and *inessential*, otherwise.

Rule 1. A dimension p homology class of $\mathbb{M}^b$ dies at the same time a dimension $p+1$ relative homology class of $(\mathbb{M}, \mathbb{M}^b)$ dies.

Rule 2. An inessential dimension p homology class of $\mathbb{M}^b$ gets born at the same time a dimension $p+1$ relative homology class of $(\mathbb{M}, \mathbb{M}^b)$ gets born.

Rule 3. An essential dimension p homology class of $\mathbb{M}^b$ gets born at the same time a dimension p relative homology class of $(\mathbb{M}, \mathbb{M}^b)$ dies.

We can prove these relationships by studying the kernels and cokernels of the maps from the homology groups of $\mathbb{M}^b$ into those of $\mathbb{M}$. Leaving this to the interested reader, we develop our intuition by considering an example.

Example. Consider the height function of the genus-2 torus in Figure VII.9. Going up, a_1 and a_2 give birth to classes in H_0, a_4, a_5, a_6, a_7, a_8 give birth to classes in H_1, and a_{10} gives birth to a class in H_2. All classes live until the end of the ascending pass, except for the dimension 0 class born at a_2, which dies at a_3, and the dimension 1 class born at a_8, which dies at a_9. These are the only two finite off-diagonal points in the persistence diagrams as we used to know them. Coming down, a_{10} kills the class in H_0 and a_9 gives birth to a class in H_1 that dies at a_8. Furthermore, a_7, a_6, a_5, a_4 kill the classes in H_1, a_3 gives birth to a class in H_2 that dies at a_2, and finally a_1 kills the class in H_2. To summarize, the pairs of critical values defining the points in the diagrams are $(a_1, a_{10}), (a_2, a_3)$ in dimension 0, $(a_4, a_7), (a_5, a_6), (a_6, a_5), (a_7, a_4), (a_8, a_9), (a_9, a_8)$ in dimension 1, and $(a_{10}, a_1), (a_3, a_2)$ in dimension 2. We show the diagrams in Figure VII.10 using different symbols for classes born and dying going up, born going up and dying coming down, and born and dying coming down. They make up the

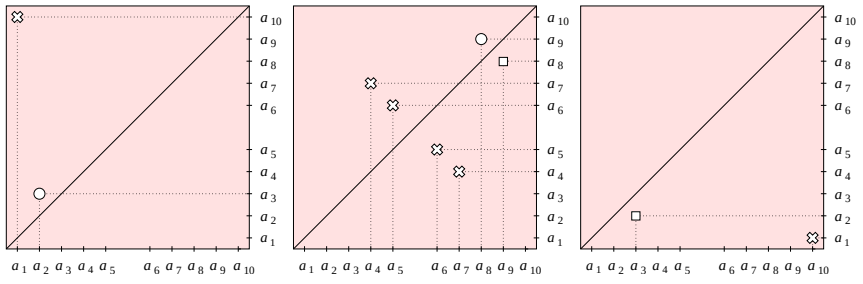

Figure VII.10: From left to right: the 0-th, 1-st, 2-nd persistence diagrams of the height function in Figure VII.9.

ordinary, the *extended*, and the *relative sub-diagrams*, which we denote as Ord, Ext, and Rel, with the dimension in the index and the function in parenthesis, as before. Note that the points of the ordinary sub-diagrams lie

above and those of the relative sub-diagrams lie below the diagonal. The points of the extended sub-diagrams can lie on either side.

Duality and symmetry. The symmetries we observe in Figure VII.10 are not coincidental. They arise as consequences of the Lefschetz duality between absolute and relative homology groups of complementary dimensions, $\mathsf{H}_p(\mathbb{M}_b) \simeq \mathsf{H}_{d-p}(\mathbb{M}, \mathbb{M}^b)$. This translates into a duality result for persistence diagrams which we state without proof. We use a superscript 'T' to indicate reflection across the main diagonal, mapping the point (a, b) to (b, a).

Persistence Duality Theorem. A tame function f on a d-manifold without boundary has persistence diagrams that are reflections of each other as follows,

$$\begin{aligned} \mathrm{Ord}_p(f) &= \mathrm{Rel}^T_{d-p}(f); \\ \mathrm{Ext}_p(f) &= \mathrm{Ext}^T_{d-p}(f); \\ \mathrm{Rel}_p(f) &= \mathrm{Ord}^T_{d-p}(f). \end{aligned}$$

Equivalently, the full p-th persistence diagram is the reflection of the full $(d-p)$-th persistence diagram, $\mathrm{Dgm}_p(f) = \mathrm{Dgm}^T_{d-p}(f)$. We have $d = 2$ for the example illustrated in Figures VII.9 and VII.10 and we indeed have diagrams that are reflections of each other as described. For $2p = d$, the extended sub-diagram is the reflection of itself and therefore symmetric across the main diagonal.

Recall that the definition of elevation requires the pairing of critical points be the same for antipodal height functions. We can use duality to prove that they are indeed the same. More specifically, we have the following structural result again expressed in terms of sub-diagrams of the persistence diagrams. We use the superscript 'R' to indicate reflection across the minor diagonal, mapping the point (a, b) to $(-b, -a)$. Similarly, we use the superscript '0' to indicate central reflection through the origin, mapping the point (a, b) to $(-a, -b)$.

Persistence Symmetry Theorem. Let f be a tame function on a d-manifold without boundary and $-f$ its negative. Then the persistence diagrams of the two functions are reflections of each other,

$$\begin{aligned} \mathrm{Ord}_p(f) &= \mathrm{Ord}^R_{d-p-1}(-f); \\ \mathrm{Ext}_p(f) &= \mathrm{Ext}^0_{d-p}(-f); \\ \mathrm{Rel}_p(f) &= \mathrm{Rel}^R_{d-p+1}(-f). \end{aligned}$$

In lieu of a full-blown proof, we just mention that each of the three equations can be obtained using the Persistence Duality Theorem together with the above three rules relating events in the parallel sequences of absolute and relative homology groups.

Lower and upper stars. To describe how we compute extended persistence, let K be a triangulation of a d-manifold $\mathbb{M}$. We assume the height function is defined at the vertices. We also assume that the height values are distinct so we can index the vertices such that $f(v_1) < f(v_2) < \ldots < f(v_n)$. Let $f : |K| \to \mathbb{R}$ be obtained by piecewise linear extension. Writing $a_i = f(v_i)$ and introducing interleaved values $b_0 < b_1 < \ldots < b_n$, we can define sublevel sets and superlevel sets as before. The set of points $x \in |K|$ with $f(x) \leq b_i$ is homeomorphic to $\mathbb{M}_{b_i}$ and thus a manifold with boundary. Similarly, the set of points with $f(x) \geq b_i$ is homeomorphic to $\mathbb{M}^{b_i}$ and a manifold with boundary. We can retract the partially used simplices and get homotopy equivalent subcomplexes of K. Specifically, let K_i be the full subcomplex defined by the first i vertices in the ordering and K^i the full subcomplex defined by the last $n-i$ vertices. The two subcomplexes of K are disjoint although together they cover all n vertices. The only simplices not in either subcomplex are the ones that connect the first i with the last $n-i$ vertices. Recall that the lower star of a vertex v_i consists of all simplices that have v_i as their highest vertex. Symmetrically, we define the *upper star* to consist of all simplices that have v_i as their lowest vertex. More formally,

$$\begin{aligned} \mathrm{St}_- v_i &= \{\sigma \in \mathrm{St}\, v_i \mid x \in \sigma \Rightarrow f(x) \leq f(v_i)\}, \\ \mathrm{St}^+ v_i &= \{\sigma \in \mathrm{St}\, v_i \mid x \in \sigma \Rightarrow f(x) \geq f(v_i)\}. \end{aligned}$$

Since every simplex has a unique highest and a unique lowest vertex, the lower stars partition K and so do the upper stars. With this notation, $K_0 = \emptyset$ and $K_i = K_{i-1} \cup \mathrm{St}_- v_i$ for $1 \leq i \leq n$. Equivalently, K_i is the union of the first i lower stars. Symmetrically, $K^n = \emptyset$, $K^i = K^{i+1} \cup \mathrm{St}^+ v_{i+1}$, and K^i is the union of the last $n-i$ upper stars.

Computation. By the Persistence Equivalence Theorem in the previous section, the K_i have the same homotopy type as the sublevel sets and the K^i have the same homotopy types as the superlevel sets of $\mathbb{M}$. We can therefore compute persistence by adding the simplices accordingly. Let A be the boundary matrix for the ascending pass, storing the simplices in blocks that correspond to the lower stars of v_1 to v_n, in this order. Within each block, we store the simplices in the order of non-decreasing dimension and break remaining

ties arbitrarily. All simplices in the same block are assigned the same value, namely the height of the vertex defining the lower star. If two simplices in the same block are paired, they define a point on the diagonal of the appropriate persistence diagram. In other words, the homology class dies as soon as it is born and therefore has zero persistence. Only pairs between blocks carry any significance.

Let B be the boundary matrix for the descending pass, storing the simplices in blocks that correspond to the upper stars of v_n to v_1, in this order. Using A and B, we form a bigger matrix by adding the zero matrix at the lower left and the permutation matrix P that translates between A and B at the upper right, as in Figure VII.11. We can think of the result as the boundary matrix of a

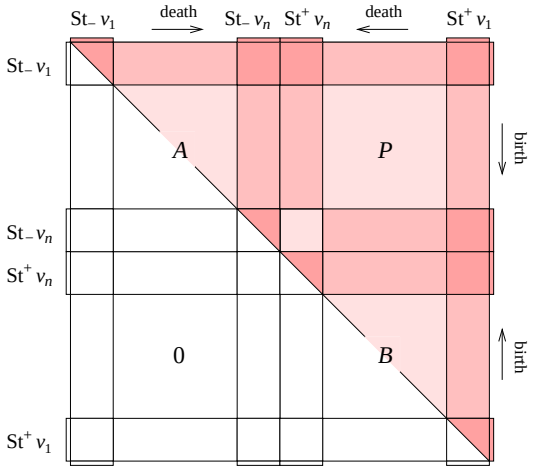

Figure VII.11: The block structure of the boundary matrix representing the construction of K going up and the subsequence destruction coming down.

new complex, namely the cone over K. We pick a new, dummy vertex, v_0, and for each i-simplex σ in K add the $(i+1)$-simplex $\sigma \cup \{v_0\}$. Adding the cone removes any non-trivial homology. This explains why reducing the big matrix works. As we move from left to right, we first construct K forming pairs by reducing A. At the halfway point, the only unpaired simplices are the ones that gave birth to the essential homology classes. As we continue, we cone off K step by step, eventually removing all non-trivial homology. In the end, the ordinary, extended, and relative sub-diagrams are given by the lowest ones in the upper-left, upper-right, and lower-right quadrants of the reduced matrix.

Indeed, we draw the diagram that corresponds to one of the three quadrants

by marking each lowest one as a point, replacing indices by function values. For A, the birth values increase downward and the death values from left to right, so we need to turn the quadrant by 90° to get the ordinary sub-diagram. Symmetrically, we turn the quadrant of B by -90° to get the relative sub-diagram and we reflect the quadrant of P across the main diagonal to get the extended sub-diagram. Since the reduced versions of A and B are upper triangular, we indeed get the ordinary sub-diagram above and the relative sub-diagram below the diagonal.

Bibliographic notes. The extension of persistence described in this section is due to Cohen-Steiner, Edelsbrunner and Harer [2]. It makes essential use of Poincaré and Lefschetz duality to obtain the desired symmetry properties for manifolds. The construction applies equally well to general topological spaces but without guarantee of duality and symmetry. The main motivation for the extension is the elevation function introduced in [1] whose primary purpose is the prediction of interactions between known protein structures [3].

[1] P. K. AGARWAL, H. EDELSBRUNNER, J. HARER AND Y. WANG. Extreme elevation on a 2-manifold. *Discrete Comput. Geom.* **36** (2006), 553–572.

[2] D. COHEN-STEINER, H. EDELSBRUNNER AND J. HARER. Extended persistence using Poincaré and Lefschetz duality. *Found. Comput. Math.*, to appear.

[3] Y. WANG, P. K. AGARWAL, P. BROWN, H. EDELSBRUNNER AND J. RUDOLPH. Coarse and reliable geometric alignment for protein docking. *In* "Proc. Pacific Sympos. Biocomput., 2005", 65–75.

VII.4 Spectral Sequences

Topologists will immediately recognize a connection between persistence and spectral sequences. We shed light on this relation by reviewing spectral sequences, first in terms of the matrix reduction algorithm and second in terms of groups and maps between them.

The matrix reduction view. As usual, we start with a filtration of a simplicial complex, $\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K$, letting $k_i = \operatorname{card} K_i$ be the number of simplices in the i-th complex. Using a compatible total ordering of the simplices, we let ∂ be the boundary matrix which we write in block form. Specifically, ∂_i consists of the rows numbered $k_{i-1}+1$ to k_i corresponding to the simplices in $K_i - K_{i-1}$ and ∂^j consists of the columns numbered $k_{j-1}+1$ to k_j corresponding to the simplices in $K_j - K_{j-1}$. The intersection of the i-th block of rows and the j-th block of columns is then ∂_i^j, which records the codimension one faces of the simplices in $K_j - K_{j-1}$ that lie in $K_i - K_{i-1}$. Since the boundary matrix is upper triangular, we have $\partial_i^j = 0$ whenever $i > j$. We reduce the boundary matrix with left-to-right column additions, as before, but instead of sweeping the matrix from left to right, we sweep it diagonally. More precisely, we work in phases and in Phase r, we reduce columns in ∂^j by adding columns in the blocks from ∂^{j-r+1} all the way to ∂^j itself. The Spectral Sequence Algorithm thus reduces the columns from the diagonal outward, as illustrated in Figure VII.12.

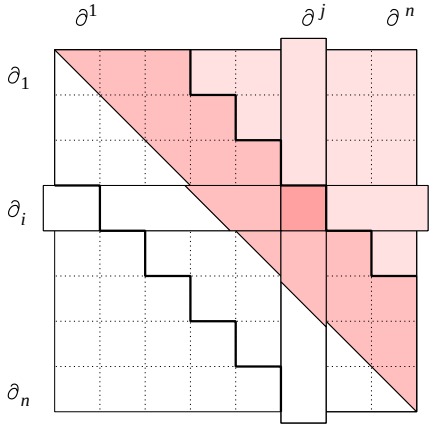

Figure VII.12: After three phases, the triple blocks along the diagonal are reduced. The highlighted blocks of rows and columns intersect in the block matrix ∂_i^j.

```
for r = 1 to n do
  for j = r to n do
    for ι = k_{j−1} + 1 to k_j do
      while ∃k_{j−r} < ι′ < ι with k_{j−r} < low(ι′) = low(ι) ≤ k_{j−r+1} do
        add column ι′ to column ι
      endwhile
    endfor
  endfor
endfor.
```

The result is the same as that of the Persistence Algorithm in the first section of this chapter, only the order in which the columns are added is different. An easy connection to persistence arises by considering the monotonic function $f : K \to \mathbb{R}$ mapping a simplex $\sigma \in K_i - K_{i-1}$ to $f(\sigma) = i$. A leftmost lowest one in ∂_i^j then belongs to a simplex pair of persistence $j - i$. The Spectral Sequence Algorithm thus computes the pairs in the order of non-decreasing persistence.

Groups and maps. We now interpret the algorithm in terms of groups that make up the spectral sequence of the filtration. Recall the chain groups and boundary maps, $\partial : \mathsf{C}_p \to \mathsf{C}_{p-1}$, which form the chain complex defined by K. For each j, we let C_p^j be the group of p-chains of $K_j - K_{j-1}$, and for each chain $c \in \mathsf{C}_p^j$, we let $\partial_i^j c$ be the sum of terms of ∂c that lie in $K_i - K_{i-1}$. Suppressing the dimension in the notation for the boundary map, we have $\partial_i^j : \mathsf{C}_p^j \to \mathsf{C}_{p-1}^i$ and

$$\partial c \quad = \quad \partial_j^j c + \partial_{j-1}^j c + \ldots + \partial_1^j c.$$

The block ∂_i^j in the boundary matrix represents the maps ∂_i^j simultaneously for all dimensions. In spectral sequences, we approximate ∂ by the sum of maps ∂_j^j to ∂_i^j and then decrease i. The spectral sequence itself consists of a collection of groups $\mathsf{E}_{p,q}^r$ and maps $\mathsf{d}_{p,q}^r$ between them. To describe them, we break with the convention of using p for the dimension. Instead, we follow the convention entrenched in the spectral sequence literature in which the first subscript, p, identifies the block of columns, the sum of subscripts, $p + q$, gives the dimension, and the superscript, r, counts the phases in the iteration.

As usual, we think of the columns of the boundary matrix as generators of the chain groups. Limiting our attention to the p-th block of columns, ∂^p, we get the groups of $(p+q)$-chains of $K_p - K_{p-1}$, for all q. If we further limit ∂^p to the blocks of rows ∂_i to ∂_p, we effectively ignore any boundary in K_{i-1}. For $i = p$,

this is equivalent to taking the relative chain groups, $\mathsf{C}_{p+q}(K_p, K_{p-1})$. For $i < p$, we have a subgroup of the relative chain group $\mathsf{C}_{p+q}(K_p, K_{i-1})$, namely the one generated by the $(p+q)$-simplices in $K_p - K_{p-1}$; see Figure VII.13. For what follows, it is important to remember that the boundary matrix, ∂,

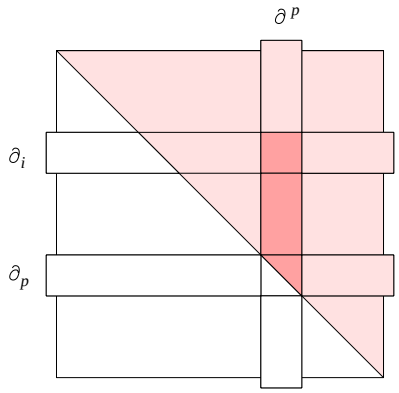

Figure VII.13: The shaded portion of the p-th block of columns represents the chains of $K_p - K_{p-1}$ and their boundaries in $K_p - K_{i-1}$.

represents simplices of all dimensions in one. Hence, each block will correspond to a sequence of groups, namely one for each dimension.

The E^0-term of the spectral sequence. To prepare for the first phase of the algorithm, we focus on the diagonal blocks of the boundary matrix. Fixing $r = 0$, we write $\mathsf{E}^0_{p,q} = \mathsf{C}^p_{p+q}$ for the group of $(p+q)$-chains of $K_p - K_{p-1}$. Fixing p and varying q, these groups are generated by the p-th block of columns. Furthermore, we let

$$\mathsf{d}^0_{p,q} : \mathsf{E}^0_{p,q} \to \mathsf{E}^0_{p,q-1}$$

be defined by the $(p+q)$-dimensional boundary map restricted to the block ∂^p_p. In other words, $\mathsf{d}^0_{p,q}$ is ∂^p_p as applied to $(p+q)$-chains. We note that $\mathsf{E}^0_{p,q}$ is isomorphic to the relative chain group $\mathsf{C}_{p+q}(K_p, K_{p-1})$ and $\mathsf{d}^0_{p,q}$ agrees with the corresponding relative boundary map. It follows that the maps satisfy the Fundamental Lemma of Homology, that is, $\mathsf{d}^0_{p,q-1} \circ \mathsf{d}^0_{p,q} = 0$. Indeed, a codimension two face of a $(p+q)$-simplex in $K_p - K_{p-1}$ either does not belong to $K_p - K_{p-1}$ or it does, but then both codimension one faces that contain it also belong to $K_p - K_{p-1}$. Hence, we get a chain complex,

$$\ldots \longrightarrow \mathsf{E}^0_{p,q+1} \longrightarrow \mathsf{E}^0_{p,q} \longrightarrow \mathsf{E}^0_{p,q-1} \longrightarrow \ldots,$$

in which the maps are implied. It is customary to draw this chain complex vertically, and adding the chain complexes for the other diagonal blocks, we get a 2-dimensional grid of groups, as shown in Figure VII.14. To reduce the clutter, we omit the arrows that connect the groups in each vertical line from top to bottom. We call this the E^0-*term* of the spectral sequence, noting that a vertical line in the grid contains all groups represented by a diagonal block of the boundary matrix.

$$\begin{array}{ccccccc}
 & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & \mathsf{E}^0_{1,1} & \mathsf{E}^0_{2,1} & \mathsf{E}^0_{3,1} & \mathsf{E}^0_{4,1} & \mathsf{E}^0_{5,1} & \cdots \\
\cdots & \mathsf{E}^0_{1,0} & \mathsf{E}^0_{2,0} & \mathsf{E}^0_{3,0} & \mathsf{E}^0_{4,0} & \mathsf{E}^0_{5,0} & \cdots \\
\cdots & \mathsf{E}^0_{1,-1} & \mathsf{E}^0_{2,-1} & \mathsf{E}^0_{3,-1} & \mathsf{E}^0_{4,-1} & \mathsf{E}^0_{5,-1} & \cdots \\
\cdots & 0 & \mathsf{E}^0_{2,-2} & \mathsf{E}^0_{3,-2} & \mathsf{E}^0_{4,-2} & \mathsf{E}^0_{5,-2} & \cdots \\
\cdots & 0 & 0 & \mathsf{E}^0_{3,-3} & \mathsf{E}^0_{4,-3} & \mathsf{E}^0_{5,-3} & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}$$

Figure VII.14: The E^0-term of the spectral sequence. We have maps going vertically downward, from $\mathsf{E}^0_{p,q}$ to $\mathsf{E}^0_{p,q-1}$ for every choice of p and q.

The E^1-term. After interpreting the diagonal blocks of the original boundary matrix in terms of relative chain groups, we now push this interpretation through the phases of the algorithm. For the first phase, we take the homology of the above vertical complexes and define $\mathsf{E}^1_{p,q} = \ker \mathsf{d}^0_{p,q} / \operatorname{im} \mathsf{d}^0_{p,q+1}$. An element of $\mathsf{E}^1_{p,q}$ is thus the equivalence class of a chain $c \in \mathsf{C}^p_{p+q}$ with $\partial^p_p c = 0$, where two chains are equivalent if their difference lies in the image of ∂^p_p, taking of course the boundary map that applies to chains of one higher dimension. In other words, the element is a relative homology class and more generally $\mathsf{E}^1_{p,q} \simeq \mathsf{H}_{p+q}(K_p, K_{p-1})$. Representatives of $\mathsf{E}^1_{p,q}$ are computed by reducing the matrix ∂^p_p, which is what the algorithm does in Phase $r = 1$. The zero columns in ∂^p_p correspond to simplices that give birth and represent cycles. Some are paired and have zero persistence since their classes come and go within $K_p - K_{p-1}$. Others are not paired and their cycles are the generators of $\mathsf{E}^1_{p,q}$. Next we let

$$\mathsf{d}^1_{p,q} : \mathsf{E}^1_{p,q} \to \mathsf{E}^1_{p-1,q}$$

be defined by the $(p+q)$-th boundary map restricted to ∂_p^{p-1}. Recall that an element in $\mathsf{E}^1_{p,q}$ is represented by a relative $(p+q)$-cycle, c. Hence, $\partial_p^p c = 0$ but $\partial_p^{p-1} c$ is possibly non-zero and represents a class in $\mathsf{E}^1_{p-1,q}$. All this sounds complicated but it is rather straightforward if interpreted in terms of the boundary matrix after one phase of the algorithm. As before, the boundary maps satisfy the Fundamental Lemma of Homology, $\mathsf{d}^1_{p-1,q} \circ \mathsf{d}^1_{p,q} = 0$, so we get again a chain complex,

$$\ldots \longrightarrow \mathsf{E}^1_{p+1,q} \longrightarrow \mathsf{E}^1_{p,q} \longrightarrow \mathsf{E}^1_{p-1,q} \longrightarrow \ldots .$$

Going back to the grid in Figure VII.14, we can see these complexes as horizontal lines going from right to left. Of course, we are now in the next phase so we need to substitute $r = 1$ for the superscript 0 everywhere. This is the E^1-*term* of the spectral sequence.

The E^2-term. We take one more step before appealing to induction, taking the homology of the horizontal complexes, $\mathsf{E}^2_{p,q} = \ker \mathsf{d}^1_{p,q} / \operatorname{im} \mathsf{d}^1_{p+1,q}$. An element of $\mathsf{E}^2_{p,q}$ is the equivalence class of the sum of a chain $c \in \mathsf{C}^p_{p+q}$ and another chain $c' \in \mathsf{C}^{p-1}_{p+q}$. The chains satisfy $\partial^p_p c = 0$ and $\partial^p_{p-1} c + \partial^{p-1}_{p-1} c' = 0$ and being equivalent means that the difference lies in $\operatorname{im} \partial^p_p + \operatorname{im} \partial^p_{p-1} + \operatorname{im} \partial^{p-1}_{p-1}$. The group $\mathsf{E}^2_{p,q}$ is not a relative homology group by itself but a subgroup of one, namely $\mathsf{E}^2_{p,q} \oplus \mathsf{E}^1_{p-1,q+1} \simeq \mathsf{H}_{p+q}(K_p, K_{p-2})$. Representatives of $\mathsf{E}^2_{p,q}$ are computed by reducing the double block of matrices ∂^p_p, ∂^{p-1}_{p-1}, ∂^{p-1}_p, ∂^p_{p-1}. The first two have already been reduced and the third is zero. Phase $r = 2$ completes the reduction of the double block for the remaining fourth matrix. Next, we let

$$\mathsf{d}^2_{p,q} : \mathsf{E}^2_{p,q} \to \mathsf{E}^2_{p-2,q+1}$$

be defined by the $(p+q)$-th boundary map restricted to ∂^p_{p-2}. By construction, an element of $\mathsf{E}^2_{p,q}$ is represented by a $(p+q)$-chain, c, whose boundary in $K_p - K_{p-2}$ is empty. Its boundary in $K_{p-2} - K_{p-3}$ is possibly non-empty and represents a class in $\mathsf{E}^2_{p-2,q+1}$, the image of the class of c in $\mathsf{E}^2_{p,q}$. Taking the thus defined boundary map twice gives again zero, so we get a chain complex,

$$\ldots \longrightarrow \mathsf{E}^2_{p+2,q-1} \longrightarrow \mathsf{E}^2_{p,q} \longrightarrow \mathsf{E}^2_{p-2,q+1} \longrightarrow \ldots ,$$

similar to before. Going back to the grid in Figure VII.14, we see this complex along a line of slope one half going from right to left. In other words, the groups are connected by knight moves in chess, two to the left and one up. Of course, we are now in the next phase, so we need to substitute $r = 2$ for the superscript 0 everywhere. This is the E^2-*term* of the spectral sequence.

Iteration. The process continues and for general phase numbers r, the maps take r steps to the left and $r-1$ steps up,

$$\mathsf{d}^r_{p,q} : \mathsf{E}^r_{p,q} \to \mathsf{E}^r_{p-r,q+r-1}.$$

This gives a set of chain complexes and we take homology to enter the next phase. Since K is finite, the maps are eventually zero and the sequence converges to a limit term, $\mathsf{E}^r = \mathsf{E}^\infty$ for r large enough. The homology groups of K are obtained by taking direct sums along the diagonal lines in the limit term for which the dimension is constant.

Before reaching the limit term, we may consider each class in $\mathsf{E}^r_{p,q}$ as generated by an "almost" cycle of dimension $p+q$. This is a chain whose boundary in $K_p - K_{p-r}$ is empty by may have non-empty boundary in K_{p-r}. It is either an essential cycle of K, or a cycle of persistence at least r, assuming the monotonic function $f : K \to \mathbb{R}$ that maps $\sigma \in K_p - K_{p-1}$ to $f(\sigma) = p$, as before. This leads to the following summary connection between persistence and spectral sequences.

SPECTRAL SEQUENCE THEOREM. The total rank of the groups of dimension $p+q$ after $r \geq 1$ phases of the Spectral Sequence Algorithm equals the number of points in the $(p+q)$-th persistence diagram of f whose persistence is r or larger, that is,

$$\sum_{p=1}^{n} \operatorname{rank} \mathsf{E}^r_{p,q} \quad = \quad \operatorname{card} \{a \in \operatorname{Dgm}_{p+q}(f) \mid \operatorname{pers}(a) \geq r\},$$

where q decreases as p increases so that the dimension remains constant.

In the limit, for r large enough, we have $\sum_{p=1}^{n} \operatorname{rank} \mathsf{E}^r_{p,q} = \operatorname{rank} \mathsf{H}_{p+q}(K)$ equal to the number of points in the $(p+q)$-th persistence diagram whose persistence is infinite.

Bibliographic notes. A comprehensive account of spectral sequences can be found in [3]. The treatment in this section follows the more concise presentation in the survey of persistent homology [2]. Similar to persistent homology, working over a field is crucial for the construction of spectral sequences. Over $\mathbb{Z}$, there are extension problems to solve because of torsion; see [1].

[1] K. S. BROWN. *Cohomology of Groups.* Springer-Verlag, New York, 1994.

[2] H. Edelsbrunner and J. Harer. Persistent homology — a survey. *Surveys on Discrete and Computational Geometry. Twenty Years Later*, eds. J. E. Goodman, J. Pach and R. Pollack, Contemporary Mathematics **453**, 257–282, Amer. Math. Soc., Providence, Rhode Island, 2008.

[3] J. McCleary. *A User's Guide to Spectral Sequences.* Second edition, Cambridge Univ. Press, England, 2001.

Exercises

The credit assignment reflects a subjective assessment of difficulty. A typical question can be answered using knowledge of the material combined with some thought and analysis.

1. **Tetrahedron complex** (one credit). Let K consist of a tetrahedron and its faces.
   (i) Apply the matrix reduction algorithm to the filtration of K obtained by adding the simplices in the order of dimension.
   (ii) Do any of the three diagrams depend on the way you order the simplices of the same dimension?

2. **Matrix reduction revisited** (two credits). Change the standard matrix reduction implementation of the persistence algorithm described in Section VII.1 by adding each j-th column to columns on its right rather than adding columns on its left to it. Specifically, consider the following implementation.

```
R = ∂;
for j = 1 to m do
  while there exists j_0 > j with low(j_0) = low(j) do
    add column j to column j_0
  endwhile
endfor.
```

   (i) Show that this implementation of the persistence algorithm generates the same lowest ones as the standard matrix reduction implementation.
   (ii) Give an example for which this and the standard implementation of the persistence algorithm compute different reduced matrices.

3. **Sublevel sets** (two credits). Let $f : |K| \to \mathbb{R}$ be a piecewise linear function defined by its values at the vertices, $f(u_1) < f(u_2) < \ldots < f(u_n)$. Let b be strictly between $f(u_i)$ and $f(u_{i+1})$, for some $1 \leq i \leq n-1$, and recall that the sublevel set defined by b is $f^{-1}(-\infty, b]$.
   (i) Prove that the sublevel sets defined by b and by $f(u_i)$ have the same homotopy type.
   (ii) Draw an example each for the cases when the sublevel sets defined by b and by $f(u_{i+1})$ have the same and different homotopy types.

4. **Graphs without branching** (three credits). Let K be a 1-dimensional simplicial complex in which each vertex belongs to one or two edges. In other words, K is a simple graph whose components are paths and closed curves. Show that the sparse matrix implementation of the persistence algorithm described in Section VII.2 takes time proportional to the number of simplices in K.
5. **Persistence diagram** (one credit). Draw a genus-3 torus, consider its height function, and draw the non-trivial persistence diagrams of the function. Distinguish between points in the ordinary, extended, and relative sub-diagrams.
6. **Breaking symmetry** (two credits). Design a topological space $\mathbb{X}$ and a continuous function $f : \mathbb{X} \to \mathbb{R}$ such that
   (i) the persistence diagrams violate the Persistence Duality Theorem in Section VII.3;
   (ii) the persistence diagrams violate the Persistence Symmetry Theorem in the same section.
7. **Matrix reduction once again** (one credit). Prove that the reduced matrix computed by the spectral sequence algorithm in Section VII.4 is the same as that generated by the persistence algorithm in Section VII.1.
8. **Parallel matrix reduction** (three credits). First, rewrite the Spectral Sequence Algorithm of Section VII.4 for the case in which each block, $K_j - K_{j-1}$, consists of a single simplex. Second, show that the thus simplified algorithm can be run on a parallel computer architecture using n processors taking time at most proportional to n^2.

Chapter VIII

Stability

Persistence is a measure theoretic concept built on top of algebraic structures. Its most important property is the stability under perturbations of the data. In other words, small changes in the data imply at most small changes in the measured persistence. This has major ramifications, including the study of time series and the comparison and classification of shapes. Of particular importance are biological shapes, which their sheer endless variety in the midst of unmistaken similarity and delicate variation. The scope of this book does not extend to this fascinating topic, but we are confident that the proper development of persistence as a measurement tool will facilitate future inroads in this direction.

VIII.1 Time Series

In this section, we study how continuous change of the data affects the measured persistence. We focus on the structural effects and on their computation. An off-shot of the analysis is a first proof of stability, but this will have to wait until the next section.

Straight-line homotopy. Let $f : K \to \mathbb{R}$ and $g : K \to \mathbb{R}$ be two monotonic functions on the same simplicial complex. We recall this means that the functions are non-decreasing along increasing chains of the face relation. We use the straight-line homotopy $F : K \times [0, 1] \to \mathbb{R}$ defined by

$$F(\sigma, t) \quad = \quad (1 - t)f(\sigma) + tg(\sigma)$$

to interpolate between f and g. Define $f_t(\sigma) = F(\sigma, t)$ and note that $f_0 = f$ and $f_1 = g$, as intended. Furthermore, f_t is monotonic for each $t \in [0, 1]$. Indeed, if σ is a face of τ then $f(\sigma) \leq f(\tau)$ and $g(\sigma) \leq g(\tau)$ and therefore $f_t(\sigma) \leq f_t(\tau)$ for every $t \in [0, 1]$. Hence, we can find a compatible ordering of the simplices, that is, a total order that extends the partial orders defined by f_t and by the face relation. Using this compatible ordering, we compute the persistence diagrams of f_t as explained in the previous chapter. However, if we somehow already have the diagrams for f then we may consider modifying them to get the diagrams for f_t. This turns out to be more efficient than recomputing the diagrams provided the two total orders are not too different. To describe what exactly this means, we plot the function values with time, giving us a straight line for each simplex; see Figure VIII.1. It is convenient to assume that f

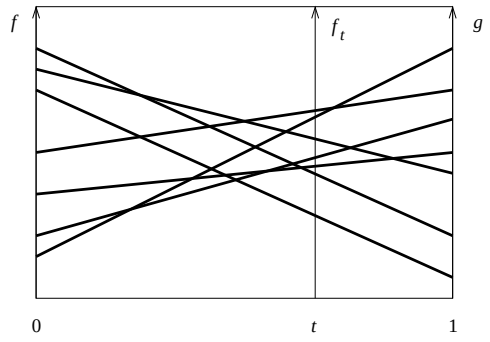

Figure VIII.1: Each line tracks the function value of a simplex as t increases. At any moment $t \in [0, 1]$, we get f_t by intersection with the corresponding vertical line.

and g are injective because this implies that f_t is injective except at finitely many moments t when two or more of the lines cross. To further simplify the situation, we may assume that no two different pairs of lines cross at the same moment. Equivalently, every f_t has at most one violation of injectivity, namely at most two simplices with the same function value. As we sweep from left to right, in the direction of increasing t, we pass through this violation by transposing the two simplices in the compatible ordering. This motivates us to study the impact of a transposition on persistence.

Matrix decomposition. We recall that we compute the persistence diagrams of $f : K \to \mathbb{R}$ by reducing the boundary matrix whose rows and columns are ordered like the simplices in a compatible ordering. Starting with $R = \partial$, we perform left-to-right column additions until R is reduced, that is, each non-zero column has its lowest one in a unique row. In other words, the mapping from non-zero columns to rows defined by low is injective. Each lowest one gives a pair of simplices, namely (σ_i, σ_j) if $i = low(j)$, and a finite off-diagonal point in the p-th persistence diagram, namely $(f(\sigma_i), f(\sigma_j))$ in $\mathrm{Dgm}_p(f)$ with $p = \dim \sigma_i$. It will be convenient to assume a bijection between the lowest ones and the off-diagonal points in the persistence diagrams. In other words, we assume there are no off-diagonal points at infinity or, equivalently, every zero column in the reduced matrix corresponds to a row with a lowest one. We get this property in reduced homology iff K is homologically trivial, that is, $\tilde{\beta}_p(K) = 0$ for every p. Assuming this property is no loss of generality since we can always add simplices at the end so that they do not alter the earlier homological evolution along the filtration. For example, we can form the cone over a given simplicial complex, which is necessarily homologically trivial.

The reduced matrix can be written as $R = \partial V$, where V keeps track of the column operations. Its j-th column stores the chain whose boundary is stored in the j-th column of R. Since we only use left-to-right column additions, the matrix V is upper triangular, with $V[i, i] = 1$ for each i and therefore invertible. Let U be the right inverse of V and note that it is again upper triangular and invertible. Multiplying from the right, we get $RU = \partial VU$ and therefore

$$\partial \quad = \quad RU.$$

We call this an *ru-decomposition* of the boundary matrix. Implicit in this definition are the requirements that U be upper triangular and invertible and that R be reduced. We get these properties from the way we compute the matrices, but there are other ru-decompositions that may be obtained by other, similar algorithms. Indeed, the ru-decomposition of ∂ is not unique but as noted in the previous chapter, the lowest ones in the reduced matrix are. The

specific question we now ask is how we can update the ru-decomposition of the boundary matrix if we transpose two simplices in contiguous positions along the compatible ordering.

Updating the decomposition. Suppose ∂ is the boundary matrix for the ordering of the simplices as $\sigma_1, \sigma_2, \ldots, \sigma_m$. We write ∂' for the boundary matrix after transposing σ_i with σ_{i+1}. Letting $P = P_i^{i+1}$ be the corresponding permutation matrix, we have $\partial' = P\partial P$. The difference between P and the unit matrix, I, is localized to the 2-by-2 submatrix for which $P[i,i] = P[i+1,i+1] = 0$ and $P[i,i+1] = P[i+1,i] = 1$. Multiplying with P from the left exchanges the two rows and multiplying with P from the right exchanges the two columns. Note also that P is its own inverse, that is, $PP = I$. We therefore get

$$\partial' \quad = \quad P\partial P \quad = \quad PRUP \quad = \quad (PRP)(PUP).$$

But this is not necessarily an ru-decomposition of the new boundary matrix. It fails to be one if $R' = PRP$ is not reduced or if $U' = PUP$ is not upper triangular. We will now show that either deficiency can be remedied with modest effort, namely a constant number of row and column operations.

The only way R' fails to be reduced is when rows i and $i+1$ of R both contain a lowest one, $i = low(k)$ and $i+1 = low(l)$, and row i has a one in column l as well. There are two cases, distinguished by $k < l$ and $l < k$. In both cases, we add the left column to the right column before we do the transposition. This fixes the deficiency, as illustrated in Figure VIII.2.

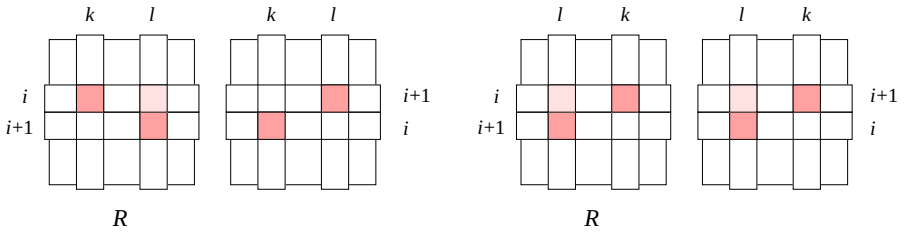

Figure VIII.2: After swapping rows i and $i+1$ in R, the matrix would be no longer reduced. We thus add the left to the right column before exchanging the two rows.

The only way U' fails to be upper triangular is when $U[i,i+1] = 1$. We fix this deficiency by adding row $i+1$ to row i in U and adding column i to column $i+1$ in R. Letting $S = S_i^{i+1}$ be the matrix whose only difference to the unit matrix is $S[i,i+1] = 1$, we thus consider SU and RS. Since $SS = I$, this does not change the matrix product, that is, $\partial' = (PRSP)(PSUP) = PRSP$,

same as before. With this modification, $PSUP$ is upper triangular, but $PRSP$ may again fail to be reduced. If column i is zero or $low(i) < low(i+1)$ then multiplying with S preserves the lowest ones and RS is reduced. In this case, we have an ru-decomposition after the transposition. On the other hand, if column $i+1$ is zero while column i is not or if $low(i) > low(i+1)$, as in Figure VIII.3, then we need to make the lowest ones unique again. We do this by adding column $i+1$ to column i, but after the transposition so that this is again a left-to-right column addition. This repairs all deficiencies and we have an ru-decomposition of ∂'.

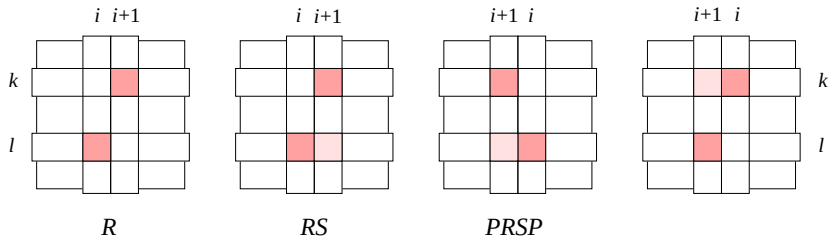

Figure VIII.3: After adding column i to $i+1$ and exchanging the two columns, the matrix $PRSP$ is no longer reduced. Adding column $i+1$ to i after the transposition finally produces a reduced matrix.

Transpositions that change the pairing. The more important changes require more work. These are the *switches* , which we define as the transpositions that alter the pairing. Recall that each lowest one establishes a correspondence between a positive simplex (a row) and a negative simplex (a column). For example, in Figure VIII.2 on the left, we have the pairs (σ_i, σ_k) and (σ_{i+1}, σ_l) which are preserved through the transposition. On the right, we have the same two pairs but they change to (σ_i, σ_l) and (σ_{i+1}, σ_k). This identifies the transposition as a switch. In Figure VIII.3, we have the pairs (σ_k, σ_{i+1}) and (σ_l, σ_i) which change to (σ_k, σ_i) and (σ_l, σ_{i+1}), again a switch.

As a rule of thumb, most transitions are not switches. For example, if σ_i and σ_{i+1} do not have the same dimension then their transposition does not require any changes other than the obligatory swapping of rows and columns. Even if they have the same dimension but if σ_i is positive and σ_{i+1} is negative, then the transposition cannot be a switch. This is because row i has no lowest one, so $R' = PRP$ is reduced and requires no further effort. Similarly, column i of R is zero so we can set $U[i, i+1] = 0$ to make sure $U' = PUP$ is upper triangular, if necessary. In words, the ru-decomposition is maintained without

any of the repair operations that change the pairing. However, the remaining three combinations of types can be switches, and we see an example each in Figure VIII.4. We get a switch between two positive vertices, v and w, when we

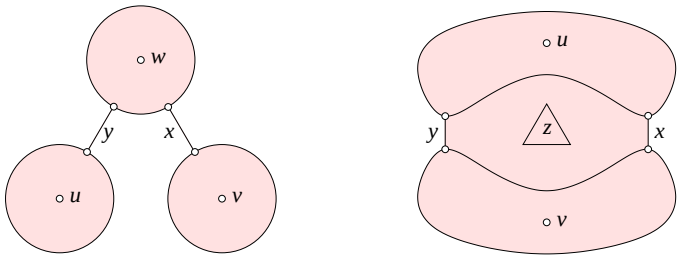

Figure VIII.4: The vertices u, v, w are the oldest in their respective components, which are eventually joined by the edges x and y. On the right, the two edges form a hole, which is eventually filled by the triangle z.

go from $uvwxy$ to $uwvxy$ on the left. Indeed, the pairs (v, y) and (w, x) before the transposition of v and w change to (v, x) and (w, y) after the transposition. We get a switch between two negative edges, x and y, when we go from $uvwxy$ to $uvwyx$, again on the left. Indeed, the transposition of x and y produces the same change between pairs as in the previous example. Finally, we get a switch between a negative edge, x, and a positive edge, y, when we go from $uvxyz$ to $uvyxz$ on the right. Indeed, the pairs (v, x) and (y, z) before the transposition of x and y change to (v, y) and (x, z) after the transposition. The last switch is the most interesting of all. Besides changing the pairing, it convinces the negative x to become positive and the positive y to become negative. The two edge thus contribute to different persistence diagrams before and after the transposition.

Summary. When we transpose σ_i and σ_{i+1}, we touch only the columns of σ_i and σ_{i+1} and of the simplices σ_k and σ_l paired with them. The changes are therefore limited to these two pairs. Furthermore, there are no changes unless the transposed simplices have the same dimension. Assuming $p = \dim \sigma_i = \dim \sigma_{i+1}$, the other two simplices have dimension $p - 1$ and $p + 1$. The only possible change is therefore that the transposed simplices trade places. We state this result for later reference.

TRANSPOSITION LEMMA. Let ∂ and ∂' be the boundary matrices for compatible orderings of two monotonic functions on a simplicial complex that differ

by a single transposition of two contiguous simplices, σ_i and σ_{i+1}. Then the pairings defined by ru-decompositions $\partial = RU$ and $\partial' = R'U'$ differ only if $\dim \sigma_i = \dim \sigma_{i+1}$, and if they differ then only by σ_i and σ_{i+1} trading places.

The computational effort for updating the ru-decomposition is modest, namely a constant number of row and column operations, each computable in time proportional to the number of simplices. Returning to our two monotonic functions, $f, g : K \to \mathbb{R}$, we have m simplices and thus at most $\binom{m}{2}$ transpositions to go from a compatible ordering for f to a compatible ordering for g. To get started, we compute the persistence diagrams of f in m^3 time using the algorithm explained in Section VII.1. Thereafter, we spend m time per transposition and therefore $m\binom{m}{2} < m^3$ time in total until we arrive at the persistence diagrams of g. This is roughly the same amount of time required to compute the diagrams of g from scratch, at least in the worst case. However, going through the transposition has the advantage that we get the interpolating diagrams for free.

Bibliographic notes. The material of this section is taken from [1], where continuous families of persistence diagrams are proposed as a tool to study time series of functions. As explained, the algorithm constructs these families by maintaining the ru-decomposition of the boundary matrix through a sequence of transpositions scheduled by sweeping an arrangement of lines. We can find these transpositions in logarithmic time each by sorting the crossings, or in constant time each by sweeping the arrangement topologically [2].

[1] D. Cohen-Steiner, H. Edelsbrunner and D. Morozov. Vines and vineyards by updating persistence in linear time. *In* "Proc. 22nd Ann. Sympos. Comput. Geom., 2006", 119–126.

[2] H. Edelsbrunner and L. J. Guibas. Topologically sweeping an arrangement. *J. Comput. System Sci.* **38** (1989), 165–194. Corrigendum. *J. Comput. System Sci.* **42** (1991), 249–251.

VIII.2 Stability Theorems

Like any good measurement device, persistence gives similar readings for similar functions. We make this statement precise for two notions of similarity between persistence diagrams. The bottleneck distance is the cruder of the two but leads to a more general result. The Wasserstein distance is more sensitive to details in the diagrams but requires additional properties to be stable.

Bottleneck distance. Recall that a persistence diagram is a multiset of points in the extended plane, $\bar{\mathbb{R}}^2$. Under the assumptions on the input functions considered in this book, the diagram consists of finitely many points above the diagonal. To this finite multiset, we add the infinitely many points on the diagonal, each with infinite multiplicity. These extra points are not essential to the diagram but their presence simplifies upcoming definitions. Let now X and Y be two persistence diagrams. To define the distance between them, we consider bijections $\eta : X \to Y$ and record the supremum of the distances between corresponding points for each. Measuring distance between points $x = (x_1, x_2)$ and $y = (y_1, y_2)$ as $\|x - y\|_\infty = \max\{|x_1 - y_1|, |x_2 - y_2|\}$ and taking the infimum over all bijections, we get the *bottleneck distance* between the diagrams,

$$W_\infty(X, Y) \quad = \quad \inf_{\eta: X \to Y} \sup_{x \in X} \|x - \eta(x)\|_\infty .$$

As illustrated in Figure VIII.5, we can draw squares of side length twice the bottleneck distance centered at the points of X so that each square contains

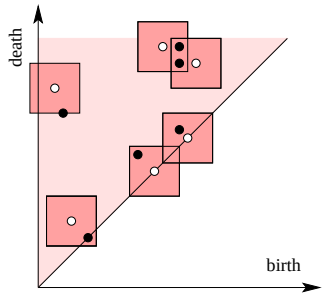

Figure VIII.5: The superposition of two persistence diagrams consisting of the white and the black points. Only the marked points on the diagonal correspond to off-diagonal points in the other diagram. The bottleneck distance is half the side length of the squares illustrating the bijection.

the corresponding point of Y. Clearly, $W_\infty(X,Y) = 0$ iff $X = Y$. Furthermore, $W_\infty(X,Y) = W_\infty(Y,X)$ and $W_\infty(X,Z) \leq W_\infty(X,Y) + W_\infty(Y,Z)$. We see that W_∞ satisfies all axioms of a metric and thus deserves to be called a distance.

Bottleneck stability. Letting $f, g : K \to \mathbb{R}$ be two monotonic functions, we consider the straight-line homotopy $f_t = (1-t)f + tg$, same as in the previous section. This gives a monotonic function f_t with a persistence diagram for each dimension p and each $t \in [0,1]$. Fixing a dimension p, the family of persistence diagrams is a multiset in $\bar{\mathbb{R}}^2 \times [0,1]$. Drawing t along a third coordinate axis, we get a three-dimensional visualization of how the persistent homology evolves as we go from $f = f_0$ to $g = f_1$. To describe this, we assume that K has no non-trivial (reduced) homology, same in the previous section. Adding the third coordinate, each off-diagonal point of $X_t = \mathrm{Dgm}_p(f_t)$ is of the form $x(t) = (f_t(\sigma), f_t(\tau), t)$, where σ and τ are simplices in K. The point represent the fact that when we construct K by adding the simplices in the order defined by f_t, then adding σ gives birth to a p-dimensional homology class and adding τ gives death to the same. There are only finitely many values at which the pairing of the simplices changes, and we denote these as $0 = t_0 < t_1 < \ldots < t_n < t_{n+1} = 1$. Within each interval (t_i, t_{i+1}), the pairing is constant and each pair σ, τ gives rise to a line segment of points $x(t)$ connecting points in the planes $t = t_i$ and $t = t_{i+1}$. If the endpoint is an off-diagonal point at t_{i+1}, then there is a unique other line segment that begins at that point. This line segment may correspond to the same simplex pair and thus continue on the same straight line, or it may correspond to a different pair created in a switch and make a turn at the shared point. It is also possible that the endpoint lies on the diagonal at $t_{i=1}$, in which case there is no continuation. In summary, the line segments form polygonal paths that monotonically increase in t. Each path begins at an off-diagonal point in $X = X_0$ or at a diagonal point in some X_{t_i} and it ends at an off-diagonal point in $Y = X_1$ or at a diagonal point in some X_{t_j}. We call each polygonal path a *vine* and the multiset of vines a *vineyard*; see Figure VIII.6.

The fact that the points in the family of persistence diagrams form connected vines is important. It is a way of saying that the persistence diagram is stable. To further quantify this notion, we differentiate $x(t) = (1-t)(f(\sigma), f(\tau), 0) + t(g(\sigma), g(\tau), 1)$ and get $\frac{\partial x}{\partial t}(t) = (g(\sigma) - f(\sigma), g(\tau) - f(\tau), 1)$. Projecting the endpoints of the line segment back into $\bar{\mathbb{R}}^2$, we get two points whose L_∞-distance is $t_{i+1} - t_i$ times the larger of the differences between f and g at the two simplices. Letting v be the simplex in K that maximizes this difference, we get the L_∞-distance between the two functions, $\|f - g\|_\infty = |f(v) - g(v)|$. This is also an upper bound on the slope of any line segment in the vineyard and

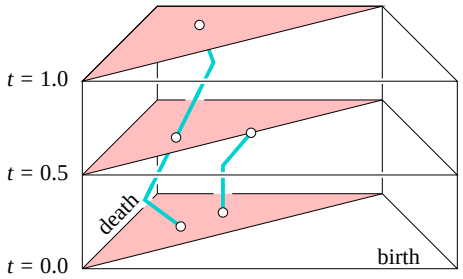

Figure VIII.6: The 1-parameter family of persistence diagrams of the straight-line homotopy between $f = f_0$ and $g = f_1$. One point traces out a vine spanning the entire interval while the other merges into the diagonal halfway through the homotopy.

therefore an upper bound on the L_∞-distance between the projected endpoints of any vine.

STABILITY THEOREM FOR FILTRATIONS. Let K be a simplicial complex and $f, g : K \to \mathbb{R}$ two monotonic functions. For each dimension p, the bottleneck distance between the diagrams $X = \mathrm{Dgm}_p(f)$ and $Y = \mathrm{Dgm}_p(g)$ is bounded from above by the L_∞-distance between the functions, $W_\infty(X, Y) \leq \|f - g\|_\infty$.

Tame functions. To apply the Stability Theorem, it is convenient to get it into a form that allows for more general functions. According to the Simplicial Approximation Theorem in Chapter III, every continuous function on a triangulable topological space can be approximated by a piecewise linear function, and as shown in Chapter VII, for every piecewise linear function there is a monotonic function that generates the same persistence diagrams. It is therefore not surprising that what we said about filtrations can indeed be generalized. We explain this for functions that satisfy a mild tameness condition.

Let $\mathbb{X}$ be triangulable and $f : \mathbb{X} \to \mathbb{R}$ continuous. Given a threshold $a \in \mathbb{R}$, the *sublevel set* consists of all points $x \in \mathbb{R}$ with function value less than or equal to a, $\mathbb{X}_a = f^{-1}(-\infty, a]$. Similar to the complexes in a filtration, the sublevel sets are nested and give rise to a sequence of homology groups connected by maps induced by inclusion, one for each dimension. Writing $f_p^{a,b} : \mathsf{H}_p(\mathbb{X}_a) \to \mathsf{H}_p(\mathbb{X}_b)$ for the map from the p-th homology group of the sublevel set at a to that at b, we call its image a *persistent homology group*, as before. The corresponding *persistent Betti number* is $\beta_p^{a,b} = \operatorname{rank} \operatorname{im} f_p^{a,b}$. As long as the topology of the sublevel set does not change, the maps between the homology groups are isomorphisms. We thus call $a \in \mathbb{R}$ a *homological critical*

value if there is no $\varepsilon > 0$ for which $f_p^{a-\varepsilon,a+\varepsilon}$ is an isomorphism for each dimension p. Finally, we call f *tame* if it has only finitely many homological critical values and all homology groups of all sublevel sets have finite rank. The main motivation for this definition is the relative ease with which we can define persistence diagrams. Letting $a_1 < a_2 < \ldots < a_n$ be the homological critical values of f, we construct interleaved values b_0 to b_n with $b_{i-1} < a_i < b_i$ for all i. Adding $b_{-1} = a_0 = -\infty$ and $a_{n+1} = b_{n+1} = \infty$, we consider the corresponding sequence of homology groups,

$$0 = \mathsf{H}_p(\mathbb{X}_{b_{-1}}) \to \mathsf{H}_p(\mathbb{X}_{b_0}) \to \ldots \to \mathsf{H}_p(\mathbb{X}_{b_n}) \to \mathsf{H}_p(\mathbb{X}_{b_{n+1}}) = \mathsf{H}_p(\mathbb{X}),$$

and the maps between them. For $0 \leq i < j \leq n+1$, the *multiplicity* of the pair a_i, a_j is now defined as $\mu_p^{a_i,a_j} = (\beta_p^{b_i,b_{j-1}} - \beta_p^{b_i,b_j}) - (\beta_p^{b_{i-1},b_{j-1}} - \beta_p^{b_{i-1},b_j})$. To get the *p-th persistence diagram* of f, we draw each point (a_i, a_j) with multiplicity $\mu_p^{a_i,a_j}$, and we add the points of the diagonal, each with infinite multiplicity. With these definitions, we have the following stability result, which we state without proof, illustrating it in Figure VIII.7.

Stability Theorem for Tame Functions. Let $\mathbb{X}$ be a triangulable topological space and $f, g : \mathbb{X} \to \mathbb{R}$ two tame functions. For each dimension p, the bottleneck distance between $X = \mathrm{Dgm}_p(f)$ and $Y = \mathrm{Dgm}_p(g)$ is bounded by the L_∞-distance between the functions, $W_\infty(X, Y) \leq \|f - g\|_\infty$.

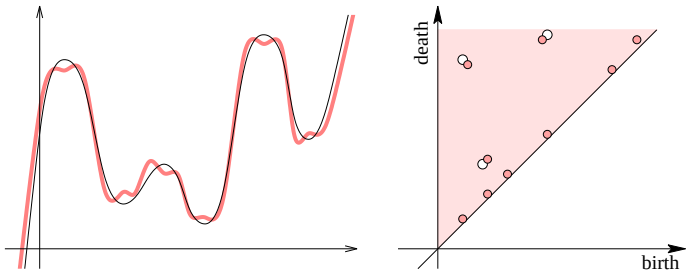

Figure VIII.7: Left: two functions with small L_∞-distance. Right: the corresponding two persistence diagrams with small bottleneck distance.

Wasserstein distance. A drawback of the bottleneck distance is its insensitivity to details of the bijection beyond the furthest pair of corresponding points. To remedy this shortcoming, we introduce the *degree q Wasserstein distance* between X and Y for any positive real number q. It takes the sum

of q-th powers of the L_∞-distances between corresponding points, again minimizing over all bijections,

$$W_q(X,Y) \quad = \quad \left[\inf_{\eta:X\to Y} \sum_{x\in X} \|x-\eta(x)\|_\infty^q\right]^{1/q}.$$

As suggested by our notation, the bottleneck distance is the limit of the Wasserstein distance for q going to infinity. Similar to the bottleneck distance, it is straightforward to verify that W_q satisfies the requirements of a metric and thus deserves to be called a distance.

It should be obvious that we cannot substitute the degree q Wasserstein distance for the bottleneck distance and expect that the Stability Theorem for Tame Functions still holds. Indeed, we can approximate a function $f:\mathbb{R}\to\mathbb{R}$ with a function g that has arbitrarily many wrinkles without deviating from f by more than some positive ε; see Figure VIII.7. Each wrinkle generates a point with persistence about 2ε in the 0-th persistence diagram. Making the wrinkles narrow we can get an arbitrarily large number and therefore an arbitrarily large Wasserstein distance between the diagrams of f and g.

Wasserstein stability. Although a general stability result like for the bottleneck distance is out of reach, we get stability under the Wasserstein distance for a reasonably large class of functions. Let $\mathbb{X}$ be a metric space, that is, a topological space for which the distance between points $x,y\in\mathbb{X}$, denoted as $\|x-y\|$, is well defined. A function $f:\mathbb{X}\to\mathbb{R}$ is *Lipschitz* if there is a constant C such that $|f(x)-f(y)|\leq\|x-y\|$ for all points $x,y\in\mathbb{X}$. Without loss of generality, we only consider Lipschitz functions with constant $C=1$. This condition prevents narrow wrinkles. Indeed, each wrinkle now requires an amount of space that relates to its persistence. It is therefore not possible to crowd arbitrarily many wrinkles together without shrinking their persistence. What we suggest here is a packing argument, the metric version of the combinatorial pigeonhole principle, but homology classes can interact so that the packing argument cannot be applied directly. Indeed, making it a rigorous proof is work which we rather skip. Instead, we introduce the precise conditions on the space $\mathbb{X}$ for which we can prove stability of persistence.

Assume $\mathbb{X}$ is triangulable and consider a triangulation, that is, a simplicial complex K together with a homeomorphism $\phi:|K|\to\mathbb{X}$. Letting its mesh be the maximum distance between the images of two points of the same simplex in K, we define $N(r)$ as the minimum number of simplices in a triangulation with mesh at most r. We say the triangulations of $\mathbb{X}$ *grow polynomially* if there are constants c and j such that $N(r)\leq c/r^j$. Finally, we define the *degree* k

total persistence of a persistence diagram X as the sum of k-th powers of the persistences of its points, $\text{Pers}_k(X) = \sum_{x \in X} \text{pers}(x)^k$. The main technical insight is that polynomial growth implies bounded total persistence. Specifically, if $\mathbb{X}$ is a metric space whose triangulations grow polynomially with constant exponent j, $f : \mathbb{X} \to \mathbb{R}$ is Lipschitz, and $X = \text{Dgm}_p(f)$, then $\text{Pers}_k(X)$ is bounded from above by a constant for every $k > j$. The proof of this implication is omitted. For example the d-dimensional sphere is triangulable and its triangulations grow polynomially, with constant exponent $j = d$. It follow that for every $k > d$, the degree k total persistence of a Lipschitz function on the sphere is bounded by a constant. Using these ingredients, we are now ready to prove an upper bound on the Wasserstein distance that implies stability for $q > k$.

Stability Theorem for Lipschitz Functions. Let $f, g : \mathbb{X} \to \mathbb{R}$ be tame Lipschitz functions on a metric space whose triangulations grow polynomially with constant exponent j. Then there are constants C and $k > j$ no smaller than one such that the degree q Wasserstein distance between $X = \text{Dgm}_p(f)$ and $Y = \text{Dgm}_p(g)$ is $W_q(X, Y) \leq C \cdot \|f - g\|_\infty^{1-k/q}$ for every $q \geq k$.

Proof. Let $\eta : X \to Y$ be a bijection that realizes the bottleneck distance, that is, $\|x - \eta(x)\|_\infty \leq \varepsilon = \|f - g\|_\infty$ for each point $x \in X$. In addition, we require that $\|x - \eta(x)\|_\infty \leq \frac{1}{2}[\text{pers}(x) + \text{pers}(\eta(x))]$. Indeed, if this inequality does not hold then $\text{pers}(x) \leq 2\varepsilon$ and $\text{pers}(\eta(x)) \leq 2\varepsilon$ and we can change the bijection by matching both with points on the diagonal within L_∞-distance ε. The q-th power of the degree q Wasserstein distance is therefore

$$\begin{aligned} W_q(X,Y)^q &\leq \sum_{x \in X} \|x - \eta(x)\|_\infty^q \\ &\leq \varepsilon^{q-k} \sum_{x \in X} \|x - \eta(x)\|_\infty^k \\ &\leq \frac{\varepsilon^{q-k}}{2^k} \sum_{x \in X} [\text{pers}(x) + \text{pers}(\eta(x))]^k \\ &\leq \frac{\varepsilon^{q-k}}{2^k} \sum_{x \in X} [(2\text{pers}(x))^k + (2\text{pers}(\eta(x)))^k]. \end{aligned}$$

The last step uses the fact that taking the k-th power is convex. The sum is 2^k times the degree k total persistences of the two diagrams, $W_q(X, Y)^q \leq \varepsilon^{q-k}[\text{Pers}_k(X) + \text{Pers}_k(Y)]$. By assumption, they are bounded by a constant. Taking the q-th root thus gives the claimed inequality. ∎

Bibliographic notes. Vineyards have been introduced as a tool to study time-series of functions in [4]. The proof that the vines in it are connected polygonal paths is equivalent to establishing the stability for monotonic functions under the bottleneck distance between diagrams. The first proof of stability goes back to Cohen-Steiner, Edelsbrunner and Harer [2] who used an algebraic argument to establish it for tame functions. A further generalization to 1-parameter families of vector spaces can be found in [1]. A proof of the Stability Theorem for Lipschitz Functions along with applications in systems biology can be found in [3]. The Wasserstein distance is named after the author of [8]. It is related to optimal transportation as studied by Monge [6] and Kantorovich [5]; see also [7].

[1] F. CHAZAL, D. COHEN-STEINER, L. J. GUIBAS AND S. Y. OUDOT. The stability of persistence diagrams revisited. Rept. 6568, Centre de recherche INRIA Saclay, Orsay, France, 2008

[2] D. COHEN-STEINER, H. EDELSBRUNNER AND J. HARER. Stability of persistence diagrams. *Discrete Comput. Geom.* **37** (2007), 103–120.

[3] D. COHEN-STEINER, H. EDELSBRUNNER, J. HARER AND Y. MILEYKO. Lipschitz functions have L_p-stable persistence. *Found. Comput. Math.*, to appear.

[4] D. COHEN-STEINER, H. EDELSBRUNNER AND D. MOROZOV. Vines and vineyards by updating persistence in linear time. *In* "Proc. 22nd Ann. Sympos. Comput. Geom., 2006", 119–126.

[5] L. V. KANTOROVICH. On the translocation of masses. *C. R. (Dokl.) Acad. Sci. USSR* **37** (1942), 199-226.

[6] G. MONGE. Mémoire sur la théorie des déblais et des remblais. In *Histoire de l'Académie Royale des Sciences de Paris* (1781), 666–704.

[7] C. VILLANI. *Topics in Optimal Transportation.* Amer. Math. Soc., Providence, Rhode Island, 2003.

[8] L. N. WASSERSTEIN. Markov processes over denumerable products of spaces describing large systems of automata. *Problems of Information Transmission* **5** (1969), 47–52.

VIII.3 Length of a Curve

In this section, we use the stability of persistence to generalize a classic result on curves, proving an inequality connecting the lengths and total curvatures of two curves.

Closed curves. We consider a closed curve $\gamma : \mathbb{S}^1 \to \mathbb{R}^2$, with or without self-intersections. Assuming γ is smooth, we have derivatives of all orders. The *speed at a point* $\gamma(s)$ is the length of the velocity vector, $\|\dot{\gamma}(s)\|$. We can use it to compute the length as the integral over the curve,

$$\mathit{length}(\gamma) \quad = \quad \int_{s \in \mathbb{S}^1} \|\dot{\gamma}(s)\| \, \mathrm{d}s.$$

It is convenient to assume a constant speed parametrization, that is, $\mathit{speed} = \|\dot{\gamma}(s)\| = \mathit{length}(\gamma)/2\pi$ for all $s \in \mathbb{S}^1$. With this assumption, the *curvature at a point* $\gamma(s)$ is the norm of the second derivative divided by the square of the speed, $\kappa(s) = \|\ddot{\gamma}(s)\|/\mathit{speed}^2$. One over the curvature is the radius of the circle that best approximates the shape of the curve at the point $\gamma(s)$. To interpret this formula geometrically, we follow the velocity vector as we trace out the curve. Since its length is constant, it sweeps out a circle of radius *speed*, as illustrated in Figure VIII.8. The curvature is the speed at which the

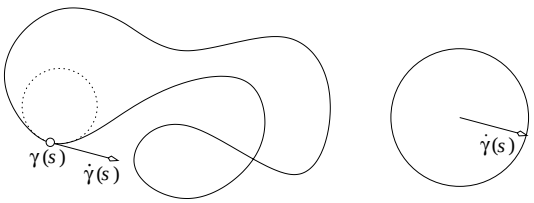

Figure VIII.8: A curve with constant speed parametrization and its velocity vector sweeping out a circle with radius equal to the speed.

unit tangent vector sweeps out the unit circle as we move the point with unit speed along the curve. This explains why we divide by the speed twice, first to compensate for the length of the velocity vector and second for the actual speed. The *total curvature* is the distance traveled by the unit tangent vector,

$$\mathit{curv}(\gamma) \quad = \quad \mathit{speed} \int_{s \in \mathbb{S}^1} \kappa(s) \, \mathrm{d}s.$$

As an example consider the constant speed parametrization of the circle with radius r, $\gamma(s) = rs$. Writing a point in terms of its angle, we get

$$s = \begin{bmatrix} \cos\varphi \\ \sin\varphi \end{bmatrix}, \quad \gamma(s) = \begin{bmatrix} r\cos\varphi \\ r\sin\varphi \end{bmatrix}, \quad \dot{\gamma}(s) = \begin{bmatrix} -r\sin\varphi \\ r\cos\varphi \end{bmatrix}.$$

We thus have $\mathit{speed} = r$ and $\mathit{length}(\gamma) = \int \mathit{speed}\,\mathrm{d}s = 2\pi r$. The curvature is $\kappa(s) = \|\ddot{\gamma}(s)\|/\mathit{speed}^2 = 1/r$, which is of course independent of the location on the circle. The total curvature is $\mathit{curv}(\gamma) = \int \frac{r}{r}\,\mathrm{d}s = 2\pi$, which is independent of the radius. Indeed, the unit tangent vector travels once around the unit circle, no matter how small or how big the parametrized circle is.

Integral geometry. The length and total curvature of a curve can also be expressed in terms of integrals of elementary quantities. We begin with the length. Take a unit length line segment in the plane. The lines that cross the line segment at an angle φ form a strip of width $\sin\varphi$. Integrating over all angles gives $\int_{\varphi=0}^{\pi} \sin\varphi\,\mathrm{d}\varphi = [-\cos\varphi]_0^{\pi} = 2$. In words, the integral of the number of intersections over all lines in the plane is twice the length of the line segment. Since we can approximate a curve by a polygon whose total length approaches that of the curve, the same holds for our curve γ. To express this result formally, we introduce $g_u : \mathbb{R}^2 \to \mathbb{R}$ defined by $g_u(x) = \langle u, x\rangle$, mapping each point $x \in \mathbb{R}^2$ to its height in the direction $u \in \mathbb{S}^1$. The preimage of a value $z \in \mathbb{R}$, $g_u^{-1}(z)$, is the line with normal direction u and offset z. The composition with the curve, $f_u = g_u \circ \gamma$, maps each $s \in \mathbb{S}^1$ to the height of the point $\gamma(s)$. The preimage of this function thus corresponds to points at which the line intersects the curve. We are now ready to formulate the length of the curve in terms of the number of intersections.

CAUCHY-CROFTON FORMULA. The length of a curve in the plane is one quarter the integral of the number of intersections with lines,

$$\mathit{length}(\gamma) \quad = \quad \frac{1}{4}\int_{u\in\mathbb{S}^1}\int_{z\in\mathbb{R}} \mathrm{card}\,(f_u^{-1}(z))\,\mathrm{d}z\,\mathrm{d}u.$$

Here we divide by two twice, once because $\int \sin\varphi\,\mathrm{d}\varphi = 2$ and again because we integrate over all $u \in \mathbb{S}^1$ and therefore over all lines twice.

To get an integral geometry expression of the total curvature, we again consider a direction $u \in \mathbb{S}^1$ and the height of the curve in that direction, $f_u : \mathbb{S}^1 \to \mathbb{R}$. For generic directions u, this height function has a finite number of minima and maxima, as illustrated in Figure VIII.9. Recall that the total

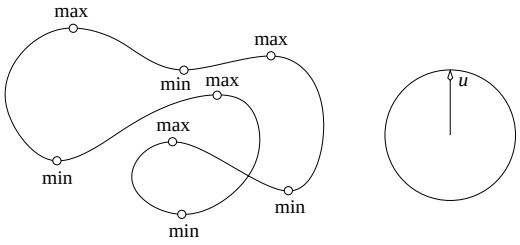

Figure VIII.9: The vertical height function defined on the curve has four local minima which alternate with the four local maxima along the curve.

curvature is the length traveled by the unit tangent vector. Equivalently, it is the length traveled by the outward unit normal vector. The number of maxima of f_u is the number of times the unit normal passes $u \in \mathbb{S}^1$ and the number of minima is the number of times it passes $-u \in \mathbb{S}^1$. Writing $\#crit(f_u)$ for the number of minima and maxima, we get the total curvature by integration.

Total Curvature Formula. The total curvature of a smooth curve in the plane is half the integral of the number of critical points over all directions,

$$curv(\gamma) \quad = \quad \frac{1}{2} \int_{u \in \mathbb{S}^1} \#crit(f_u) \, \mathrm{d}u.$$

The integral in the above formula can be interpreted as 2π times the average number of critical points, where the average is taken over all directions. Hence the total curvature is π times this average.

Theorems relating length with total curvature. Suppose the image of γ fits inside the unit disk in the plane, $\operatorname{im}\gamma \subseteq \mathbb{B}^2$. Then γ must turn to avoid crossing the boundary circle of the disk. We can therefore expect that the total curvature is bounded from below by some constant times the length. A classic result in geometry asserts that this constant is one.

Fáry Theorem. Let $\gamma : \mathbb{S}^1 \to \mathbb{R}$ be a smooth closed curve with $\operatorname{im}\gamma \subseteq \mathbb{B}^2$. Then its length is at most its total curvature, $length(\gamma) \leq curv(\gamma)$.

To generalize this result, we consider two curves, $\gamma, \gamma_0 : \mathbb{S}^1 \to \mathbb{R}^2$, and the 'shortest leash distance' between them. Specifically, we trace out both curves simultaneously and connect the two moving points by a leash so that their

distance can never exceed the length of that leash. Formally, this concept is known as the *Fréchet distance* between the curves. To define it, we record the leash length for a homeomorphism $\eta : \mathbb{S}^1 \to \mathbb{S}^1$ and take the infimum over all homeomorphisms, $F(\gamma, \gamma_0) = \inf_\eta \max_s \|\gamma(s) - \gamma_0(\eta(s))\|$. This notion of distance does not depend on the parametrizations of the two curves.

GENERALIZED FÁRY THEOREM. Let $\gamma, \gamma_0 : \mathbb{S}^1 \to \mathbb{R}^2$ be two smooth closed curves. Then $|\mathit{length}(\gamma) - \mathit{length}(\gamma_0)| \leq [\mathit{curv}(\gamma) + \mathit{curv}(\gamma_0) - 2\pi]\, F(\gamma, \gamma_0)$.

To see that Fáry's Theorem is indeed a special case, let the image of γ be contained in the unit disk and let the image of γ_0 be a tiny circle centered at the origin, as in Figure VIII.10. Since γ_0 is a circle, its total curvature is 2π. Furthermore, we can make it arbitrarily small so its length approaches zero. While for some curves γ, the Fréchet distance to γ_0 exceeds one, it approaches the maximum distance from the origin, which is at most one. Substituting 0 for $\mathit{length}(\gamma_0)$, 2π for $\mathit{curv}(\gamma_0)$, and 1 for $F(\gamma, \gamma_0)$ in the Generalized Fáry Theorem gives the original Fáry Theorem.

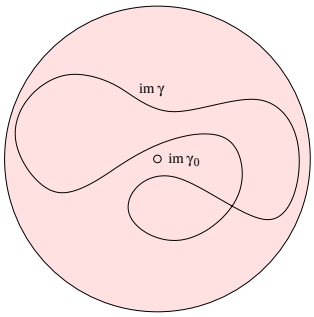

Figure VIII.10: Two curves inside the unit disk. The Fréchet distance between the tiny circle and the other curve approaches a constant at most one as the circle shrinks toward the origin.

Length and total curvature in terms of persistence. A first step toward proving the Generalized Fáry Theorem is a re-interpretation of the length and the total curvature. Fix a direction $u \in \mathbb{S}^1$ and consider $f_u = g_u \circ \gamma$, the height function of the first curve. Almost all level sets, $f_u^{-1}(z)$, consist of an even number of points, decomposing γ into the same number of arcs, half of which belong to the sublevel set, $f_u^{-1}(-\infty, z]$. The number of arcs in the sublevel set is equal to the number of components that are born at or before z and are

still alive at z. To be precise, this is true as long as z does not exceed the height of the global maximum of f_u. To make it true for all height values, we declare that the component born at the global minimum dies at the global maximum; see Figure VIII.11. This is incidentally what we would get with

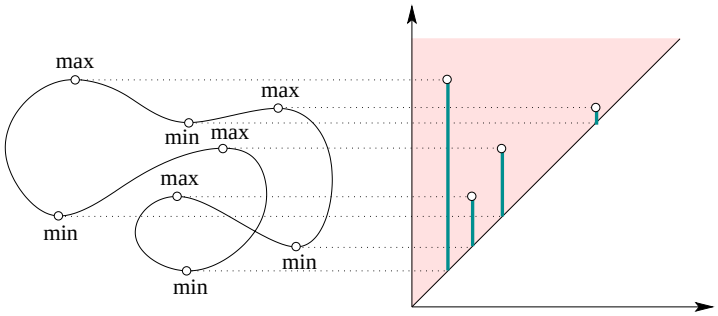

Figure VIII.11: The zeroth persistence diagram of the height function on the curve. We simplify the situation by pairing the global minimum with the global maximum so that all the pair information is contained in this one diagram.

extended persistence as described in the previous chapter. Drawing the vertical lines from the off-diagonal points in the persistence diagram down to the the diagonal gives a set of line segments with total length $\text{Pers}_0(f_u) = \sum \text{pers}(a)$, where the sum is over all points $a \in \text{Dgm}_0(f_u)$. We refer to this quantity as the *zeroth total persistence* of f_u. By what we said above, the number of line segments that intersect the horizontal line at height z is equal to half the number of points in $f_u^{-1}(z)$. Integrating the number of intersections between γ and lines with normal direction u thus gives twice the total persistence,

$$\int_{z\in\mathbb{R}} f_u^{-1}(z) \quad = \quad 2\text{Pers}_0(f_u).$$

The relationship between total curvature and the persistence diagram is even more straightforward. Assuming f_u is Morse, we have a finite number of critical points. This number is even, with equally many minima and maxima paired up to give half the number of off-diagonal points in the persistence diagram. We get similar relationships for the height function of the second curve.

Bounding the difference and integrating. To relate the quantities for the two curves, we write $\varepsilon = F(\gamma, \gamma_0)$ for the Fréchet distance and assume that γ and γ_0 are parametrized such that $\|\gamma(s) - \gamma_0(s)\| \leq \varepsilon$, for all s. It follows

that $|f_u(s) - f_{0,u}(s)| \leq \varepsilon$, for all s. The Stability Theorem for Tame Functions then implies that there is a bijection between the points of $\mathrm{Dgm}_0(f_u)$ and of $\mathrm{Dgm}_0(f_{0,u})$ such that corresponding points have L_∞-distance at most ε. It follows that the difference in persistence between two corresponding points is at most 2ε. If both are off-diagonal points then we have four critical points (two of f_u and two of $f_{0,u}$) we can hold responsible for the difference. However, if an off-diagonal point is matched with a point on the diagonal then we have only two critical points to take responsibility for the 2ε difference. This is indeed the worse of the two possibilities, but we can guarantee that at least two off-diagonal points can be matched within L_∞-distance ε, namely the two points formed by the global min-max pairs. This is because these critical points correspond to points at infinity in the ordinary persistence diagrams, and being at infinity they cannot be matched to points on the diagonal. In summary, the difference in total persistence between f_u and $f_{0,u}$ is at most ε times the number of critical points of f_u and $f_{0,u}$ minus two. We are now ready to integrate over all directions $u \in \mathbb{S}^1$ to get the final result. Specifically,

$$\begin{aligned} |\mathit{length}(\gamma) - \mathit{length}(\gamma_0)| &\leq \frac{1}{2}\int_{u\in\mathbb{S}^1} |\mathrm{Pers}_0(f_u) - \mathrm{Pers}_0(f_{0,u})|\,\mathrm{d}u \\ &\leq \frac{\varepsilon}{2}\int_{u\in\mathbb{S}^1} [\#\mathit{crit}(f) + \#\mathit{crit}(f_0) - 2]\,\mathrm{d}u \\ &= \varepsilon[\mathit{curv}(\gamma) + \mathit{curv}(\gamma_0) - 2\pi], \end{aligned}$$

using first the Cauchy-Crofton Formula, second the re-interpretations in terms of persistence, third the inequality implied by the Stability Theorem for Functions, and fourth the Total Curvature Formula. This completes the proof of the Generalized Fáry Theorem.

Bibliographic notes. The inequality that connects the length with the total curvature of a closed curve is due to Fáry [2]. The generalization that compares the lengths of curves that are close in the Fréchet distance sense is more recent [1]. Both results have generalizations to curves in dimensions beyond two. The integral geometry interpretations of length and total curvature can be found in Santaló [3].

[1] D. Cohen-Steiner and H. Edelsbrunner. Inequalities for the curvature of curves and surfaces. *Found. Comput. Math.* **7** (2007), 391–404.

[2] I. Fáry. Sur certaines inégalités géométriques. *Acta Sci. Math. Szeged* **12** (1950), 117–124.

[3] L. Santaló. *Integral Geometry and Geometric Probability.* Addison-Wesley, 1976, reprinted by Cambridge Univ. Press, England, 2004.

VIII.4 Bipartite Graph Matching

In this section, we consider algorithms for the bottleneck and Wasserstein distances between persistence diagrams. Both problems reduce to constructing optimal matchings in bipartite graphs.

Distance from matching. We begin by reducing the computation of distance to constructing a matching. Let X and Y be two persistence diagrams. We assume both consist of finitely many points above the diagonal and infinitely many points on the diagonal. Letting X_0 be the finite multiset of off-diagonal points in X and X_0' the orthogonal projection of X_0 onto the diagonal, we construct a complete bipartite graph $G = (U \dot\cup V, E)$ with $U = X_0 \dot\cup Y_0'$, $V = Y_0 \dot\cup X_0'$, and $E = U \times V$. For each $q > 0$, we introduce the cost function $c = c^q : E \to \mathbb{R}$ defined by mapping the edge $uv \in E$ to the q-th power of the L_∞-distance between the points,

$$c(uv) \quad = \quad \begin{cases} \|u - v\|_\infty^q & \text{if } u \in X_0 \text{ or } v \in Y_0; \\ 0 & \text{if } u \in X_0' \text{ and } v \in Y_0'. \end{cases}$$

By construction, the minimum cost edge connecting an off-diagonal point u to a point on the diagonal is the edge uu', where u' is the orthogonal projection of u. For $q = 1$, the cost of this edge is half the persistence of u.

A *matching* of G is a subset of vertex disjoint edges, $M \subseteq E$. It is *maximum* if there is no matching with more edges and *perfect* if every vertex is endpoint of an edge in M. Since G is complete with equally many vertices on the two sides, every maximum matching is also a perfect matching. We will also consider matchings for graphs $G(\varepsilon) = (U \dot\cup V, E_\varepsilon)$ obtained from G by removing all edges $uv \in E$ with cost $c(uv) > \varepsilon$. Of course, every perfect matching of $G(\varepsilon)$ is a maximum matching but not necessarily the other way round. A *minimum cost matching* is a maximum matching that minimizes the sum of costs of the edges in the matching. We refer to this sum as the *total cost* of the matching. It is not difficult to prove the following relation between distance and matching.

Reduction Lemma. Let X and Y be two persistence diagrams and $G = (U \dot\cup V, E)$ the corresponding complete bipartite graph. Then

(i) the bottleneck distance between X and Y is the smallest $\varepsilon \geq 0$ such that the subgraph $G(\varepsilon)$ of G with cost function $c = c^1$ has a perfect matching;

(ii) the q-th Wasserstein distance between X and Y is the q-th root of the total cost of the minimum cost matching of G with cost function $c = c^q$.

We are therefore interested in recognizing bipartite graphs that have perfect matchings and in constructing minimum cost matchings.

Augmenting paths. We begin by considering the algorithmic problem of constructing a maximum matching of the bipartite graph $G(\varepsilon) = (U \dot\cup V, E_\varepsilon)$. The algorithm is iterative, improving the matching in each round, until no further improvement is possible. Let M_i be the matching after i iterations. The crucial concept is a path that alternates between edges in and out of M_i. To explain this, we introduce a directed graph D_i that depends on $G(\varepsilon)$ and M_i. For the most part, it is the same as $G(\varepsilon)$ except that each edge is drawn with a direction, namely from V to U, if the edge belongs to M_i, and from U to V, if the edge does not belong to M_i. In addition to the vertices in $G(\varepsilon)$, the directed graph contains two new vertices, the source s with an edge from s to every unmatched vertex $u \in U$, and the target t with an edge from every unmatched vertex $v \in V$ to t; see Figure VIII.12. An *augmenting path* is a

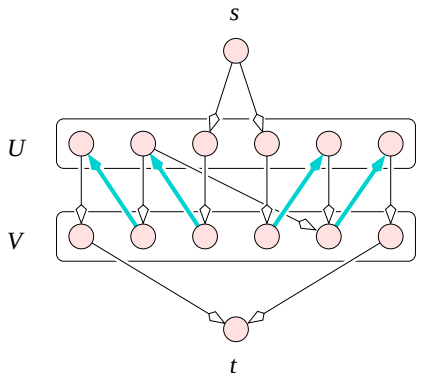

Figure VIII.12: A bipartite graph with six plus six vertices and a matching with four edges giving rise to a directed graph with three paths from s to t.

directed path from s to t that visits every vertex at most once. By construction, an augmenting path consists of $2k + 1$ edges, one from s to U, an interleaved sequence of k edges not in M_i and $k - 1$ edges in M_i, and finally an edge from V to t. Clearly, if we have an augmenting path, we can improve the matching by substituting the k edges not in M_i for the $k - 1$ edges in M_i. When we make this improvement, we say we *augment* the matching using the path. To get an algorithm, we also need the existence of an augmenting path unless M_i is maximum. To construct such a path, draw the edges of an assumed maximum matching from U to V and those of M_i from V to U. Each vertex is

incident to at most two edges, one incoming and the other outgoing, so we can partition the edges into maximal, vertex disjoint paths and closed curves that interleave edges from the two matchings. A path in this partition extends to an augmenting path from s to t iff it contains one more edge from the maximum matching than from M_i. Since M_i is smaller, there is at least one such path. We use this fact to give an algorithm for constructing a maximum matching of $G(\varepsilon)$.

```
M_0 = ∅;  i = 0;
while there exists an augmenting path in D_i do
  augment M_i using this path to get M_{i+1};
  i = i + 1
endwhile.
```

Each iteration increases the size of the matching by one. The number of edges in the maximum matching is at most $n = \operatorname{card} U = \operatorname{card} V$, which implies that the algorithm terminates after at most n iterations. We can use Depth-first Search or Breadth-first Search to find an alternating path in time proportional to the number of edges, $m_\varepsilon = \operatorname{card} E_\varepsilon$. In either case, we have an algorithm that runs in time at most proportional to $m_\varepsilon n \leq n^3$.

Shortest augmenting paths. The running time of the algorithm can be improved if we use multiple augmenting paths at a time. Specifically, we use a maximal set of edge disjoint, shortest, augmenting paths. To find them, we use Breadth-first Search to label all vertices by their distance from the source, and Depth-first Search to construct a maximal set of paths in the thus labeled directed graph. Since Depth-first Search has been explained in detail in Section II.2, we focus on the first step.

```
S_0 = {s};  label s with 0;  j = 0;
while S_j ≠ ∅ do
  forall vertices x ∈ S_j do
    forall unlabeled successors y of x do
      label y with j + 1 and add y to S_{j+1}
    endfor
  endfor;  j = j + 1
endwhile.
```

Assuming suitable data structures, we can iterate through the vertices in the sets S_j and their successors in constant time per vertex. Using repeated Depth-first Search in the labeled graph D_i, we construct a maximal set of edge disjoint

paths from s to t. If we remove edges and vertices as they become useless, we get an algorithm that computes the paths in time proportional to m_ε. For example, if we start with the directed graph in Figure VIII.12, we get either two paths of length seven, one on the left and the other on the right, or just one path of the same length, as shown in Figure VIII.13. Finally, we augment the matching using all paths in the maximal set.

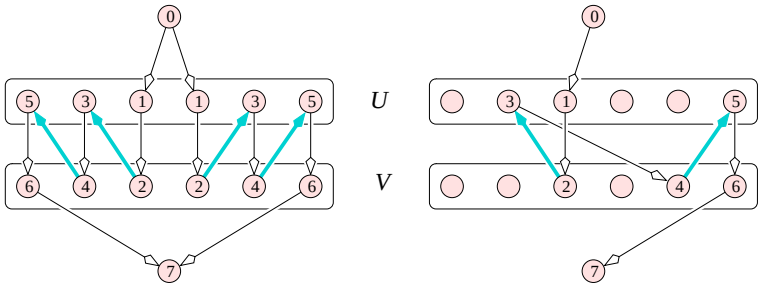

Figure VIII.13: The two maximal sets of edge disjoint, shortest, augmenting paths in the directed graph of Figure VIII.12.

Analysis. We now show that the new strategy leads to a substantially smaller number of iterations. In a nut-shell, the reason is that there cannot be many augmenting paths that are all long. Playing off length against number, we get a bound of some constant times the square root of the number of vertices.

ITERATION BOUND. Starting with the empty matching and augmenting the matching of $G(\varepsilon)$ using a maximal set of edge disjoint, shortest, augmenting paths each time, we reach a maximum matching in fewer than $2\sqrt{2n}$ iterations.

PROOF. We first show that the length of the shortest path from s to t increases from one iteration to the next. Let $\ell_i(x)$ be the length of the shortest path from s to the vertex x in D_i; it is the label assigned to x by Breadth-first Search. We prove that $\ell_{i+1}(t)$ is strictly larger than $\ell_i(t)$, assuming both are defined. Consider a shortest path π from s to t in D_{i+1}. It is also a path in D_i iff none of its edges belongs to the paths selected in the i-th round. If π is a path in D_i then it cannot be shortest else it would have been added to the maximal set. On the other hand, if π is not in D_i then it has at least one edge xy that is reversed in D_i. Since yx belongs to a shortest path in D_i, we have $\ell_i(y) = \ell_i(x) - 1$. For an edge xy of π that is not reversed in D_i, we have

$\ell_i(y) \leq \ell_i(x) + 1$ by definition of ℓ_i. As we walk along the path, ℓ_{i+1} grows by one at each step while ℓ_i grows by at most one and at least once it shrinks. Hence $\ell_i(t) < \ell_{i+1}(t)$, as required.

For the second part of the proof, we note that two edge disjoint paths from s to t share no vertices other than the source and the target. This is because each vertex of U has only one incoming edge and each vertex of V has only one outgoing edge. Let $\bar{m}_i$ be the size deficit of M_i, that is, the number of edges it is short of being a maximum matching. Since M_i can be improved by this much, there are at least $\bar{m}_i$ augmenting paths from s to t in D_i. Using the construction of augmenting paths given earlier in this section, we find $\bar{m}_i$ augmenting paths that share no vertices other than s and t. By the pigeonhole principle, the shortest of these paths contains at most a fraction of $1/\bar{m}_i$ of the vertices of $G(\varepsilon)$. Equivalently, $\ell_i(t) \leq 2n/\bar{m}_i + 1$. Since the distance of t from s begins at three and grows with increasing i, this implies $i \leq 2n/\bar{m}_i - 2$. To increase M_i by another $\bar{m}_i$ edges takes at most $\bar{m}_i$ additional iterations. The total number of iterations is therefore bounded from above by $2n/\bar{m}_i - 2 + \bar{m}_i$. Setting $\bar{m}_i$ to the smallest integer no smaller than $\sqrt{2n}$ implies the claimed bound. □

Recall that each iteration takes time at most proportional to the number of edges. The bound on the number of iterations thus implies that the algorithm runs in time at most proportional to $m_\varepsilon\sqrt{n} \leq n^{5/2}$.

Minimum cost matching. To compute the smallest ε for which $G(\varepsilon)$ has a perfect matching, we do binary search in the list of edges sorted by cost, constructing a maximum matching at every step. Similarly, constructing a minimum cost matching of G is done by iterating the maximum matching algorithm, but the iteration is different. There are two easy structural insights that show the way.

1. If the subgraph $G(0)$ consisting of the cost zero edges in G has a perfect matching then this is a minimum cost matching. Indeed, its total cost is zero which is as small as it gets.
2. Subtracting the same amount from the cost of all edges incident to a vertex in G affects all perfect matchings the same way. In particular, a perfect matching minimizes the total cost before the subtractions iff it does so after the subtractions.

To compute a minimum cost matching of G, we begin with all zero cost edges and construct a maximum matching of $G(0)$. If the matching is perfect, we are done. Otherwise, we change the costs of the edges in G while preserving the

ordering of the perfect matchings by total cost. To describe how this is done, we introduce *deduction maps* $d_i : U \cup V \to \mathbb{R}$. Starting with the zero map, $d_0(x) = 0$ for all vertices x, the algorithm will change the map and this way modify the costs. Writing $c(xy)$ for the original cost of the edge xy in G, the *modified cost* after i iterations is

$$c_i(xy) \quad = \quad c(xy) - d_i(x) - d_i(y).$$

It is important for the efficiency but also the correctness of the algorithm that all modified costs are always non-negative. This will be an invariant of the algorithm. Letting G_i be the graph G with costs modified using d_i, the algorithm iterates the construction of a maximum matching of $G_i(0)$, the graph G_i with edges of positive modified cost removed. Increasing the maximum matching by one edge each time, we get a perfect matching after n iterations. By construction, all edges in this matching have zero modified cost.

Minimum cost paths. We now show how to change the deduction map so that the maximum matching increases. Let M_i be a maximum matching of $G_i(0)$ and let $D_i(0)$ be the directed graph defined by $G_i(0)$ and M_i. Because M_i is maximum, $D_i(0)$ has no directed path from s to t. Let D_i be the directed graph defined by G_i and the same matching M_i and note that it contains $D_i(0)$ as a subgraph. Assuming M_i is not perfect, it is not maximum for G_i which implies that D_i has directed paths from s to t. Each such path is an augmenting path and we define its *total cost* as the sum of modified costs of its edges. By definition, the modified cost of the first edge, from s to U, is zero, and so is the modified cost of the last edge, from V to t. Let π be the augmenting path in D_i that minimizes the total cost. It can be computed by an algorithm similar to Breadth-first Search. Indeed, the only difference is it visits the vertices in a particular ordering that depends on the modified costs of the edges. At every moment during the construction, we have a set of visited vertices forming a tree rooted at s, and a set of unvisited vertices. For each unvisited vertex, y, we consider the minimum cost path that starts at s, goes to a vertex x using edges in the tree, and ends with the edge from x to y. The next vertex visited by the algorithm is the unvisited vertex y that minimizes this cost and we add y together with the last edge of its path to the tree. This is known as Dijkstra's Single Source Shortest Path Algorithm, or Dijkstra's Algorithm for short. We compute the minimizing vertex y and update the costs of all yet unvisited vertices in time proportional to n. Iterating this step n times, we find the minimum cost path π in time proportional to n^2.

We augment the matching M_i using π to get M_{i+1}. This increases the matching, but to be sure that we made progress toward computing a minimum

cost matching, we have to show that it is possible to change the deduction map so that all edges in M_{i+1} have zero modified costs. To this end, let $\gamma_i(x)$ be the minimum total cost of a path from s to x; it is the total cost of the path from s to x within the tree computed by Dijkstra's Algorithm. Using these quantities, we update the deduction map to

$$d_{i+1}(x) = \begin{cases} d_i(x) - \gamma_i(x) & \text{if } x \in U; \\ d_i(x) + \gamma_i(x) & \text{if } x \in V. \end{cases}$$

For vertices $u \in U$ and $v \in V$, the new modified cost of the edge connecting u with v is

$$\begin{aligned} c_{i+1}(uv) &= c(uv) - d_{i+1}(u) - d_{i+1}(v) \\ &= c(uv) - d_i(u) - d_i(v) + \gamma_i(u) - \gamma_i(v). \end{aligned}$$

In words, it is the old modified cost plus $\gamma_i(u) - \gamma_i(v)$, no matter whether in D_i the edge goes from u to v or from v to u. If $\gamma_i(u) \geq \gamma_i(v)$ we use induction to get $c_{i+1}(uv) \geq 0$ from $c_i(uv) \geq 0$. Otherwise, $\gamma_i(v) - \gamma_i(u) \leq c_i(uv)$, else we get a contradiction to $\gamma_i(v)$ being the minimum total cost of a path from s to v. If follows that all new modified costs are non-negative. But we need more, namely zero new modified cost for all edges of the new matching. There are two kinds of such edges uv, those that belong to M_i and those that belong to the path π. For the first kind, we have $\gamma_i(v) = \gamma_i(u)$ because $c_i(uv) = 0$ and the only way to reach u is along the directed edge from v to u. For the second kind, we have $\gamma_i(v) - \gamma_i(u) = c_i(uv)$ by definition of γ_i. In both cases, we have $c_{i+1}(uv) = 0$, as required.

This completes the proof that the iteration ends with a perfect matching minimizing the total cost. The maximum matching gains one edge per iteration. We thus have n iterations each taking time proportional to n^2. Our algorithm thus constructs a minimum cost matching in time at most proportional to n^3.

Bibliographic notes. Computing a maximum matching of a bipartite graph is a classic optimization problem discussed in operations research texts [2]. As explained in [9], it is a special case of the more general maximum flow problem in networks. Indeed, Dinic's maximum flow algorithm for so-called unit networks [4] specializes to the $n^{5/2}$ time algorithm for maximum matching independently discovered by Hopcroft and Karp [6] and explained in this section. The Minimum Cost Matching Algorithm, is a variant of what is known as the Hungarian method [8]. Following [7], we describe a version that uses Dijkstra's Algorithm for finding shortest paths in a weighted graph as a subroutine [3]. Using the geometry of the persistence diagrams, the Maximum Matching Algorithm can be improved to run in time at most proportional to $n^{3/2} \log_2 n$ [5]

and the Minimum Cost Matching Algorithm can be improved to run in time at most proportional to $n^{2+\varepsilon}$ [1].

[1] P. K. AGARWAL, A. EFRAT AND M. SHARIR. Vertical decomposition of shallow levels in 3-dimensional arrangements and its applications. *SIAM J. Comput.* **29** (2000), 912–953.

[2] R. AHUJA, T. MAGNANTI AND J. ORLIN. *Network Flows.* Prentice Hall, 1993.

[3] E. W. DIJKSTRA. A note on two problems in connexion with graphs. *Numerische Mathematik* **1** (1959), 269–271.

[4] E. A. DINIC. Algorithm for solution of a problem of maximum flow in a network with power estimation. *Soviet Math. Doklady* **11** (1970), 1277–1280.

[5] A. EFRAT, A. ITAI AND M. J. KATZ. Geometry helps in bottleneck matching and related problems. *Algorithmica* **31** (2001), 1–28.

[6] J. E. HOPCROFT AND R. M. KARP. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM J. Comput.* **2** (1973), 225–231.

[7] J. KLEINBERG AND E. TARDOS. *Algorithm Design.* Pearson Education, Boston, Massachusetts, 2006.

[8] H. W. KUHN. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly* **2** (1955), 83–97.

[9] R. E. TARJAN. *Data Structures and Network Algorithms.* SIAM, Philadelphia, Pennsylvania, 1983.

Exercises

The credit assignment reflects a subjective assessment of difficulty. A typical question can be answered using knowledge of the material combined with some thought and analysis.

1. **Examples of switches** (two credits). Given examples for the types of switches analogous to the ones shown in Figure VIII.4 but one dimension up in each of the three types.
2. **Matrix maintenance** (two credits). Formulate the algorithm that maintains the reduced boundary matrix under transpositions for $\partial = RV$, that is, maintain the matrix V instead of its inverse, U.
3. **Sparse matrix representation** (two credits). Give a sparse matrix representation that allows an implementation of the maintenance algorithm running in time proportional to the number of ones in the changed columns and rows of $R = \partial U$.
4. **Measuring vineyards** (two credits). Let $f, g : K \to \mathbb{R}$ be two monotonic functions on a simplicial complex and $f_t = (1-t)f + tg$ for $t \in [0,1]$ the straight-line homotopy between them. Each vine of the homotopy is a map $x : [a,b] \to \bar{\mathbb{R}}^2$ with $0 \leq a < b \leq 1$. Let

$$\mu(x) \quad = \quad \int_{s=a}^{b} \|x(s) - x(a)\| \, \mathrm{d}s$$

   and define a measure by summing the integrals over all vines in the p-th vineyard, $\mu_p(f,g) = \sum_x \mu(x)$. Give examples that show that μ_p and the first Wasserstein distance are incomparable, that is, there are monotonic functions f, g, f_0, g_0 such that $\mu_p(f,g) < W_1(\mathrm{Dgm}_p(f), \mathrm{Dgm}_p(g))$ and $\mu_p(f_0,g_0) > W_1(\mathrm{Dgm}_p(f_0), \mathrm{Dgm}_p(g_0))$.
5. **Cauchy-Crofton** (two credits). Generalize the Cauchy-Crofton formula for curves in the plane given in Section V.3 to

   (i) curves in three-dimensional Euclidean space;
   (ii) surfaces in three-dimensional Euclidean space.

6. **Mean and Gaussian curvatures** (three credits). Use the structure of the proof of the Generalized Fáry Theorem to show the following relationship between the total mean curvature and the total absolute Gaussian curvature of two homeomorphic closed surfaces embedded in $\mathbb{R}^3$,

$$|\mathit{mean}(S) - \mathit{mean}(S_0)| \quad \leq \quad [\mathit{gauss}(S) + \mathit{gauss}(S_0) - 4\pi(1+g)]F(\bar{S}, \bar{S}_0),$$

where g is the common genus of S and of S_0, $\bar{S}$ and $\bar{S}_0$ are the solid bodies bounded by the two surfaces, and $F(\bar{S}, \bar{S}_0)$ is the Fréchet distance between them.

7. **Breadth-first search** (one credit). Reformulate the breadth-first search algorithm for labeling the vertices of D_i using a single queue to represent all sets of vertices S_j in one data structure. As suggested by the name, this is a data structure that supports adding an element at the end and removing it from the front, both in constant time.

8. **Incremental matching** (three credits). Recall that the maximum matching of a bipartite graph with n vertices can be constructed in time at most proportional to $n^{5/2}$. Running this algorithm within a binary search routine, we find the perfect matching of a complete bipartite graph that minimizes the largest cost of any of its edges in time at most proportional to $n^{5/2} \log_2 n$. Show that the two algorithms can be integrated to avoid the $\log_2 n$ overhead, constructing the perfect matching in time at most proportional to $n^{5/2}$. [[Can this really be done?]]

Chapter IX

Applications

The primary application of the mathematical and computational tools introduced in the previous chapters is in data analysis, and activity that reaches into every discipline in science and engineering. The data may comprise the readings of an array of sensors, the pixes of an image, the accumulation of observations, or what have you. Invariably, there is noise in the data, which may be systematic, or random. It may also reflect genuine properties of the measured phenomenon but at a scale level that is outside the window of interest. The traditional approach to noise is to 'smooth' or 'regularize' the data, which invariably means we change the data. This is in sharp contrast to the approach we advocate here, namely measuring the noise and not change the data. What is new is the measurement and the additional level of rationality it affords us. The four case studies selected to illustrate the possibilities all start with biological data.

IX.1 Simplification for Gene Expression Data

Background. [[Introduce somitogenesis, the fact that this is a rhythmic process, and mention cyclic gene expression driving the process.]]

Technology. [[Explain that micro-arrays are used to look at all (known) genes of the organism, in this case a mouse, at several stages during the process. Can we talk about noise in micro-array experiments?]]

Mathematics. [[Introduce the concept of simplification of a function. For a function on the circle it is easy to see that such simplifications exist.]]

[[Explain the series of integrals, relate them to the moments of total persistence, and mention that we have stability for $i \geq 2$ but not for $i = 0, 1$.]]

Wrap-up. [[We apply our methods to ranked data, both time (horizontal) and expression value (vertical).]]

[[Present the ranking of the genes by μ_2, and compare the placement of the verified genes with the ranking for other measures.]]

Bibliographic notes.

[1] M.-L. DEQUÈANT, S. AHNERT, H. EDELSBRUNNER, T. M. A. FINK, E. F. GLYNN, G. HATTEM, A. KUDLICKI, Y. MILEYKO, J. MORTON, A. R. MUSHEGIAN, L. PACHTER, M. ROWICKA, A. SHIU, B. STURMFELS AND O. POURQUIÉ. Comparison of pattern detection methods in microarray time series of the segmentation clock. *PLoS ONE* **3** (2008).

IX.2 Elevation for Protein Docking

Background. [[Talk about the importance of protein interaction, or more generally the interaction of biomolecules. Protein docking is the computational approach to predicting interactions.]]

Technology. [[This work starts with molecular structures describes a pdb-files (protein data bank) and obtained mostly but not exclusively through x-ray cristallography.]]

Mathematics. [[Introduce the idea of matching protrusions with cavities, first described in [2].]]

[[Explain elevation first for a smooth curve embedded in $\mathbb{R}^2$ and second for a piecewise linear curve. This case is fairly elementary and we can already address the types of maxima.]]

Suppose that C is a smooth curve embedded in $\mathbb{R}^2$. We define the height function

$$H : S^1 \times C \to \mathbb{R}$$

by $H(x, u) =< x, u >$. For each u this map associates to each point $x \in C$ the height of x in the u direction. For generic C and general direction u, $H_u(x) = H(x, u)$ is a Morse function. There are, however, a finite number of directions in which this fails. The failure is one of two types:

- H_u has a single degenerate critical point which is a "birth-death" point.
- H_u has two critical points that share the same critical value.

A birth death point is modeled by the family of functions $f_t(x) = x^3 - tx$. For $t = 0$, we have a degenerate critical point, for $t > 0$ a pair of non-degenerate critical points, one a local max and one a local min, and for for $t < 0$, no critical point. In our case, this variation of t corresponds to varying u so that the critical points move along C.

Now, when u is a general direction, we can use the Morse function H_u to define an extended persistence pairing on C. The points for which u is normal to C are paired, and we associate to each the persistence of the pair to which it belongs. Since every point $p \in C$ has a normal direction, this defines a function E on C, except at the special directions of the two types above.

For the first type, we can set E to 0 at p, since the pair of points that die are paired by persistence, and the result is continuous. For the second type,

however, we have an ambiguity of how to define E at p, and a corresponding discontinuity in E. This is illustrated in Figure .

[[Explain elevation for a smooth surface embedded in $\mathbb{R}^3$ and for a piecewise linear surface. Explain the types of maxima, one-, two-, three-, and four-legged; see [1].]]

[[The algorithm for extended persistence for 2-manifolds uses splitting and cutting trees [3].]]

Wrap-up. [[Discuss the experimental results presented in [4].]]

Bibliographic notes.

[1] P. K. Agarwal, H. Edelsbrunner, J. Harer and Y. Wang. Extreme elevation on a 2-manifold. *Discrete Comput. Geom.* **36** (2006), 553–572.

[2] M. L. Connolly. Shape complementarity at the hemo-globin albl subunit interface. *Biopolymers* **25** (1986), 1229–1247.

[3] L. Georgiadis, R. E. Tarjan and R. F. Werneck. Design of data structures for mergeable trees. *In* "Proc. 17th Ann. ACM-SIAM Sympos. Discrete Alg., 2006", 394–403.

[4] Y. Wang, P. K. Agarwal, P. Brown, H. Edelsbrunner and J. Rudolph. Coarse and reliable geometric alignment for protein docking. *In* "Proc. Pacific Sympos. Biocomput., 2005", 65–75.

IX.3 Image Segmentation

Background. [[We explore organisms and medical conditions.]]

Technology. [[We start with images, which are often 3-dimensional (MRI etc) but sometimes 2-dimensional (confocal microscopy, etc).]]

Mathematics. The first application of computational topology methods we will give is to the segmentation of images. The segmentation problem is to identify regions of interest in an image. We usually try to draw curves around these regions, and when possible we do this "automatically", i.e. without help from the user of the software. This is an imperfect art at best, and every type of image provides a different set of challenges.

The *watershed* method fits nicely into the framework of computational topology. It is widely used [], but always has a problem in that it tends to overdo the segmentation, surrounding more features that are usually wanted. For this reason there is always a clean-up step, sometimes done systematically and sometimes in a an ad hoc or even manual way. Here we will use persistence for this step.

Watersheds Talk about the idea of filling up with water. Watershed lines are built to keep the water from overflowing.

An image I is a matrix of $m \times n$ values, the intensity of the pixels. Values can be bytes (integers between 0 and 255), ints, longs, etc, call the range space of values V. Color images consist of three separate images R, G and B, together with a blending that renders the color image. Let $D = [1, m] \times [1, n]$, we will think of I as samples of a continuous function $f : D \to V$. Choose some triangulation of D so that its vertices are the interger lattice points $\{0, \ldots, m\} \times \{0, \ldots, n\}$, which we think of as the centers of the image pixels, edges are straight segments and triangles are the region they cut out. We should always assume that the border $\{0, m\} \times \{0, \ldots, n\} \cup \{0, \ldots, m\} \times \{0, n\}$ is a union of edges and we usually will take edges to be vertical, horizontal and diagonal edges between neighboring vertices.

Morse Complex We now take $\mathbb{M}$ to be any 2-dimensional manifold and f to be a function defined on $\mathbb{M}$. Recall that in section $VI.2$ we constructed a complex whose vertices were the minima of f, edges were the stable 1-manifolds of f and faces were the stable 2-manifolds, the resulting complex is called the

Morse Complex of f. Since $\mathbb{M}$ is dimension 2, the edges of the complex divide $\mathbb{M}$ into regions, these are the elements of the segmentation we are looing for.

Since our image is really given by a piecewise linear function, we need to discuss how we construct the Morse complex, or at least an approximation to it, in this case.

Clean up with Persistence

Fine Tuning the Segmentation

Segmentation in Three Dimensions [[We focus on the watershed algorithm for segmentation. It may be called the Morse complex consisting of all unstable manifolds.]]

[[A common drawback is 'over-segmentation'. This is caused by noise in the image and can be removed using persistence. We explain this is 2D and comment on extensions to 3D.]]

[[Survey on watershed algorithms [3]. Oldest paper on the topic [1]; Extension from 2D to 3D [4], using a diffusion filter to cope with the oversegmentation.]]

Wrap-up. [[Give some 2D cell images and the segmentation we get.]]

Bibliographic notes.

[1] S. Beucher. Watersheds of functions and picture segmentation. *In* "Proc. IEEE Intl. Conf. Acoustic, Speech, Signal Process, 1982", 1928–1931.

[2] H. Edelsbrunner and J. Harer. The persistent Morse complex segmentation of a 3-manifold. Report rgi-tech-04-066, Geomagic, Research Triangle Park, North Carolina, 2004.

[3] J. Roerdink and A. Meijster. The watershed transform: definitions, algorithms, and parallelization strategies. *Fundamenta Informaticae* **41** (2000), 187–228.

[4] J. Sijbers, P. Scheunders, M. Verhoye, A. Van der Linden, D. van Dyck and E. Raman. Watershed-based segmentation of 3D MR data for volume quantization. *Magn. Reson. Imag.* **15** (1997), 679–688.

IX.4 Local Homology for Root Architecture

Background. [[The general need to classify phenotypes to study the connection between genotype and phenotype. We focus on agricultural plants, in particular rice.]]

Technology. [[We grow the rice in laboratory conditions so we can take 2D pictures.]]

Mathematics. [[There is the problem of reconstructing the 3D root from the 2D pictures. We can then use 3D methods to characterize the shape of the root. Alternatively, we can analyze the 2D images.]]

[[We use local homology to analyze the images, aiming at recognizing and counting tips of roots, branches, and crossings.]]

Wrap-up. [[Images of rice roots and colorings of the sought features.]]

[[We could also talk about the classification we got already using simple geometric descriptors.]]

Bibliographic notes.

[1] W. A. Cannon. A tentative classification of root systems. *Ecology* **30** (1947), 452–458.

[2] D. Cohen-Steiner, H. Edelsbrunner, J. Harer and D. Morozov. Persistent homology for kernels and images *In* "Proc. 20th Ann. ACM-SIAM Sympos. Discrete Alg., 2009", to appear.

[3] P. Bendich, D. Cohen-Steiner, H. Edelsbrunner, J. Harer and D. Morozov. Inferring local homology from sampled stratified spaces *In* "Proc. 48th Ann. Sympos. Found. Comput. Sci., 2007", 536–546.

Exercises

The credit assignment reflects a subjective assessment of difficulty. A typical question can be answered using knowledge of the material combined with some thought and analysis.

Chapter X

Open Problems

[[There are additional questions we might use to add new problems or replace some of the old ones:

- Simplification of a PL function on $\mathbb{S}^3$. For general 3-manifolds, the Poincaré Theorem is an obstacle but knowing that we have the 3-sphere makes sense in practice and removes the obstacle.

]]

X.1 Complexity of Reidemeister Moves

Recall the definition of Reidemeister moves from Chapter I. As proved by Reidemeister in the first half of the twentieth century, any generic projection of a knot can be transformed into any other generic projection of the same knot by a sequence of such moves [1]. In particular, for any generic projection of the unknot there is a sequence of Reidemeister moves that eliminates all crossings. This suggests a graph search algorithm to decide whether or not two generic projections describe the same knot. The only difficulty with this approach is that we do not know how long such a sequence of moves may get. We also do not know how many crossings we can expect for intermediate projections. For example, the know in Figure X.1 is the unknot but to get it into a crossing-free projection we need to first increase the number of crossings beyond the seven in the drawing.

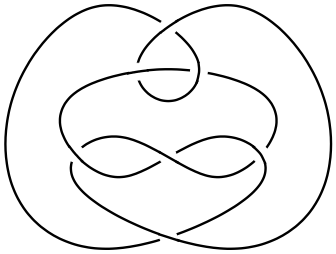

Figure X.1: A generic projection of the unknot.

Given generic projections P and Q of the same knot, let $x(P, Q)$ be the minimum, over all Reidemeister moves transforming P to Q, of the maximum number of crossings of any projection in the sequence. Let now $x(n)$ be the maximum $x(P, Q)$, over all pairs P and Q in which P and Q have at most n crossings each. In other words, we can transform P into Q while staying below $x(n) + 1$ crossings at all times.

Question. Is there a positive constant c such that $x(n) \leq n + c$ for all c? Or less ambitiously, is $x(n)$ bounded from above by a polynomial in n?

It would be rather surprising if the answer to the first question were in the affirmative but perhaps it is to the second question.

[1] K. Reidemeister. Knotentheorie. In *Ergebnisse der Mathematik und ihrer Grenzgebiete*, Springer, Berlin, Germany, 1932.

X.2 Shelling a 3-ball

Let K be a triangulation of a 3-ball, that is, a collection of tetrahedra sharing triangles, edges, and vertices whose union is homeomorphic to $\mathbb{B}^3$. No other, improper intersections between the tetrahedra are permitted. A *shelling* of K is an ordering of the tetrahedra such that each prefix of the ordering defines a triangulation of $\mathbb{B}^3$, and K is *shellable* if it has a shelling. It is not difficult to prove that every triangulation of $\mathbb{B}^2$ has a shelling, but the following example taken from Bing [1] shows that the same is not true for 3-balls.

The house-with-two-rooms is sketched in Figure X.2. There are two rooms, one above the other. The only way to access the lower room is through a chimney and the only way to access the upper room is through an underground tunnel. The chimney and the tunnel are connected to the side of the house by a screen each. Now we thicken each wall, floor, ceiling and screen to one layer of bricks. All vertices belong the the boundary but edges and triangles may be on the boundary or in the interior. For a given cube, we refer to the connected component of faces that belong to the boundary as *exposures*. By construction, each cube has two exposures. The union of the cubes is a 3-ball but removing any one cube destroys this property. Indeed, removing any one cube either creates a hole in a wall (a tunnel through the 3-ball), or it pinches the 3-ball at a point or along an edge.

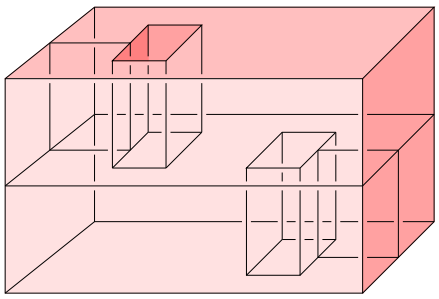

Figure X.2: House-with-two-rooms. We can construct it from a solid block of clay by a continuous deformation, without tearing or gluing.

Since we prefer to work with simplicial as opposed to cubical complexes, we still need to decompose the cubes into tetrahedra, but this is an afterthought that does not distract from the essential idea of the construction. For this purpose, we decompose each cube into six tetrahedra in such a way that removing any one tetrahedron destroys the property of their union begin a 3-ball. In

other words, no tetrahedron can be last in the shelling, which implies that the triangulation has no shelling. In order to avoid improper intersections, we first decompose each square into two triangles and then each cube into six tetrahedra in a compatible fashion. Let the *type* of a vertex be the minimum dimension of any exposure of a cube that contains the vertex. For example, vertices at corners inside the rooms are type 0, vertices along edges inside the rooms are type 1, and the rest are type 2. Order the vertices such that type-0 vertices precede type-1 vertices which precede type-2 vertices. Now decompose each square by connecting its first vertex in the ordering to the opposite two edges. Similarly decompose each cube by connecting its first vertex in the ordering to the opposite six triangles. Again by construction each tetrahedron has two exposures, a vertex and its opposite triangle or an edge and its opposite edge. This completes the construction of the triangulation of the house-with-two-rooms that is not shellable. Since not every triangulation of $\mathbb{B}^3$ has a shelling it makes sense to ask for a decision procedure.

Question. Is there a polynomial-time algorithm that decides whether or not a given triangulation of $\mathbb{B}^3$ has a shelling?

The construction of a shelling for a triangulated 2-ball is straightforward because every partial shelling is extendable and can therefore be completed [2]. This is no longer the case for the 3-ball. In other words, there are shellable triangulations of $\mathbb{B}^3$ that have non-extendable partial shellings [3]. Without a way to recognize such dead-ends we are forced into back-tracking, which takes time.

[1] R. H. Bing. Some aspects of the topology of 3-manifolds related to the Poincaré conjecture. In *Lectures on Modern Mathematics II*, T. L. Saaty (ed.), Wiley, New York, 1964, 93–128.

[2] G. Danaraj and V. Klee. Which spheres are shellable? In *Algorithmic Aspects of Combinatorics*, B. Alspach et al. (eds.), *Ann. Discrete Math.* **2** (1978), 33–52.

[3] G. M. Ziegler. Shelling polyhedral 3-balls and 4-polytopes. *Discrete Comput. Geom.* **19** (1998), 159–174.

X.3 Geometric Realization of 2-manifolds

Recall that a geometric realization of a simplicial complex K is an embedding in which each vertex maps to a point and each (abstract) simplex maps to the (geometric) simplex spanned by the images of its vertices. The existence of a geometric realization in $\mathbb{R}^d$ can be decided using Tarski's theory of real closed fields [7]. The question is therefore decidable but Tarski's quantifier elimination method is far from practical even for small problem instances. As a special case we consider simplicial complexes K that triangulate orientable 2-manifolds and ask for geometric realizations in $\mathbb{R}^3$. Not every such K can be geometrically realized in $\mathbb{R}^3$. For example, there is a twelve-vertex triangulation of the genus-six torus that is not [1]. There is also a twelve-vertex triangulation of the genus-five torus that is not geometrically realizable even after removing one of the triangles. We can therefore take the connected sum and form arbitrarily large triangulations that have no geometric realization in $\mathbb{R}^3$ [6]. Perhaps five is the smallest genus for which this works.

QUESTION. For $1 \leq g \leq 4$, does every triangulation of the genus-g torus have a geometric realization in $\mathbb{R}^3$?

There have been attempts to prove this in the affirmative for $g = 1$ but the answer is still outstanding. The question of geometric realizability for triangulated 2-manifolds has been mentioned by Császár [3] and Grünbaum [4, Chapter 13.2]. A first serious approach to the question is described in [2]. Enumeration results can be found on Frank Lutz' web-pages [5].

[1] J. BOKOWSKI AND A. GUEDES DE OLIVEIRA. On the generation of oriented matroids. *Discrete Comput. Geom.* **24** (2000), 197–208.

[2] J. BOKOWSKI AND B. STURMFELS. *Computational Synthetic Geometry.* Springer-Verlag, Berlin, Germany, 1980.

[3] A. CSÁSZÁR. A polyhedron without diagonals. *Acta Sci. Math. (Szeged)* **13** (1949), 140–142.

[4] B. GRÜNBAUM. *Convex Polytopes.* John Wiley & Sons, London, England, 1967.

[5] F. LUTZ. The manifold page, 1999–2006. `www.math.tu-berlin.de/diskregeom/-stellar`.

[6] L. SCHEWE. Nonrealizability of triangulated surfaces. Oberwolfach Reports **3** (2006), 707–708.

[7] A. TARSKI. *A Decision Method for Elementary Algebra and Geometry.* Second edition, Univ. California Press, 1951.

X.4 Embedding in Three Dimensions

Recall that a simple graph is an abstract simplicial complex of dimension 1 and a straight-line embedding is a geometric realization. We have seen in Section I.4 that every simple graph that has an embedding also has a straight-line embedding in $\mathbb{R}^2$. This property does not extend to higher dimensions. To construct an abstract simplicial complex of dimension 2 that has an embedding but not a geometric realization in $\mathbb{R}^3$, we use a non-trivial knot, such as the trefoil knot in Figure I.7 in the middle. Every generic projection of this knot has at least three crossing. A polygonal cycle forming the knot in $\mathbb{R}^3$ thus necessarily consists of more than three line segments. Now take a sufficiently fine simplicial complex whose underlying space is the unit cube in $\mathbb{R}^3$. Remove from this complex a tunnel in the form of the mentioned knot. Draw a closed curve running along the tunnel and decompose it into three edges, which are necessarily curved. Finally, repair the triangulation by connecting the three edges to the boundary of the tunnel. The 2-skeleton of this triangulation has an embedding but no geometric realization in $\mathbb{R}^3$.

As mentioned in Section X.3, we can decide whether a simplicial complex of dimension 2 has a geometric realization in $\mathbb{R}^3$ using Tarski's theory of real closed fields. It is not clear whether we can also decide embeddability. Besides asking whether a simplicial complex has a geometric realization or an embedding in $\mathbb{R}63$, we can also ask whether it has a subdivision that has a geometric realization. Such a realization is sometimes called a PL embedding of the complex [2].

Question. Are there simplicial complexes that have embeddings but no PL embeddings in $\mathbb{R}^3$? Is the recognition of simplicial complexes that have embeddings or PL embeddings in $\mathbb{R}^3$ decidable?

The algorithmic problem of embeddability provides a striking example of a dramatic complexity increase by adding just one dimension. Indeed, the time it takes to decide whether or not a simple graph with n vertices has an embedding in $\mathbb{R}^2$ is only proportional to n, see e.g. Hopcroft and Tarjan [1].

[1] J. E. Hopcroft and R. E. Tarjan. Efficient planarity testing. *J. ACM* **21** (1974), 549–568.

[2] J. Matoušek, M. Tancer and U. Wagner. Hardness of embedding simplicial complexes in $\mathbb{R}^d$. Manuscript, 2008.

X.5 Equipartition in Four Dimensions

We define a *density* in d-dimensional Euclidean space as a Borel measure $f : \mathbb{R}^d \to \mathbb{R}$ with unit mass, $\int f(x)\,\mathrm{d}x = 1$. Let f_i be a *density* in $\mathbb{R}^d$ for each $1 \le i \le d$. The Ham Sandwich Theorem asserts that there is a $(d-1)$-plane that simultaneously bisects all d densities, that is, each f_i has exactly half of its mass on each side of the plane, see e.g. [3, Chapter 3]. In $\mathbb{R}^2$ this implies that every density can be decomposed by two lines into four quadrants each a quarter of the mass. More generally, we say that d $(d-1)$-planes form an *equipartition* of a density in $\mathbb{R}^d$ if they define 2^d orthants each containing an equal share of the mass. Hadwiger showed that equipartitions also exist in $\mathbb{R}^3$ [2]. The situation is different in five and higher dimensions [1]. We count degrees of freedom to see that a negative result is to be expected. A $(d-1)$-plane in $\mathbb{R}^d$ has d degrees of freedom, and since we can choose d $(d-1)$-planes we have a total of d^2 degrees at our disposal. To use the degrees, we specify the $(d-1)$-planes in sequence and consume a degree for each density we bisect. The total number of consumed degrees is $1+2+\ldots+2^{d-1} = 2^d-1$. For $d \ge 5$, we therefore consume more than we have, which leads to the negative result. For $d = 4$, we have 16 degrees of freedom but we need only 15. It thus seems that there should be an equipartion for each density in $\mathbb{R}^4$, but it is not known whether this is indeed the case.

QUESTION. Does every density in $\mathbb{R}^4$ have an equipartition?

A host of results related to this question but not answering it can be found in Ramos [4]. He generalizes the Borsuk-Ulam Theorem and proves among other things that every density in $\mathbb{R}^4$ has an equipartition by four 3-spheres. As another step towards resolving the question, Živaljević proved the existence of an equipartition provided the density in $\mathbb{R}^4$ has a 2-plane of symmetry [5].

[1] D. AVIS. Non-partitionable point sets. *Inform. Process. Lett.* **19** (1984), 125–129.

[2] H. HADWIGER. Simultane Vierteilung zweier Körper. *Arch. Math. (Basel)* **17** (1966), 274–278.

[3] J. MATOUŠEK. *Using the Borsuk-Ulam Theorem.* Springer-Verlag, Berlin, 2003.

[4] E. A. RAMOS. Equipartition of mass distributions by hyperplanes. *Discrete Comput. Geom.* **15** (1996), 147–167.

[5] R. T. ŽIVALJEVIĆ. Equipartitions of measures in $\mathbb{R}^4$. *Trans. Amer. Math. Soc.* **360** (2008), 153–169.

X.6 Running-time of Matrix Reduction

[[Summarize what is known about the running time of the matrix reduction algorithm, including the work on integer coefficients. State the problem of proving the cubic lower bound.]]

[1] R. Kannan and A. Bachem. Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix. *SIAM J. Comput.* **8** (1979), 499–507.

[2] D. Morozov. Persistence algorithm takes cubic time in worst case. BioGeometry News, Dept. Comput. Sci., Duke Univ., Durham, North Carolina, 2005.

X.7 Multi-parameter Persistence

[[Summarize the results known on multi-parameter persistence, citing Carlsson and Zomorodian as well as Frosini et al.]]

[1] G. Carlsson and A. Zomorodian. The theory of multidmensional persistence. Manuscript, Dept. Math., Stanford Univ., California, 2006.

[2] A. Cerri, P. Frosini and C. Landi. Stability in multidimensional size theory. Manuscript, Dept. Math., Univ. di Bologna, Italy, 2006.

X.8 Unfolding PL Critical Points

[[Explain the method for 3-manifolds.]]

[[State the problem for d-manifolds, $d \geq 4$, as an open question.]]

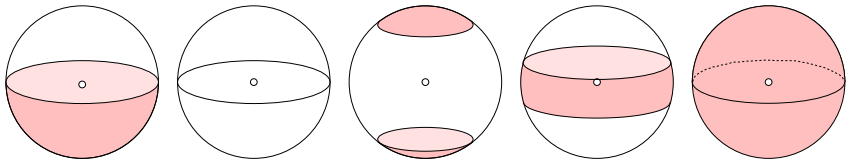

Figure X.3: From left to right: the lower link of a regular point, a minimum, a 1-saddle, a 2-saddle, and a maximum.

[1] H. Edelsbrunner, J. Harer, V. Natarajan and V. Pascucci. Hierarchy of Morse-Smale complexes for piecewise linear 3-manifolds. *In* "Proc. 19th Ann. Sympos. Comput. Geom., 2003", 361–370.

X.9 PL in the Limit

[[Explain the method for a piecewise linear function on a 2-manifold, or perhaps for a PL 2-manifold itself.]]

[[Formulate the question for general PL d-manifolds.]]

[1] H. Edelsbrunner, J. Harer and A. Zomorodian. Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete Comput. Geom.* **30** (2003), 87–107.

X.10 Counting Halving Sets

[[Review the best results: upper and lower bound in $\mathbb{R}^2$, same in $\mathbb{R}^3$, upper bound in higher dimensions.]]

[1] T. K. Dey. Improved bounds on planar k-sets and related problems. *Discrete Comput. Geom.* **19** (1998), 373–383.

[2] P. Erdős, L. Lovász, A. Simmons and E. G. Straus. Dissection graphs of planar point sets. In *A Survey of Combinatorial Theory*, eds. J. N. Srivastava et al., North-Holland, Amsterdam (1973), 139–149.

[3] L. Lovász. On the number of halving lines. *Ann. Univ. Sci. Budapest Eőtvős Sect. Math.* **14** (1971), 107–108.

[4] M. Sharir, S. Smorodinsky and G. Tardos. An improved bound for k-sets in three dimensions. *In* "16th Ann. Sympos. Comput. Geom., 2000".

[5] G. Tóth. Point sets with many k-sets. *Discrete Comput. Geom.* **26** (2001), 187–194.

[6] R. T. Živaljević and S. T. Vrećica. The colored Tverberg's problem and complexes of injective functions. *J. Combin. Theory, Ser. A* **61** (1992), 309–318.

Index

Author Index

Appendix A

$G = (V, E)$	graph, vertex set, edges set
K_5, $K_{3,3}$	complete graph, complete bipartite graph
n, m, ℓ	number of vertices, edges, faces
u, v, y, z, u_i	vertices
$f : V \to \mathbb{R}^2$	vertex mapping
$h : \mathbb{R}^2 \to \mathbb{R}$	linear map
t, t_i, t_{uv}	coefficients
$\Delta(x, a, b)$	left-turn matrix
$\varepsilon_1 \ll \varepsilon_2$	small, positive indeterminants
$[n] = \{1, \ldots, n\}$	first n positive integers
$V[1..n]$	linear array
$V[i].parent, V[i].size$	parent pointer, cardinality
$x \in \mathbb{X}$, $\mathbb{Y}$	point, topological spaces
$\mathcal{U}$, $\mathcal{B}$	topology, basis
$\gamma : [0, 1] \to \mathbb{X}$	path
$\gamma : \mathbb{S}^1 \to \mathbb{R}^2$	closed curve
$\kappa, \lambda : \mathbb{S}^1 \to \mathbb{R}^3$	knots
$T(s), N(s), B(s)$	unit tangent, nomal, binormal
$u \in \mathbb{S}^2$	direction
$W(\gamma, x)$	winding number
$Wr(\kappa), DWr(\kappa, u)$	writhing, directional writhing number
$Lk(\kappa, \lambda), Tw(\kappa, \lambda)$	linking, twisting number

Table X.1: Notation in Chapter I.

$\mathbb{M}$, $\mathbb{N}$, K, L	2-manifolds, triangulations
$\mathbb{S}^2, \mathbb{T}^2, \mathbb{P}^2$	sphere, torus, projective plane
D	open disk
n, m, ℓ	number of vertices, edges, triangles
χ, g	Euler characteristic, genus
$V[1..n], \mu, abc$	vertex array, node, triangle
$(\mu, \iota), -.fnext, -.org$	ordered triangle, next triangle, origin
b_x, b_y, b_z	boolean variables
$f : \mathbb{M} \to \mathbb{R}^3$	mapping, immersion
$J, \frac{f_i}{s_j}$	Jacobian, partial derivative
$\operatorname{sign} \det \Delta(a, x, y, z)$	orientation of four points in space
a, b, x, y	vertices
$\varphi : \operatorname{Vert} K \to \operatorname{Vert} L$	contraction
$\operatorname{Lk} a, \operatorname{Lk} ab$	vertex, edge link
$h, u, \delta, x, y, d(h, x)$	plane, normal, offset, points, signed distance
$H, E_H(x)$	set of planes, sum of square distances
$\mathbf{x}, \mathbf{u}, \mathbf{Q}, \mathbf{Q}_i$	4D point, 4D normal, fundamental quadric, column

Table X.2: Notation in Chapter II.

$\sigma = \operatorname{conv}\{u_0, \dots, u_k\}$	k-dimensional simplex
$x = \sum_i \lambda_i u_i$	linear, affine, convex combination
$\tau \leq \sigma, \operatorname{bd}\sigma, \operatorname{int}\sigma$	face, boundary, interior
$K, L, \lvert K\rvert$	simplicial complexes, underlying space
$\operatorname{Vert} K, K^{(j)}, \operatorname{Sd} K$	vertex set, j-skeleton, barycentric subdivision
$\operatorname{St}\sigma, \overline{\operatorname{St}}\,\sigma, \operatorname{Lk}\sigma$	star, closed star, link
α, β, A, B	abstract simplices, abstract simplicial complexes
$\varphi : \operatorname{Vert} K \to \operatorname{Vert} L$	vertex map
$b_i(x), f : \lvert K\rvert \to \lvert L\rvert$	barycentric coordinates, simplicial map
$F, \operatorname{Nrv} F, \bigcup F$	set system, nerve, union
$f, g, r : \mathbb{X} \to \mathbb{Y}$	continuous maps, (deformation) retraction
$H : \mathbb{X} \times [0,1] \to \mathbb{Y}$	homotopy
$f \simeq g, \mathbb{X} \simeq \mathbb{Y}$	homotopic, homotopy equivalent
$\mathrm{id}_{\mathbb{X}} : \mathbb{X} \to \mathbb{X}$	identity map
$\check{\mathrm{C}}\mathrm{ech}(r)$, Vietoris-Rips$(r)$	Čech, Vietoris-Rips complex
$t_j(n)$	expected number of tests
$\sigma \subseteq S, \operatorname{diam}\sigma$	simplex, point set, diameter
$\iota : \mathbb{R}^{d+1} - \{0\} \to \mathbb{R}^{d+1} - \{0\}$	inversion, origin
$\varsigma : \mathbb{S}^d - \{N\} \to \mathbb{R}^d$	stereographic projection, north-pole
$V_u, V_v, \Sigma, \Sigma_u, \Pi, \Pi_u$	Voronoi cells, spheres, planes
Delaunay	Delaunay complex
S, u, v, w_u, w_v	finite point set, points, weights
x, y, z, p, r	points, radius
$K_i = \operatorname{Alpha}(r_i)$	i-th alpha complex
$R_u(r) = B_u(r) \cap V_u$	intersection of ball with Voronoi cell
$\alpha, \beta, \beta_0, \sigma, \tau, \upsilon$	abstract, geometric simplices

Table X.3: Notation in Chapter III.

K, L	simplicial complexes
$a_i, b_i, \sigma_i, \tau_i, c, c', c'', d$	coefficients, simplices, chains
$\partial_p : \mathsf{C}_p \to \mathsf{C}_{p-1}$	chain group, dimension, boundary map
$\mathsf{Z}_p = \ker \partial_p, \mathsf{B}_p = \operatorname{im} \partial_{p+1}$	cycle, boundry groups
$\mathsf{H}_p = \mathsf{Z}_p / \mathsf{B}_p, \beta_p = \operatorname{rank} \mathsf{H}_p$	homology group, Betti number
$\epsilon : \mathsf{C}_0 \to \mathbb{Z}_2, \tilde{\mathsf{H}}_p, \tilde{\beta}_p$	augmentation map, reduced group, Betti number
n_p, z_p, b_p	ranks of the chains, cycles, boundaries
$\partial_p = [a_i^j]$	boundary matrix
$N_p = U_{p-1} \partial_p V_p$	normal form matrix
i, k, j, l	row indices, column indices
$(K, K_0), \mathsf{H}_p(K, K_0)$	pair of complexes, homology group
$f : K \to L$	simplicial map
$f_\# : \mathsf{C}(K) \to \mathsf{C}(L)$	induced map on chains
$f_* : \mathsf{H}(K) \to \mathsf{H}(L)$	induced map on homology
$A, B, g : \mathbb{S}^p \to \mathbb{S}^p$	continuous maps
$f : \mathbb{B}^{p+1} \to \mathbb{B}^{p+1}$	continuous map
α	modulo 2 degree
$\mathsf{U}, \mathsf{V}, \mathsf{W}$	vector spaces
$\mathsf{U} \oplus \mathsf{V}$	direct sum
$\mathcal{U} = (\mathsf{U}_p, u_p), \mathcal{V}, \mathcal{W}$	chain complexes
$\mathcal{C}(K) = (\mathsf{C}_p(K), \partial_p)$	chain complex
$\phi : \mathcal{U} \to \mathcal{V}, \psi : \mathcal{V} \to \mathcal{W}$	chain maps
$D : \mathsf{H}_p(\mathcal{W}) \to \mathsf{H}_{p-1}(\mathcal{U})$	connecting homomorphism
$\alpha, \alpha_0, \beta, \beta_0, \gamma, \gamma_0, \mu, \mu', \nu, \varrho$	chains and cycles
$i', i'' : A \to (K', K'')$	inclusions
$j : (K', K'') \to K$	inclusion
$\mathbb{P}^d$	d-dimensional real projective space

Table X.4: Notation in Chapter IV.

$\mathsf{G} = \mathbb{Z}_2$	coefficient group
A, B	groups of the form G^n
$\varphi, \varphi_0, \psi, \psi_0$	homomorphisms
$\mathrm{Hom}(\mathsf{A}, \mathsf{G})$	group of homomorphisms
$f : \mathsf{A} \to \mathsf{B}$	homomorphism
$\tilde{f} : \mathrm{Hom}(\mathsf{B}, \mathsf{G}) \to \mathrm{Hom}(\mathsf{A}, \mathsf{G})$	dual homomorphism
$\varphi(c) = \langle \varphi, c \rangle$	cochain evaluating chain
$\mathsf{C}^p, \mathsf{Z}^p, \mathsf{B}^p, \mathsf{H}^p$	cochain, cocycle, coboundary, cohomology groups
$\delta^{p-1} : \mathsf{C}^{p-1} \to \mathsf{C}^p$	coboundary map

Table X.5: Notation in Chapter V.

$\mathbb{M}, f : \mathbb{M} \to \mathbb{R}$	manifold, Morse function
$\mathbb{M}_a = f^{-1}(-\infty, a]$	sublevel set
u, v, w, z	critical points
$Df_x : \mathrm{T}\mathbb{M}_x \to \mathrm{T}\mathbb{R}_{f(x)}$	derivative at x
$(x_1, x_2, \ldots, x_d)$	local coordinate system
$H(x) = [\frac{\partial^2 f}{\partial x_i \partial x_j}(x)]$	Hessian
$\mathrm{index}(u)$	index of critical point
$X : \mathbb{M} \to \mathrm{T}\mathbb{M}$	vector fields
$X[f], \nabla f$	derivative in direction $X(x)$, gradient
$\varphi : \mathbb{R} \times \mathbb{M} \to \mathbb{M}$	1-paramenter family of diffeomorphisms
$g : \mathbb{S}^{q-1} \to \mathrm{bd}\,\mathbb{M}_a$	gluing function to attach q-handle
$\gamma_x : \mathbb{R} \to \mathbb{M}$	integral line through x
$\mathrm{org}(\gamma), \mathrm{dest}(\gamma)$	origin, destination
$S(u), U(u)$	stable, unstable manifold
$\sigma, \upsilon : \mathbb{R}^p \to \mathbb{M}$	immersions
c_q	number of index q critical points
$f : \lvert K\rvert \to \mathbb{R}$	PL Morse function
$\lvert K\rvert_a = f^{-1}(-\infty, a]$	sublevel set
$f(u_1) < \ldots < f(u_n)$	ordered vertices
$K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n$	lower star filtration
$\mathrm{St}_- u_i, \mathrm{Lk}_- u_i$	lower star, lower link
$\psi_p, k_p = \mathrm{rank}\ker\psi_p$	homomorphism, rank of kernel
$\varphi_p, k^p = \mathrm{rank}\,\mathrm{cok}\,\varphi_p$	homomorphism, rank of cokernel
$\mathbb{X}, \mathbb{M}$	space, manifold
$R(f)$	Reeb graph, quotient space
$\psi \circ \pi : \mathbb{X} \to R(f) \to \mathbb{R}$	maps decomposing f
$n = \sum n_i, m, c_q$	number of nodes, arcs, critical points

Table X.6: Notation in Chapter VI.

$f : K \to \mathbb{R}$	function on simplices
$K_0 \subseteq \ldots \subseteq K_n$	filtration
$f_p^{i,j} : \mathsf{H}_p(K_i) \to \mathsf{H}_p(K_j)$	induced homomorphism
$\mathsf{H}_p^{i,j} = \operatorname{im} f_p^{i,j}$	persistent homology group
$\beta_p^{i,j}, \mu_p^{i,j}$	persistent Betti number, multiplicity
$\mathrm{Dgm}_p(f)$	persistence diagram
$R = \partial V, R_i^j$	matrices, submatrices
$\#Zero_p, \#Low_p$	#zero columns, #lowest ones
$f : \lvert K\rvert \to \mathbb{R}$, $g : K \to \mathbb{R}$	PL function, montonic function
$\sigma_1, \ldots, \sigma_m$	compatible ordering
$K_0 \subseteq \ldots \subseteq K_n$	lower star filtration
$\partial, R[1..m]$	linear arrays
$L = R[j].cycle, L_i$	linked lists
u, v, w	vertices
$\lvert K\rvert_a, K_a$	sublevel sets
$\phi_i : \mathsf{U}_i \to \mathsf{V}_i$	homomorphism between vector spaces
$f_u : \mathbb{M} \to \mathbb{R}$	height function
$\mathbb{M}_a, \mathbb{M}^a$	sublevel set, superlevel set
$a_1 < \ldots < a_n$	homological critical values
$b_0 < \ldots < b_n$	interleaved values
$\mathrm{Ord}_p(f), \mathrm{Ext}_p(f), \mathrm{Rel}_p(f)$	sub-diagrams
$R, T, 0$	reflections
$\mathrm{St}_- v_i, \mathrm{St}^+ v_i$	lower, upper star
K_i, K^i	lower, upper star filtration
$\partial^j, \partial_i, \partial_i^j$	block of columns, rows, intersection
$K_j - K_{j-1}$, $k_j = \operatorname{card} K_j$	block in filtration, size
$\mathsf{C}_p^j, \partial_i^j$	chain group, boundary map
$\mathsf{E}_{p,q}^r, \mathsf{d}_{p,q}^r$	groups, maps in spectral sequence
E^r-term	r-th term of spectral sequence

Table X.7: Notation in Chapter VII.

$f, g : K \to \mathbb{R}$	monotonic function on simplices
$f_t = (1-t)f + tg$	straight-line homotopy
$\sigma_1, \ldots, \sigma_m$	compatible ordering of simplices
∂, R, V, U	matrices
$i = low(j)$	row of lowest one in column
$P = P_i^{i+1}, S = S_i^{i+1}$	permutation, addition matrix
$R' = PRSP, U' = PSUP$	modified matrices
$u, v, w; x, y; z$	vertices, edges, triangle
X, Y, X_t	persistence diagrams
$W_\infty(X, Y)$	bottleneck distance
W_q	degree q Wasserstein distance
$\mathrm{Pers}_k(X)$	degree k total persistence
$\eta : X \to Y$	bijection between persistence diagrams
$\|x-y\|_\infty, \|f-g\|_\infty$	L_∞-distance
$f_t = (1-t)f + tg$	straight-line homotopy
$x(t) = (f_t(\sigma), f_t(\tau), t)$	point of vineyard
$b_{i-1} < a_i < b_i$	interleaved values, homological critical values
$\mu_p^{a_i, a_j}$	multiplicity of point
$\operatorname{diam}\sigma$	diameter
$r = \operatorname{mesh} K = \max_\sigma \operatorname{diam}\sigma$	mesh
$N(r) \leq c/r^j$	size of triangulation
$c, C, j < k \leq q$	constants
$\gamma, \gamma_0 : \mathbb{S}^1 \to \mathbb{R}^2$	smooth curves
$length(\gamma) = \int \|\dot\gamma(s)\|$	length
$curv(\gamma) = \int \kappa(s)$	total curvature
$\mathrm{Pers}_0(f_u) = \sum \mathrm{pers}(a)$	zeroth total persistence
$f_u = g_u \circ \gamma, f_{0,u} = g_u \circ \gamma_0$	height on curves
$F(\gamma, \gamma_0)$	Frechet distance
X, Y	persistence diagrams
X_0, X_0', Y_0, Y_0'	off-diagonal points, projections
$G, G(\varepsilon), G_i, G_i(\varepsilon)$	bipartite graphs
$c = c^q : E \to \mathbb{R}$	cost function
$d_i : U \cup V \to \mathbb{R}$	deduction map
$M, M_i \subseteq E$	matchings
$n+n, m, m_\varepsilon, \bar m_i$	number of vertices, edges
s, t, u, v, x, y	source, target, other vertices
$D_i, \ell_i(x), pi$	directed graph, distance, path
$mean(S), gauss(S)$	total mean, absolute Gaussian curvature

Table X.8: Notation in Chapter VIII.

$f : K \to \mathbb{R}$	function on simplices
$K_0 \subseteq \ldots \subseteq K_m$	filtration

Table X.9: Notation in Chapter IX.

$f : K \to \mathbb{R}$	function on simplices
$K_0 \subseteq \ldots \subseteq K_m$	filtration

Table X.10: Notation in Chapter X.

Appendix B

Algorithms in Chapter I:
Union-find to determine connectedness.
Parity Algorithm (for point-in-polygon test).
Straight-line Embedding Algorithm (of planar graphs).

Algorithms in Chapter II:
Depth-first Search (in abstract graphs).
Recognizing orientability of triangulated, compact 2-manifolds.
Classifying a triangulated, compact 2-manifold.
Recognizing crossing triangles in space.
Surface Simplifation Algorithm (by repeated edge contraction).

Algorithms in Chapter III:
Miniball Algorithm (for finite sets of points).

Algorithms in Chapter IV:
SNF Reduction for Homology.
SNF Reduction for Relative Homology.

Algorithms in Chapter V:
SNF Reduction for Cohomology.
Incremental Betti Number Algorithm (for complexes in $\mathbb{S}^3$).

Algorithms in Chapter VI:
Marching Cube Algorithm.
Reeb Graph Algorithm (for 2-manifolds).

Algorithms in Chapter VII:
Persistence Algorithm (matrix reduction version).
Persistence Algorithm (sparse matrix version).
Persistence Algorithm (for components).
Persistence Aglrotihm (for 2-manifolds).
Extended Persistence Algorithm.
Spectral Sequence Algorithm.

Algorithms in Chapter VIII:
Maintaining an ru-decomposition.
Breadth-first Search.
Maximum Matching Algorithm (for bipartite graphs).
Minimum Cost Matching Algorithm (for bipartite graphs).
Dijkstra's Single Source Shortest Path Algorithm.

Algorithms in Chapter IX:
Contours and silhouettes.
Jacobi sets.

Algorithms in Chapter X:

Appendix C

Theorems in Chapter I:
- Jordan Curve Theorem.
- Euler Relation for Planar Graphs.
- Kuratowski Theorem.
- White-Calugareanu Formula.
- Tutte's Theorem.

Theorems in Chapter II:
- Classification Theorem for Compact 2-manifolds.
- Euler Characteristic of Compact 2-manifolds.
- Link Condition.

Theorems in Chapter III:
- Geometric Realization Theorem.
- Mesh Lemma.
- Simplicial Approximation Theorem.
- Helly's Theorem.
- Nerve Theorem.
- Vietoris-Rips Lemma.
- Inversion Lemma.
- Stereographic Projection Lemma.
- First Sphere/Plane Lemma.

Theorems in Chapter IV:
- Fundamental Lemma of Homology.
- Euler-Poincare Theorem.
- Brouwer's Fixed Point Theorem.
- Exact Homology Sequence of a Pair.
- Snake Lemma.
- Mayer-Vietoris Sequence.

Theorems in Chapter V:
- Universal Coefficient Theorem.
- Block Complex Lemma.
- Poincare Duality Theorem (first form).
- Poincare Duality Theorem (second form).
- Lefschetz Duality Theorem (first form).
- Lefschetz Duality Theorem (second form).
- Alexander Duality Theorem.

Theorems in Chapter VI:
- Morse Lemma.
- Morse Inequalities, weak and strong.
- PL Morse Inequalities, weak and strong.
- Loop Lemma for 2-manifolds.

Theorems in Chapter VII:
- Elder Rule.
- Fundamental Lemma of Persistent Homology.
- Equivalence of Persistence.
- Pairing Lemma.
- Persistence Equivalence Theorem.
- Persistence Duality Theorem.
- Persistence Symmetry Theorem.
- Spectral Sequence Theorem.

Theorems in Chapter VIII:

Transposition Lemma.
Stability Theorem for Filtrations.
Stability Theorem for Tame Functions.
Stability Theorem for Lipschitz Functions.
Cauchy-Crofton Formula.
Total Curvature Formula.
Fary Theorem.
Generalized Fary Theorem.
Reduction Lemma.
Iteration Bound.

Theorems in Chapter IX:

Theorems in Chapter X:

Appendix Z

Here we list things that should still be done to complete or improve the book. The first list contains specific items, short or spread out.

- Can we say something about the history of topological duality in Chapter V? I thought there was a book on the topic but now I can't find it.
- Right now there is no discussion of handle slides and cancellations in Chapter VI. It would be useful to add this material as it foreshadows the concept of persistence discussed in Chapter VII.
- Is the reference to Edwards about the Hauptvermutung correct? (It used to be in Section IV.4 but has no home at the moment.)
- Double check the definition of 1-parameter family of diffeomorphisms in Section VI.1, which seems a bit odd as we approach a critical point.
- Can we add the proof that pairing is perfect iff the implied natural homomorphism is an isomorphism in Section V.3?

The second list contains chapters and sections that still need work.

- Finish Chapter V on Duality.
  - Finish Section on Cohomology.
  - Revise Section on Poincare Duality.
  - Revise Section on Intersection Theory.
  - Write Section on Alexander Duality.
  - Formulate exercises.
- Write Chapter IX on Application. Rethink the structure of sections, which currently is IX.1 Simplification for Gene Expression, IX.2 Elevation for Protein Docking, IX.3 Image Segmentation, IX.4 Local Homology for Root Architecture.
- Finish Chapter X on Open Problems.
  - Write Problem 6 on Running-time of Matrix Reduction, adding the recent upper bound of matrix multiplication time.
  - Write Problem 7 on Multi-parameter Persistence.
  - Write Problem 8 on Unfolding PL Critical Points.
  - Write Problem 9 on PL in the Limit.
  - Write Problem 10 on Counting Halving Sets.
- Beautify the index and the glossary.