Advanced Geometric Methods In Computer Science CIS610

Jean Gallier
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
e-mail: jean@saul.cis.upenn.edu

November 21, 2014

Chapter 1

Introduction: Questions, Motivations, Problems

What's the difference between an affine space and a linear space (vector space)?

What are affine maps?

What's Euclidean space?

What's Euclidean geometry?

What's an isometry?

What's a rotation?

What's the difference between a rotation and an orthogonal matrix (with det = 1)?

What's a self-adjoint map?

Why do self-adjoint maps have real eigenvalues? Why can they be diagonalized using orthogonal matrices?

How does one solve an "inconsistent" linear system

$$Ax = b,$$

e.g., when there are more equations than variables?

Some Answers

The set of rotations of Euclidean n-space $\mathbb{E}^n$ forms a group $\mathbf{SO}(n)$.

The group $\mathbf{SO}(n)$ is generated by the hyperplanes reflections. In fact, n reflections suffice.

Rotations in $\mathbf{SO}(3)$ can be "represented" by a quaternion.

Rotations in $\mathbf{SO}(4)$ can be "represented" by two quaternions.

The group $\mathbf{SO}(n)$ is a (path-connected) topological space.

The group $\mathbf{SO}(n)$ has a (smooth) differential structure. The notion of tangent space at a point makes sense.

This makes $\mathbf{SO}(n)$ into a *Lie group*.

The tangent space $\mathfrak{so}(n)$ (at the origin) has some additional structure: it is a *Lie algebra*.

There is a map

$$\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$$

called the *exponential map*. It provides a local representation of the space.

Similarly, the rigid motions of Euclidean affine space form a Lie group $\mathbf{SE}(n)$, etc.

The exponential map can be used to *interpolate* in $\mathbf{SO}(n)$ or $\mathbf{SE}(n)$. There are applications to motion interpolations.

Lie groups pop up naturally in certain problems in robotics and computer vision.

Often, it is necessary to find out the number of "degrees of freedom" and this turns out to be the dimension of the Lie group.

Chris Geyer (Ph.D. under Kostas Daniilidis) used Lie groups in an interesting way in his dissertation on cata-dioptric sensors (and mirrors).

Every $n \times n$-matrix A can be written as

$$A = QS$$

where Q is orthogonal and S is symmetric with non-negative eigenvalues (the *polar form*).

Every $n \times n$-matrix A can be written as

$$A = VDU^\top$$

where U and V are orthogonal and D is a diagonal matrix with non-negative entries (*singular value decomposition, or SVD*).

The SVD can be used solve an "inconsistent" linear system

$$Ax = b.$$

We solve the *least squares problem*: Minimize $\|Ax - b\|$.

It can be shown that there is a vector x of smallest norm minimizing $\|Ax - b\|$. It is given by the (Penrose) *pseudo-inverse* (itself given by the SVD).

All this suggests studying some basic of
Euclidean Geometry and
Lie groups and Lie algebras.

I. Some problems in 3D-Mesh Generation

There is a class of problems, including 3D-mesh generation in medical imaging, notably brain imaging, where it is necessary to construct a 3D-tetrahedral mesh (a "tetrahedrization") from an input polyhedron (image data) possibly with boundaries.

One of the major difficulties is that such a polyhedron is usually **not** convex. If the input is convex, a 3D-Delaunay triangulation (i.e., a decomposition of the convex body into tetrahedra, a "tetrahedrization") is a very satisfactory answer.

Furthermore, a nonconvex polyhedron **cannot always** be triangulated by tetrahedra without adding extra vertices!

How do we cope with these problems?

There are at least two possible attacks:

(1) Conformal Delaunay triangulations. This means, Delaunay tetrahedral triangulations whose vertices and faces include the vertices and faces of the original input polyhedron.

    Their construction requires the addition of new vertices and faces. No upper bound known. Quality of the tetrahedra not always good.

(2) Use the medial axis concept (see below).

These problems are pretty much open.

II. Dirichlet-Voronoi Diagrams and Delaunay Triangulations in Euclidean Space. Applications to Tetrahedral Mesh Generation

In a Euclidean space, given a finite set $P = \{p_1, ..., p_m\}$ of points, the Voronoi region $V(p_i)$ of p_i consists of all points that are closer to p_i than to any $p_j \neq p_i$. The Voronoi region $V(p_i)$ is the intersection of the half planes containing p_i defined by the bisector hyperplanes of the pairs of points (p_i, p_j).

Voronoi diagrams and their duals, Delaunay triangulations, have many applications, as we will see below.

The figure below shows the Voronoi diagram of a set of twelve points.

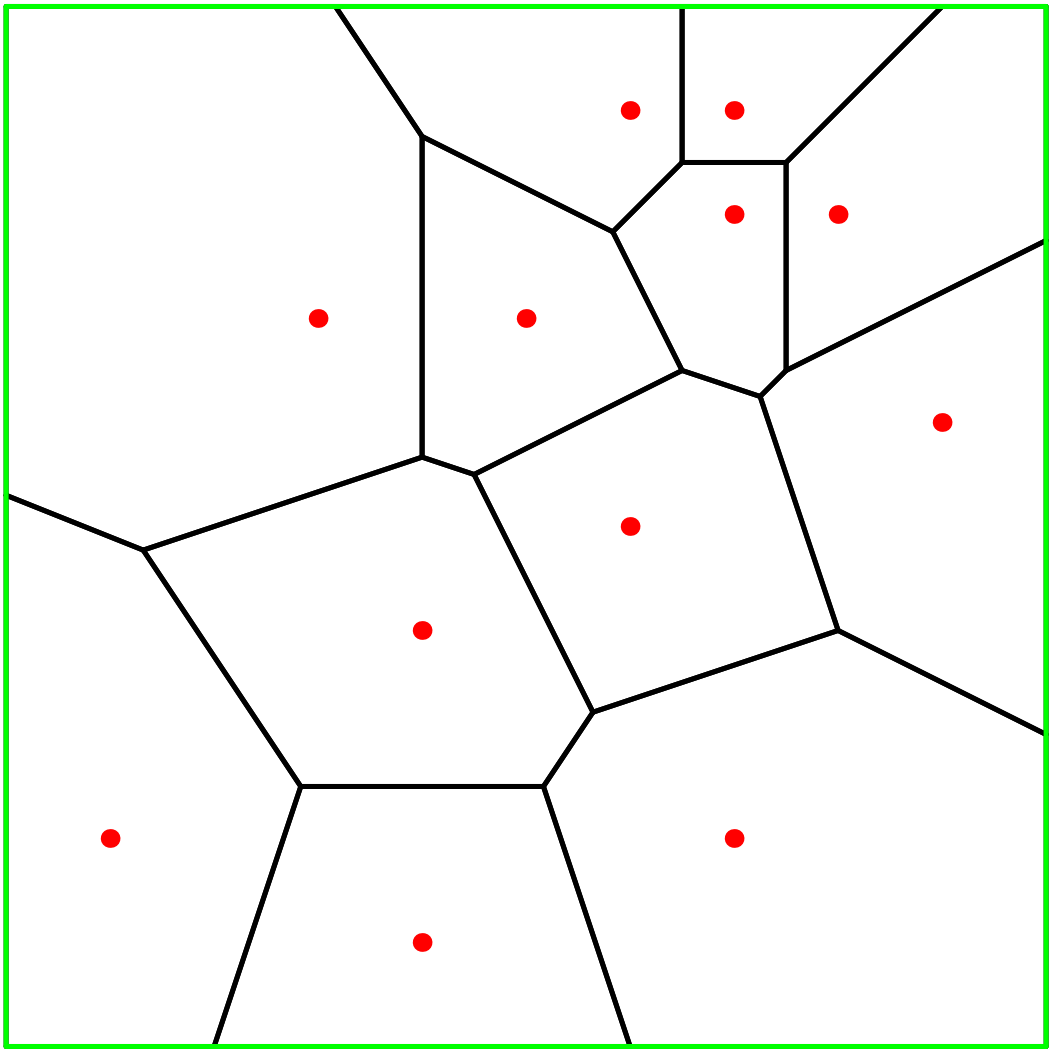

Figure 1.1: A Voronoi diagram

The figure below shows the Delaunay triangulation associated with the earlier Voronoi diagram.

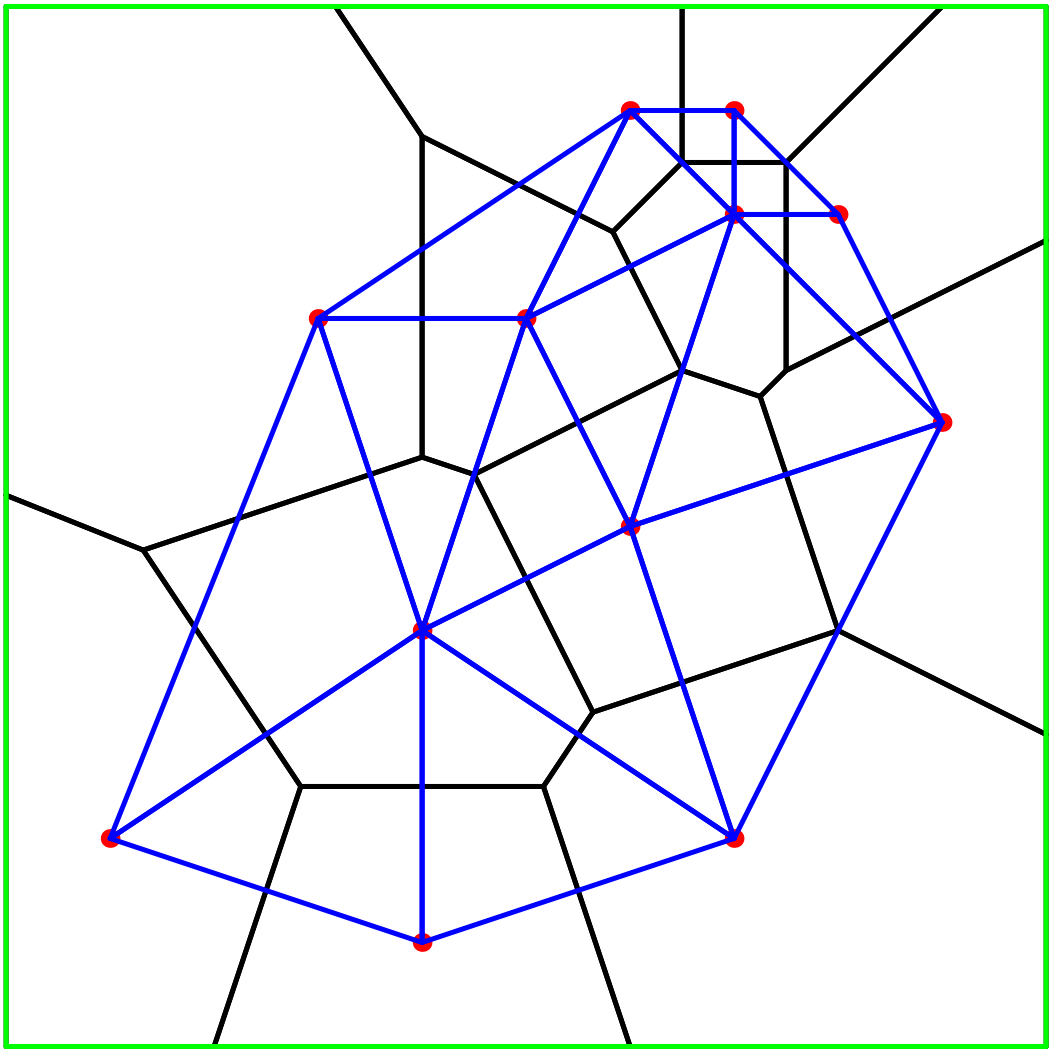

Figure 1.2: Delaunay triangulation associated with a Voronoi diagram

Delaunay trianguations are used to provide "nice" triangulations for defining complicated shapes, or in the method of finite elements.

III. Medial Axis Representation

We can generalize the notion of a medial line. This leads to the notion of medial axis. The basic idea is to represent a complex shape in terms of a kind of *skeleton* and a boundary representation.

The Blum *medial axis* of a planar shape S is defined (roughly) as the locus of centers of all circles bitangent to the boundary of S in two distinct points.

A similar notion can be defined for 3D-shapes, using spheres instead of circles.

Medial axis representations are useful in medical applications, for instance: Geometric representation of the kidneys, liver, prostate, etc.

How are medial axes computed? For instance, as pruned trees of Voronoi diagrams.

Chapter 2

Basics of Affine Geometry

2.1 Affine Spaces

For simplicity, it is assumed that all vector spaces under consideration are defined over the field $\mathbb{R}$ of real numbers.

It is also assumed that all families $(\lambda_i)_{i \in I}$ of scalars have finite support. Recall that a family $(\lambda_i)_{i \in I}$ of scalars has *finite support* if

$\lambda_i = 0$ for all $i \in I - J$,

where J is a finite subset of I.

Obviously, finite families of scalars have finite support, and for simplicity, the reader may assume that all families are finite.

Suppose we have a particle moving in 3-space and that we want to describe the trajectory of this particle.

If one looks up a good textbook on dynamics, such as Greenwood [**?**], one finds out that the particle is modeled as a point, and that the position of this point x is determined with respect to a "frame" in $\mathbb{R}^3$ by a vector.

A frame is a pair

$$(O, (e_1, e_2, e_3))$$

consisting of an origin O (which is a point) together with a basis of three vectors (e_1, e_2, e_3).

For example, the standard frame in $\mathbb{R}^3$ has origin $O = (0, 0, 0)$ and the basis of three vectors $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, and $e_3 = (0, 0, 1)$.

The position of a point x is then defined by the "unique vector" from O to x.

But wait a minute, this definition seems to be defining frames and the position of a point without defining what a point is!

Well, let us identify points with elements of $\mathbb{R}^3$.

If so, given any two points $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$, there is a unique *free vector* denoted $\mathbf{ab}$ from a to b, the vector $\mathbf{ab} = (b_1 - a_1, b_2 - a_2, b_3 - a_3)$.

Note that

$$b = a + \mathbf{ab},$$

addition being understood as addition in $\mathbb{R}^3$.

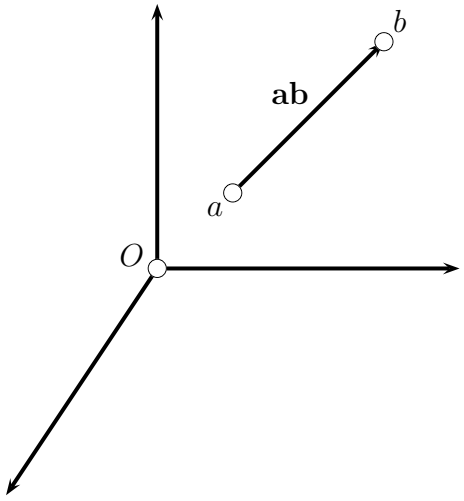

Figure 2.1: Points and free vectors

Then, in the standard frame, given a point $x = (x_1, x_2, x_3)$,

the position of x is the vector $\mathbf{Ox} = (x_1, x_2, x_3)$, which coincides with the point itself.

What if we pick a frame with a different origin, say $\Omega = (\omega_1, \omega_2, \omega_3)$, but the same basis vectors (e_1, e_2, e_3)?

This time, the point $x = (x_1, x_2, x_3)$ is defined by two position vectors:

$\mathbf{Ox} = (x_1, x_2, x_3)$ in the frame $(O, (e_1, e_2, e_3))$, and

$\mathbf{\Omega x} = (x_1 - \omega_1, x_2 - \omega_2, x_3 - \omega_3)$ in the frame $(\Omega, (e_1, e_2, e_3))$.

This is because

$$\mathbf{Ox} = \mathbf{O\Omega} + \mathbf{\Omega x} \quad \text{and} \quad \mathbf{O\Omega} = (\omega_1, \omega_2, \omega_3).$$

We note that in the second frame $(\Omega, (e_1, e_2, e_3))$, points and position vectors are no longer identified.

This gives us evidence that **points are not vectors**. Inspired by physics, it is important to define points and properties of points that are frame invariant.

An undesirable side-effect of the present approach shows up if we attempt to define linear combinations of points.

If we consider the change of frame from the frame

$$(O, (e_1, e_2, e_3))$$

to the frame

$$(\Omega, (e_1, e_2, e_3)),$$

where

$$\mathbf{O\Omega} = (\omega_1, \omega_2, \omega_3),$$

given two points a and b of coordinates (a_1, a_2, a_3) and (b_1, b_2, b_3) with respect to the frame $(O, (e_1, e_2, e_3))$ and of coordinates (a'_1, a'_2, a'_3) and (b'_1, b'_2, b'_3) of with respect to the frame $(\Omega, (e_1, e_2, e_3))$, since

$$(a'_1, a'_2, a'_3) = (a_1 - \omega_1, a_2 - \omega_2, a_3 - \omega_3)$$

and

$$(b'_1, b'_2, b'_3) = (b_1 - \omega_1, b_2 - \omega_2, b_3 - \omega_3),$$

the coordinates of $\lambda a + \mu b$ with respect to the frame $(O, (e_1, e_2, e_3))$ are

$$(\lambda a_1 + \mu b_1, \lambda a_2 + \mu b_2, \lambda a_3 + \mu b_3),$$

but the coordinates

$$(\lambda a'_1 + \mu b'_1, \lambda a'_2 + \mu b'_2, \lambda a'_3 + \mu b'_3)$$

of $\lambda a + \mu b$ with respect to the frame $(\Omega, (e_1, e_2, e_3))$ are

$$\begin{aligned}(&\lambda a_1 + \mu b_1 - (\lambda + \mu)\omega_1,\\ &\lambda a_2 + \mu b_2 - (\lambda + \mu)\omega_2,\\ &\lambda a_3 + \mu b_3 - (\lambda + \mu)\omega_3)\end{aligned}$$

which are different from

$$(\lambda a_1 + \mu b_1 - \omega_1, \lambda a_2 + \mu b_2 - \omega_2, \lambda a_3 + \mu b_3 - \omega_3),$$

unless $\lambda + \mu = 1$.

Thus, we discovered a major difference between vectors and points: the notion of linear combination of vectors is basis independent, but the notion of linear combination of points is frame dependent.

In order to salvage the notion of linear combination of points, some restriction is needed: the scalar coefficients must add up to 1.

A clean way to handle the problem of frame invariance and to deal with points in a more intrinsic manner is to make a clearer distinction between points and vectors.

We duplicate $\mathbb{R}^3$ into two copies, the first copy corresponding to points, where we forget the vector space structure, and the second copy corresponding to free vectors, where the vector space structure is important.

Furthermore, we make explicit the important fact that the vector space $\mathbb{R}^3$ acts on the set of points $\mathbb{R}^3$: Given any **point** $a = (a_1, a_2, a_3)$ and any **vector** $v = (v_1, v_2, v_3)$, we obtain the **point**

$$a + v = (a_1 + v_1, a_2 + v_2, a_3 + v_3),$$

which can be thought of as the result of translating a to b using the vector v.

This action $+: \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ satisfies some crucial properties. For example,

$$a + 0 = a,$$
$$(a + u) + v = a + (u + v),$$

and for any two points a, b, there is a unique free vector $\mathbf{ab}$ such that

$$b = a + \mathbf{ab}.$$

It turns out that the above properties, although trivial in the case of $\mathbb{R}^3$, are all that is needed to define the abstract notion of affine space (or affine structure).

Definition 2.1.1 An *affine space* is either the empty set, or a triple $\langle E, \overrightarrow{E}, + \rangle$ consisting of a nonempty set E (of *points*), a vector space $\overrightarrow{E}$ (of *translations, or free vectors*), and an action $+: E \times \overrightarrow{E} \to E$, satisfying the following conditions:

(A1) $a + 0 = a$, for every $a \in E$;

(A2) $(a + u) + v = a + (u + v)$, for every $a \in E$, and every $u, v \in \overrightarrow{E}$;

(A3) For any two points $a, b \in E$, there is a unique $u \in \overrightarrow{E}$ such that $a + u = b$.

The unique vector $u \in \overrightarrow{E}$ such that $a + u = b$ is denoted as $\mathbf{ab}$, or sometimes as $b - a$. Thus, we also write

$$b = a + \mathbf{ab}$$

(or even $b = a + (b - a)$).

The *dimension of the affine space* $\langle E, \overrightarrow{E}, + \rangle$ is the dimension $\dim(\overrightarrow{E})$ of the vector space $\overrightarrow{E}$. For simplicity, it is denoted by $\dim(E)$.

Conditions (A1) and (A2) say that the (abelian) group $\overrightarrow{E}$ acts on E, and condition (A3) says that $\overrightarrow{E}$ acts transitively and faithfully on E.

Note that

$$\mathbf{a}(\mathbf{a} + \mathbf{v}) = v$$

for all $a \in E$ and all $v \in \overrightarrow{E}$, since $\mathbf{a}(\mathbf{a} + \mathbf{v})$ is the unique vector such that $a + v = a + \mathbf{a}(\mathbf{a} + \mathbf{v})$.

Thus, $b = a + v$ is equivalent to $\mathbf{ab} = v$.

It is natural to think of all vectors as having the same origin, the null vector.

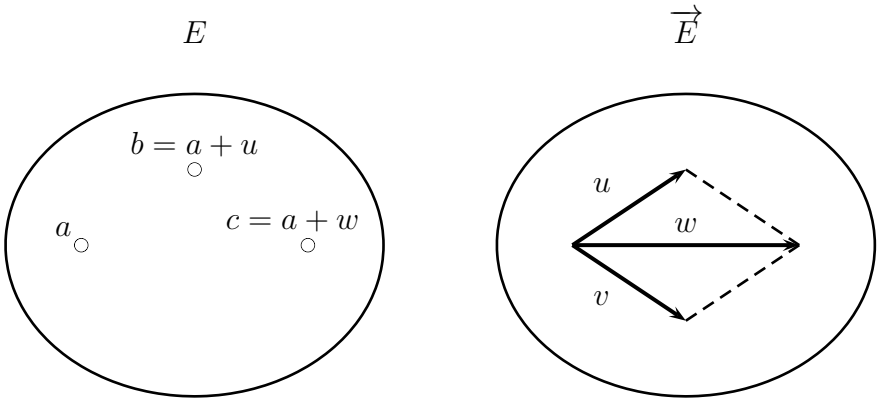

Figure 2.2: Intuitive picture of an affine space

For every $a \in E$, consider the mapping from $\overrightarrow{E}$ to E:

$$u \mapsto a + u,$$

where $u \in \overrightarrow{E}$, and consider the mapping from E to $\overrightarrow{E}$:

$$b \mapsto \mathbf{ab},$$

where $b \in E$.

The composition of the first mapping with the second is

$$u \mapsto a + u \mapsto \mathbf{a}(\mathbf{a} + \mathbf{u}),$$

which, in view of (A3), yields u.

The composition of the second with the first mapping is

$$b \mapsto \mathbf{ab} \mapsto a + \mathbf{ab},$$

which, in view of (A3), yields b.

Thus, these compositions are the identity from $\overrightarrow{E}$ to $\overrightarrow{E}$ and the identity from E to E, and the mappings are both bijections.

When we identify E to $\overrightarrow{E}$ via the mapping $b \mapsto \mathbf{ab}$, we say that we consider E as the vector space obtained *by taking a as the origin in E*, and we denote it as E_a. Thus, an affine space $\langle E, \overrightarrow{E}, + \rangle$ is a way of defining a vector space structure on a set of points E, without making a commitment to a **fixed** origin in E.

For notational simplicity, we will often denote an affine space $\langle E, \overrightarrow{E}, + \rangle$ as $(E, \overrightarrow{E})$, or even as E. The vector space $\overrightarrow{E}$ is called the *vector space associated with E*.

One should be careful about the overloading of the addition symbol $+$. Addition is well-defined on vectors, as in $u + v$, the translate $a + u$ of a point $a \in E$ by a vector $u \in \overrightarrow{E}$ is also well-defined, but addition of points $a + b$ **does not make sense**.

In this respect, the notation $b - a$ for the unique vector u such that $b = a + u$, is somewhat confusing, since it suggests that points can be substracted (but not added!).

Any vector space $\overrightarrow{E}$ has an affine space structure specified by choosing $E = \overrightarrow{E}$, and letting $+$ be addition in the vector space $\overrightarrow{E}$. We will refer to the affine structure $\langle \overrightarrow{E}, \overrightarrow{E}, + \rangle$ on a vector space as the *canonical (or natural) affine structure on* $\overrightarrow{E}$.

In particular, the vector space $\mathbb{R}^n$ can be viewed as the affine space $\langle \mathbb{R}^n, \mathbb{R}^n, + \rangle$ denoted as $\mathbb{A}^n$. In order to distinguish between the double role played by members of $\mathbb{R}^n$, points and vectors, we will denote points as row vectors, and vectors as column vectors. Thus, the action of the vector space $\mathbb{R}^n$ over the set $\mathbb{R}^n$ simply viewed as a set of points, is given by

$$(a_1, \ldots, a_n) + \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = (a_1 + u_1, \ldots, a_n + u_n).$$

We will also use the convention that if $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, then the column vector associated with x is denoted as $\mathbf{x}$ (in boldface notation). Abusing the notation slightly, if $a \in \mathbb{R}^n$ is a point, we also write $a \in \mathbb{A}^n$.

The affine space $\mathbb{A}^n$ is called the *real affine space of dimension* n. In most cases, we will consider $n = 1, 2, 3$.

For a slightly wilder example, consider the subset P of $\mathbb{A}^3$ consisting of all points (x, y, z) satisfying the equation

$$x^2 + y^2 - z = 0.$$

The set P is a paraboloid of revolution, with axis Oz.

The surface P can be made into an official affine space by defining the action

$$+: P \times \mathbb{R}^2 \to P$$

of $\mathbb{R}^2$ on P defined such that for every point (x, y, x^2+y^2) on P and any $\begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^2$,

$$(x, y, x^2+y^2)+\begin{pmatrix} u \\ v \end{pmatrix} = (x+u, y+v, (x+u)^2+(y+v)^2).$$

Affine spaces not already equipped with an obvious vector space structure arise in projective geometry. Indeed, we will see in section **??** that the complement of a hyperplane in a projective space has an affine structure.

Given any three points $a, b, c \in E$, since $c = a + \mathbf{ac}$, $b = a + \mathbf{ab}$, and $c = b + \mathbf{bc}$, we get

$$c = b + \mathbf{bc} = (a + \mathbf{ab}) + \mathbf{bc} = a + (\mathbf{ab} + \mathbf{bc})$$

by (A2), and thus, by (A3),

$$\mathbf{ab} + \mathbf{bc} = \mathbf{ac},$$

which is known as *Chasles' identity*.

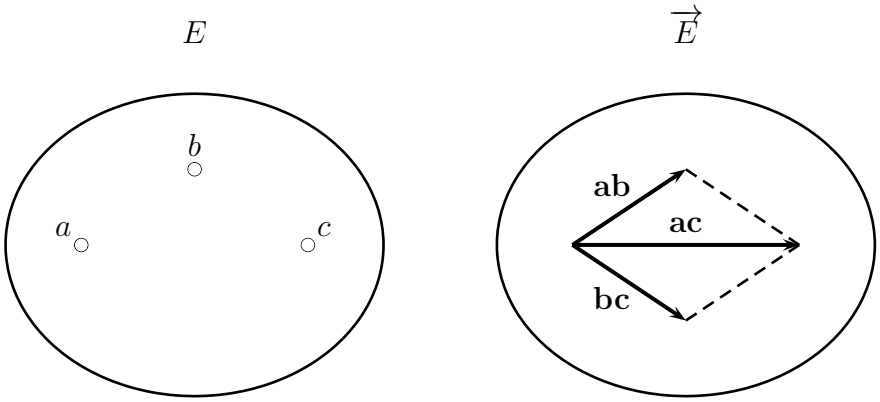

Figure 2.3: Points and corresponding vectors in affine geometry

2.2 Affine Combinations, Barycenters

A fundamental concept in linear algebra is that of a linear combination. The corresponding concept in affine geometry is that of an affine combination, also called a barycenter.

However, there is a problem with the naive approach involving a coordinate system. The problem is that the sum $a + b$ may correspond to two different points depending on which coordinate system is used for its computation!

Thus, some extra condition is needed in order for affine combinations to make sense. It turns out that if the scalars sum up to 1, the definition is intrinsic, as the following lemma shows.

Lemma 2.2.1 *Given an affine space E, let $(a_i)_{i\in I}$ be a family of points in E, and let $(\lambda_i)_{i\in I}$ be a family of scalars. For any two points $a, b \in E$, the following properties hold:*

(1) If $\sum_{i\in I} \lambda_i = 1$, then

$$a + \sum_{i\in I} \lambda_i \mathbf{aa_i} = b + \sum_{i\in I} \lambda_i \mathbf{ba_i}.$$

(2) If $\sum_{i\in I} \lambda_i = 0$, then

$$\sum_{i\in I} \lambda_i \mathbf{aa_i} = \sum_{i\in I} \lambda_i \mathbf{ba_i}.$$

Thus, by lemma 2.2.1, for any family of points $(a_i)_{i\in I}$ in E, for any family $(\lambda_i)_{i\in I}$ of scalars such that $\sum_{i\in I} \lambda_i = 1$, the point

$$x = a + \sum_{i\in I} \lambda_i \mathbf{aa_i}$$

is independent of the choice of the origin $a \in E$.

The unique point x is called the *barycenter (or barycentric combination, or affine combination) of the points a_i assigned the weights λ_i.* and it is denoted as

$$\sum_{i \in I} \lambda_i a_i.$$

In dealing with barycenters, it is convenient to introduce the notion of a *weighted point*, which is just a pair (a, λ), where $a \in E$ is a point, and $\lambda \in \mathbb{R}$ is a scalar.

Then, given a family of weighted points $((a_i, \lambda_i))_{i \in I}$, where $\sum_{i \in I} \lambda_i = 1$, we also say that the point

$$\sum_{i \in I} \lambda_i a_i$$

is the *barycenter of the family of weighted points* $((a_i, \lambda_i))_{i \in I}$.

Note that the barycenter x of the family of weighted points $((a_i, \lambda_i))_{i \in I}$ is also the unique point such that

$$\mathbf{ax} = \sum_{i \in I} \lambda_i \mathbf{aa_i} \quad \text{for every } a \in E,$$

and setting $a = x$, the point x is the unique point such that

$$\sum_{i \in I} \lambda_i \mathbf{xa_i} = 0.$$

In physical terms, the barycenter is the *center of mass* of the family of weighted points $((a_i, \lambda_i))_{i \in I}$ (where the masses have been normalized, so that $\sum_{i \in I} \lambda_i = 1$, and negative masses are allowed).

The figure below illustrates the geometric construction of the barycenters g_1 and g_2 of the weighted points $\left(a, \frac{1}{4}\right)$, $\left(b, \frac{1}{4}\right)$, and $\left(c, \frac{1}{2}\right)$, and $(a, -1)$, $(b, 1)$, and $(c, 1)$.

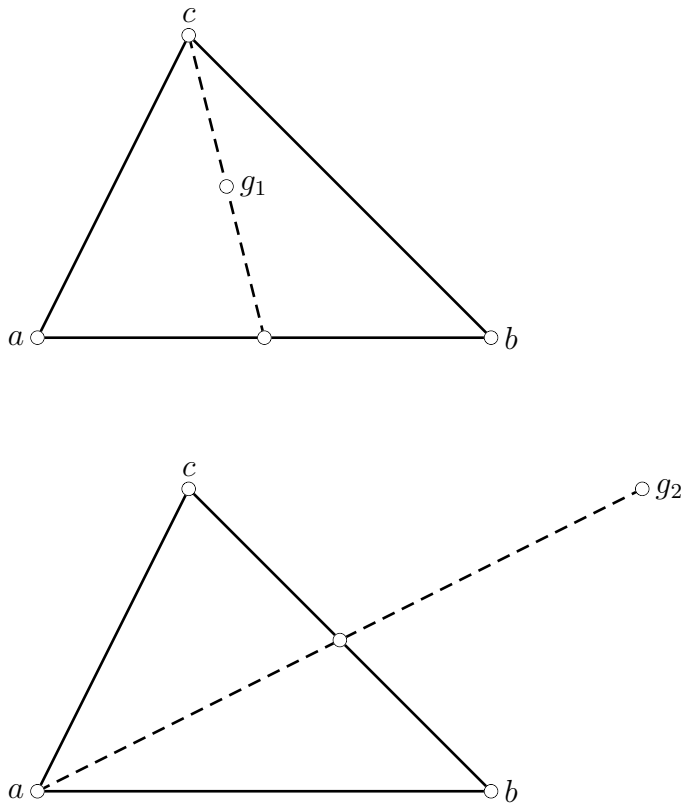

Figure 2.4: Barycenters, $g_1 = \frac{1}{4}a + \frac{1}{4}b + \frac{1}{2}c, \quad g_2 = -a + b + c.$

2.3 Affine Subspaces

In linear algebra, a (linear) subspace can be characterized as a nonempty subset of a vector space closed under linear combinations. In affine spaces, the notion corresponding to the notion of (linear) subspace is the notion of affine subspace.

It is natural to define an affine subspace as a subset of an affine space closed under affine combinations.

Definition 2.3.1 Given an affine space $\langle E, \overrightarrow{E}, + \rangle$, a subset V of E is an *affine subspace (of* $\langle E, \overrightarrow{E}, + \rangle$*)* if for every family of points $(a_i)_{i \in I}$ in V, for any family $(\lambda_i)_{i \in I}$ of scalars such that $\sum_{i \in I} \lambda_i = 1$, the barycenter $\sum_{i \in I} \lambda_i a_i$ belongs to V.

An affine subspace is also called a *flat* by some authors.

According to definition 2.3.1, the empty set is trivially an affine subspace, and every intersection of affine subspaces is an affine subspace.

As an example, consider the subset U of $\mathbb{R}^2$ defined by

$$U = \{(x, y) \in \mathbb{R}^2 \mid ax + by = c\},$$

i.e. the set of solutions of the equation

$$ax + by = c,$$

where it is assumed that $a \neq 0$ or $b \neq 0$.

Given any m points $(x_i, y_i) \in U$ and any m scalars λ_i such that $\lambda_1 + \cdots + \lambda_m = 1$, we claim that

$$\sum_{i=1}^{m} \lambda_i (x_i, y_i) \in U.$$

Thus, U is an affine subspace of $\mathbb{A}^2$. In fact, it is just a usual line in $\mathbb{A}^2$.

It turns out that U is closely related to the subset of $\mathbb{R}^2$ defined by

$$\overrightarrow{U} = \{(x, y) \in \mathbb{R}^2 \mid ax + by = 0\},$$

i.e. the set of solution of the homogeneous equation

$$ax + by = 0$$

obtained by setting the right-hand side of $ax + by = c$ to zero.

Indeed, for any m scalars λ_i, the same calculation as above yields that

$$\sum_{i=1}^{m} \lambda_i (x_i, y_i) \in \overrightarrow{U},$$

this time **without any restriction on the** λ_i, since the right-hand side of the equation is null.

Thus, $\overrightarrow{U}$ is a subspace of $\mathbb{R}^2$. In fact, $\overrightarrow{U}$ is one-dimensional, and it is just a usual line in $\mathbb{R}^2$.

This line can be identified with a line passing through the origin of $\mathbb{A}^2$, line which is parallel to the line U of equation $ax + by = c$.

Now, if (x_0, y_0) is any point in U, we claim that

$$U = (x_0, y_0) + \overrightarrow{U},$$

where

$$(x_0, y_0) + \overrightarrow{U} = \{(x_0 + u_1, y_0 + u_2) \mid (u_1, u_2) \in \overrightarrow{U}\}.$$

The above example shows that the affine line U defined by the equation

$$ax + by = c$$

is obtained by "translating" the parallel line $\overrightarrow{U}$ of equation

$$ax + by = 0$$

passing through the origin.

In fact, given any point $(x_0, y_0) \in U$,

$$U = (x_0, y_0) + \overrightarrow{U}.$$

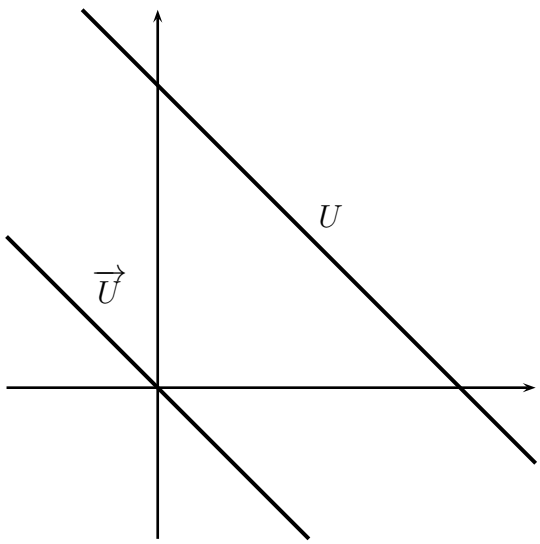

Figure 2.5: An affine line U and its direction

More generally, it is easy to prove the following fact. Given any $m \times n$ matrix A and any vector $b \in \mathbb{R}^m$, the subset U of $\mathbb{R}^n$ defined by

$$U = \{x \in \mathbb{R}^n \mid Ax = b\}$$

is an affine subspace of $\mathbb{A}^n$.

Actually, observe that $Ax = b$ should really be written as $Ax^\top = b$, to be consistent with our convention that points are represented by row vectors.

We can also use the boldface notation for column vectors, in which case the equation is written as $A\mathbf{x} = b$.

If we consider the corresponding homogeneous equation $Ax = 0$, the set

$$\overrightarrow{U} = \{x \in \mathbb{R}^n \mid Ax = 0\}$$

is a subspace of $\mathbb{R}^n$, and for any $x_0 \in U$, we have

$$U = x_0 + \overrightarrow{U}.$$

This is a general situation. Affine subspaces can also be characterized in terms of subspaces of $\overrightarrow{E}$.

Given any point $a \in E$ and any subset $\overrightarrow{V}$ of $\overrightarrow{E}$, let $a + \overrightarrow{V}$ denote the following subset of E:

$$a + \overrightarrow{V} = \{a + v \mid v \in \overrightarrow{V}\}.$$

Lemma 2.3.2 *Let $\langle E, \overrightarrow{E}, + \rangle$ be an affine space.*

(1) A nonempty subset V of E is an affine subspace iff, for every point $a \in V$, the set

$$\overrightarrow{V_a} = \{\mathbf{ax} \mid x \in V\}$$

is a subspace of $\overrightarrow{E}$. Consequently, $V = a + \overrightarrow{V_a}$. Furthermore,

$$\overrightarrow{V} = \{\mathbf{xy} \mid x, y \in V\}$$

is a subspace of $\overrightarrow{E}$ and $\overrightarrow{V_a} = \overrightarrow{V}$ for all $a \in E$. Thus, $V = a + \overrightarrow{V}$.

(2) For any subspace $\overrightarrow{V}$ of $\overrightarrow{E}$, for any $a \in E$, the set $V = a + \overrightarrow{V}$ is an affine subspace.

The subspace $\overrightarrow{V}$ associated with an affine subspace V is called the *direction of V*.

It is clear that the map $+: V \times \overrightarrow{V} \to V$ induced by $+: E \times \overrightarrow{E} \to E$ confers to $\langle V, \overrightarrow{V}, + \rangle$ an affine structure.

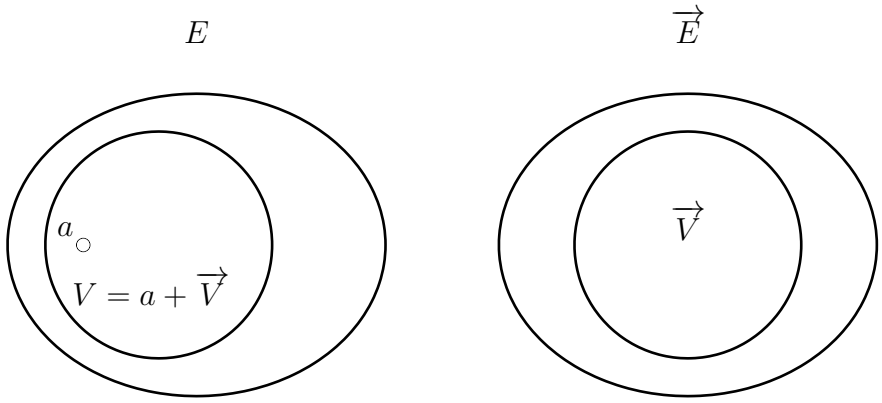

Figure 2.6: An affine subspace V and its direction $\overrightarrow{V}$

By the dimension of the subspace V, we mean the dimension of $\overrightarrow{V}$.

An affine subspace of dimension 1 is called a *line*, and an affine subspace of dimension 2 is called a *plane*.

An affine subspace of codimension 1 is called an *hyperplane*.

We say that two affine subspaces U and V are *parallel* if their directions are identical. Equivalently, since $\overrightarrow{U} = \overrightarrow{V}$, we have $U = a + \overrightarrow{U}$, and $V = b + \overrightarrow{U}$, for any $a \in U$ and any $b \in V$, and thus, V is obtained from U by the translation $\mathbf{ab}$.

In general, when we talk about n points $a_1, \ldots, a_n$, we mean the sequence $(a_1, \ldots, a_n)$, and not the set $\{a_1, \ldots, a_n\}$ (the a_i's need not be distinct).

We say that three points a, b, c are *collinear*, if the vectors $\mathbf{ab}$ and $\mathbf{ac}$ are linearly dependent.

If two of the points a, b, c are distinct, say $a \neq b$, then there is a unique $\lambda \in \mathbb{R}$, such that $\mathbf{ac} = \lambda\mathbf{ab}$, and we define the ratio $\frac{\mathbf{ac}}{\mathbf{ab}} = \lambda$.

We say that four points a, b, c, d are *coplanar*, if the vectors $\mathbf{ab}, \mathbf{ac}$, and $\mathbf{ad}$, are linearly dependent.

Lemma 2.3.3 *Given an affine space $\langle E, \overrightarrow{E}, + \rangle$, for any family $(a_i)_{i \in I}$ of points in E, the set V of barycenters $\sum_{i \in I} \lambda_i a_i$ (where $\sum_{i \in I} \lambda_i = 1$) is the smallest affine subspace containing $(a_i)_{i \in I}$.*

Given a nonempty subset S of E, the smallest affine subspace of E generated by S is often denoted as $\langle S \rangle$. For example, a line specified by two distinct points a and b is denoted as $\langle a, b \rangle$, or even (a, b), and similarly for planes, etc.

Remarks: Since it can be shown that the barycenter of n weighted points can be obtained by repeated computations of barycenters of two weighted points, a nonempty subset V of E is an affine subspace iff for every two points $a, b \in V$, the set V contains all barycentric combinations of a and b.

If V contains at least two points, V is an affine subspace iff for any two distinct points $a, b \in V$, the set V contains the line determined by a and b, that is, the set of all points $(1 - \lambda)a + \lambda b$, $\lambda \in \mathbb{R}$.

2.4 Affine Independence and Affine Frames

Corresponding to the notion of linear independence in vector spaces, we have the notion of affine independence.

Given a family $(a_i)_{i\in I}$ of points in an affine space E, we will reduce the notion of (affine) independence of these points to the (linear) independence of the families $(\mathbf{a_i a_j})_{j\in(I-\{i\})}$ of vectors obtained by chosing any a_i as an origin.

First, the following lemma shows that it sufficient to consider only one of these families.

Lemma 2.4.1 *Given an affine space $\langle E, \overrightarrow{E}, +\rangle$, let $(a_i)_{i\in I}$ be a family of points in E. If the family $(\mathbf{a_i a_j})_{j\in(I-\{i\})}$ is linearly independent for some $i \in I$, then $(\mathbf{a_i a_j})_{j\in(I-\{i\})}$ is linearly independent for every $i \in I$.*

Definition 2.4.2 Given an affine space $\langle E, \overrightarrow{E}, + \rangle$, a family $(a_i)_{i \in I}$ of points in E is *affinely independent* if the family $(\mathbf{a_i a_j})_{j \in (I - \{i\})}$ is linearly independent for some $i \in I$.

Definition 2.4.2 is reasonable, since by Lemma 2.4.1, the independence of the family $(\mathbf{a_i a_j})_{j \in (I - \{i\})}$ does not depend on the choice of a_i.

A crucial property of linearly independent vectors $(u_1, \ldots, u_m)$ is that if a vector v is a linear combination

$$v = \sum_{i=1}^{m} \lambda_i u_i$$

of the u_i, then the λ_i are unique. A similar result holds for affinely independent points.

Lemma 2.4.3 *Given an affine space* $\langle E, \overrightarrow{E}, + \rangle$, *let* $(a_0, \ldots, a_m)$ *be a family of* $m+1$ *points in* E. *Let* $x \in E$, *and assume that* $x = \sum_{i=0}^{m} \lambda_i a_i$, *where* $\sum_{i=0}^{m} \lambda_i = 1$. *Then, the family* $(\lambda_0, \ldots, \lambda_m)$ *such that* $x = \sum_{i=0}^{m} \lambda_i a_i$ *is unique iff the family* $(\mathbf{a_0a_1}, \ldots, \mathbf{a_0a_m})$ *is linearly independent.*

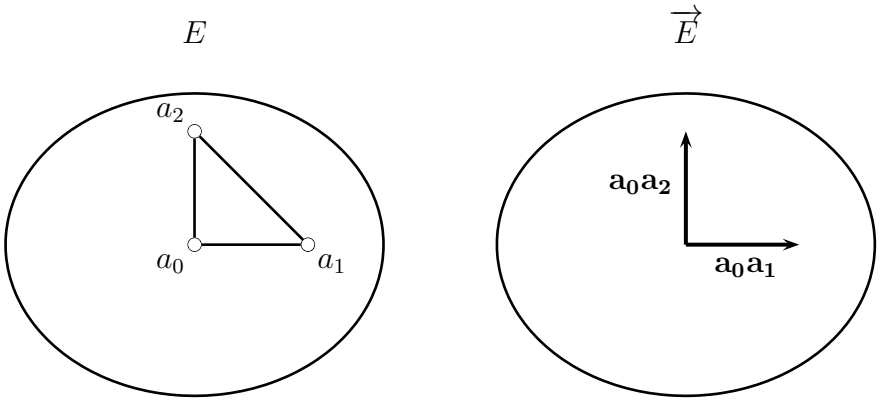

Figure 2.7: Affine independence and linear independence

Lemma 2.4.3 suggests the notion of affine frame.

Let $\langle E, \overrightarrow{E}, + \rangle$ be a nonempty affine space, and let $(a_0, \ldots, a_m)$ be a family of $m+1$ points in E. The family $(a_0, \ldots, a_m)$ determines the family of m vectors $(\mathbf{a_0a_1}, \ldots, \mathbf{a_0a_m})$ in $\overrightarrow{E}$.

Conversely, given a point a_0 in E and a family of m vectors $(u_1, \ldots, u_m)$ in $\overrightarrow{E}$, we obtain the family of $m+1$ points $(a_0, \ldots, a_m)$ in E, where $a_i = a_0 + u_i$, $1 \leq i \leq m$.

Thus, for any $m \geq 1$, it is equivalent to consider a family of $m+1$ points $(a_0, \ldots, a_m)$ in E, and a pair $(a_0, (u_1, \ldots, u_m))$, where the u_i are vectors in $\overrightarrow{E}$.

When $(\mathbf{a_0a_1}, \ldots, \mathbf{a_0a_m})$ is a basis of $\overrightarrow{E}$, then, for every $x \in E$, since $x = a_0 + \mathbf{a_0x}$, there is a unique family $(x_1, \ldots, x_m)$ of scalars, such that

$$x = a_0 + x_1\mathbf{a_0a_1} + \cdots + x_m\mathbf{a_0a_m}.$$

The scalars $(x_1, \ldots, x_m)$ are coordinates with respect to $(a_0, (\mathbf{a_0a_1}, \ldots, \mathbf{a_0a_m}))$. Since

$$x = a_0 + \sum_{i=1}^{m} x_i\mathbf{a_0a_i} \quad \text{iff} \quad x = (1 - \sum_{i=1}^{m} x_i)a_0 + \sum_{i=1}^{m} x_i a_i,$$

$x \in E$ can also be expressed uniquely as

$$x = \sum_{i=0}^{m} \lambda_i a_i$$

with $\sum_{i=0}^{m} \lambda_i = 1$, and where $\lambda_0 = 1 - \sum_{i=1}^{m} x_i$, and $\lambda_i = x_i$ for $1 \leq i \leq m$.

The scalars $(\lambda_0, \ldots, \lambda_m)$ are also certain kinds of coordinates with respect to $(a_0, \ldots, a_m)$.

Definition 2.4.4 Given an affine space $\langle E, \overrightarrow{E}, + \rangle$, an *affine frame with origin* a_0 is a family $(a_0, \ldots, a_m)$ of $m+1$ points in E such that $(\mathbf{a_0a_1}, \ldots, \mathbf{a_0a_m})$ is a basis of $\overrightarrow{E}$. The pair $(a_0, (\mathbf{a_0a_1}, \ldots, \mathbf{a_0a_m}))$ is also called an *affine frame with origin* a_0.

Then, every $x \in E$ can be expressed as

$$x = a_0 + x_1\mathbf{a_0a_1} + \cdots + x_m\mathbf{a_0a_m}$$

for a unique family $(x_1, \ldots, x_m)$ of scalars, called the *coordinates of x w.r.t. the affine frame* $(a_0, (\mathbf{a_0a_1}, \ldots, \mathbf{a_0a_m}))$.

Furthermore, every $x \in E$ can be written as

$$x = \lambda_0 a_0 + \cdots + \lambda_m a_m$$

for some unique family $(\lambda_0, \ldots, \lambda_m)$ of scalars such that $\lambda_0 + \cdots + \lambda_m = 1$ called the *barycentric coordinates of x with respect to the affine frame* $(a_0, \ldots, a_m)$.

The coordinates $(x_1, \ldots, x_m)$ and the barycentric coordinates $(\lambda_0, \ldots, \lambda_m)$ are related by the equations $\lambda_0 = 1 - \sum_{i=1}^{m} x_i$ and $\lambda_i = x_i$, for $1 \leq i \leq m$.

An affine frame is called an *affine basis* by some authors. The figure below shows affine frames and their convex hulls for $|I| = 0, 1, 2, 3$.

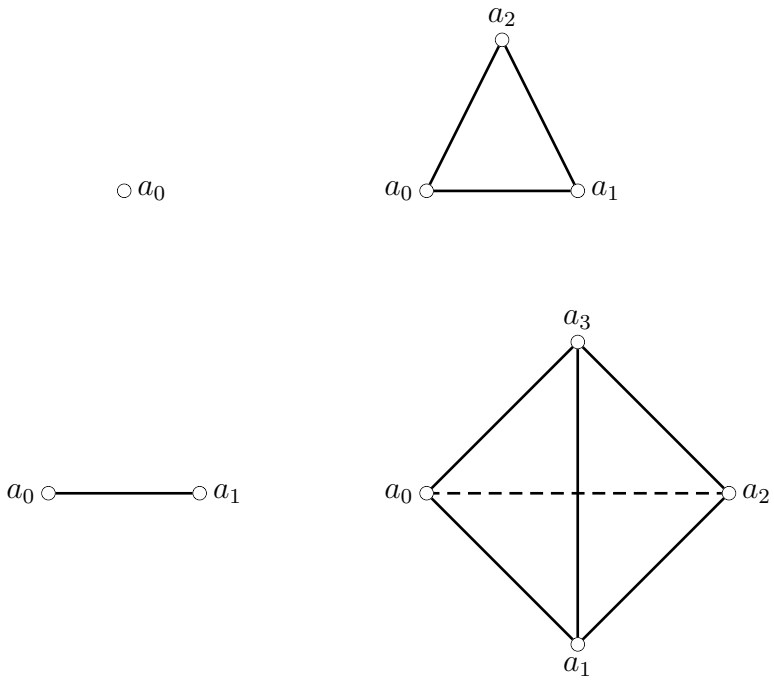

Figure 2.8: Examples of affine frames and their convex hulls.

A family of two points (a, b) in E is affinely independent iff $\mathbf{ab} \neq 0$, iff $a \neq b$. If $a \neq b$, the affine subspace generated by a and b is the set of all points $(1-\lambda)a+\lambda b$, which is the unique line passing through a and b.

A family of three points (a, b, c) in E is affinely independent iff $\mathbf{ab}$ and $\mathbf{ac}$ are linearly independent, which means that a, b, and c are not on a same line (they are not collinear). In this case, the affine subspace generated by (a, b, c) is the set of all points $(1 - \lambda - \mu)a + \lambda b + \mu c$, which is the unique plane containing a, b, and c.

A family of four points (a, b, c, d) in E is affinely independent iff $\mathbf{ab}$, $\mathbf{ac}$, and $\mathbf{ad}$ are linearly independent, which means that a, b, c, and d are not in a same plane (they are not coplanar). In this case, a, b, c, and d, are the vertices of a tetrahedron.

Given $n+1$ affinely independent points $(a_0, \ldots, a_n)$ in E, we can consider the set of points $\lambda_0 a_0 + \cdots + \lambda_n a_n$, where $\lambda_0 + \cdots + \lambda_n = 1$ and $\lambda_i \geq 0$, $\lambda_i \in \mathbb{R}$. Such affine combinations are called *convex combinations*. This set is called the *convex hull* of $(a_0, \ldots, a_n)$ (or *n-simplex spanned by* $(a_0, \ldots, a_n)$).

When $n = 1$, we get the segment between a_0 and a_1, including a_0 and a_1.

When $n = 2$, we get the interior of the triangle whose vertices are a_0, a_1, a_2, including boundary points (the edges).

When $n = 3$, we get the interior of the tetrahedron whose vertices are a_0, a_1, a_2, a_3, including boundary points (faces and edges).

The set

$$\{a_0+\lambda_1\mathbf{a_0a_1}+\cdots+\lambda_n\mathbf{a_0a_n} \mid \text{where } 0 \leq \lambda_i \leq 1 \ (\lambda_i \in \mathbb{R})\},$$

is called the *parallelotope spanned by* $(a_0, \ldots, a_n)$. When E has dimension 2, a parallelotope is also called a *parallelogram*, and when E has dimension 3, a *parallelepiped*.

A parallelotope is shown in figure 2.9: it consists of the points inside of the parallelogram (a_0, a_1, a_2, d), including its boundary.

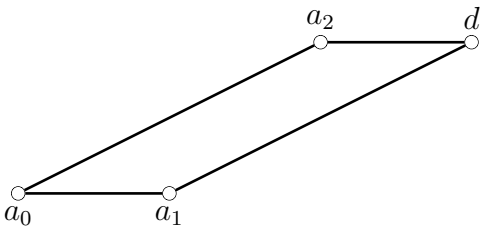

Figure 2.9: A parallelotope

More generally, we say that a subset V of E is *convex*, if for any two points $a, b \in V$, we have $c \in V$ for every point $c = (1 - \lambda)a + \lambda b$, with $0 \leq \lambda \leq 1$ $(\lambda \in \mathbb{R})$.

2.5 Affine Maps

Corresponding to linear maps, we have the notion of an affine map.

Definition 2.5.1 Given two affine spaces $\langle E, \overrightarrow{E}, + \rangle$ and $\langle E', \overrightarrow{E'}, +' \rangle$, a function $f: E \to E'$ is an *affine map* iff for every family $(a_i)_{i \in I}$ of points in E, for every family $(\lambda_i)_{i \in I}$ of scalars such that $\sum_{i \in I} \lambda_i = 1$, we have

$$f(\sum_{i \in I} \lambda_i a_i) = \sum_{i \in I} \lambda_i f(a_i).$$

In other words, f preserves affine combinations (barycenters).

Affine maps can be obtained from linear maps as follows. For simplicity of notation, the same symbol $+$ is used for both affine spaces (instead of using both $+$ and $+'$).

Given any point $a \in E$, any point $b \in E'$, and any linear map $h\colon \overrightarrow{E} \to \overrightarrow{E'}$, the map $f\colon E \to E'$ defined such that

$$f(a + v) = b + h(v)$$

is an affine map.

As a more concrete example, the map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

defines an affine map in $\mathbb{A}^2$. It is a "shear" followed by a translation. The effect of this shear on the square (a, b, c, d) is shown in figure 2.10. The image of the square (a, b, c, d) is the parallelogram (a', b', c', d').

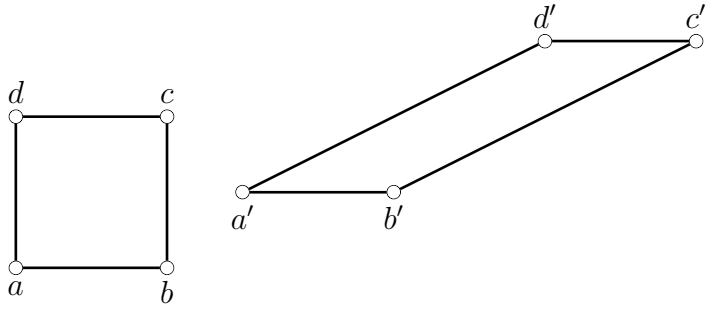

Figure 2.10: The effect of a shear

Let us consider one more example.

The map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \end{pmatrix}$$

is an affine map.

Since we can write

$$\begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix} = \sqrt{2} \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix},$$

this affine map is the composition of a shear, followed by a rotation of angle $\pi/4$, followed by a magnification of ratio $\sqrt{2}$, followed by a translation. The effect of this map on the square (a, b, c, d) is shown in figure 2.11. The image of the square (a, b, c, d) is the parallelogram (a', b', c', d').

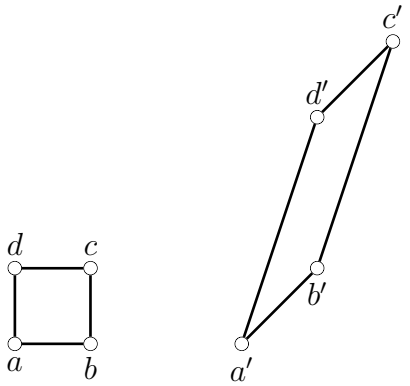

Figure 2.11: The effect of an affine map

The following lemma shows the converse of what we just showed. Every affine map is determined by the image of any point and a linear map.

Lemma 2.5.2 *Given an affine map $f: E \to E'$, there is a unique linear map $\overrightarrow{f}: \overrightarrow{E} \to \overrightarrow{E'}$, such that*

$$f(a + v) = f(a) + \overrightarrow{f}(v),$$

for every $a \in E$ and every $v \in \overrightarrow{E}$.

The unique linear map $\overrightarrow{f}: \overrightarrow{E} \to \overrightarrow{E'}$ given by lemma 2.5.2 is the *linear map associated with the affine map* f.

Note that the condition

$$f(a+v) = f(a) + \overrightarrow{f}(v),$$

for every $a \in E$ and every $v \in \overrightarrow{E}$, can be stated equivalently as

$$f(x) = f(a) + \overrightarrow{f}(\mathbf{ax}), \quad \text{or} \quad \mathbf{f(a)f(x)} = \overrightarrow{f}(\mathbf{ax}),$$

for all $a, x \in E$.

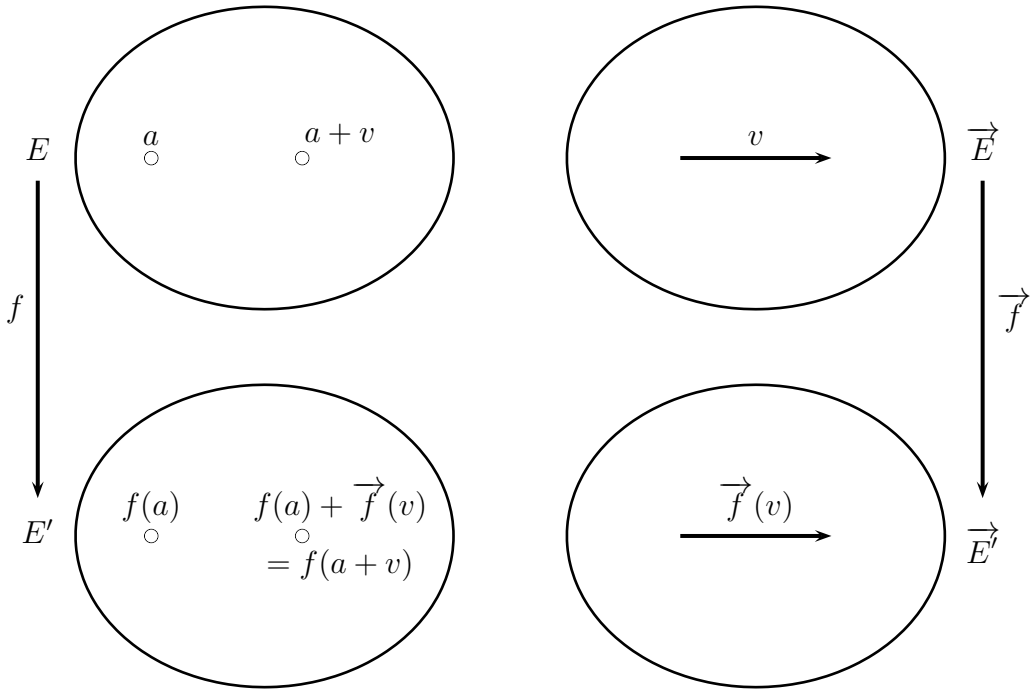

Figure 2.12: An affine map f and its associated linear map $\overrightarrow{f}$

Lemma 2.5.2 shows that for any affine map $f\colon E \to E'$, there are points $a \in E$, $b \in E'$, and a unique linear map $\overrightarrow{f}\colon \overrightarrow{E} \to \overrightarrow{E'}$, such that

$$f(a + v) = b + \overrightarrow{f}(v),$$

for all $v \in \overrightarrow{E}$ (just let $b = f(a)$, for any $a \in E$).

Since an affine map preserves barycenters, and since an affine subspace V is closed under barycentric combinations, the image $f(V)$ of V is an affine subspace in E'.

So, for example, the image of a line is a point or a line, the image of a plane is either a point, a line, or a plane.

Affine maps for which $\overrightarrow{f}$ is the identity map are called *translations*. Indeed, if $\overrightarrow{f} = \mathrm{id}$, it is easy to show that for any two points $a, x \in E$,

$$f(x) = x + \mathbf{af}(\mathbf{a}).$$

It is easily verified that the composition of two affine maps is an affine map.

Also, given affine maps $f\colon E \to E'$ and $g\colon E' \to E''$, we have

$$g(f(a+v)) = g(f(a) + \overrightarrow{f}(v)) = g(f(a)) + \overrightarrow{g}(\overrightarrow{f}(v)),$$

which shows that $\overrightarrow{(g \circ f)} = \overrightarrow{g} \circ \overrightarrow{f}$.

It is easy to show that an affine map $f\colon E \to E'$ is injective iff $\overrightarrow{f} : \overrightarrow{E} \to \overrightarrow{E'}$ is injective, and that $f\colon E \to E'$ is surjective iff $\overrightarrow{f} : \overrightarrow{E} \to \overrightarrow{E'}$ is surjective.

An affine map $f\colon E \to E'$ is constant iff $\overrightarrow{f} : \overrightarrow{E} \to \overrightarrow{E'}$ is the null (constant) linear map equal to 0 for all $v \in \overrightarrow{E}$.

If E is an affine space of dimension m, and $(a_0, a_1, \ldots, a_m)$ is an affine frame for E, for any other affine space F, for any sequence $(b_0, b_1, \ldots, b_m)$ of $m+1$ points in F, there is a unique affine map $f: E \to F$ such that $f(a_i) = b_i$, for $0 \leq i \leq m$.

The following diagram illustrates the above result when $m = 2$.

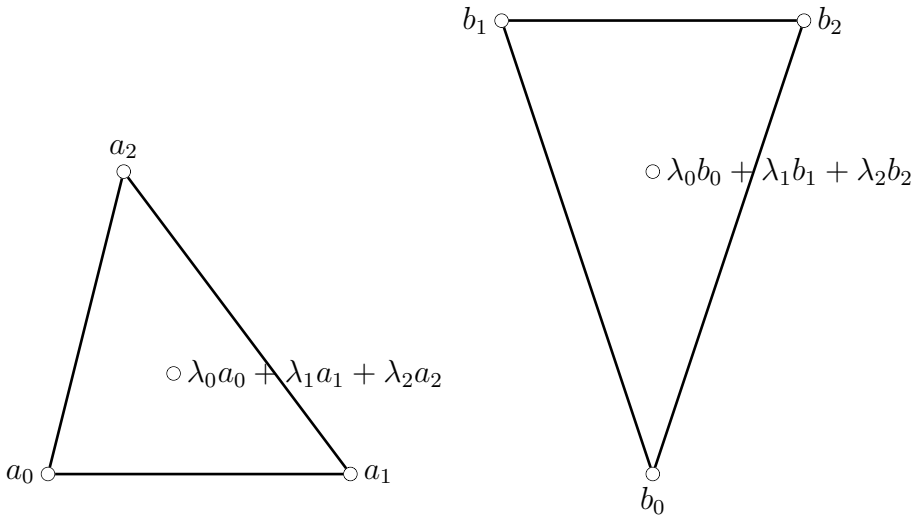

Figure 2.13: An affine map mapping a_0, a_1, a_2 to b_0, b_1, b_2.

Using affine frames, affine maps can be represented in terms of matrices.

We explain how an affine map $f\colon E \to E$ is represented with respect to a frame $(a_0, \ldots, a_n)$ in E.

Since

$$f(a_0 + x) = f(a_0) + \overrightarrow{f}(x)$$

for all $x \in \overrightarrow{E}$, we have

$$\mathbf{a_0 f}(\mathbf{a_0} + \mathbf{x}) = \mathbf{a_0 f}(\mathbf{a_0}) + \overrightarrow{f}(x).$$

Since x, $\mathbf{a_0 f}(\mathbf{a_0})$, and $\mathbf{a_0 f}(\mathbf{a_0} + \mathbf{x})$, can be expressed as

$$\begin{aligned} x &= x_1 \mathbf{a_0 a_1} + \cdots + x_n \mathbf{a_0 a_n}, \\ \mathbf{a_0 f}(\mathbf{a_0}) &= b_1 \mathbf{a_0 a_1} + \cdots + b_n \mathbf{a_0 a_n}, \\ \mathbf{a_0 f}(\mathbf{a_0} + \mathbf{x}) &= y_1 \mathbf{a_0 a_1} + \cdots + y_n \mathbf{a_0 a_n}, \end{aligned}$$

if $A = (a_{i\,j})$ is the $n \times n$-matrix of the linear map $\overrightarrow{f}$ over the basis $(\mathbf{a_0 a_1}, \ldots, \mathbf{a_0 a_n})$, letting x, y, and b denote the column vectors of components $(x_1, \ldots, x_n)$, $(y_1, \ldots, y_n)$, and $(b_1, \ldots, b_n)$,

$$\mathbf{a_0 f}(\mathbf{a_0} + \mathbf{x}) = \mathbf{a_0 f}(\mathbf{a_0}) + \overrightarrow{f}(x)$$

is equivalent to

$$y = Ax + b.$$

Note that $b \neq 0$ unless $f(a_0) = a_0$. Thus, f is generally not a linear transformation, unless it has a *fixed point*, i.e., there is a point a_0 such that $f(a_0) = a_0$. The vector b is the "translation part" of the affine map.

Affine maps do not always have a fixed point. Obviously, nonnull translations have no fixed point. A less trivial example is given by the affine map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

This map is a reflection about the x-axis followed by a translation along the x-axis. The affine map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 1 & -\sqrt{3} \\ \frac{\sqrt{3}}{4} & \frac{1}{4} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

can also be written as

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

which shows that it is the composition of a rotation of angle $\pi/3$, followed by a stretch (by a factor of 2 along the

x-axis, and by a factor of $1/2$ along the y-axis), followed by a translation. It is easy to show that this affine map has a unique fixed point.

On the other hand, the affine map

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} \frac{8}{5} & -\frac{6}{5} \\ \frac{3}{10} & \frac{2}{5} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

has no fixed point, even though

$$\begin{pmatrix} \frac{8}{5} & -\frac{6}{5} \\ \frac{3}{10} & \frac{2}{5} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{4}{5} & -\frac{3}{5} \\ \frac{3}{5} & \frac{4}{5} \end{pmatrix},$$

and the second matrix is a rotation of angle θ such that $\cos\theta = \frac{4}{5}$ and $\sin\theta = \frac{3}{5}$.

There is a useful trick to convert the equation $y = Ax+b$ into what looks like a linear equation. The trick is to consider an $(n+1) \times (n+1)$-matrix. We add 1 as the $(n+1)$th component to the vectors x, y, and b, and form the $(n+1) \times (n+1)$-matrix

$$\begin{pmatrix} A & b \\ 0 & 1 \end{pmatrix}$$

so that $y = Ax + b$ is equivalent to

$$\begin{pmatrix} y \\ 1 \end{pmatrix} = \begin{pmatrix} A & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}.$$

This trick is very useful in kinematics and dynamics, where A is a rotation matrix. Such affine maps are called *rigid motions*.

If $f: E \to E'$ is a bijective affine map, given any three collinear points a, b, c in E, with $a \neq b$, where say, $c = (1 - \lambda)a + \lambda b$, since f preserves barycenters, we have $f(c) = (1-\lambda)f(a)+\lambda f(b)$, which shows that $f(a), f(b), f(c)$ are collinear in E'.

There is a converse to this property, which is simpler to state when the ground field is $K = \mathbb{R}$.

The converse states that given any bijective function $f: E \to E'$ between two real affine spaces of the same dimension $n \geq 2$, if f maps any three collinear points to collinear points, then f is affine. The proof is rather long (see Berger [**?**] or Samuel [**?**]).

Given three collinear points where a, b, c, where $a \neq c$, we have $b = (1 - \beta)a + \beta c$ for some unique β, and we define the *ratio of the sequence* $a, b,$ c, as

$$\text{ratio}(a, b, c) = \frac{\beta}{(1 - \beta)} = \frac{\mathbf{ab}}{\mathbf{bc}},$$

provided that $\beta \neq 1$, i.e. that $b \neq c$. When $b = c$, we agree that $\text{ratio}(a, b, c) = \infty$.

We warn our readers that other authors define the ratio of a, b, c as $-\text{ratio}(a, b, c) = \frac{\mathbf{ba}}{\mathbf{bc}}$. Since affine maps preserves barycenters, it is clear that affine maps preserve the ratio of three points.

2.6 Affine Groups

We now take a quick look at the bijective affine maps.

Given an affine space E, the set of affine bijections $f: E \to E$ is clearly a group, called the *affine group of* E, and denoted as $\mathrm{GA}(E)$.

Recall that the group of bijective linear maps of the vector space $\overrightarrow{E}$ is denoted as $\mathrm{GL}(\overrightarrow{E})$. Then, the map $f \mapsto \overrightarrow{f}$ defines a group homomorphism $L: \mathrm{GA}(E) \to \mathrm{GL}(\overrightarrow{E})$. The kernel of this map is the set of translations on E.

The subset of all linear maps of the form $\lambda \mathrm{id}_{\overrightarrow{E}}$, where $\lambda \in \mathbb{R} - \{0\}$, is a subgroup of $\mathrm{GL}(\overrightarrow{E})$, and is denoted as $\mathbb{R}^* \mathrm{id}_{\overrightarrow{E}}$.

The subgroup $\mathrm{DIL}(E) = L^{-1}(\mathbb{R}^* \mathrm{id}_{\overrightarrow{E}})$ of $\mathrm{GA}(E)$ is particularly interesting. It turns out that it is the disjoint union of the translations and of the dilatations of ratio $\lambda \neq 1$.

The elements of $\mathrm{DIL}(E)$ are called *affine dilatations (or dilations)*.

Given any point $a \in E$, and any scalar $\lambda \in \mathbb{R}$, a *dilatation (or central dilatation, or magnification, or homothety) of center a and ratio λ*, is a map $H_{a,\lambda}$ defined such that

$$H_{a,\lambda}(x) = a + \lambda \mathbf{ax},$$

for every $x \in E$.

Observe that $H_{a,\lambda}(a) = a$, and when $\lambda \neq 0$ and $x \neq a$, $H_{a,\lambda}(x)$ is on the line defined by a and x, and is obtained by "scaling" $\mathbf{ax}$ by λ. When $\lambda = 1$, $H_{a,1}$ is the identity.

Note that $\overrightarrow{H_{a,\lambda}} = \lambda \, \mathrm{id}_{\overrightarrow{E}}$. When $\lambda \neq 0$, it is clear that $H_{a,\lambda}$ is an affine bijection.

It is immediately verified that

$$H_{a,\lambda} \circ H_{a,\mu} = H_{a,\lambda\mu}.$$

We have the following useful result.

Lemma 2.6.1 *Given any affine space E, for any affine bijection $f \in GA(E)$, if $\overrightarrow{f} = \lambda \, \mathrm{id}_{\overrightarrow{E}}$, for some $\lambda \in \mathbb{R}^*$ with $\lambda \neq 1$, then there is a unique point $c \in E$ such that $f = H_{c,\lambda}$.*

Clearly, if $\overrightarrow{f} = \mathrm{id}_{\overrightarrow{E}}$, the affine map f is a translation.

Thus, the group of affine dilatations $\mathrm{DIL}(E)$ is the disjoint union of the translations and of the dilatations of ratio $\lambda \neq 0, 1$. Affine dilatations can be given a purely geometric characterization.

2.7 Affine Geometry, a Glimpse

In this section, we state and prove three fundamental results of affine geometry.

Roughly speaking, affine geometry is the study of properties invariant under affine bijections. We now prove one of the oldest and most basic results of affine geometry, the theorem of Thalés.

Lemma 2.7.1 *Given any affine space E, if H_1, H_2, H_3 are any three distinct parallel hyperplanes, and A and B are any two lines not parallel to H_i, letting $a_i = H_i \cap A$ and $b_i = H_i \cap B$, then the following ratios are equal:*

$$\frac{\mathbf{a_1a_3}}{\mathbf{a_1a_2}} = \frac{\mathbf{b_1b_3}}{\mathbf{b_1b_2}} = \rho.$$

Conversely, for any point d on the line A, if $\frac{\mathbf{a_1d}}{\mathbf{a_1a_2}} = \rho$, then $d = a_3$.

The diagram below illustrates the theorem of Thalés.

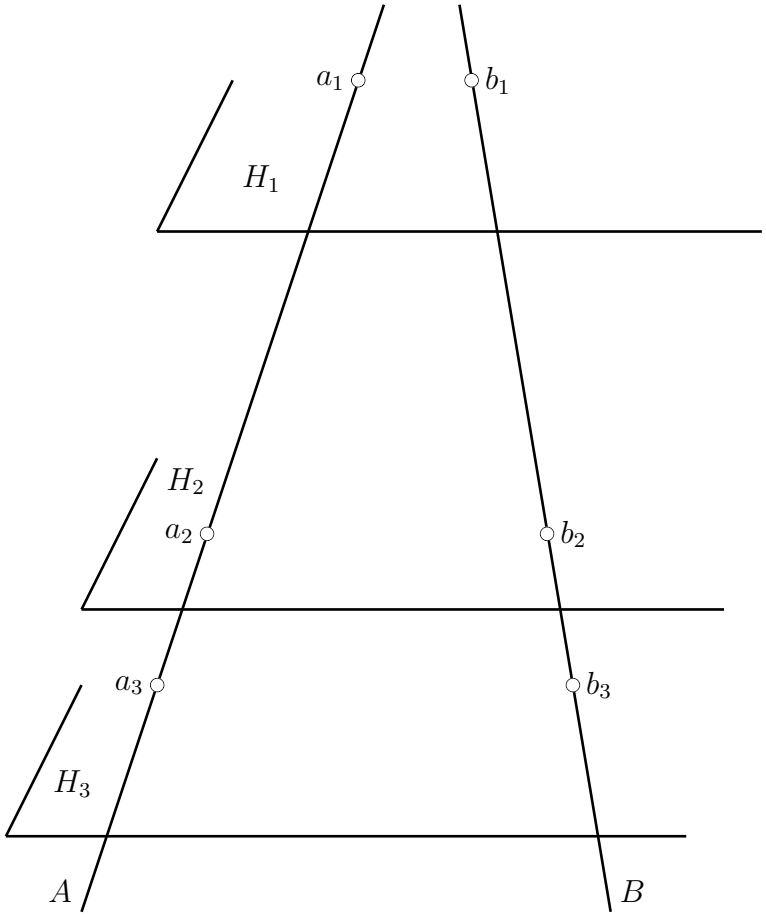

Figure 2.14: The theorem of Thalés

Lemma 2.7.2 *Given any affine space E, given any two distinct points $a, b \in E$, for any affine dilatation f different from the identity, if $a' = f(a)$, $D = \langle a, b\rangle$ is the line passing through a and b, and D' is the line parallel to D and passing through a', the following are equivalent:*

(i) $b' = f(b)$;

(ii) If f is a translation, then b' is the intersection of D' with the line parallel to $\langle a, a'\rangle$ passing through b;

If f is a dilatation of center c, then $b' = D' \cap \langle c, b\rangle$.

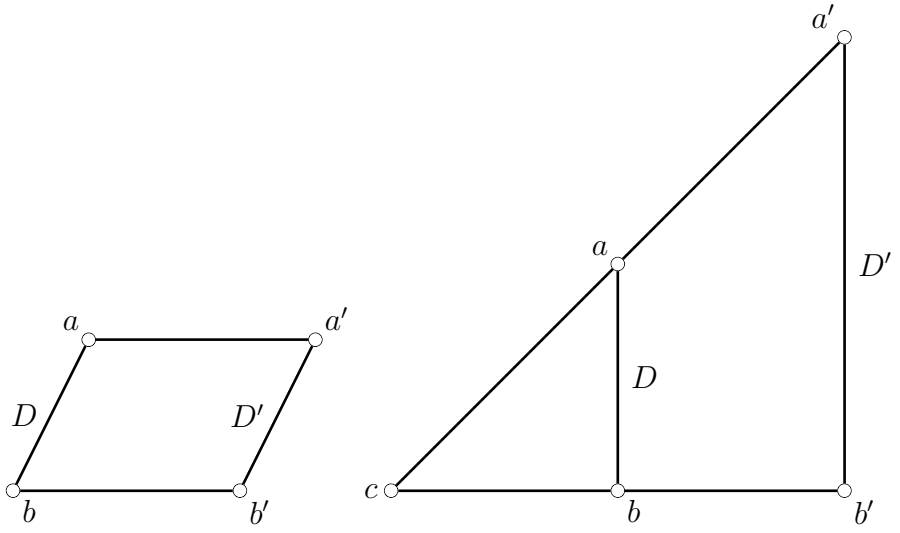

Figure 2.15: Affine Dilatations

The first case is the parallelogram law, and the second case follows easily from Thalés' theorem.

We are now ready to prove two classical results of affine geometry, Pappus' theorem and Desargues' theorem. Actually, these results are theorem of projective geometry, and we are stating affine versions of these important results. There are stronger versions which are best proved using projective geometry.

There is a converse to Pappus' theorem, which yields a fancier version of Pappus' theorem, but it is easier to prove it using projective geometry.

Lemma 2.7.3 *Given any affine plane E, given any two distinct lines D and D', for any distinct points a, b, c on D, and a', b', c' on D', if a, b, c, a', b', c' are distinct from the intersection of D and D' (if D and D' intersect) and if the lines $\langle a, b' \rangle$ and $\langle a', b \rangle$ are parallel, and the lines $\langle b, c' \rangle$ and $\langle b', c \rangle$ are parallel, then the lines $\langle a, c' \rangle$ and $\langle a', c \rangle$ are parallel.*

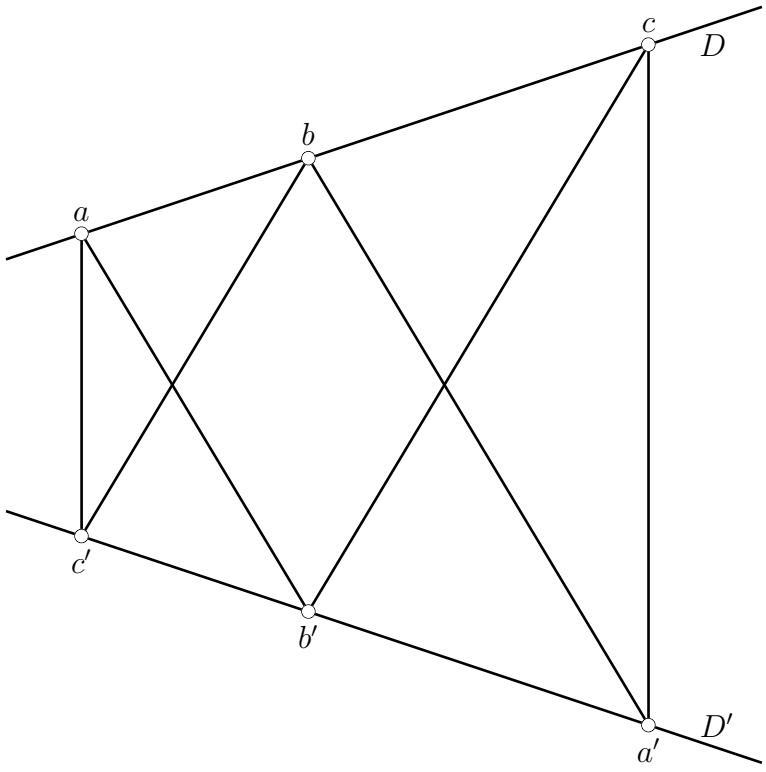

Figure 2.16: Pappus' theorem (affine version)

We now prove an affine version of Desargues' theorem.

Lemma 2.7.4 *Given any affine space E, given any two triangles (a, b, c) and (a', b', c'), where a, b, c, a', b', c' are all distinct, if $\langle a, b\rangle$ and $\langle a', b'\rangle$ are parallel and $\langle b, c\rangle$ and $\langle b', c'\rangle$ are parallel, then $\langle a, c\rangle$ and $\langle a', c'\rangle$ are parallel iff the lines $\langle a, a'\rangle$, $\langle b, b'\rangle$, and $\langle c, c'\rangle$, are either parallel or concurrent (i.e., intersect in a common point).*

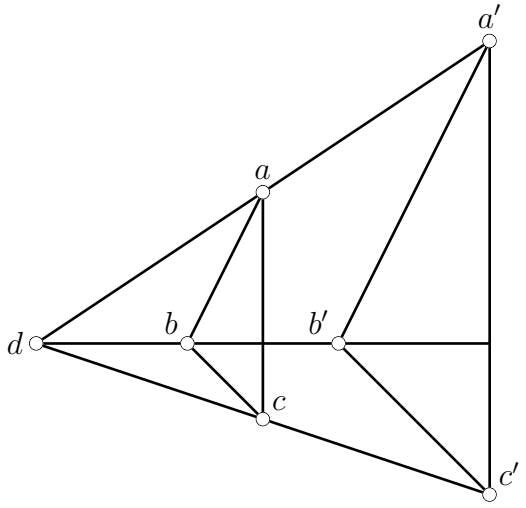

Figure 2.17: Desargues' theorem (affine version)

There is a fancier version of Desargues' theorem, but it is easier to prove it using projective geometry.

Desargues' theorem yields a geometric characterization of the affine dilatations. An affine dilatation f on an affine space E is a bijection that maps every line D to a line $f(D)$ parallel to D.

2.8 Affine Hyperplanes

In section 2.3, we observed that the set L of solutions of an equation

$$ax + by = c$$

is an affine subspace of $\mathbb{A}^2$ of dimension 1, in fact a line (provided that a and b are not both null).

It would be equally easy to show that the set P of solutions of an equation

$$ax + by + cz = d$$

is an affine subspace of $\mathbb{A}^3$ of dimension 2, in fact a plane (provided that a, b, c are not all null).

More generally, the set H of solutions of an equation

$$\lambda_1 x_1 + \cdots + \lambda_m x_m = \mu$$

is an affine subspace of $\mathbb{A}^m$, and if $\lambda_1, \ldots, \lambda_m$ are not all null, it turns out that it is a subspace of dimension $m-1$ called a *hyperplane*.

We can interpret the equation

$$\lambda_1 x_1 + \cdots + \lambda_m x_m = \mu$$

in terms of the map $f\colon \mathbb{R}^m \to \mathbb{R}$ defined such that

$$f(x_1, \ldots, x_m) = \lambda_1 x_1 + \cdots + \lambda_m x_m - \mu$$

for all $(x_1, \ldots, x_m) \in \mathbb{R}^m$.

It is immediately verified that this map is affine, and the set H of solutions of the equation

$$\lambda_1 x_1 + \cdots + \lambda_m x_m = \mu$$

is the *null set, or kernel,* of the affine map $f\colon \mathbb{A}^m \to \mathbb{R}$, in the sense that

$$H = f^{-1}(0) = \{x \in \mathbb{A}^m \mid f(x) = 0\},$$

where $x = (x_1, \ldots, x_m)$.

Thus, it is interesting to consider *affine forms*, which are just affine maps $f\colon E \to \mathbb{R}$ from an affine space to $\mathbb{R}$.

Unlike linear forms f^*, for which $\operatorname{Ker} f^*$ is never empty (since it always contains the vector 0), it is possible that $f^{-1}(0) = \emptyset$, for an affine form f.

Recall the characterization of hyperplanes in terms of linear forms.

Given a vector space E over a field K, a linear map $f\colon E \to K$ is called a *linear form*. The set of all linear forms $f\colon E \to K$ is a vector space called the *dual space of* E, and denoted as E^*.

Hyperplanes are precisely the Kernels of nonnull linear forms.

Lemma 2.8.1 *Let E be a vector space. The following properties hold:*

(a) Given any nonnull linear form $f \in E^*$, its kernel $H = \operatorname{Ker} f$ is a hyperplane.

(b) For any hyperplane H in E, there is a (nonnull) linear form $f \in E^*$ such that $H = \operatorname{Ker} f$.

(c) Given any hyperplane H in E and any (nonnull) linear form $f \in E^*$ such that $H = \operatorname{Ker} f$, for every linear form $g \in E^*$, $H = \operatorname{Ker} g$ iff $g = \lambda f$ for some $\lambda \neq 0$ in K.

Going back to an affine space E, given an affine map $f\colon E \to \mathbb{R}$, we also denote $f^{-1}(0)$ as $\operatorname{Ker} f$, and we call it the *kernel* of f. Recall that an (affine) hyperplane is an affine subspace of codimension 1.

Affine hyperplanes are precisely the Kernels of nonconstant affine forms.

Lemma 2.8.2 *Let E be an affine space. The following properties hold:*

(a) Given any nonconstant affine form $f: E \to \mathbb{R}$, its kernel $H = \operatorname{Ker} f$ is a hyperplane.

(b) For any hyperplane H in E, there is a nonconstant affine form $f: E \to \mathbb{R}$ such that $H = \operatorname{Ker} f$. For any other affine form $g: E \to \mathbb{R}$ such that $H = \operatorname{Ker} g$, there is some $\lambda \in \mathbb{R}$ such that $g = \lambda f$ (with $\lambda \neq 0$).

(c) Given any hyperplane H in E and any (nonconstant) affine form $f: E \to \mathbb{R}$ such that $H = \operatorname{Ker} f$, every hyperplane H' parallel to H is defined by a nonconstant affine form g such that $g(a) = f(a) - \lambda$, for all $a \in E$, for some $\lambda \in \mathbb{R}$.

2.9 Intersection of Affine Spaces

In this section, we take a closer look at the intersection of affine subspaces.

First, we need a result of linear algebra.

Lemma 2.9.1 *Given a vector space E and any two subspaces M and N, we have the Grassmann relation:*

$$\dim(M) + \dim(N) = \dim(M + N) + \dim(M \cap N).$$

We now prove a simple lemma about the intersection of affine subspaces.

Lemma 2.9.2 *Given any affine space E, for any two nonempty affine subspaces M and N, the following facts hold:*

(1) $M \cap N \neq \emptyset$ iff $\mathbf{ab} \in \overrightarrow{M} + \overrightarrow{N}$ for some $a \in M$ and some $b \in N$.

(2) $M \cap N$ consists of a single point iff $\mathbf{ab} \in \overrightarrow{M} + \overrightarrow{N}$ for some $a \in M$ and some $b \in N$, and $\overrightarrow{M} \cap \overrightarrow{N} = \{0\}$.

(3) If S is the least affine subspace containing M and N, then $\overrightarrow{S} = \overrightarrow{M} + \overrightarrow{N} + K\mathbf{ab}$ (the vector space $\overrightarrow{E}$ is defined over the field K).

Remarks: (1) The proof of Lemma 2.9.2 shows that if $M \cap N \neq \emptyset$ then $\mathbf{ab} \in \overrightarrow{M} + \overrightarrow{N}$ for all $a \in M$ and all $b \in N$.

(2) Lemma 2.9.2 (2) implies that for any two nonempty affine subspaces M and N, if $\overrightarrow{E} = \overrightarrow{M} \oplus \overrightarrow{N}$, then $M \cap N$ consists of a single point.

Lemma 2.9.3 *Given an affine space E and any two nonempty affine subspaces M and N, if S is the least affine subspace containing M and N, then the following properties hold:*

(1) If $M \cap N = \emptyset$, then

$$\dim(M) + \dim(N) < \dim(E) + \dim(\overrightarrow{M} + \overrightarrow{N}),$$

and

$$\dim(S) = \dim(M) + \dim(N) + 1 - \dim(\overrightarrow{M} \cap \overrightarrow{N}).$$

(2) If $M \cap N \neq \emptyset$, then

$$\dim(S) = \dim(M) + \dim(N) - \dim(M \cap N).$$

Chapter 3

Properties of Convex Sets: A Glimpse

3.1 Convex Sets

Convex sets play a very important role in geometry. In this chapter, we state some of the "classics" of convex affine geometry: Carathéodory's theorem, Radon's theorem, and Helly's theorem.

These theorems share the property that they are easy to state, but they are deep, and their proof, although rather short, requires a lot of creativity.

Given an affine space E, recall that a subset V of E is *convex* if for any two points $a, b \in V$, we have $c \in V$ for every point $c = (1 - \lambda)a + \lambda b$, with $0 \leq \lambda \leq 1$ ($\lambda \in \mathbb{R}$).

The notation $[a, b]$ is often used to denote the line segment between a and b, that is,

$$[a, b] = \{c \in E \mid c = (1 - \lambda)a + \lambda b,\ 0 \leq \lambda \leq 1\},$$

and thus, a set V is convex if $[a, b] \subseteq V$ for any two points $a, b \in V$ ($a = b$ is allowed).

The empty set is trivially convex, every one-point set $\{a\}$ is convex, and the entire affine space E is of course convex.

It is obvious that the intersection of any family (finite or infinite) of convex sets is convex.

Then, given any (nonempty) subset S of E, there is a smallest convex set containing S denoted as $\mathcal{C}(S)$ (or $\mathrm{conv}(S)$) and called the *convex hull of* S (namely, the intersection of all convex sets containing S).

Lemma 3.1.1 *Given an affine space* $\langle E, \overrightarrow{E}, + \rangle$*, for any family* $(a_i)_{i \in I}$ *of points in* E*, the set* V *of convex combinations* $\sum_{i \in I} \lambda_i a_i$ *(where* $\sum_{i \in I} \lambda_i = 1$ *and* $\lambda_i \geq 0$*) is the convex hull of* $(a_i)_{i \in I}$*.*

In view of lemma 3.1.1, it is obvious that any affine subspace of E is convex.

Convex sets also arise in terms of hyperplanes. Given a hyperplane H, if $f\colon E \to \mathbb{R}$ is any nonconstant affine form defining H (i.e., $H = \operatorname{Ker} f$), we can define the two subsets

$$H_+(f) = \{a \in E \mid f(a) \geq 0\},$$
$$H_-(f) = \{a \in E \mid f(a) \leq 0\},$$

called *(closed) half spaces associated with* f.

Observe that if $\lambda > 0$, then $H_+(\lambda f) = H_+(f)$, but if $\lambda < 0$, then $H_+(\lambda f) = H_-(f)$, and similarly for $H_-(\lambda f)$.

However, the set $\{H_+(f), H_-(f)\}$ only depends on the hyperplane H, and the choice of a specific f defining H amounts to the choice of one of the two half-spaces.

For this reason, we will also say that $H_+(f)$ and $H_-(f)$ are the (closed) half spaces associated with H.

Clearly,

$$H_+(f) \cup H_-(f) = E \quad \text{and} \quad H_+(f) \cap H_-(f) = H.$$

It is immediately verified that $H_+(f)$ and $H_-(f)$ are convex.

Bounded convex sets arising as the intersection of a finite family of half-spaces associated with hyperplanes play a major role in convex geometry and topology (they are called *convex polytopes*).

It is natural to wonder whether lemma 3.1.1 can be sharpened in two directions:

(1) is it possible have a fixed bound on the number of points involved in the convex combinations?

(2) Is it necessary to consider convex combinations of all points, or is it possible to only consider a subset with special properties?

The answer is yes in both cases. In case 1, assuming that the affine space E has dimension m, *Carathéodory's theorem* asserts that it is enough to consider convex combinations of $m+1$ points.

In case 2, the theorem of Krein and Milman asserts that a convex set which is also compact is the convex hull of its extremal points (see Berger [**?**] or Lang [**?**]).

First, we will prove Carathéodory's theorem. The following technical (and dull!) lemma plays a crucial role in the proof.

Lemma 3.1.2 *Given an affine space $\langle E, \overrightarrow{E}, + \rangle$, let $(a_i)_{i \in I}$ be a family of points in E. The family $(a_i)_{i \in I}$ is affinely dependent iff there is a family $(\lambda_i)_{i \in I}$ such that $\lambda_j \neq 0$ for some $j \in I$, $\sum_{i \in I} \lambda_i = 0$, and $\sum_{i \in I} \lambda_i \mathbf{xa_i} = 0$ for every $x \in E$.*

Theorem 3.1.3 *Given any affine space E of dimension m, for any (nonempty) family $S = (a_i)_{i \in L}$ in E, the convex hull $\mathcal{C}(S)$ of S is equal to the set of convex combinations of families of $m+1$ points of S.*

Proof. By lemma 3.1.1,

$$\mathcal{C}(S) = \{\sum_{i \in I} \lambda_i a_i \mid a_i \in S, \sum_{i \in I} \lambda_i = 1, \ \lambda_i \geq 0, \\ I \subseteq L, \ I \text{ finite}\}.$$

We would like to prove that

$$\mathcal{C}(S) = \{\sum_{i \in I} \lambda_i a_i \mid a_i \in S, \sum_{i \in I} \lambda_i = 1, \ \lambda_i \geq 0, \\ I \subseteq L, \ |I| = m+1\}.$$

We proceed by contradiction. If the theorem is false, there is some point $b \in \mathcal{C}(S)$ such that b can be expressed as a convex combination $b = \sum_{i \in I} \lambda_i a_i$, where $I \subseteq L$ is a finite set of cardinality $|I| = q$ with $q \geq m+2$, and b cannot be expressed as any convex combination $b = \sum_{j \in J} \mu_j a_j$ of strictly less than q points in S (with $|J| < q$).

We shall prove that b can be written as a convex combination of $q-1$ of the a_i. Since E has dimension m and $q \geq m+2$, the points $a_1, \ldots, a_q$ must be affinely dependent, and we use lemma 3.1.2. □

If S is a finite (of infinite) set of points in the affine plane $\mathbb{A}^2$, theorem 3.1.3 confirms our intuition that $\mathcal{C}(S)$ is the union of triangles (including interior points) whose vertices belong to S.

Similarly, the convex hull of a set S of points in $\mathbb{A}^3$ is the union of tetrahedra (including interior points) whose vertices belong to S.

We get the feeling that triangulations play a crucial role, which is of course true!

We conlude this short section by stating two other classics of convex geometry. We begin with *Radon's theorem*.

Theorem 3.1.4 *Given any affine space E of dimension m, for every subset X of E, if X has at least $m+2$ points, then there is a partition of X into two nonempty disjoint subsets X_1 and X_2 such that the convex hulls of X_1 and X_2 have a nonempty intersection.*

Finally, we state a version of *Helly's theorem*.

Theorem 3.1.5 *Given any affine space E of dimension m, for every family $\{K_1, \ldots, K_n\}$ of n convex subsets of E, if $n \geq m+2$ and the intersection $\bigcap_{i \in I} K_i$ of any $m+1$ of the K_i is nonempty (where $I \subseteq \{1, \ldots, n\}$, $|I| = m+1$), then $\bigcap_{i=1}^{n} K_i$ is nonempty.*

An amusing corollary of Helly's theorem is the following result. Consider $n \geq 4$ parallel line segments in the affine plane $\mathbb{A}^2$. If every three of these line segments meet a line, then all of these line segments meet a common line.

3.2 Separation Theorems

It seems intuitively rather obvious that if A and B are two nonempty disjoint convex sets in $\mathbb{A}^2$, then there is a line, H, separating them, in the sense that A and B belong to the two (disjoint) open half–planes determined by H.

However, this is not always true! For example, this fails if both A and B are closed and unbounded (find an example).

Nevertheless, the result is true if both A and B are open, or if the notion of separation is weakened a little bit.

The key result, from which most separation results follow, is a geometric version of the *Hahn-Banach theorem*.

In the sequel, we restrict our attention to real affine spaces of finite dimension. Then, if X is an affine space of dimension d, there is an affine bijection f between X and $\mathbb{A}^d$.

Now, $\mathbb{A}^d$ is a topological space, under the usual topology on $\mathbb{R}^d$ (in fact, $\mathbb{A}^d$ is a metric space).

Recall that if $a = (a_1, \ldots, a_d)$ and $b = (b_1, \ldots, b_d)$ are any two points in $\mathbb{A}^d$, their **Euclidean distance**, $d(a, b)$, is given by

$$d(a, b) = \sqrt{(b_1 - a_1)^2 + \cdots + (b_d - a_d)^2},$$

which is also the *norm*, $\|\mathbf{ab}\|$, of the vector $\mathbf{ab}$ and that for any $\epsilon > 0$, the *open ball of center a and radius ϵ*, $B(a, \epsilon)$, is given by

$$B(a, \epsilon) = \{b \in \mathbb{A}^d \mid d(a, b) < \epsilon\}.$$

A subset $U \subseteq \mathbb{A}^d$ is *open* (in the *norm topology*) if either U is empty or for every point, $a \in U$, there is some (small) open ball, $B(a, \epsilon)$, contained in U.

A subset $C \subseteq \mathbb{A}^d$ is *closed* iff $\mathbb{A}^d - C$ is open. For example, the *closed balls*, $\overline{B(a, \epsilon)}$, where

$$\overline{B(a, \epsilon)} = \{b \in \mathbb{A}^d \mid d(a, b) \leq \epsilon\},$$

are closed.

A subset $W \subseteq \mathbb{A}^d$ is *bounded* iff there is some ball (open or closed), B, so that $W \subseteq B$.

A subset $W \subseteq \mathbb{A}^d$ is *compact* iff every family, $\{U_i\}_{i \in I}$, that is an open cover of W (which means that $W = \bigcup_{i \in I}(W \cap U_i)$, with each U_i an open set) possesses a finite subcover (which means that there is a finite subset, $F \subseteq I$, so that $W = \bigcup_{i \in F}(W \cap U_i)$).

In $\mathbb{A}^d$, it can be shown that a subset W is compact iff W is closed and bounded.

Given a function, $f\colon \mathbb{A}^m \to \mathbb{A}^n$, we say that f is *continuous* if $f^{-1}(V)$ is open in $\mathbb{A}^m$ whenever V is open in $\mathbb{A}^n$.

If $f\colon \mathbb{A}^m \to \mathbb{A}^n$ is a continuous function, although it is generally **false** that $f(U)$ is open if $U \subseteq \mathbb{A}^m$ is open, it is easily checked that $f(K)$ is compact if $K \subseteq \mathbb{A}^m$ is compact.

An affine space X of dimension d becomes a topological space if we give it the topology for which the open subsets are of the form $f^{-1}(U)$, where U is any open subset of $\mathbb{A}^d$ and $f\colon X \to \mathbb{A}^d$ is an affine bijection.

Given any subset, A, of a topological space X, the smallest closed set containing A is denoted by $\overline{A}$, and is called the *closure* or *adherence of* A.

A subset, A, of X, is *dense in* X if $\overline{A} = X$.

The largest open set contained in A is denoted by $\overset{\circ}{A}$, and is called the *interior of* A.

The set $\text{Fr}\, A = \overline{A} \cap \overline{X - A}$, is called the *boundary* (or *frontier*) of A. We also denote the boundary of A by ∂A.

In order to prove the Hahn-Banach theorem, we will need two lemmas.

Given any two distinct points $x, y \in X$, we let

$$]x, y[= \{(1 - \lambda)x + \lambda y \in X \mid 0 < \lambda < 1\}.$$

Lemma 3.2.1 *Let S be a nonempty convex set, and let $x \in \overset{\circ}{S}$ and $y \in \overline{S}$. Then, we have $]x, y[\subseteq \overset{\circ}{S}$.*

Corollary 3.2.2 *If S is convex, then $\overset{\circ}{S}$ is also convex and we have $\overset{\circ}{S} = \overline{\overset{\circ}{S}}$. Further, if $\overset{\circ}{S} \neq \emptyset$, then $\overline{S} = \overline{\overset{\circ}{S}}$.*

There is a simple criterion to test whether a convex set has an empty interior, based on the notion of dimension of a convex set.

Definition 3.2.3 The *dimension* of a nonempty convex subset, S, of X, denoted by $\dim S$, is the dimension of the smallest affine subset $\langle S \rangle$ containing S.

Proposition 3.2.4 *A nonempty convex set S has a nonempty interior iff* $\dim S = \dim X$.

Proposition 3.2.4 is false in infinite dimension.

Proposition 3.2.5 *If S is convex, then $\overline{S}$ is also convex.*

One can also easily prove that convexity is preserved under direct image and inverse image by an affine map.

The next lemma, which seems intuitively obvious, is the core of the proof of the Hahn-Banach theorem. This is the case where the affine space has dimension two.

First, we need to define what is a convex cone.

Definition 3.2.6 A convex set, C, is a *convex cone with vertex* x if C is invariant under all central magnifications $H_{x,\lambda}$ of center x and ratio λ, with $\lambda > 0$ (i.e., $H_{x,\lambda}(C) = C$).

Given a convex set, S, and a point $x \notin S$, we can define

$$\mathrm{cone}_x(S) = \bigcup_{\lambda>0} H_{x,\lambda}(S).$$

It is easy to check that this is a convex cone.

Lemma 3.2.7 *Let B be a nonempty open and convex subset of $\mathbb{A}^2$, and let O be a point of $\mathbb{A}^2$ so that $O \notin B$. Then, there is some line, L, through O, so that $L \cap B = \emptyset$.*

Finally, we come to the Hahn-Banach theorem.

Theorem 3.2.8 *(Hahn-Banach theorem, geometric form) Let X be a (finite-dimensional) affine space, A be a nonempty open and convex subset of X and L be an affine subspace of X so that $A \cap L = \emptyset$. Then, there is some hyperplane, H, containing L, that is disjoint from A.*

Proof. The case where $\dim X = 1$ is trivial. Thus, we may assume that $\dim X \geq 2$. We reduce the proof to the case where $\dim X = 2$. □

Remark: The geometric form of the Hahn-Banach theorem also holds when the dimension of X is infinite, but a more sophisticated proof is required (it uses Zorn's lemma).

Theorem 3.2.8 is false if we omit the assumption that A is open. For a counter-example, let $A \subseteq \mathbb{A}^2$ be the union of the half space $y < 0$ with the close segment $[0, 1]$ on the x-axis and let L be the point $(2, 0)$ on the boundary of A. It is also false if A is closed! (Find a counter-example).

Theorem 3.2.8 has many important corollaries. First, we define the notion of separation. For this, recall the definition of the closed (or open) half–spaces determined by a hyperplane.

Given a hyperplane H, if $f\colon E \to \mathbb{R}$ is any nonconstant affine form defining H (i.e., $H = \operatorname{Ker} f$), we define the *closed half-spaces associated with* f by

$$H_+(f) = \{a \in E \mid f(a) \geq 0\},$$
$$H_-(f) = \{a \in E \mid f(a) \leq 0\}.$$

Observe that if $\lambda > 0$, then $H_+(\lambda f) = H_+(f)$, but if $\lambda < 0$, then $H_+(\lambda f) = H_-(f)$, and similarly for $H_-(\lambda f)$.

Thus, the set $\{H_+(f), H_-(f)\}$ only depends on the hyperplane H, and the choice of a specific f defining H amounts to the choice of one of the two half-spaces.

We also define the *open half–spaces associated with* f as the two sets

$$\overset{\circ}{H}_+(f) = \{a \in E \mid f(a) > 0\},$$
$$\overset{\circ}{H}_-(f) = \{a \in E \mid f(a) < 0\}.$$

The set $\{\overset{\circ}{H}_+(f), \overset{\circ}{H}_-(f)\}$ only depends on the hyperplane H.

Clearly, $\overset{\circ}{H}_+(f) = H_+(f) - H$ and $\overset{\circ}{H}_-(f) = H_-(f) - H$.

Definition 3.2.9 Given an affine space, X, and two nonempty subsets, A and B, of X, we say that a hyperplane H *separates (resp. strictly separates)* A *and* B if A is in one and B is in the other of the two half–spaces (resp. open half–spaces) determined by H.

We will eventually prove that for any two nonempty disjoint convex sets A and B there is a hyperplane separating A and B, but this will take some work.

We begin with the following version of the Hahn-Banach theorem:

Theorem 3.2.10 *(Hahn-Banach, second version) Let X be a (finite-dimensional) affine space, A be a nonempty convex subset of X with nonempty interior and L be an affine subspace of X so that $A \cap L = \emptyset$. Then, there is some hyperplane, H, containing L and separating L and A.*

Corollary 3.2.11 *Given an affine space, X, let A and B be two nonempty disjoint convex subsets and assume that A has nonempty interior ($\overset{\circ}{A} \neq \emptyset$). Then, there is a hyperplane separating A and B.*

Remark: Theorem 3.2.10 and Corollary 3.2.11 also hold in the infinite case.

Corollary 3.2.12 *Given an affine space, X, let A and B be two nonempty disjoint open and convex subsets. Then, there is a hyperplane strictly separating A and B.*

Beware that Corollary 3.2.12 *fails* for *closed* convex sets. However, Corollary 3.2.12 holds if we also assume that A (or B) is compact.

We need to review the notion of distance from a point to a subset.

Let X be a metric space with distance function d. Given any point $a \in X$ and any nonempty subset B of X, we let

$$d(a, B) = \inf_{b \in B} d(a, b)$$

(where inf is the notation for least upper bound).

Now, if X is an affine space of dimension d, it can be given a metric structure by giving the corresponding vector space a metric structure, for instance, the metric induced by a Euclidean structure.

We have the following important property: For any nonempty closed subset, $S \subseteq X$ (not necessarily convex), and any point, $a \in X$, there is some point $s \in S$ "achieving the distance from a to S," i.e., so that

$$d(a, S) = d(a, s).$$

Corollary 3.2.13 *Given an affine space, X, let A and B be two nonempty disjoint closed and convex subsets, with A compact. Then, there is a hyperplane strictly separating A and B.*

Finally, we have the separation theorem announced earlier for arbitrary nonempty convex subsets. (For a different proof, see Berger [**?**], Corollary 11.4.7.)

Corollary 3.2.14 *Given an affine space, X, let A and B be two nonempty disjoint convex subsets. Then, there is a hyperplane separating A and B.*

Remarks:

(1) The reader should compare the proof from Valentine [**?**], Chapter II with Berger's proof using compactness of the projective space $\mathbb{P}^d$ [**?**] (Corollary 11.4.7).

(2) Rather than using the Hahn-Banach theorem to deduce separation results, one may proceed differently and use the following intuitively obvious lemma, as in Valentine [**?**] (Theorem 2.4):

Lemma 3.2.15 *If A and B are two nonempty convex sets such that $A \cup B = X$ and $A \cap B = \emptyset$, then $V = \overline{A} \cap \overline{B}$ is a hyperplane.*

One can then deduce Corollaries 3.2.11 and 3.2.14. Yet another approach is followed in Barvinok [**?**].

(3) How can some of the above results be generalized to infinite dimensional affine spaces, especially Theorem 3.2.8 and Corollary 3.2.11?

One approach is to simultaneously relax the notion of interior and tighten a little the notion of closure, in a more "linear and less topological" fashion, as in Valentine [**?**].

Given any subset $A \subseteq X$ (where X may be infinite dimensional, but is a Hausdorff topological vector space), say that a point $x \in X$ is *linearly accessible from* A iff there is some $a \in A$ with $a \neq x$ and $]a, x[\subseteq A$. We let lina A be the set of all points linearly accessible from A and lin $A = A \cup$ lina A.

A point $a \in A$ is a *core point of* A iff for every $y \in X$, with $y \neq a$, there is some $z \in]a, y[$, such that $[a, z] \subseteq A$. The set of all core points is denoted core A.

It is not difficult to prove that lin $A \subseteq \overline{A}$ and $\overset{\circ}{A} \subseteq$ core A. If A has nonempty interior, then lin $A = \overline{A}$ and $\overset{\circ}{A} =$ core A.

Also, if A is convex, then core A and lin A are convex. Then, Lemma 3.2.15 still holds (where X is not necessarily finite dimensional) if we redefine V as $V = \text{lin } A \cap \text{lin } B$ and allow the possibility that V could be X itself.

Corollary 3.2.11 also holds in the general case if we assume that core A is nonempty. For details, see Valentine [**?**], Chapter I and II.

(4) Yet another approach is to define the notion of an algebraically open convex set, as in Barvinok [**?**].

A convex set, A, is *algebraically open* iff the intersection of A with every line, L, is an open interval, possibly empty or infinite at either end (or all of L).

An open convex set is algebraically open. Then, the Hahn-Banach theorem holds provided that A is an algebraically open convex set and similarly, Corollary 3.2.11 also holds provided A is algebraically open.

For details, see Barvinok [**?**], Chapter 2 and 3. We do not know how the notion "algebraically open" relates to the concept of core.

(5) Theorems 3.2.8, 3.2.10 and Corollary 3.2.11 are proved in Lax using the notion of *gauge function* in the more general case where A has some core point (but beware that Lax uses the terminology *interior point* instead of core point!).

An important special case of separation is the case where A is convex and $B = \{a\}$ for some point a in A.

3.3 Supporting Hyperplanes

Definition 3.3.1 Let X be an affine space and let A be any nonempty subset of X. A *supporting hyperplane of* A is any hyperplane, H, containing some point, a, of A, and separating $\{a\}$ and A. We say that H is a *supporting hyperplane of* A *at* a.

Observe that if H is a supporting hyperplane of A at a, then we must have $a \in \partial A$.

Also, if A is convex, then $H \cap \overset{\circ}{A} = \emptyset$.

One should experiment with various pictures and realize that supporting hyperplanes at a point may not exist (for example, if A is not convex), may not be unique, and may have several distinct supporting points!

However, we have the following important proposition first proved by Minkowski (1896):

Proposition 3.3.2 *(Minkowski) Let A be a nonempty closed and convex subset. Then, for every point, $a \in \partial A$, there is a supporting hyperplane to A through a.*

Beware that Proposition 3.3.2 is false when the dimension X of A is infinite and when $\overset{\circ}{A} = \emptyset$.

The proposition below gives a sufficient condition for a closed subset to be convex.

Proposition 3.3.3 *Let A be a closed subset with nonempty interior. If there is a supporting hyperplane for every point $a \in \partial A$, then A is convex.*

The condition that A has nonempty interior is crucial!

The proposition below characterizes closed convex sets in terms of (closed) half–spaces. It is another intuitive fact whose rigorous proof is nontrivial.

Proposition 3.3.4 *Let A be a nonempty closed and convex subset. Then, A is the intersection of all the closed half–spaces containing it.*

Next, we consider various types of boundary points of closed convex sets.

3.4 Boundary of a Convex Set: Vertices and Extremal Points

Definition 3.4.1 Let X be an affine space of dimension d. For any nonempty closed and convex subset, A, of dimension d, a point $a \in \partial A$ has *order* $k(a)$ if the intersection of all the supporting hyperplanes of A at a is an affine subspace of dimension $k(a)$. We say that $a \in \partial A$ is a *vertex* if $k(a) = 0$; we say that a is *smooth* if $k(a) = d - 1$, i.e., if the supporting hyperplane at a is unique.

A vertex is a boundary point a such that there are d independent supporting hyperplanes at a.

A d-simplex has boundary points of order $0, 1, \ldots, d-1$. The following proposition is shown in Berger [**?**] (Proposition 11.6.2):

Proposition 3.4.2 *The set of vertices of a closed and convex subset is countable.*

Another important concept is that of an extremal point.

Definition 3.4.3 Let X be an affine space. For any nonempty convex subset A, a point $a \in \partial A$ is *extremal* (or *extreme*) if $A - \{a\}$ is still convex.

It is fairly obvious that a point $a \in \partial A$ is extremal if it does not belong to any closed nontrivial line segment $[x, y] \subseteq A$ $(x \neq y)$.

Observe that a vertex is extremal, but the converse is false.

Also, if $\dim X \geq 3$, the set of extremal points of a compact convex may not be closed.

Actually, it is not at all obvious that a nonempty compact convex possesses extremal points.

In fact, a stronger results holds (Krein and Milman's theorem).

In preparation for the proof of this important theorem, observe that any compact (nontrivial) interval of $\mathbb{A}^1$ has two extremal points, its two endpoints.

Lemma 3.4.4 *Let X be an affine space of dimension n, and let A be a nonempty compact and convex set. Then, $A = \mathcal{C}(\partial A)$, i.e., A is equal to the convex hull of its boundary.*

The following important theorem shows that only extremal points matter as far as determining a compact and convex subset from its boundary.

Theorem 3.4.5 *(Krein and Milman) Let X be an affine space of dimension n. Every compact and convex nonempty subset A is equal to the convex hull of its set of extremal points.*

Observe that Krein and Milman's theorem implies that any nonemty compact and convex set has a nonempty subset of extremal points. This is intuitively obvious, but hard to prove!

Krein and Milman's theorem also holds for infinite dimensional affine spaces, provided that they are locally convex.

Chapter 4

Basics of Euclidean Geometry

4.1 Inner Products, Euclidean Spaces

In Affine geometry, it is possible to deal with ratios of vectors and barycenters of points, but there is no way to express the notion of length of a line segment, or to talk about orthogonality of vectors.

A Euclidean structure will allow us to deal with *metric notions* such as orthogonality and length (or distance).

We begin by defining inner products and Euclidean Spaces. The Cauchy-Schwarz inequality and the Minkovski inequality are shown.

We define othogonality of vectors and of subspaces, othogonal families, and orthonormal families. We offer a glimpse at Fourier series in terms of the orthogonal families $(\sin px)_{p\geq 1} \cup (\cos qx)_{q\geq 0}$ and $(e^{ikx})_{k\in\mathbb{Z}}$.

We prove that every finite dimensional Euclidean space has orthonormal bases.

The first proof uses duality, and the second one the Gram-Schmidt procedure. The QR-decomposition of matrices is shown as an application.

Linear isometries (also called orthogonal transformations) are defined and studied briefly.

The orthogonal group and orthogonal matrices are studied briefly.

First, we define a Euclidean structure on a vector space, and then, on an affine space.

Definition 4.1.1 A real vector space E is a *Euclidean space* iff it is equipped with a symmetric bilinear form $\varphi\colon E \times E \to \mathbb{R}$ which is also *positive definite*,which means that

$$\varphi(u, u) > 0, \quad \text{for every } u \neq 0.$$

More explicitly, $\varphi\colon E \times E \to \mathbb{R}$ satisfies the following axioms:

$$\begin{aligned}
\varphi(u_1 + u_2, v) &= \varphi(u_1, v) + \varphi(u_2, v),\\
\varphi(u, v_1 + v_2) &= \varphi(u, v_1) + \varphi(u, v_2),\\
\varphi(\lambda u, v) &= \lambda\varphi(u, v),\\
\varphi(u, \lambda v) &= \lambda\varphi(u, v),\\
\varphi(u, v) &= \varphi(v, u),\\
u \neq 0 \text{ implies that } \varphi(u, u) &> 0.
\end{aligned}$$

The real number $\varphi(u, v)$ is also called the *inner product (or scalar product) of u and v.*

We also define the *quadratic form associated with* φ as the function $\Phi\colon E \to \mathbb{R}_+$ such that

$$\Phi(u) = \varphi(u, u),$$

for all $u \in E$.

Since φ is bilinear, we have $\varphi(0, 0) = 0$, and since it is positive definite, we have the stronger fact that

$$\varphi(u, u) = 0 \quad \text{iff} \quad u = 0,$$

that is $\Phi(u) = 0$ iff $u = 0$.

Given an inner product $\varphi\colon E \times E \to \mathbb{R}$ on a vector space E, we also denote $\varphi(u, v)$ by

$$u \cdot v, \quad \text{or} \quad \langle u, v \rangle, \quad \text{or} \quad (u|v),$$

and $\sqrt{\Phi(u)}$ as $\|u\|$.

Example 1. The standard example of a Euclidean space is $\mathbb{R}^n$, under the inner product $\cdot$ defined such that

$$(x_1, \ldots, x_n) \cdot (y_1, \ldots, y_n) = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

Example 2. Let E be a vector space of dimension 2, and let (e_1, e_2) be a basis of E.

If $a > 0$ and $b^2 - ac < 0$, the bilinear form defined such that

$$\varphi(x_1e_1+y_1e_2,\, x_2e_1+y_2e_2) = ax_1x_2+b(x_1y_2+x_2y_1)+cy_1y_2$$

yields a Euclidean structure on E.

In this case,

$$\Phi(xe_1 + ye_2) = ax^2 + 2bxy + cy^2.$$

Example 3. Let $\mathcal{C}[a,b]$ denote the set of continuous functions $f\colon [a,b] \to \mathbb{R}$. It is easily checked that $\mathcal{C}[a,b]$ is a vector space of infinite dimension.

Given any two functions $f, g \in \mathcal{C}[a,b]$, let

$$\langle f, g\rangle = \int_a^b f(t)g(t)dt.$$

We leave as an easy exercise that $\langle -, -\rangle$ is indeed an inner product on $\mathcal{C}[a,b]$.

When $[a,b] = [-\pi, \pi]$ (or $[a,b] = [0, 2\pi]$, this makes basically no difference), one should compute

$$\langle \sin px, \sin qx\rangle, \quad \langle \sin px, \cos qx\rangle,$$
$$\text{and} \quad \langle \cos px, \cos qx\rangle,$$

for all natural numbers $p, q \geq 1$. The outcome of these calculations is what makes Fourier analysis possible!

Let us observe that φ can be recovered from Φ. Indeed, by bilinearity and symmetry, we have

$$\begin{aligned}\Phi(u+v) &= \varphi(u+v,\, u+v)\\ &= \varphi(u,\, u+v) + \varphi(v,\, u+v)\\ &= \varphi(u,\, u) + 2\varphi(u,\, v) + \varphi(v,\, v)\\ &= \Phi(u) + 2\varphi(u,\, v) + \Phi(v).\end{aligned}$$

Thus, we have

$$\varphi(u,\, v) = \frac{1}{2}[\Phi(u+v) - \Phi(u) - \Phi(v)].$$

We also say that φ is the *polar form of* Φ.

One of the very important properties of an inner product φ is that the map $u \mapsto \sqrt{\Phi(u)}$ is a norm.

Lemma 4.1.2 *Let E be a Euclidean space with inner product φ and quadratic form Φ. For all $u, v \in E$, we have the* Cauchy-Schwarz inequality*:*

$$\varphi(u, v)^2 \leq \Phi(u)\Phi(v),$$

the equality holding iff u and v are linearly dependent.

We also have the Minkovski inequality*:*

$$\sqrt{\Phi(u + v)} \leq \sqrt{\Phi(u)} + \sqrt{\Phi(v)},$$

the equality holding iff u and v are linearly dependent, where in addition if $u \neq 0$ and $v \neq 0$, then $u = \lambda v$ for some $\lambda > 0$.

Sketch of proof. Define the function $T\colon \mathbb{R} \to \mathbb{R}$, such that

$$T(\lambda) = \Phi(u + \lambda v),$$

for all $\lambda \in \mathbb{R}$. Using bilinearity and symmetry, we can show that

$$\Phi(u + \lambda v) = \Phi(u) + 2\lambda\varphi(u,\, v) + \lambda^2\Phi(v).$$

Since φ is positive definite, we have $T(\lambda) \geq 0$ for all $\lambda \in \mathbb{R}$.

If $\Phi(v) = 0$, then $v = 0$, and we also have $\varphi(u,\, v) = 0$. In this case, the Cauchy-Schwarz inequality is trivial,

If $\Phi(v) > 0$, then

$$\lambda^2\Phi(v) + 2\lambda\varphi(u,\, v) + \Phi(u) = 0$$

can't have distinct roots, which means that its discriminant

$$\Delta = 4(\varphi(u,\, v)^2 - \Phi(u)\Phi(v))$$

is zero or negative, which is precisely the Cauchy-Schwarz inequality.

The Minkovski inequality can then be shown.

Let us review the definition of a normed vector space.

Definition 4.1.3 Let E be a vector space over a field K, where K is either the field $\mathbb{R}$ of reals, or the field $\mathbb{C}$ of complex numbers. A *norm on* E is a function $\| \| : E \to \mathbb{R}_+$, assigning a nonnegative real number $\|u\|$ to any vector $u \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(N1) $\|x\| \geq 0$ and $\|x\| = 0$ iff $x = 0$. (positivity)

(N2) $\|\lambda x\| = |\lambda| \, \|x\|$. (scaling)

(N3) $\|x + y\| \leq \|x\| + \|y\|$. (triangle inequality)

A vector space E together with a norm $\| \|$ is called a *normed vector space*.

From (N3), we easily get

$$| \, \|x\| - \|y\| \, | \leq \|x - y\| .$$

The Minkovski inequality

$$\sqrt{\Phi(u+v)} \leq \sqrt{\Phi(u)} + \sqrt{\Phi(v)}$$

shows that the map $u \mapsto \sqrt{\Phi(u)}$ satisfies the *triangle inequality*, condition (N3) of definition 4.1.3, and since φ is bilinear and positive definite, it also satisfies conditions (N1) and (N2) of definition 4.1.3, and thus, it is a *norm* on E.

The norm induced by φ is called the *Euclidean norm induced by* φ.

Note that the Cauchy-Schwarz inequality can be written as

$$|u \cdot v| \leq \|u\| \, \|v\| \, ,$$

and the Minkovski inequality as

$$\|u + v\| \leq \|u\| + \|v\| \, .$$

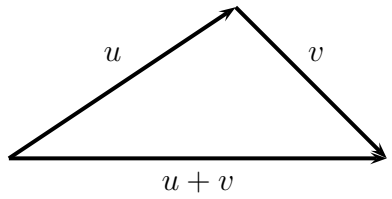

Figure 4.1: The triangle inequality

We now define orthogonality.

4.2 Orthogonality

Definition 4.2.1 Given a Euclidean space E, any two vectors $u, v \in E$ are *orthogonal, or perpendicular* iff $u \cdot v = 0$. Given a family $(u_i)_{i \in I}$ of vectors in E, we say that $(u_i)_{i \in I}$ is *orthogonal* iff $u_i \cdot u_j = 0$ for all $i, j \in I$, where $i \neq j$. We say that the family $(u_i)_{i \in I}$ is *orthonormal* iff $u_i \cdot u_j = 0$ for all $i, j \in I$, where $i \neq j$, and $\|u_i\| = u_i \cdot u_i = 1$, for all $i \in I$. For any subset F of E, the set

$$F^{\perp} = \{v \in E \mid u \cdot v = 0, \text{ for all } u \in F\},$$

of all vectors orthogonal to all vectors in F, is called the *orthogonal complement of* F.

Since inner products are positive definite, observe that for any vector $u \in E$, we have

$$u \cdot v = 0 \quad \text{for all } v \in E \quad \text{iff} \quad u = 0.$$

It is immediately verified that the orthogonal complement $F^{\perp}$ of F is a subspace of E.

Example 4. Going back to example 3, and to the inner product

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(t)g(t)dt$$

on the vector space $\mathcal{C}[-\pi, \pi]$, it is easily checked that

$$\langle \sin px, \sin qx\rangle = \begin{cases} \pi & \text{if } p = q,\ p, q \geq 1, \\ 0 & \text{if } p \neq q,\ p, q \geq 1 \end{cases}$$

$$\langle \cos px, \cos qx\rangle = \begin{cases} \pi & \text{if } p = q,\ p, q \geq 1, \\ 0 & \text{if } p \neq q,\ p, q \geq 0 \end{cases}$$

and

$$\langle \sin px, \cos qx\rangle = 0,$$

for all $p \geq 1$ and $q \geq 0$, and of course,
$\langle 1, 1\rangle = \int_{-\pi}^{\pi} dx = 2\pi$.

As a consequence, the family $(\sin px)_{p\geq 1} \cup (\cos qx)_{q\geq 0}$ is orthogonal.

It is not orthonormal, but becomes so if we divide every trigonometric function by $\sqrt{\pi}$, and 1 by $\sqrt{2\pi}$.

Remark: Observe that if we allow complex valued functions, we obtain simpler proofs. For example, it is immediately checked that

$$\int_{-\pi}^{\pi} e^{ikx} dx = \begin{cases} 2\pi & \text{if } k = 0; \\ 0 & \text{if } k \neq 0, \end{cases}$$

because the derivative of e^{ikx} is ike^{ikx}.

However, beware that something strange is going on!

Indeed, unless $k = 0$, we have

$$\langle e^{ikx}, e^{ikx} \rangle = 0,$$

since

$$\langle e^{ikx}, e^{ikx} \rangle = \int_{-\pi}^{\pi} (e^{ikx})^2 dx = \int_{-\pi}^{\pi} e^{i2kx} dx = 0.$$

The inner product $\langle e^{ikx}, e^{ikx} \rangle$ should be strictly positive. What went wrong?

The problem is that we are using the wrong inner product. When we use complex-valued functions, we must use the *Hermitian inner product*

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx,$$

where $\overline{g(x)}$ is the *conjugate* of $g(x)$.

The Hermitian inner product is not symmetric. Instead,

$$\langle g, f\rangle = \overline{\langle f, g\rangle}.$$

(Recall that if $z = a+ib$, where $a, b \in \mathbb{R}$, then $\overline{z} = a-ib$. Also $e^{i\theta} = \cos\theta + i\sin\theta$).

With the Hermitian inner product, everthing works out beautifully! In particular, the family $(e^{ikx})_{k\in\mathbb{Z}}$ is orthogonal.

Lemma 4.2.2 *Given a Euclidean space E, for any family $(u_i)_{i\in I}$ of nonnull vectors in E, if $(u_i)_{i\in I}$ is orthogonal, then it is linearly independent.*

Lemma 4.2.3 *Given a Euclidean space E, any two vectors $u, v \in E$ are orthogonal iff*

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2.$$

One of the most useful features of orthonormal bases is that they afford a very simple method for computing the coordinates of a vector over any basis vector.

Indeed, assume that $(e_1, \ldots, e_m)$ is an orthonormal basis. For any vector

$$x = x_1 e_1 + \cdots + x_m e_m,$$

if we compute the inner product $x \cdot e_i$, we get

$$x \cdot e_i = x_1 e_1 \cdot e_i + \cdots + x_i e_i \cdot e_i + \cdots + x_m e_m \cdot e_i = x_i,$$

since

$$e_i \cdot e_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

is the property characterizing an orthonormal family.

Thus,

$$x_i = x \cdot e_i,$$

which means that $x_i e_i = (x \cdot e_i) e_i$ is the orthogonal projection of x onto the subspace generated by the basis vector e_i.

If the basis is orthogonal but not necessarily orthonormal, then

$$x_i = \frac{x \cdot e_i}{e_i \cdot e_i} = \frac{x \cdot e_i}{\|e_i\|^2}.$$

All this is true even for an infinite orthonormal (or orthogonal) basis $(e_i)_{i \in I}$.

However, remember that every vector x is expressed as a linear combination

$$x = \sum_{i \in I} x_i e_i$$

where the family of scalars $(x_i)_{i \in I}$ has **finite support**, which means that $x_i = 0$ for all $i \in I - J$, where J is a finite set.

Thus, even though the family $(\sin px)_{p \geq 1} \cup (\cos qx)_{q \geq 0}$ is orthogonal (it is not orthonormal, but becomes one if we divide every trigonometric function by $\sqrt{\pi}$, and 1 by $\sqrt{2\pi}$; we won't because it looks messy!), the fact that a function $f \in \mathcal{C}^0[-\pi, \pi]$ can be written as a Fourier series as

$$f(x) = a_0 + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

does not mean that $(\sin px)_{p \geq 1} \cup (\cos qx)_{q \geq 0}$ is a basis of this vector space of functions, because in general, the families (a_k) and (b_k) **do not** have finite support!

In order for this infinite linear combination to make sense, it is necessary to prove that the partial sums

$$a_0 + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx)$$

of the series converge to a limit when n goes to infinity.

This requires a topology on the space.

Still, a small miracle happens. If $f \in \mathcal{C}[-\pi, \pi]$ can indeed be expressed as a Fourier series

$$f(x) = a_0 + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx),$$

the coefficients a_0 and a_k, b_k, $k \geq 1$, can be computed by projecting f over the basis functions, i.e. by taking inner products with the basis functions in $(\sin px)_{p \geq 1} \cup (\cos qx)_{q \geq 0}$.

Indeed, for all $k \geq 1$, we have

$$a_0 = \frac{\langle f, 1\rangle}{\|1\|^2},$$

and

$$a_k = \frac{\langle f, \cos kx\rangle}{\|\cos kx\|^2}, \qquad b_k = \frac{\langle f, \sin kx\rangle}{\|\sin kx\|^2},$$

that is

$$a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)dx,$$

and

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx\, dx, \qquad b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx\, dx.$$

If we allow f to be complex-valued and use the family $(e^{ikx})_{k\in\mathbb{Z}}$, which is is indeed orthogonal w.r.t. the Hermitian inner product

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx,$$

we consider functions $f \in \mathcal{C}[-\pi, \pi]$ that can be expressed as the sum of a series

$$f(x) = \sum_{k\in\mathbb{Z}} c_k e^{ikx}.$$

Note that the index k is allowed to be a negative integer. Then, the formula giving the c_k is very nice:

$$c_k = \frac{\langle f, e^{ikx}\rangle}{\left\|e^{ikx}\right\|^2},$$

that is

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}dx.$$

Note the presence of the negative sign in e^{-ikx}, which is due to the fact that the inner product is Hermitian.

Of course, the real case can be recovered from the complex case. If f is a real-valued function, then we must have

$$a_k = c_k + c_{-k} \quad \text{and} \quad b_k = i(c_k - c_{-k}).$$

Also note that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} dx$$

is not only defined for all discrete values $k \in \mathbb{Z}$, but for all $k \in \mathbb{R}$, and that if f is continuous over $\mathbb{R}$, the integral makes sense.

This suggests defining

$$\widehat{f}(k) = \int_{-\infty}^{\infty} f(x)e^{-ikx}dx,$$

called the *Fourier transform* of f. It analyses the function f in the "frequency domain" in terms of its spectrum of harmonics.

Amazingly, there is an inverse Fourier transform (change e^{-ikx} to e^{+ikx} and divide by the scale factor 2π) which reconstructs f (under certain assumptions on f).

A very important property of Euclidean spaces of finite dimension is that the inner product induces a canonical bijection (i.e., independent of the choice of bases) between the vector space E and its dual E^*.

Given a Euclidean space E, for any vector $u \in E$, let $\varphi_u \colon E \to \mathbb{R}$ be the map defined such that

$$\varphi_u(v) = u \cdot v,$$

for all $v \in E$.

Since the inner product is bilinear, the map φ_u is a linear form in E^*.

Thus, we have a map $\flat\colon E \to E^*$, defined such that

$$\flat(u) = \varphi_u.$$

Lemma 4.2.4 *Given a Euclidean space E, the map $\flat\colon E \to E^*$, defined such that*

$$\flat(u) = \varphi_u,$$

is linear and injective. When E is also of finite dimension, the map $\flat\colon E \to E^*$ is a canonical isomorphism.

The inverse of the isomorphism $\flat\colon E \to E^*$ is denoted by $\sharp\colon E^* \to E$.

As a consequence of lemma 4.2.4, if E is a Euclidean space of finite dimension, every linear form $f \in E^*$ corresponds to a unique $u \in E$, such that

$$f(v) = u \cdot v,$$

for every $v \in E$.

In particular, if f is not the null form, the kernel of f, which is a hyperplane H, is precisely the set of vectors that are orthogonal to u.

Lemma 4.2.4 allows us to define the adjoint of a linear map on a Euclidean space.

Let E be a Euclidean space of finite dimension n, and let $f: E \to E$ be a linear map.

For every $u \in E$, the map

$$v \mapsto u \cdot f(v)$$

is clearly a linear form in E^*, and by lemma 4.2.4, there is a unique vector in E denoted as $f^*(u)$, such that

$$f^*(u) \cdot v = u \cdot f(v),$$

for every $v \in E$.

Lemma 4.2.5 *Given a Euclidean space E of finite dimension, for every linear map $f: E \to E$, there is a unique linear map $f^*: E \to E$, such that*

$$f^*(u) \cdot v = u \cdot f(v),$$

for all $u, v \in E$. The map f^* is called the adjoint of f (w.r.t. to the inner product).

Linear maps $f: E \to E$ such that $f = f^*$ are called *self-adjoint* maps.

They play a very important role because they have real eigenvalues, and because orthonormal bases arise from their eigenvectors.

Furthermore, many physical problems lead to self-adjoint linear maps (in the form of symmetric matrices).

Linear maps such that $f^{-1} = f^*$, or equivalently

$$f^* \circ f = f \circ f^* = \mathrm{id},$$

also play an important role. They are *isometries*. Rotations are special kinds of isometries.

Another important class of linear maps are the linear maps satisfying the property

$$f^* \circ f = f \circ f^*,$$

called *normal linear maps*.

We will see later on that normal maps can always be diagonalized over orthonormal bases of eigenvectors, but this will require using a Hermitian inner product (over $\mathbb{C}$).

Given two Euclidean spaces E and F, where the inner product on E is denoted as $\langle -, - \rangle_1$ and the inner product on F is denoted as $\langle -, - \rangle_2$, given any linear map $f\colon E \to F$, it is immediately verified that the proof of lemma 4.2.5 can be adapted to show that there is a unique linear map $f^*\colon F \to E$ such that

$$\langle f(u), v \rangle_2 = \langle u, f^*(v) \rangle_1$$

for all $u \in E$ and all $v \in F$. The linear map f^* is also called the adjoint of f.

Remark: Given any basis for E and any basis for F, it is possible to characterize the matrix of the adjoint f^* of f in terms of the matrix of f, and the symmetric matrices defining the inner products. We will do so with respect to orthonormal bases.

We can also use lemma 4.2.4 to show that any Euclidean space of finite dimension has an orthonormal basis.

Lemma 4.2.6 *Given any nontrivial Euclidean space E of finite dimension $n \geq 1$, there is an orthonormal basis $(u_1, \ldots, u_n)$ for E.*

There is a more constructive way of proving lemma 4.2.6, using a procedure known as the *Gram–Schmidt orthonormalization procedure*.

Among other things, the Gram–Schmidt orthonormalization procedure yields the so-called *QR-decomposition for matrices*, an important tool in numerical methods.

Lemma 4.2.7 *Given any nontrivial Euclidean space E of dimension $n \geq 1$, from any basis $(e_1, \ldots, e_n)$ for E, we can construct an orthonormal basis $(u_1, \ldots, u_n)$ for E, with the property that for every k, $1 \leq k \leq n$, the families $(e_1, \ldots, e_k)$ and $(u_1, \ldots, u_k)$ generate the same subspace.*

Proof. We proceed by induction on n. For $n = 1$, let

$$u_1 = \frac{e_1}{\|e_1\|}.$$

For $n \geq 2$, we define the vectors u_k and u'_k as follows.

$$u'_1 = e_1, \qquad u_1 = \frac{u'_1}{\|u'_1\|},$$

and for the inductive step

$$u'_{k+1} = e_{k+1} - \sum_{i=1}^{k} (e_{k+1} \cdot u_i)\, u_i, \qquad u_{k+1} = \frac{u'_{k+1}}{\left\|u'_{k+1}\right\|}.$$

We need to show that u'_{k+1} is nonzero, and we conclude by induction.

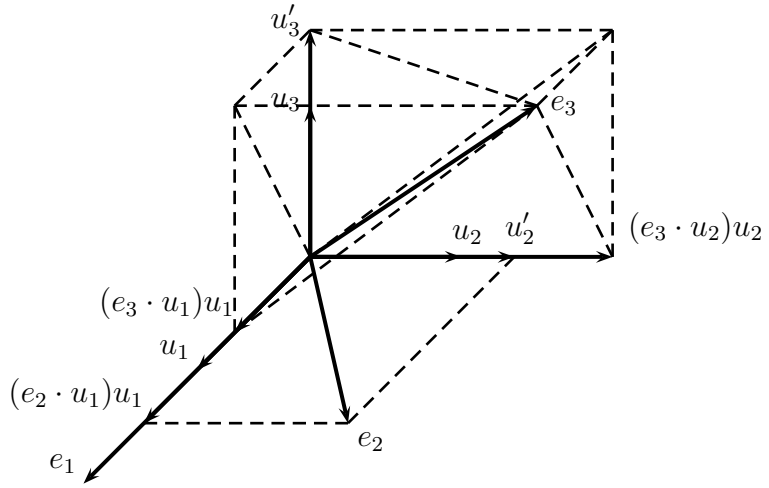

Figure 4.2: The Gram-Schmidt orthonormalization procedure

Remarks:

(1) Note that u'_{k+1} is obtained by subtracting from e_{k+1} the projection of e_{k+1} itself onto the orthonormal vectors $u_1, \ldots, u_k$ that have already been computed. Then, we normalize u'_{k+1}.

The QR-decomposition can now be obtained very easily. We will do this in section 4.4.

(2) We could compute u'_{k+1} using the formula

$$u'_{k+1} = e_{k+1} - \sum_{i=1}^{k} \left(\frac{e_{k+1} \cdot u'_i}{\|u'_i\|^2} \right) u'_i,$$

and normalize the vectors u'_k at the end.

This time, we are subtracting from e_{k+1} the projection of e_{k+1} itself onto the orthogonal vectors $u'_1, \ldots, u'_k$.

This might be preferable when writing a computer program.

(3) The proof of lemma 4.2.7 also works for a countably infinite basis for E, producing a countably infinite orthonormal basis.

Example 5. If we consider polynomials and the inner product

$$\langle f, g\rangle = \int_{-1}^{1} f(t)g(t)dt,$$

applying the Gram–Schmidt orthonormalization procedure to the polynomials

$$1, x, x^2, \ldots, x^n, \ldots,$$

which form a basis of the polynonials in one variable with real coefficients, we get a family of orthonormal polynomials $Q_n(x)$ related to the *Legendre polynomials*.

The Legendre polynomials $P_n(x)$ have many nice properties. They are orthogonal, but their norm is not always 1. The Legendre polynomials $P_n(x)$ can be defined as follows:

If we let f_n be the function

$$f_n(x) = (x^2 - 1)^n,$$

we define $P_n(x)$ as follows:

$$P_0(x) = 1, \quad \text{and} \quad P_n(x) = \frac{1}{2^n n!} f_n^{(n)}(x),$$

where $f_n^{(n)}$ is the nth derivative of f_n.

They can also be defined inductively as follows:

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ P_{n+1}(x) &= \frac{2n+1}{n+1}\, xP_n(x) - \frac{n}{n+1}\, P_{n-1}(x). \end{aligned}$$

It turns out that the polynomials Q_n are related to the Legendre polynomials P_n as follows:

$$Q_n(x) = \frac{2^n (n!)^2}{(2n)!} P_n(x).$$

As a consequence of lemma 4.2.6 (or lemma 4.2.7), given any Euclidean space of finite dimension n, if $(e_1, \ldots, e_n)$ is an orthonormal basis for E, then for any two vectors $u = u_1 e_1 + \cdots + u_n e_n$ and $v = v_1 e_1 + \cdots + v_n e_n$, the inner product $u \cdot v$ is expressed as

$$u \cdot v = (u_1 e_1 + \cdots + u_n e_n) \cdot (v_1 e_1 + \cdots + v_n e_n) = \sum_{i=1}^{n} u_i v_i,$$

and the norm $\|u\|$ as

$$\|u\| = \|u_1 e_1 + \cdots + u_n e_n\| = \sqrt{\sum_{i=1}^{n} u_i^2}.$$

We can also prove the following lemma regarding orthogonal spaces.

Lemma 4.2.8 *Given any nontrivial Euclidean space E of finite dimension $n \geq 1$, for any subspace F of dimension k, the orthogonal complement $F^{\perp}$ of F has dimension $n - k$, and $E = F \oplus F^{\perp}$. Furthermore, we have $F^{\perp\perp} = F$.*

Definition 4.2.9 An affine space $(E, \overrightarrow{E})$ is a *Euclidean affine space* iff its underlying vector space $\overrightarrow{E}$ is a Euclidean vector space. Given any two points $a, b \in E$, we define the *distance between a and b, or length of the segment* (a, b), as $\|\mathbf{ab}\|$, the Euclidean norm of $\mathbf{ab}$. Given any two pairs of points (a, b) and (c, d), we define their inner product as $\mathbf{ab} \cdot \mathbf{cd}$. We say that (a, b) and (c, d) are *orthogonal, or perpendicular* iff $\mathbf{ab} \cdot \mathbf{cd} = 0$. We say that two affine subspaces F_1 and F_2 of E are *orthogonal* iff their directions $\overrightarrow{F_1}$ and $\overrightarrow{F_2}$ are orthogonal.

Note that a Euclidean affine space is a normed affine space, in the sense of definition 4.2.10 below.

Definition 4.2.10 Given an affine space $(E, \overrightarrow{E})$, where the space of translations $\overrightarrow{E}$ is a vector space over $\mathbb{R}$ or $\mathbb{C}$, we say that $(E, \overrightarrow{E})$ is a *normed affine space* iff $\overrightarrow{E}$ is a normed vector space with norm $\| \|$.

We denote as $\mathbb{E}^m$ the Euclidean affine space obtained from the affine space $\mathbb{A}^m$ by defining on the vector space $\mathbb{R}^m$ the standard inner product

$$(x_1, \ldots, x_m) \cdot (y_1, \ldots, y_m) = x_1 y_1 + \cdots + x_m y_m.$$

The corresponding Euclidean norm is

$$\|(x_1, \ldots, x_m)\| = \sqrt{x_1^2 + \cdots + x_m^2}.$$

We now consider linear maps between Euclidean spaces that preserve the Euclidean norm. These transformations sometimes called *rigid motions* play an important role in geometry.

4.3 Linear Isometries (Orthogonal Transformations)

In this section, we consider linear maps between Euclidean spaces that preserve the Euclidean norm.

Definition 4.3.1 Given any two nontrivial Euclidean spaces E and F of the same finite dimension n, a function $f: E \to F$ is *an orthogonal transformation, or a linear isometry* iff it is linear and

$$\|f(u)\| = \|u\|,$$

for all $u \in E$.

Thus, a linear isometry is a linear map that preserves the norm.

Remarks: (1) A linear isometry is often defined as a linear map such that

$$\|f(v) - f(u)\| = \|v - u\|,$$

for all $u, v \in E$. Since the map f is linear, the two definitions are equivalent. The second definition just focuses on preserving the distance between vectors.

(2) Sometimes, a linear map satisfying the condition of definition 4.3.1 is called a *metric map*, and a linear isometry is defined as a *bijective* metric map.

Also, an isometry (without the word linear) is sometimes defined as a function $f: E \to F$ (not necessarily linear) such that

$$\|f(v) - f(u)\| = \|v - u\|,$$

for all $u, v \in E$, i.e., as a function that preserves the distance.

This requirement turns out to be very strong. Indeed, the next lemma shows that all these definitions are equivalent when E and F are of finite dimension, and for functions such that $f(0) = 0$.

Lemma 4.3.2 *Given any two nontrivial Euclidean spaces E and F of the same finite dimension n, for every function $f: E \to F$, the following properties are equivalent:*

(1) f is a linear map and $\|f(u)\| = \|u\|$, for all $u \in E$;

(2) $\|f(v) - f(u)\| = \|v - u\|$, for all $u, v \in E$, and $f(0) = 0$;

(3) $f(u) \cdot f(v) = u \cdot v$, for all $u, v \in E$.

Furthermore, such a map is bijective.

For (2), we shall prove a slightly stronger result. We prove that if

$$\|f(v) - f(u)\| = \|v - u\|$$

for all $u, v \in E$, for any vector $\tau \in E$, the function $g\colon E \to F$ defined such that

$$g(u) = f(\tau + u) - f(\tau)$$

for all $u \in E$ is a linear map such that $g(0) = 0$ and (3) holds.

Remarks: (i) The dimension assumption is only needed to prove that (3) implies (1) when f is not known to be linear, and to prove that f is surjective, but the proof shows that (1) implies that f is injective.

(ii) In (2), when f does not satisfy the condition $f(0) = 0$, the proof shows that f is an affine map.

Indeed, taking any vector τ as an origin, the map g is linear, and

$$f(\tau + u) = f(\tau) + g(u)$$

for all $u \in E$, proving that f is affine with associated linear map g.

(iii) The implication that (3) implies (1) holds if we also assume that f is surjective, even if E has infinite dimension.

In view of lemma 4.3.2, we will drop the word "linear" in "linear isometry", unless we wish to emphasize that we are dealing with a map between vector spaces.

We are now going to take a closer look at the isometries $f: E \to E$ of a Euclidean space of finite dimension.

4.4 The Orthogonal Group, Orthogonal Matrices

In this section, we explore some of the fundamental properties of the orthogonal group and of orthogonal matrices.

As an immediate corollary of the Gram–Schmidt orthonormalization procedure, we obtain the QR-decomposition for invertible matrices.

We prove an important structure theorem for the isometries, namely that they can always be written as a composition of reflections (Theorem 5.2.1).

Lemma 4.4.1 *Let E be any Euclidean space of finite dimension n, and let $f: E \to E$ be any linear map. The following properties hold:*

(1) The linear map $f: E \to E$ is an isometry iff

$$f \circ f^* = f^* \circ f = \mathrm{id}.$$

(2) For every orthonormal basis $(e_1, \ldots, e_n)$ of E, if the matrix of f is A, then the matrix of f^* is the transpose $A^\top$ of A, and f is an isometry iff A satisfies the identities

$$A\,A^\top = A^\top A = I_n,$$

where I_n denotes the identity matrix of order n, iff the columns of A form an orthonormal basis of E, iff the rows of A form an orthonormal basis of E.

Lemma 4.4.1 shows that the inverse of an isometry f is its adjoint f^*. Lemma 4.4.1 also motivates the following definition:

Definition 4.4.2 A real $n \times n$ matrix is an *orthogonal matrix* iff

$$A\,A^\top = A^\top A = I_n.$$

Remarks: It is easy to show that the conditions $A\,A^\top = I_n$, $A^\top A = I_n$, and $A^{-1} = A^\top$, are equivalent.

Given any two orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$, if P is the change of basis matrix from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$ since the columns of P are the coordinates of the vectors v_j with respect to the basis $(u_1, \ldots, u_n)$, and since $(v_1, \ldots, v_n)$ is orthonormal, the columns of P are orthonormal, and by lemma 4.4.1 (2), the matrix P is orthogonal.

The proof of lemma 4.3.2 (3) also shows that if f is an isometry, then the image of an orthonormal basis $(u_1, \ldots, u_n)$ is an orthonormal basis.

Recall that the determinant $\det(f)$ of an endomorphism $f: E \to E$ is independent of the choice of a basis in E.

Also, for every matrix $A \in \mathrm{M}_n(\mathbb{R})$, we have $\det(A) = \det(A^\top)$, and for any two $n \times n$-matrices A and B, we have $\det(AB) = \det(A)\det(B)$ (for all these basic results, see Lang [**?**]).

Then, if f is an isometry, and A is its matrix with respect to any orthonormal basis, $A\,A^\top = A^\top A = I_n$ implies that $\det(A)^2 = 1$, that is, either $\det(A) = 1$, or $\det(A) = -1$.

It is also clear that the isometries of a Euclidean space of dimension n form a group, and that the isometries of determinant $+1$ form a subgroup.

Definition 4.4.3 Given a Euclidean space E of dimension n, the set of isometries $f: E \to E$ forms a group denoted as $\mathbf{O}(E)$, or $\mathbf{O}(n)$ when $E = \mathbb{R}^n$, called the *orthogonal group (of E)*.

For every isometry, f, we have $\det(f) = \pm 1$, where $\det(f)$ denotes the determinant of f. The isometries such that $\det(f) = 1$ are called *rotations, or proper isometries, or proper orthogonal transformations*, and they form a subgroup of the special linear group $\mathbf{SL}(E)$ (and of $\mathbf{O}(E)$), denoted as $\mathbf{SO}(E)$, or $\mathbf{SO}(n)$ when $E = \mathbb{R}^n$, called the *special orthogonal group (of E)*.

The isometries such that $\det(f) = -1$ are called *improper isometries, or improper orthogonal transformations, or flip transformations*.

4.5 QR-Decomposition for Invertible Matrices

Now that we have the definition of an orthogonal matrix, we can explain how the Gram–Schmidt orthonormalization procedure immediately yields the QR-decomposition for matrices.

Lemma 4.5.1 *Given any $n \times n$ real matrix A, if A is invertible then there is an orthogonal matrix Q and an upper triangular matrix R with positive diagonal entries such that $A = QR$.*

Proof. We can view the columns of A as vectors $A_1, \ldots, A_n$ in $\mathbb{E}^n$.

If A is invertible, then they are linearly independent, and we can apply lemma 4.2.7 to produce an orthonormal basis using the Gram–Schmidt orthonormalization procedure.

Recall that we construct vectors Q_k and Q'_k as follows:

$$Q'_1 = A_1, \qquad Q_1 = \frac{Q'_1}{\|Q'_1\|},$$

and for the inductive step

$$Q'_{k+1} = A_{k+1} - \sum_{i=1}^{k} (A_{k+1} \cdot Q_i)\, Q_i, \qquad Q_{k+1} = \frac{Q'_{k+1}}{\left\|Q'_{k+1}\right\|},$$

where $1 \leq k \leq n-1$.

If we express the vectors A_k in terms of the Q_i and Q'_i, we get a triangular system

$$\begin{aligned}
A_1 &= \|Q'_1\|\, Q_1,\\
&\cdots\\
A_j &= (A_j \cdot Q_1)\, Q_1 + \cdots + (A_j \cdot Q_i)\, Q_i + \cdots + (A_j \cdot Q_{j-1})\, Q_{j-1} + \left\|Q'_j\right\| Q_j,\\
&\cdots\\
A_n &= (A_n \cdot Q_1)\, Q_1 + \cdots + (A_n \cdot Q_{n-2})\, Q_{n-2} + (A_n \cdot Q_{n-1})\, Q_{n-1} + \|Q'_n\|\, Q_n.
\end{aligned}$$

Remarks: (1) Because the diagonal entries of R are positive, it can be shown that Q and R are unique.

(2) The QR-decomposition holds even when A is not invertible. In this case, R has some zero on the diagonal. However, a different proof is needed. We will give a nice proof using Householder matrices (see also Strang [**?**]).

Example 6. Consider the matrix

$$A = \begin{pmatrix} 0 & 0 & 5 \\ 0 & 4 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

We leave as an exercise to show that $A = QR$ with

$$Q = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad R = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 4 & 1 \\ 0 & 0 & 5 \end{pmatrix}$$

Another example of QR-decomposition is

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

where

$$Q = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{pmatrix}$$

and

$$R = \begin{pmatrix} \sqrt{2} & 1/\sqrt{2} & \sqrt{2} \\ 0 & 1/\sqrt{2} & \sqrt{2} \\ 0 & 0 & 1 \end{pmatrix}$$

The QR-decomposition yields a rather efficient and numerically stable method for solving systems of linear equations.

Indeed, given a system $Ax = b$, where A is an $n \times n$ invertible matrix, writing $A = QR$, since Q is orthogonal, we get

$$Rx = Q^\top b,$$

and since R is upper triangular, we can solve it by Gaussian elimination, by solving for the last variable x_n first, substituting its value into the system, then solving for x_{n-1}, etc.

The QR-decomposition is also very useful in solving least squares problems (we will come back to this later on), and for finding eigenvalues.

It can be easily adapted to the case where A is a rectangular $m \times n$ matrix with independent columns (thus, $n \leq m$).

In this case, Q is not quite orthogonal. It is an $m \times n$ matrix whose columns are orthogonal, and R is an invertible $n \times n$ upper diagonal matrix with positive diagonal entries. For more on QR, see Strang [**?**].

It should also be said that the Gram–Schmidt orthonormalization procedure that we have presented is not very stable numerically, and instead, one should use the *modified Gram–Schmidt method*.

To compute Q'_{k+1}, instead of projecting A_{k+1} onto $Q_1, \ldots, Q_k$ in a single step, it is better to perform k projections.

We compute $Q^1_{k+1}, Q^2_{k+1}, \ldots, Q^k_{k+1}$ as follows:

$$\begin{aligned} Q^1_{k+1} &= A_{k+1} - (A_{k+1} \cdot Q_1)\, Q_1, \\ Q^{i+1}_{k+1} &= Q^i_{k+1} - (Q^i_{k+1} \cdot Q_{i+1})\, Q_{i+1}, \end{aligned}$$

where $1 \leq i \leq k-1$.

It is easily shown that $Q'_{k+1} = Q^k_{k+1}$. The reader is urged to code this method.

Chapter 5

The Cartan–Dieudonné Theorem

5.1 Orthogonal Reflections

Orthogonal symmetries are a very important example of isometries. First let us review the definition of projections.

Given a vector space E, let F and G be subspaces of E that form a direct sum $E = F \oplus G$.

Since every $u \in E$ can be written uniquely as $u = v + w$, where $v \in F$ and $w \in G$, we can define the two *projections* $p_F \colon E \to F$ and $p_G \colon E \to G$, such that

$$p_F(u) = v \quad \text{and} \quad p_G(u) = w.$$

It is immediately verified that p_G and p_F are linear maps, and that $p_F^2 = p_F$, $p_G^2 = p_G$, $p_F \circ p_G = p_G \circ p_F = 0$, and $p_F + p_G = \mathrm{id}$.

Definition 5.1.1 Given a vector space E, for any two subspaces F and G that form a direct sum $E = F \oplus G$, the *symmetry with respect to F and parallel to G, or reflection about F* is the linear map $s\colon E \to E$, defined such that

$$s(u) = 2p_F(u) - u,$$

for every $u \in E$.

Because $p_F + p_G = \mathrm{id}$, note that we also have

$$s(u) = p_F(u) - p_G(u)$$

and

$$s(u) = u - 2p_G(u),$$

$s^2 = \mathrm{id}$, s is the identity on F, and $s = -\mathrm{id}$ on G.

We now assume that E is a Euclidean space of finite dimension.

Definition 5.1.2 Let E be a Euclidean space of finite dimension n. For any two subspaces F and G, if F and G form a direct sum $E = F \oplus G$ and F and G are orthogonal, i.e. $F = G^{\perp}$, the *orthogonal symmetry with respect to F and parallel to G, or orthogonal reflection about F* is the linear map $s\colon E \to E$, defined such that

$$s(u) = 2p_F(u) - u,$$

for every $u \in E$.

When F is a hyperplane, we call s an *hyperplane symmetry with respect to F (or reflection about F)*, and when G is a plane, we call s a *flip about F*.

It is easy to show that s is an isometry.

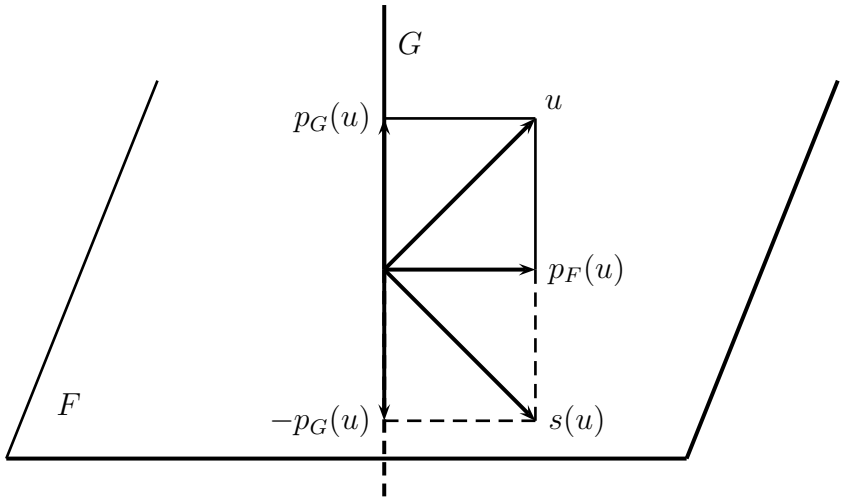

Figure 5.1: A reflection about a hyperplane F

Using lemma 4.2.7, it is possible to find an orthonormal basis $(e_1, \ldots, e_n)$ of E consisting of an orthonormal basis of F and an orthonormal basis of G.

Assume that F has dimension p, so that G has dimension $n - p$.

With respect to the orthonormal basis $(e_1, \ldots, e_n)$, the symmetry s has a matrix of the form

$$\begin{pmatrix} I_p & 0 \\ 0 & -I_{n-p} \end{pmatrix}$$

Thus, $\det(s) = (-1)^{n-p}$, and s is a rotation iff $n - p$ is even.

In particular, when F is a hyperplane H, we have $p = n - 1$, and $n - p = 1$, so that s is an improper orthogonal transformation.

When $F = \{0\}$, we have $s = -\mathrm{id}$, which is called the *symmetry with respect to the origin*. The symmetry with respect to the origin is a rotation iff n is even, and an improper orthogonal transformation iff n is odd.

When n is odd, we observe that every improper orthogonal transformation is the composition of a rotation with the symmetry with respect to the origin.

When G is a plane, $p = n - 2$, and $\det(s) = (-1)^2 = 1$, so that a flip about F is a rotation.

In particular, when $n = 3$, F is a line, and a flip about the line F is indeed a rotation of measure π.

When $F = H$ is a hyperplane, we can give an explicit formula for $s(u)$ in terms of any nonnull vector w orthogonal to H.

We get

$$s(u) = u - 2\,\frac{(u \cdot w)}{\|w\|^2}\,w.$$

Such reflections are represented by matrices called *Householder matrices*, and they play an important role in numerical matrix analysis. Householder matrices are symmetric and orthogonal.

Over an orthonormal basis $(e_1, \ldots, e_n)$, a hyperplane reflection about a hyperplane H orthogonal to a nonnull vector w is represented by the matrix

$$H = I_n - 2\frac{WW^\top}{\|W\|^2} = I_n - 2\frac{WW^\top}{W^\top W},$$

where W is the column vector of the coordinates of w.

Since

$$p_G(u) = \frac{(u \cdot w)}{\|w\|^2}\, w,$$

the matrix representing p_G is

$$\frac{WW^\top}{W^\top W},$$

and since $p_H + p_G = \mathrm{id}$, the matrix representing p_H is

$$I_n - \frac{WW^\top}{W^\top W}.$$

The following fact is the key to the proof that every isometry can be decomposed as a product of reflections.

Lemma 5.1.3 *Let E be any nontrivial Euclidean space. For any two vectors $u, v \in E$, if $\|u\| = \|v\|$, then there is an hyperplane H such that the reflection s about H maps u to v, and if $u \neq v$, then this reflection is unique.*

5.2 The Cartan–Dieudonné Theorem for Linear Isometries

The fact that the group $\mathbf{O}(n)$ of linear isometries is generated by the reflections is a special case of a theorem known as the Cartan–Dieudonné theorem.

Elie Cartan proved a version of this theorem early in the twentieth century. A proof can be found in his book on spinors [**?**], which appeared in 1937 (Chapter I, Section 10, pages 10–12).

Cartan's version applies to nondegenerate quadratic forms over $\mathbb{R}$ or $\mathbb{C}$. The theorem was generalized to quadratic forms over arbitrary fields by Dieudonné [**?**].

One should also consult Emil Artin's book [**?**], which contains an in-depth study of the orthogonal group and another proof of the Cartan–Dieudonné theorem.

First, let us recall the notions of eigenvalues and eigenvectors.

Recall that given any linear map $f\colon E \to E$, a vector $u \in E$ is called an *eigenvector, or proper vector, or characteristic vector of f* iff there is some $\lambda \in K$ such that

$$f(u) = \lambda u.$$

In this case, we say that $u \in E$ is an *eigenvector associated with λ*.

A scalar $\lambda \in K$ is called an *eigenvalue, or proper value, or characteristic value of f* iff there is some nonnull vector $u \neq 0$ in E such that

$$f(u) = \lambda u,$$

or equivalently if $\mathrm{Ker}\,(f - \lambda \mathrm{id}) \neq \{0\}$.

Given any scalar $\lambda \in K$, the set of all eigenvectors associated with λ is the subspace $\mathrm{Ker}\,(f - \lambda \mathrm{id})$, also denoted as $E_\lambda(f)$ or $E(\lambda, f)$, called the *eigenspace associated with λ, or proper subspace associated with λ*.

Theorem 5.2.1 *Let E be a Euclidean space of dimension $n \geq 1$. Every isometry $f \in \mathbf{O}(E)$ which is not the identity is the composition of at most n reflections. For $n \geq 2$, the identity is the composition of any reflection with itself.*

Remarks.

(1) The proof of theorem 5.2.1 shows more than stated.

If 1 is an eigenvalue of f, for any eigenvector w associated with 1 (i.e., $f(w) = w$, $w \neq 0$), then f is the composition of $k \leq n-1$ reflections about hyperplanes F_i, such that $F_i = H_i \oplus L$, where L is the line $\mathbb{R}w$, and the H_i are subspaces of dimension $n-2$ all orthogonal to L.

If 1 is not an eigenvalue of f, then f is the composition of $k \leq n$ reflections about hyperplanes $H, F_1, \ldots, F_{k-1}$, such that $F_i = H_i \oplus L$, where L is a line intersecting H, and the H_i are subspaces of dimension $n-2$ all orthogonal to L.

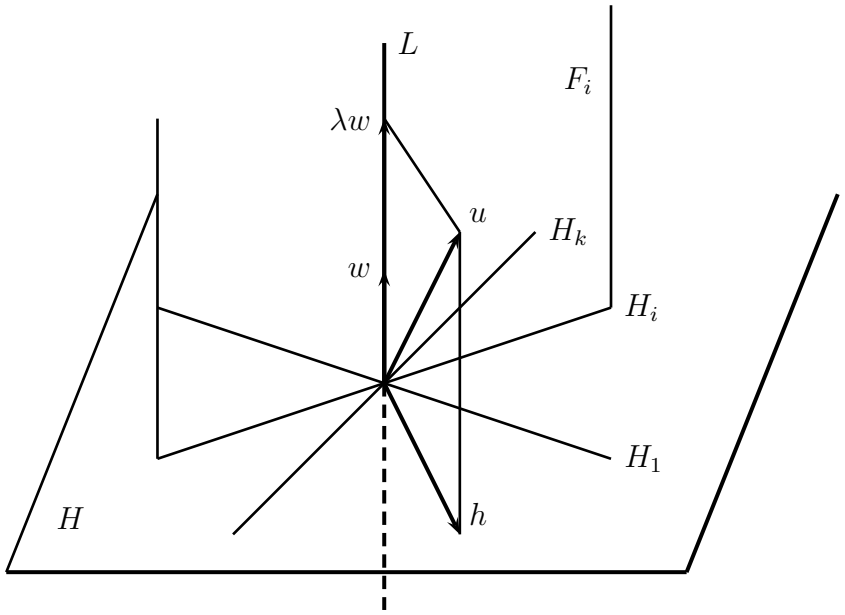

Figure 5.2: An Isometry f as a composition of reflections, when 1 is an eigenvalue of f

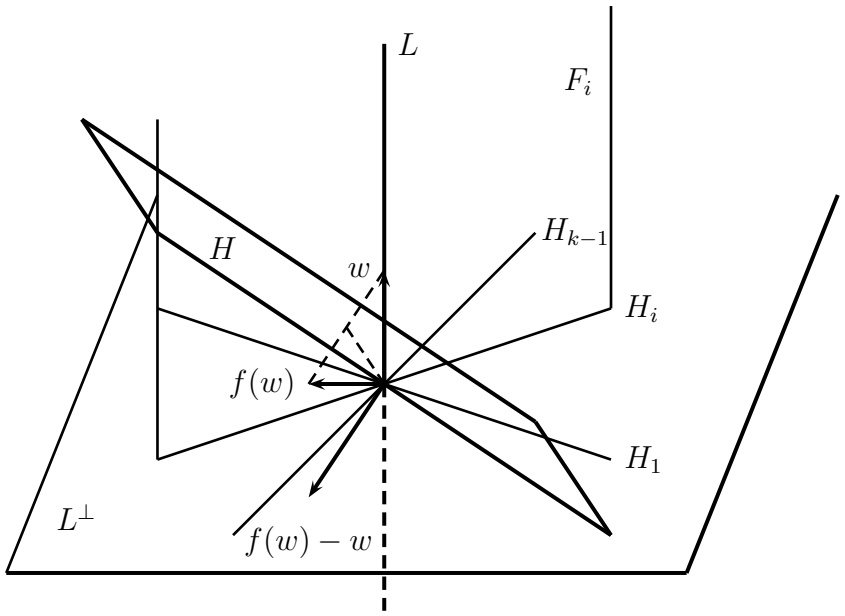

Figure 5.3: An Isometry f as a composition of reflections, when 1 is not an eigenvalue of f

(2) It is natural to ask what is the minimal number of hyperplane reflections needed to obtain an isometry f.

This has to do with the dimension of the eigenspace $\mathrm{Ker}\,(f - \mathrm{id})$ associated with the eigenvalue 1.

We will prove later that every isometry is the composition of k hyperplane reflections, where

$$k = n - \dim(\mathrm{Ker}\,(f - \mathrm{id})),$$

and that this number is minimal (where $n = \dim(E)$).

When $n = 2$, a reflection is a reflection about a line, and theorem 5.2.1 shows that every isometry in $\mathbf{O}(2)$ is either a reflection about a line or a rotation, and that every rotation is the product of two reflections about some lines.

In general, since $\det(s) = -1$ for a reflection s, when $n \geq 3$ is odd, every rotation is the product of an even number $\leq n-1$ of reflections, and when n is even, every improper orthogonal transformation is the product of an odd number $\leq n-1$ of reflections.

In particular, for $n = 3$, every rotation is the product of two reflections about planes.

If E is a Euclidean space of finite dimension and $f\colon E \to E$ is an isometry, if λ is any eigenvalue of f and u is an eigenvector associated with λ, then

$$\|f(u)\| = \|\lambda u\| = |\lambda| \, \|u\| = \|u\| \,,$$

which implies $|\lambda| = 1$, since $u \neq 0$.

Thus, the real eigenvalues of an isometry are either $+1$ or -1.

When n is odd, we can say more about improper isometries. This is because they admit -1 as an eigenvalue. When n is odd, an improper isometry is the composition of a reflection about a hyperplane H with a rotation consisting of reflections about hyperplanes $F_1, \ldots, F_{k-1}$ containing a line, L, orthogonal to H.

Lemma 5.2.2 *Let E be a Euclidean space of finite dimension n, and let $f: E \to E$ be an isometry. For any subspace F of E, if $f(F) = F$, then $f(F^\perp) \subseteq F^\perp$ and $E = F \oplus F^\perp$.*

Lemma 5.2.2 is the starting point of the proof that every orthogonal matrix can be diagonalized over the field of complex numbers.

Indeed, if λ is any eigenvalue of f, then
$f(E_\lambda(f)) = E_\lambda(f)$, and thus the orthogonal $E_\lambda(f)^\perp$ is closed under f, and

$$E = E_\lambda(f) \oplus E_\lambda(f)^\perp.$$

The problem over $\mathbb{R}$ is that there may not be any real eigenvalues.

However, when n is odd, the following lemma shows that every rotation admits 1 as an eigenvalue (and similarly, when n is even, every improper orthogonal transformation admits 1 as an eigenvalue).

Lemma 5.2.3 *Let E be a Euclidean space.*

(1) If E has odd dimension $n = 2m + 1$, then every rotation f admits 1 as an eigenvalue and the eigenspace F of all eigenvectors left invariant under f has an odd dimension $2p + 1$. Furthermore, there is an orthonormal basis of E, in which f is represented by a matrix of the form

$$\begin{pmatrix} R_{2(m-p)} & 0 \\ 0 & I_{2p+1} \end{pmatrix}$$

where $R_{2(m-p)}$ is a rotation matrix that does not have 1 as an eigenvalue.

(2) If E has even dimension $n = 2m$, then every improper orthogonal transformation f admits 1 as an eigenvalue and the eigenspace F of all eigenvectors left invariant under f has an odd dimension $2p+1$. Furthermore, there is an orthonormal basis of E, in which f is represented by a matrix of the form

$$\begin{pmatrix} S_{2(m-p)-1} & 0 \\ 0 & I_{2p+1} \end{pmatrix}$$

where $S_{2(m-p)-1}$ is an improper orthogonal matrix that does not have 1 as an eigenvalue.

An example showing that lemma 5.2.3 fails for n even is the following rotation matrix (when $n = 2$):

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

The above matrix does not have real eigenvalues if $\theta \neq k\pi$.

It is easily shown that for $n = 2$, with respect to any chosen orthonormal basis (e_1, e_2), every rotation is represented by a matrix of form

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

where $\theta \in [0, 2\pi[$, and that every improper orthogonal transformation is represented by a matrix of the form

$$S = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$$

In the first case, we call $\theta \in [0, 2\pi[$ the *measure* of the angle of rotation of R w.r.t. the orthonormal basis (e_1, e_2).

In the second case, we have a reflection about a line, and it is easy to determine what this line is. It is also easy to see that S is the composition of a reflection about the x-axis with a rotation (of matrix R).

We refrained from calling θ "the angle of rotation", because there are some subtleties involved in defining rigorously the notion of angle of two vectors (or two lines).

For example, note that with respect to the "opposite basis" (e_2, e_1), the measure θ must be changed to $2\pi - \theta$ (or $-\theta$ if we consider the quotient set $\mathbb{R}/2\pi$ of the real numbers modulo 2π).

We will come back to this point after having defined the notion of orientation (see Section 5.8).

It is easily shown that the group $\mathbf{SO}(2)$ of rotations in the plane is abelian.

We can perform the following calculation, using some elementary trigonometry:

$$\begin{pmatrix} \cos\varphi & \sin\varphi \\ \sin\varphi & -\cos\varphi \end{pmatrix} \begin{pmatrix} \cos\psi & \sin\psi \\ \sin\psi & -\cos\psi \end{pmatrix} = \begin{pmatrix} \cos(\varphi+\psi) & \sin(\varphi+\psi) \\ \sin(\varphi+\psi) & -\cos(\varphi+\psi) \end{pmatrix}.$$

The above also shows that the inverse of a rotation matrix

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is obtained by changing θ to $-\theta$ (or $2\pi - \theta$).

Incidently, note that in writing a rotation r as the product of two reflections $r = s_2 s_1$, the first reflection s_1 can be chosen arbitrarily, since $s_1^2 = \mathrm{id}$, $r = (rs_1)s_1$, and rs_1 is a reflection.

For $n = 3$, the only two choices for p are $p = 1$, which corresponds to the identity, or $p = 0$, in which case, f is a rotation leaving a line invariant.

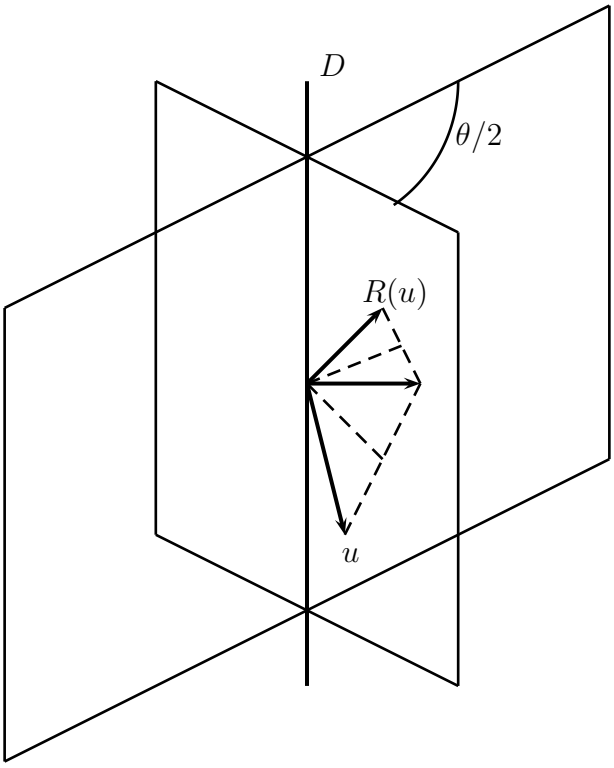

Figure 5.4: $3D$ rotation as the composition of two reflections

This line is called the *axis of rotation*. The rotation R behaves like a two dimentional rotation around the axis of rotation.

The measure of the angle of rotation θ can be determined through its cosine via the formula

$$\cos\theta = u \cdot R(u),$$

where u is any unit vector orthogonal to the direction of the axis of rotation.

However, this does not determine $\theta \in [0, 2\pi[$ uniquely, since both θ and $2\pi - \theta$ are possible candidates.

What is missing is an orientation of the plane (through the origin) orthogonal to the axis of rotation. We will come back to this point in Section 5.8.

In the orthonormal basis of the lemma, a rotation is represented by a matrix of the form

$$R = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Remark: For an arbitrary rotation matrix A, since

$$a_{1\,1} + a_{2\,2} + a_{3\,3}$$

(the *trace* of A) is the sum of the eigenvalues of A, and since these eigenvalues are $\cos\theta + i\sin\theta$, $\cos\theta - i\sin\theta$, and 1, for some $\theta \in [0, 2\pi[$, we can compute $\cos\theta$ from

$$1 + 2\cos\theta = a_{1\,1} + a_{2\,2} + a_{3\,3}.$$

It is also possible to determine the axis of rotation (see the problems).

An improper transformation is either a reflection about a plane, or the product of three reflections, or equivalently the product of a reflection about a plane with a rotation, and a closer look at theorem 5.2.1 shows that the axis of rotation is orthogonal to the plane of the reflection.

Thus, an improper transformation is represented by a matrix of the form

$$S = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

When $n \geq 3$, the group of rotations $\mathbf{SO}(n)$ is not only generated by hyperplane reflections, but also by flips (about subspaces of dimension $n-2$).

We will also see in Section 5.4 that every proper affine rigid motion can be expressed as the composition of at most n flips, which is perhaps even more surprising!

The proof of these results uses the following key lemma.

Lemma 5.2.4 *Given any Euclidean space E of dimension $n \geq 3$, for any two reflections h_1 and h_2 about some hyperplanes H_1 and H_2, there exist two flips f_1 and f_2 such that $h_2 \circ h_1 = f_2 \circ f_1$.*

Using lemma 5.2.4 and the Cartan-Dieudonné theorem, we obtain the following characterization of rotations when $n \geq 3$.

Theorem 5.2.5 *Let E be a Euclidean space of dimension $n \geq 3$. Every rotation $f \in \mathbf{SO}(E)$ is the composition of an even number of flips $f = f_{2k} \circ \cdots \circ f_1$, where $2k \leq n$. Furthermore, if $u \neq 0$ is invariant under f (i.e. $u \in \mathrm{Ker}\,(f - \mathrm{id})$), we can pick the last flip f_{2k} such that $u \in F_{2k}^{\perp}$, where F_{2k} is the subspace of dimension $n - 2$ determining f_{2k}.*

Remarks:

(1) It is easy to prove that if f is a rotation in $\mathbf{SO}(3)$, if D is its axis and θ is its angle of rotation, then f is the composition of two flips about lines D_1 and D_2 orthogonal to D and making an angle $\theta/2$.

(2) It is natural to ask what is the minimal number of flips needed to obtain a rotation f (when $n \geq 3$). As for arbitrary isometries, we will prove later that every rotation is the composition of k flips, where

$$k = n - \dim(\mathrm{Ker}\,(f - \mathrm{id})),$$

and that this number is minimal (where $n = \dim(E)$).

Hyperplane reflections can be used to obtain another proof of the QR-decomposition.

5.3 QR-Decomposition Using Householder Matrices

First, we state the result geometrically. When translated in terms of Householder matrices, we obtain the fact advertised earlier that every matrix (not necessarily invertible) has a QR-decomposition.

Lemma 5.3.1 *Let E be a nontrivial Euclidean space of dimension n. Given any orthonormal basis $(e_1, \ldots, e_n)$, for any n-tuple of vectors $(v_1, \ldots, v_n)$, there is a sequence of n isometries $h_1, \ldots, h_n$, such that h_i is a hyperplane reflection or the identity, and if $(r_1, \ldots, r_n)$ are the vectors given by*

$$r_j = h_n \circ \cdots \circ h_2 \circ h_1(v_j),$$

then every r_j is a linear combination of the vectors $(e_1, \ldots, e_j)$, $(1 \leq j \leq n)$. Equivalently, the matrix R whose columns are the components of the r_j over the basis $(e_1, \ldots, e_n)$ is an upper triangular matrix. Furthermore, the h_i can be chosen so that the diagonal entries of R are nonnegative.

Remarks. (1) Since every h_i is a hyperplane reflection or the identity,

$$\rho = h_n \circ \cdots \circ h_2 \circ h_1$$

is an isometry.

(2) If we allow negative diagonal entries in R, the last isometry h_n may be omitted.

(3) Instead of picking $r_{k,k} = \|u''_k\|$, which means that

$$w_k = r_{k,k}\, e_k - u''_k,$$

where $1 \leq k \leq n$, it might be preferable to pick $r_{k,k} = -\|u''_k\|$ if this makes $\|w_k\|^2$ larger, in which case

$$w_k = r_{k,k}\, e_k + u''_k.$$

Indeed, since the definition of h_k involves division by $\|w_k\|^2$, it is desirable to avoid division by very small numbers.

Lemma 5.3.1 immediately yields the QR-decomposition in terms of Householder transformations.

Lemma 5.3.2 *For every real $n \times n$-matrix A, there is a sequence $H_1, \ldots, H_n$ of matrices, where each H_i is either a Householder matrix or the identity, and an upper triangular matrix R, such that*

$$R = H_n \cdots H_2 H_1 A.$$

As a corollary, there is a pair of matrices Q, R, where Q is orthogonal and R is upper triangular, such that $A = QR$ (a QR-decomposition of A). Furthermore, R can be chosen so that its diagonal entries are non-negative.

Remarks. (1) Letting

$$A_{k+1} = H_k \cdots H_2 H_1 A,$$

with $A_1 = A$, $1 \leq k \leq n$, the proof of lemma 5.3.1 can be interpreted in terms of the computation of the sequence of matrices $A_1, \ldots, A_{n+1} = R$.

The matrix A_{k+1} has the shape

$$A_{k+1} = \begin{pmatrix} \times & \times & \times & u_1^{k+1} & \times & \times & \times & \times \\ 0 & \times & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \times & u_k^{k+1} & \times & \times & \times & \times \\ 0 & 0 & 0 & u_{k+1}^{k+1} & \times & \times & \times & \times \\ 0 & 0 & 0 & u_{k+2}^{k+1} & \times & \times & \times & \times \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & u_{n-1}^{k+1} & \times & \times & \times & \times \\ 0 & 0 & 0 & u_n^{k+1} & \times & \times & \times & \times \end{pmatrix}$$

where the $(k+1)$th column of the matrix is the vector

$$u_{k+1} = h_k \circ \cdots \circ h_2 \circ h_1(v_{k+1}),$$

and thus

$$u'_{k+1} = (u_1^{k+1}, \ldots, u_k^{k+1}),$$

and

$$u''_{k+1} = (u_{k+1}^{k+1}, u_{k+2}^{k+1}, \ldots, u_n^{k+1}).$$

If the last $n-k-1$ entries in column $k+1$ are all zero, there is nothing to do and we let $H_{k+1} = I$.

Otherwise, we kill these $n-k-1$ entries by multiplying A_{k+1} on the left by the Householder matrix H_{k+1} sending $(0,\ldots,0,u_{k+1}^{k+1},\ldots,u_n^{k+1})$ to $(0,\ldots,0,r_{k+1,k+1},0,\ldots,0)$, where

$$r_{k+1,k+1} = \left\|(u_{k+1}^{k+1},\ldots,u_n^{k+1})\right\|.$$

(2) If we allow negative diagonal entries in R, the matrix H_n may be omitted ($H_n = I$).

(3) If A is invertible and the diagonal entries of R are positive, it can be shown that Q and R are unique.

(4) The method allows the computation of the determinant of A. We have

$$\det(A) = (-1)^m r_{1,1} \cdots r_{n,n},$$

where m is the number of Householder matrices (not the identity) among the H_i.

(5) The "condition number" of the matrix A is preserved (see Strang [**?**]). This is very good for numerical stability.

We conclude our discussion of isometries with a brief discussion of affine isometries.

5.4 Affine Isometries (Rigid Motions)

Definition 5.4.1 Given any two nontrivial Euclidean affine spaces E and F of the same finite dimension n, a function $f: E \to F$ is *an affine isometry (or rigid map)* iff it is an affine map and

$$\|\mathbf{f(a)f(b)}\| = \|\mathbf{ab}\|,$$

for all $a, b \in E$. When $E = F$, an affine isometry $f: E \to E$ is also called a *rigid motion*.

Thus, an affine isometry is an affine map that preserves the distance. This is a rather strong requirement.

In fact, we will show that for any function $f: E \to F$, the assumption that

$$\|\mathbf{f(a)f(b)}\| = \|\mathbf{ab}\|$$

for all $a, b \in E$, forces f to be an affine map.

Remark: Sometimes, an affine isometry is defined as a *bijective* affine isometry. When E and F are of finite dimension, the definitions are equivalent.

Lemma 5.4.2 *Given any two nontrivial Euclidean affine spaces E and F of the same finite dimension n, an affine map $f: E \to F$ is an affine isometry iff its associated linear map $\overrightarrow{f}: \overrightarrow{E} \to \overrightarrow{F}$ is an isometry. An affine isometry is a bijection.*

Let us now consider affine isometries $f: E \to E$. If $\overrightarrow{f}$ is a rotation, we call f a *proper (or direct) affine isometry*, and if $\overrightarrow{f}$ is a an improper linear isometry, we call f a *an improper (or skew) affine isometry*.

It is easily shown that the set of affine isometries $f: E \to E$ forms a group denoted as $\mathbf{Is}(E)$ (or $\mathbf{Mo}(E)$), and those for which $\overrightarrow{f}$ is a rotation is a subgroup denoted as $\mathbf{SE}(E)$.

The translations are the affine isometries f for which $\overrightarrow{f} = \mathrm{id}$, the identity map on $\overrightarrow{E}$.

The following lemma is the counterpart of lemma 4.3.2 for isometries between Euclidean vector spaces:

Lemma 5.4.3 *Given any two nontrivial Euclidean affine spaces E and F of the same finite dimension n, for every function $f: E \to F$, the following properties are equivalent:*

(1) f is an affine map and $\|\mathbf{f(a)f(b)}\| = \|\mathbf{ab}\|$, for all $a, b \in E$.

(2) $\|\mathbf{f(a)f(b)}\| = \|\mathbf{ab}\|$, for all $a, b \in E$.

In order to understand the structure of affine isometries, it is important to investigate the fixed points of an affine map.

5.5 Fixed Points of Affine Maps

Recall that $E(1, \overrightarrow{f})$ denotes the eigenspace of the linear map $\overrightarrow{f}$ associated with the scalar 1, that is, the subspace consisting of all vectors $u \in \overrightarrow{E}$ such that $\overrightarrow{f}(u) = u$.

Clearly, $\mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id}) = E(1, \overrightarrow{f})$.

Given some origin $\Omega \in E$, since

$$f(a) = f(\Omega + \mathbf{\Omega a}) = f(\Omega) + \overrightarrow{f}(\mathbf{\Omega a}),$$

we get

$$\mathbf{\Omega f(a)} - \mathbf{\Omega a} = \mathbf{\Omega f(\Omega)} + \overrightarrow{f}(\mathbf{\Omega a}) - \mathbf{\Omega a}.$$

Using this, we show the following lemma which holds for arbitrary affine spaces of finite dimension and for arbitrary affine maps.

Lemma 5.5.1 *Let E be any affine space of finite dimension. For every affine map $f\colon E \to E$, let $Fix(f) = \{a \in E \mid f(a) = a\}$ be the set of fixed points of f. The following properties hold.*

(1) If f has some fixed point a, so that $Fix(f) \neq \emptyset$, then $Fix(f)$ is an affine subspace of E such that

$$Fix(f) = a + E(1, \overrightarrow{f}) = a + \mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id}),$$

where $E(1, \overrightarrow{f})$ is the eigenspace of the linear map $\overrightarrow{f}$ for the eigenvalue 1.

(2) The affine map f has a unique fixed point iff

$$E(1, \overrightarrow{f}) = \mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id}) = \{0\}.$$

Remark: The fact that E has finite dimension is only used to prove (2), and (1) holds in general.

If an isometry f leaves some point fixed, we can take such a point Ω as the origin, and then $f(\Omega) = \Omega$ and we can view f as a rotation or an improper orthogonal transformation, depending on the nature of $\overrightarrow{f}$.

Note that it is quite possible that $Fix(f) = \emptyset$. For example, nontrivial translations have no fixed points.

A more interesting example is provided the composition of a plane reflection about a line composed with a a nontrivial translation parallel to this line.

Otherwise, we will see in lemma 5.6.2 that every affine isometry is the (commutative) composition of a translation with an isometry that always has a fixed point.

5.6 Affine Isometries and Fixed Points

Given any two affine subspaces F, G of E such that $\overrightarrow{F}$ and $\overrightarrow{G}$ are orthogonal subspaces of $\overrightarrow{E}$ such that $\overrightarrow{E} = \overrightarrow{F} \oplus \overrightarrow{G}$, for any point $\Omega \in F$, we define $q\colon E \to \overrightarrow{G}$, such that

$$q(a) = p_{\overrightarrow{G}}(\mathbf{\Omega a}).$$

Note that $q(a)$ is independent of the choice of $\Omega \in F$.

Then, the map $g\colon E \to E$ such that $g(a) = a - 2q(a)$, or equivalently

$$\mathbf{ag(a)} = -2q(a) = -2p_{\overrightarrow{G}}(\mathbf{\Omega a})$$

does not depend on the choice of $\Omega \in F$.

If we identify E to $\overrightarrow{E}$ by choosing any origin Ω in F, we note that g is identified with the symmetry with respect to $\overrightarrow{F}$ and parallel to $\overrightarrow{G}$.

Thus, the map g is an affine isometry, and it is called the *orthogonal symmetry about* F.

Since

$$g(a) = \Omega + \mathbf{\Omega a} - 2p_{\overrightarrow{G}}(\mathbf{\Omega a})$$

for all $\Omega \in F$ and for all $a \in E$, we note that the linear map $\overrightarrow{g}$ associated with g is the (linear) symmetry about the subspace $\overrightarrow{F}$ (the direction of F)

The following amusing lemma shows the extra power afforded by affine orthogonal symmetries: Translations are subsumed!

Lemma 5.6.1 *Given any affine space E, if $f\colon E \to E$ and $g\colon E \to E$ are orthogonal symmetries about parallel affine subspaces F_1 and F_2, then $g \circ f$ is a translation defined by the vector $2\mathbf{ab}$, where $\mathbf{ab}$ is any vector perpendicular to the common direction $\overrightarrow{F}$ of F_1 and F_2 such that $\|\mathbf{ab}\|$ is the distance between F_1 and F_2, with $a \in F_1$ and $b \in F_2$. Conversely, every translation by a vector τ is obtained as the composition of two orthogonal symmetries about parallel affine subspaces F_1 and F_2 whose common direction is orthogonal to $\tau = \mathbf{ab}$, for some $a \in F_1$ and some $b \in F_2$ such that the distance betwen F_1 and F_2 is $\|\mathbf{ab}\|/2$.*

The following result is a generalization of Chasles' theorem about the rigid motions in $\mathbb{R}^3$.

Lemma 5.6.2 *Let E be a Euclidean affine space of finite dimension n. For every affine isometry $f: E \to E$, there is a unique isometry $g: E \to E$ and a unique translation $t = t_\tau$, with $\overrightarrow{f}(\tau) = \tau$ (i.e., $\tau \in \mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id})$), such that the set*

$$Fix(g) = \{a \in E \mid g(a) = a\}$$

of fixed points of g is a nonempty affine subspace of E of direction

$$\overrightarrow{G} = \mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id}) = E(1, \overrightarrow{f}),$$

and such that

$$f = t \circ g \quad \textit{and} \quad t \circ g = g \circ t.$$

Furthermore, we have the following additional properties:

(a) $f = g$ *and* $\tau = 0$ *iff* f *has some fixed point, i.e., iff* $Fix(f) \neq \emptyset$.

(b) If f *has no fixed points, i.e.,* $Fix(f) = \emptyset$, *then* $\dim(\mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id})) \geq 1$.

The proof rests on the following two key facts:

(1) If we can find some $x \in E$ such that $\mathbf{xf}(\mathbf{x}) = \tau$ belongs to $\mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id})$, we get the existence of g and τ.

(2) $\overrightarrow{E} = \mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id}) \oplus \mathrm{Im}\,(\overrightarrow{f} - \mathrm{id})$, and $\mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id})$ and $\mathrm{Im}\,(\overrightarrow{f} - \mathrm{id})$ are orthogonal. This implies the uniqueness of g and τ.

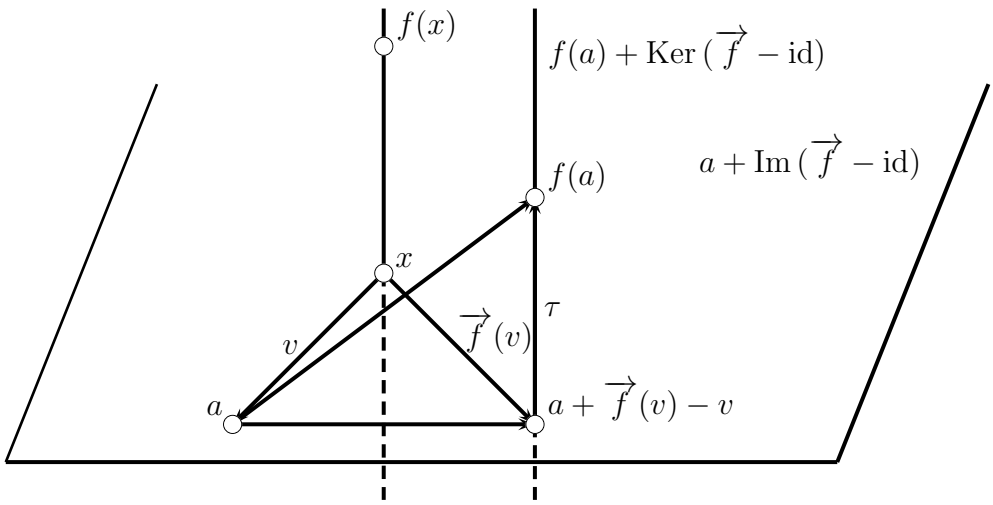

Figure 5.5: Rigid motion as $f = t \circ g$, where g has some fixed point x

Remarks. (1) Note that $\mathrm{Ker}\,(\overrightarrow{f} - \mathrm{id}) = \{0\}$ iff $\tau = 0$, iff $Fix(g)$ consists of a single element, which is the unique fixed point of f.

However, even if f is not a translation, f may not have any fixed points.

(2) The fact that E has finite dimension is only used to prove (b).

(3) It is easily checked that $Fix(g)$ consists of the set of points x such that $\|\mathbf{xf}(\mathbf{x})\|$ is minimal.

In the affine Euclidean plane, it is easy to see that the affine isometries are classified as follows.

An isometry f which has a fixed point is a rotation if it is a direct isometry, else a reflection about a line.

If f has no fixed point, then either it is a nontrivial translation or the composition of a reflection about a line with a nontrivial translation parallel to this line.

In an affine space of dimension 3, it is easy to see that the affine isometries are classified as follows.

A proper isometry with a fixed point is a rotation around a line D (its set of fixed points), as illustrated in figure 5.6.

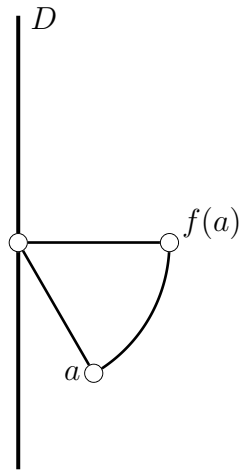

Figure 5.6: $3D$ proper rigid motion with line D of fixed points (rotation)

An improper isometry with a fixed point is either a reflection about a plane H (the set of fixed points), or the composition of a rotation followed by a reflection about a plane H orthogonal to the axis of rotation D, as illustrated in figures 5.7 and 5.8. In the second case, there is a single fixed point $O = D \cap H$.

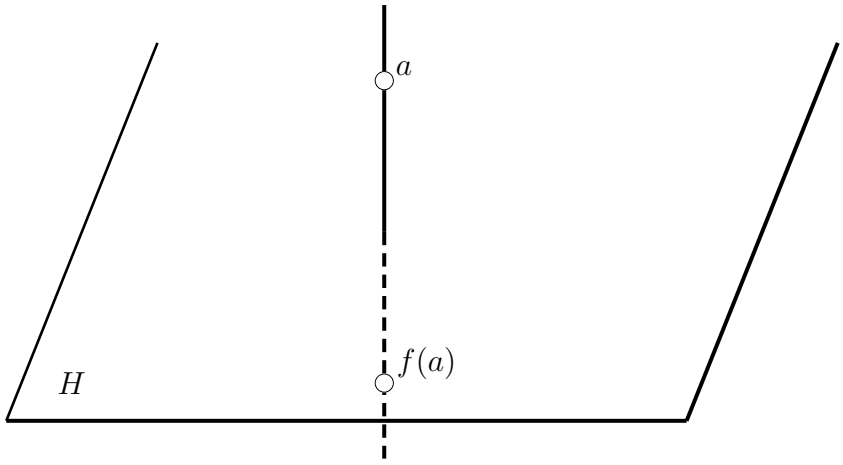

Figure 5.7: $3D$ improper rigid motion with a plane H of fixed points (reflection)

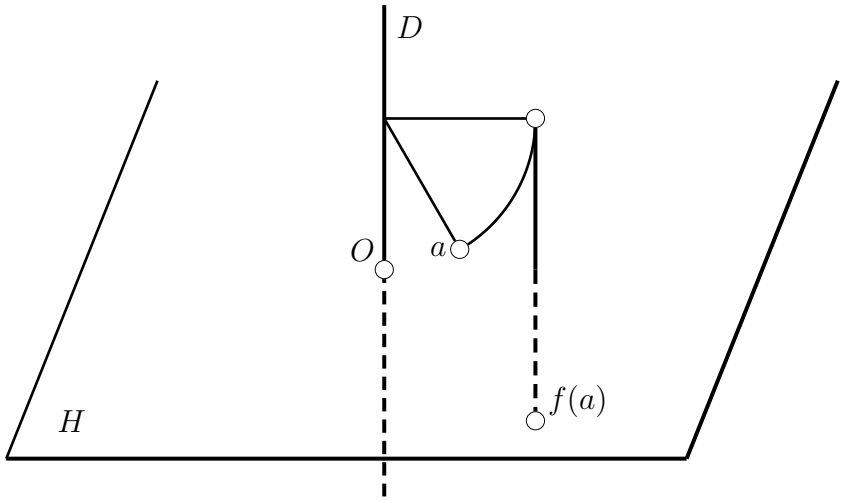

Figure 5.8: $3D$ improper rigid motion with a unique fixed point

There are three types of isometries with no fixed point. The first kind is a nontrivial translation. The second kind is the composition of a rotation followed by a nontrivial translation parallel to the axis of rotation D. Such a rigid motion is proper, and is called a *screw motion*. A screw motion is illustrated in figure 5.9.

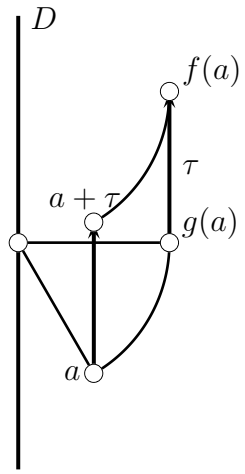

Figure 5.9: $3D$ proper rigid motion with no fixed point (screw motion)

The third kind is the composition of a reflection about a plane followed by a nontrivial translation by a vector parallel to the direction of the plane of the reflection, as illustrated in figure 5.10. It is an improper isometry.

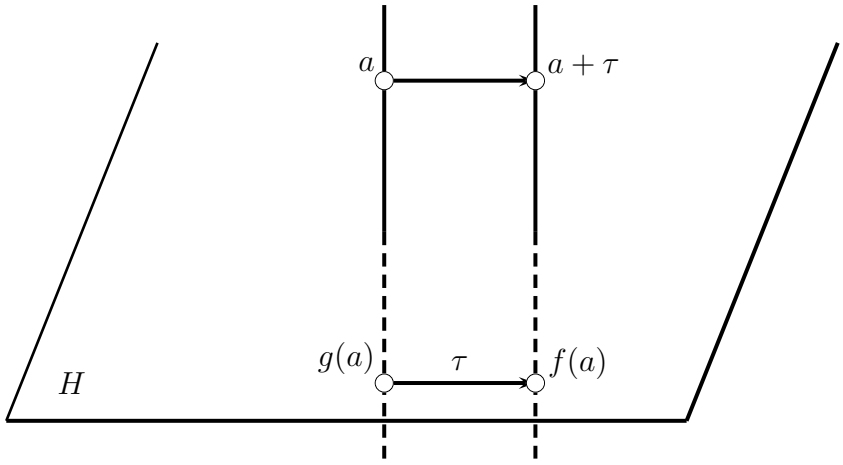

Figure 5.10: $3D$ improper rigid motion with no fixed points

The Cartan-Dieudonné also holds for affine isometries, with a small twist due to translations.

5.7 The Cartan–Dieudonné Theorem for Affine Isometries

Theorem 5.7.1 *Let E be an affine Euclidean space of dimension $n \geq 1$. Every isometry $f \in \mathbf{Is}(E)$ which has a fixed point and is not the identity is the composition of at most n reflections. Every isometry $f \in \mathbf{Is}(E)$ which has no fixed point is the composition of at most $n+2$ reflections. For $n \geq 2$, the identity is the composition of any reflection with itself.*

When $n \geq 3$, we can also characterize the affine isometries in $\mathbf{SE}(n)$ in terms of flips.

Remarkably, not only we can do without translations, but we can even bound the number of flips by n.

Theorem 5.7.2 *Let E be a Euclidean affine space of dimension $n \geq 3$. Every rigid motion $f \in \mathbf{SE}(E)$ is the composition of an even number of flips $f = f_{2k} \circ \cdots \circ f_1$, where $2k \leq n$.*

Remark. It is easy to prove that if f is a screw motion in $\mathbf{SE}(3)$, if D is its axis, θ is its angle of rotation, and τ is the translation along the direction of D, then f is the composition of two flips about lines D_1 and D_2 orthogonal to D, at a distance $\|\tau\|/2$, and making an angle $\theta/2$.

There is one more topic that we would like to cover since it is often useful in practice, the concept of *cross-product of vectors*, also called vector-product. But first, we need to discuss the question of orientation of bases.

5.8 Orientations of a Euclidean Space, Angles

In order to deal with the notion of orientation correctly, it is important to assume that every family $(u_1, \ldots, u_n)$ of vectors is ordered (by the natural ordering on $\{1, 2, \ldots, n\}$).

We will assume that all families $(u_1, \ldots, u_n)$ of vectors, in particular, bases and orthonormal bases are ordered.

Let E be a vector space of finite dimension n over $\mathbb{R}$, and let $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ be any two bases for E.

Recall that the change of basis matrix from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$ is the matrix P whose columns are the coordinates of the vectors v_j over the basis $(u_1, \ldots, u_n)$.

It is immediately verified that the set of alternating n-linear forms on E is a vector space that we denote as $\Lambda(E)$ (see Lang [**?**]).

It is easy to show that $\Lambda(E)$ has dimension 1.

We now define an equivalence relation on $\Lambda(E) - \{0\}$ (where we let 0 denote the null alternating n-linear form):

φ and ψ are equivalent iff $\psi = \lambda\varphi$ for some $\lambda > 0$.

It is immediately verified that the above relation is an equivalence relation. Furthermore, it has exactly two equivalence classes O_1 and O_2.

The first way of defining an *orientation of* E is to pick one of these two equivalence classes, say O ($O \in \{O_1, O_2\}$).

Given such a choice of a class O, we say that a basis $(w_1, \ldots, w_n)$ has *positive orientation* iff

$$\varphi(w_1, \ldots, w_n) > 0$$

for any alternating n-linear form $\varphi \in O$.

Note that this makes sense, since for any other $\psi \in O$, $\varphi = \lambda\psi$ for some $\lambda > 0$.

According to the previous definition, two bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ have the same orientation iff $\varphi(u_1, \ldots, u_n)$ and $\varphi(v_1, \ldots, v_n)$ have the same sign for all $\varphi \in \Lambda(E) - \{0\}$.

From

$$\varphi(v_1, \ldots, v_n) = \det(P)\varphi(u_1, \ldots, u_n),$$

we must have $\det(P) > 0$.

Conversely, if $\det(P) > 0$, the same argument shows that $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ have the same orientation.

This leads us to an equivalent and slightly less contorted definition of the notion of orientation. We define a relation between bases of E as follows:

Two bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ are related iff $\det(P) > 0$, where P is the change of basis matrix from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$.

Since $\det(PQ) = \det(P)\det(Q)$, and since change of basis matrices are invertible, the relation just defined is indeed an equivalence relation, and it has two equivalence classes.

Furthermore, from the discussion above, any nonnull alternating n-linear form φ will have the same sign on any two equivalent bases.

The above discussion motivates the following definition.

Definition 5.8.1 Given any vector space E of finite dimension over $\mathbb{R}$, we define an *orientation of* E as the choice of one of the two equivalence classes of the equivalence relation on the set of bases defined such that $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ have the same orientation iff $\det(P) > 0$, where P is the change of basis matrix from $(u_1, \ldots, u_n)$ to $(v_1, \ldots, v_n)$. A basis in the chosen class is said to have *positive orientation, or to be positive*. An *orientation of a Euclidean affine space* E is an orientation of its underlying vector space $\overrightarrow{E}$

In practice, to give an orientation, one simply picks a fixed basis considered as having positive orientation. The orientation of every other basis is determined by the sign of the determinant of the change of basis matrix.

Having the notation of orientation at hand, we wish to go back briefly to the concept of (oriented) angle.

Let E be a Euclidean space of dimension $n = 2$, and assume a given orientation. In any given positive orthonormal basis for E, every rotation r is represented by a matrix

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Actually, we claim that the matrix R representing the rotation r is the same in *all* orthonormal positive bases.

This is because the change of basis matrix from one positive orthonormal basis to another positive orthonormal basis is a rotation represented by some matrix of the form

$$P = \begin{pmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{pmatrix}$$

and that we have

$$P^{-1} = \begin{pmatrix} \cos(-\psi) & -\sin(-\psi) \\ \sin(-\psi) & \cos(-\psi) \end{pmatrix}$$

and after calculations, we find that PRP^{-1} is the rotation matrix associated with $\psi + \theta - \psi = \theta$.

We can choose $\theta \in [0, 2\pi[$, and we call θ the *measure of the angle of rotation of r (and R)*. If the orientation is changed, the measure changes to $2\pi - \theta$.

We now let E be a Euclidean space of dimension $n = 2$, but we do not assume any orientation.

It is easy to see that given any two unit vectors $u_1, u_2 \in E$ (unit means that $\|u_1\| = \|u_2\| = 1$), there is a unique rotation r such that

$$r(u_1) = u_2.$$

It is also possible to define an equivalence relation of pairs of unit vectors, such that

$$\langle u_1, u_2\rangle \equiv \langle u_3, u_4\rangle$$

iff there is some rotation r such that $r(u_1) = u_3$ and $r(u_2) = u_4$.

Then, the equivalence class of $\langle u_1, u_2\rangle$ can be taken as the definition of the (oriented) *angle of* $\langle u_1, u_2\rangle$, which is denoted as $\widehat{u_1 u_2}$.

Furthermore, it can be shown that there is a rotation mapping the pair $\langle u_1, u_2\rangle$ to the pair $\langle u_3, u_4\rangle$, iff there is a rotation mapping the pair $\langle u_1, u_3\rangle$ to the pair $\langle u_2, u_4\rangle$ (all vectors being unit vectors).

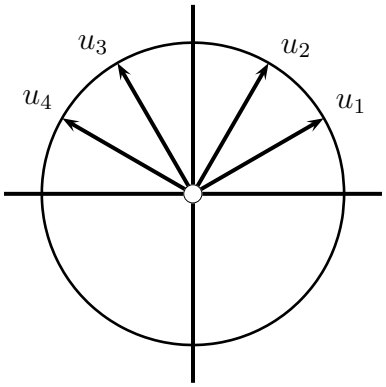

Figure 5.11: Defining Angles

As a consequence of all this, since for any pair $\langle u_1, u_2\rangle$ of unit vectors, there is a unique rotation r mapping u_1 to u_2, the angle $\widehat{u_1u_2}$ of $\langle u_1, u_2\rangle$ corresponds bijectively to the rotation r, and there is a bijection between the set of angles of pairs of unit vectors and the set of rotations in the plane.

As a matter of fact, the set of angles forms an abelian groups isomorphic to the (abelian) group of rotations in the plane.

Thus, even though we can consider angles as oriented, note that the notion of orientation is not necessary to define angles.

However, to define the *measure of an angle*, the notion of orientation is needed.

If we now assume that an orientation of E (still a Euclidean plane) is given, the unique rotation r associated with an angle $\widehat{u_1u_2}$ corresponds to a unique matrix

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

The number θ is defined up to $2k\pi$ (with $k \in \mathbb{Z}$) and is called a *measure of the angle* $\widehat{u_1u_2}$.

There is a unique $\theta \in [0, 2\pi[$ which is a measure of the angle $\widehat{u_1u_2}$.

It is also immediately seen that

$$\cos\theta = u_1 \cdot u_2.$$

In fact, since $\cos\theta = \cos(2\pi - \theta) = \cos(-\theta)$, the quantity $\cos\theta$ does not depend on the orientation.

Now still considering a Euclidean plane, given any pair $\langle u_1, u_2\rangle$ of nonnull vectors, we define their angle as the angle of the unit vectors $\frac{u_1}{\|u_1\|}$ and $\frac{u_2}{\|u_2\|}$, and if E is oriented, we define the measure θ of this angle as the measure of the angle of these unit vectors.

Note that

$$\cos\theta = \frac{u_1 \cdot u_2}{\|u_1\|\,\|u_2\|},$$

and this independently of the orientation.

Finally if E is a Euclidean space of dimension $n \geq 2$, we define the angle of a pair $\langle u_1, u_2\rangle$ of nonnull vectors as the angle of this pair in the Euclidean plane spanned by $\langle u_1, u_2\rangle$ if they are linearly independent, or any Euclidean plane containing u_1 if their are collinear.

If E is an affine Euclidean space of dimension $n \geq 2$, for any two pairs $\langle a_1, b_1\rangle$ and $\langle a_2, b_2\rangle$ of points in E, where $a_1 \neq b_1$ and $a_2 \neq b_2$, we define the angle of the pair $\langle\langle a_1, b_1\rangle, \langle a_2, b_2\rangle\rangle$ as the angle of the pair $\langle \mathbf{a_1b_1}, \mathbf{a_2b_2}\rangle$.

As for the issue of measure of an angle when $n \geq 3$, all we can do is to define the measure of the angle $\widehat{u_1u_2}$ as either θ or $2\pi - \theta$, where $\theta \in [0, 2\pi[$.

In particular, when $n = 3$, one should note that it is not enough to give a line D through the origin (the axis of rotation) and an angle θ to specify a rotation!

The problem is that depending on the orientation of the plane H (through the origin) orthogonal to D, we get two different rotations: one of angle θ, the other of angle $2\pi - \theta$.

Thus, to specify a rotation, we also need to give an orientation of the plane orthogonal to the axis of rotation.

This can be done by specifying an orientation of the axis of rotation by some unit vector ω, and chosing the basis (e_1, e_2, ω) (where (e_1, e_2) is a basis of H) such that it has positive orientation w.r.t. the chosen orientation of E.

We now return to alternating multilinear forms on a Euclidean space.

When E is a Euclidean space, we have an interesting situation regarding the value of determinants over orthornormal bases described by the following lemma.

Given any basis $B = (u_1, \ldots, u_n)$ for E, for any sequence $(w_1, \ldots, w_n)$ of n vectors, we denote as $\det_B(w_1, \ldots, w_n)$ the determinant of the matrix whose columns are the coordinates of the w_j over the basis $B = (u_1, \ldots, u_n)$.

Lemma 5.8.2 *Let E be a Euclidean space of finite dimension n, and assume that an orientation of E has been chosen. For any sequence $(w_1, \ldots, w_n)$ of n vectors, for any two orthonormal bases $B_1 = (u_1, \ldots, u_n)$ and $B_2 = (v_1, \ldots, v_n)$ of positive orientation, we have*

$$\det_{B_1}(w_1, \ldots, w_n) = \det_{B_2}(w_1, \ldots, w_n).$$

By lemma 5.8.2, the determinant $\det_B(w_1, \ldots, w_n)$ is independent of the base B, provided that B is orthonormal and of positive orientation.

Thus, lemma 5.8.2 suggests the following definition.

5.9 Volume Forms, Cross-Products

Definition 5.9.1 Given any Euclidean space E of finite dimension n over $\mathbb{R}$ and any orientation of E, for any sequence $(w_1, \ldots, w_n)$ of n vectors in E, the common value $\lambda_E(w_1, \ldots, w_n)$ of the determinant $\det_B(w_1, \ldots, w_n)$ over all positive orthonormal bases B of E is called the *mixed product (or volume form) of* $(w_1, \ldots, w_n)$.

The mixed product $\lambda_E(w_1, \ldots, w_n)$ will also be denoted as $(w_1, \ldots, w_n)$, even though the notation is overloaded.

- The mixed product $\lambda_E(w_1, \ldots, w_n)$ changes sign when the orientation changes.
- The mixed product $\lambda_E(w_1, \ldots, w_n)$ is a scalar, and definition 5.9.1 really defines an alternating multilinear form from E^n to $\mathbb{R}$.
- $\lambda_E(w_1, \ldots, w_n) = 0$ iff $(w_1, \ldots, w_n)$ is linearly dependent.

- A basis $(u_1, \ldots, u_n)$ is positive or negative iff $\lambda_E(u_1, \ldots, u_n)$ is positive or negative.
- $\lambda_E(w_1, \ldots, w_n)$ is invariant under every isometry f such that $\det(f) = 1$.

The terminology "volume form" is justified by the fact that $\lambda_E(w_1, \ldots, w_n)$ is indeed the volume of some geometric object.

Indeed, viewing E as an affine space, the *parallelotope defined by* $(w_1, \ldots, w_n)$ is the set of points

$$\{\lambda_1 w_1 + \cdots + \lambda_n w_n \mid 0 \leq \lambda_i \leq 1,\ 1 \leq i \leq n\}.$$

Then, it can be shown (see Berger [**?**], Section 9.12) that the volume of the parallelotope defined by $(w_1, \ldots, w_n)$ is indeed $\lambda_E(w_1, \ldots, w_n)$.

If $(E, \overrightarrow{E})$ is a Euclidean affine space of dimension n, given any $n+1$ affinely independent points $(a_0, \ldots, a_n)$, the set

$$\{a_0 + \lambda_1 \mathbf{a_0a_1} + \cdots + \lambda_n \mathbf{a_0a_n} \mid 0 \leq \lambda_i \leq 1,\ 1 \leq i \leq n\},$$

is called the *parallelotope spanned by* $(a_0, \ldots, a_n)$.

Then, the volume of the parallelotope spanned by $(a_0, \ldots, a_n)$ is $\lambda_{\overrightarrow{E}}(\mathbf{a_0a_1}, \ldots, \mathbf{a_0a_n})$.

It can also be shown that the volume $vol(a_0, \ldots, a_n)$ of the n-simplex $(a_0, \ldots, a_n)$ is

$$vol(a_0, \ldots, a_n) = \frac{1}{n!}\lambda_{\overrightarrow{E}}(\mathbf{a_0a_1}, \ldots, \mathbf{a_0a_n}).$$

Now, given a sequence $(w_1, \ldots, w_{n-1})$ of $n-1$ vectors in E, the map

$$x \mapsto \lambda_E(w_1, \ldots, w_{n-1}, x)$$

is a linear form.

Thus, by lemma 4.2.4, there is a unique vector $u \in E$ such that

$$\lambda_E(w_1, \ldots, w_{n-1}, x) = u \cdot x$$

for all $x \in E$.

The vector u has some interesting properties which motivate the next definition.

Definition 5.9.2 Given any Euclidean space E of finite dimension n over $\mathbb{R}$, for any orientation of E, for any sequence $(w_1, \ldots, w_{n-1})$ of $n-1$ vectors in E, the unique vector $w_1 \times \cdots \times w_{n-1}$ such that

$$\lambda_E(w_1, \ldots, w_{n-1}, x) = w_1 \times \cdots \times w_{n-1} \cdot x$$

for all $x \in E$, is called the *cross-product, or vector product, of* $(w_1, \ldots, w_{n-1})$.

The following properties hold.

- The cross-product $w_1 \times \cdots \times w_{n-1}$ changes sign when the orientation changes.
- The cross-product $w_1 \times \cdots \times w_{n-1}$ is a vector, and definition 5.9.2 really defines an alternating multilinear map from E^{n-1} to E.

- $w_1 \times \cdots \times w_{n-1} = 0$ iff $(w_1, \ldots, w_{n-1})$ is linearly dependent. This is because,

$$w_1 \times \cdots \times w_{n-1} = 0$$

iff

$$\lambda_E(w_1, \ldots, w_{n-1}, x) = 0$$

for all $x \in E$, and thus, if $(w_1, \ldots, w_{n-1})$ was linearly independent, we could find a vector $x \in E$ to complete $(w_1, \ldots, w_{n-1})$ into a basis of E, and we would have

$$\lambda_E(w_1, \ldots, w_{n-1}, x) \neq 0.$$

- The cross-product $w_1 \times \cdots \times w_{n-1}$ is orthogonal to each of the w_j.

- If $(w_1, \ldots, w_{n-1})$ is linearly independent, then the sequence

$$(w_1, \ldots, w_{n-1}, w_1 \times \cdots \times w_{n-1})$$

is a positive basis of E.

We now show how to compute the coordinates of $u_1 \times \cdots \times u_{n-1}$ over an orthonormal basis.

Given an orthonormal basis $(e_1, \ldots, e_n)$, for any sequence $(u_1, \ldots, u_{n-1})$ of $n-1$ vectors in E, if

$$u_j = \sum_{i=1}^{n} u_{i,j} e_i,$$

where $1 \leq j \leq n-1$, for any $x = x_1 e_1 + \cdots + x_n e_n$, consider the determinant

$$\lambda_E(u_1, \ldots, u_{n-1}, x) = \begin{vmatrix} u_{1\,1} & \ldots & u_{1\,n-1} & x_1 \\ u_{2\,1} & \ldots & u_{2\,n-1} & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ u_{n\,1} & \ldots & u_{n\,n-1} & x_n \end{vmatrix}.$$

Calling the underlying matrix above as A, we can expand $\det(A)$ according to the last column, using the Laplace formula (see Strang [**?**]), where $A_{i\,j}$ is the $(n-1)\times(n-1)$-matrix obtained from A by deleting row i and column j, and we get:

$$\begin{vmatrix} u_{1\,1} & \dots & u_{1\,n-1} & x_1 \\ u_{2\,1} & \dots & u_{2\,n-1} & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ u_{n\,1} & \dots & u_{n\,n-1} & x_n \end{vmatrix} =$$

$$(-1)^{n+1} x_1 \det(A_{1\,n}) + \cdots + (-1)^{n+n} x_n \det(A_{n\,n}).$$

Each $(-1)^{i+n} \det(A_{i\,n})$ is called the *cofactor of* x_i.

We note that $\det(A)$ is in fact the inner product

$$\det(A) = ((-1)^{n+1} \det(A_{1\,n}) e_1 + \cdots + (-1)^{n+n} \det(A_{n\,n}) e_n) \cdot x.$$

Since the cross-product $u_1 \times \cdots \times u_{n-1}$ is the unique vector u such that

$$u \cdot x = \lambda_E(u_1, \dots, u_{n-1}, x),$$

for all $x \in E$, the coordinates of the cross-product $u_1 \times \cdots \times u_{n-1}$ must be

$$((-1)^{n+1} \det(A_{1\,n}), \dots, (-1)^{n+n} \det(A_{n\,n})),$$

the sequence of cofactors of the x_i in the determinant $\det(A)$.

For example, when $n = 3$, the coordinates of the cross-product $u \times v$ are given by the cofactors of x_1, x_2, x_3, in the determinant

$$\begin{vmatrix} u_1 & v_1 & x_1 \\ u_2 & v_2 & x_2 \\ u_3 & v_3 & x_3 \end{vmatrix}$$

or more explicitly, by

$$(-1)^{3+1}\begin{vmatrix} u_2 & v_2 \\ u_3 & v_3 \end{vmatrix}, (-1)^{3+2}\begin{vmatrix} u_1 & v_1 \\ u_3 & v_3 \end{vmatrix}, (-1)^{3+3}\begin{vmatrix} u_1 & v_1 \\ u_2 & v_2 \end{vmatrix},$$

that is,

$$(u_2v_3 - u_3v_2,\ u_3v_1 - u_1v_3,\ u_1v_2 - u_2v_1).$$

It is also useful to observe that if we let U be the matrix

$$U = \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix}$$

then the coordinates of the cross-product $u \times v$ are given by

$$\begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} u_2v_3 - u_3v_2 \\ u_3v_1 - u_1v_3 \\ u_1v_2 - u_2v_1 \end{pmatrix}$$

We finish our discussion of cross-products by mentioning without proof a few more of their properties, in the case $n = 3$.

Firstly, the following so-called *Lagrange identity* holds:

$$(u \cdot v)^2 + \|u \times v\|^2 = \|u\|^2 \|v\|^2 .$$

If u and v are linearly independent, and if θ (or $2\pi - \theta$) is a measure of the angle $\widehat{uv}$, then

$$|\sin\theta| = \frac{\|u \times v\|}{\|u\| \|v\|}.$$

Chapter 6

Polar Duality, Polyhedra and Polytopes

6.1 Polarity and Duality

In this section, we apply the intrinsic duality afforded by a Euclidean structure to the study of convex sets and, in particular, polytopes.

Let $E = \mathbb{E}^n$ be a Euclidean space of dimension n. Pick any origin, O, in $\mathbb{E}^n$ (we may assume $O = (0, \ldots, 0)$).

We know that the inner product on $E = \mathbb{E}^n$ induces a duality between E and its dual E^*, namely, $u \mapsto \varphi_u$, where φ_u is the linear form defined by $\varphi_u(v) = u \cdot v$, for all $v \in E$.

For geometric purposes, it is more convenient to recast this duality as a correspondence between points and hyperplanes, using the notion of polarity with respect to the unit sphere, $S^{n-1} = \{a \in \mathbb{E}^n \mid \|\mathbf{Oa}\| = 1\}$.

First, we need the following simple fact: For every hyperplane, H, not passing through O, there is a *unique* point, h, so that

$$H = \{a \in \mathbb{E}^n \mid \mathbf{Oh} \cdot \mathbf{Oa} = 1\}.$$

Using the above, we make the following definition:

Definition 6.1.1 Given any point, $a \neq O$, the *polar hyperplane of a (w.r.t. S^{n-1})* or *dual of a* is the hyperplane, $a^\dagger$, given by

$$a^\dagger = \{b \in \mathbb{E}^n \mid \mathbf{Oa} \cdot \mathbf{Ob} = 1\}.$$

Given a hyperplane, H, not containing O, the *pole of H (w.r.t S^{n-1})* or *dual of H* is the (unique) point, $H^\dagger$, so that

$$H = \{a \in \mathbb{E}^n \mid \mathbf{OH}^\dagger \cdot \mathbf{Oa} = 1\}.$$

We often abbreviate polar hyperplane to polar.

We immediately check that $a^{\dagger\dagger} = a$ and $H^{\dagger\dagger} = H$, so, we obtain a bijective correspondence between $\mathbb{E}^n - \{O\}$ and the set of hyperplanes not passing through O.

When a is outside the sphere S^{n-1}, there is a nice geometric interpetation for the polar hyperplane, $H = a^\dagger$. Indeed, in this case, since

$$H = a^\dagger = \{b \in \mathbb{E}^n \mid \mathbf{Oa} \cdot \mathbf{Ob} = 1\}$$

and $\|\mathbf{Oa}\| > 1$, the hyperplane H intersects S^{n-1} (along an $(n-2)$-dimensional sphere) and if b is any point on $H \cap S^{n-1}$, we claim that $\mathbf{Ob}$ and $\mathbf{ba}$ are orthogonal.

This means that $H \cap S^{n-1}$ is the set of points on S^{n-1} where the lines through a and tangent to S^{n-1} touch S^{n-1} (they form a cone tangent to S^{n-1} with apex a).

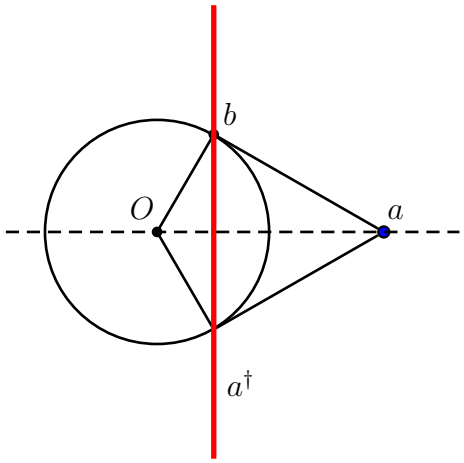

Figure 6.1: The polar, $a^\dagger$, of a point, a, outside the sphere S^{n-1}

Also, observe that for any point, $a \neq O$, and any hyperplane, H, not passing through O, if $a \in H$, then, $H^\dagger \in a^\dagger$, i.e, the pole, $H^\dagger$, of H belongs to the polar, $a^\dagger$, of a.

If $a = (a_1, \ldots, a_n)$, the equation of the polar hyperplane, $a^\dagger$, is

$$a_1 X_1 + \cdots + a_n X_n = 1.$$

Now, we would like to extend this correspondence to subsets of $\mathbb{E}^n$, in particular, to convex sets.

Given a hyperplane, H, not containing O, we denote by H_- the closed half-space containing O.

Definition 6.1.2 Given any subset, A, of $\mathbb{E}^n$, the set

$$A^* = \{b \in \mathbb{E}^n \mid \mathbf{Oa}\cdot\mathbf{Ob} \leq 1, \quad \text{for all } a \in A\} = \bigcap_{\substack{a \in A \\ a \neq O}} (a^\dagger)_-,$$

is called the *polar dual* or *reciprocal* of A.

To simplify notation we write $a^\dagger_-$ for $(a^\dagger)_-$. Note that $\{O\}^* = \mathbb{E}^n$, so it is convenient to set $O^\dagger_- = \mathbb{E}^n$, even though $O^\dagger$ is undefined.

We use a different notation, $a^\dagger$ and $H^\dagger$, for polar hyperplanes and poles, as opposed to A^*, for polar duals, to avoid confusion. Indeed, $H^\dagger$ and H^*, where H is a hyperplane (resp. $a^\dagger$ and $\{a\}^*$, where a is a point) are very different things!

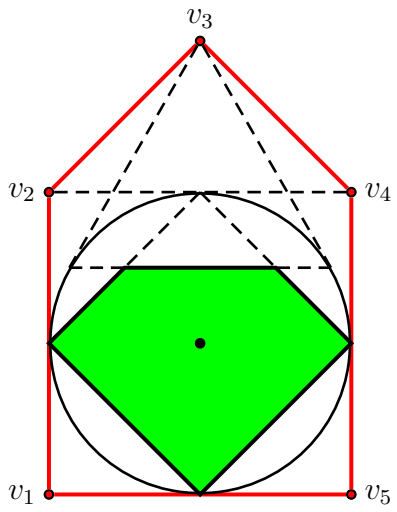

Figure 6.2: The polar dual of a polygon

In Figure 6.2, the polar dual of the polygon $(v_1, v_2, v_3, v_4, v_5)$ is the polygon shown in green.

This polygon is cut out by the half-planes determined by the polars of the vertices $(v_1, v_2, v_3, v_4, v_5)$ and containing the center of the circle.

By definition, A^* is convex even if A is not.

Furthermore, note that

(1) $A \subseteq A^{**}$.

(2) If $A \subseteq B$, then $B^* \subseteq A^*$.

(3) If A is convex and closed, then $A^* = (\partial A)^*$.

It follows immediately from (1) and (2) that $A^{***} = A^*$. Also, if $B^n(r)$ is the (closed) ball of radius $r > 0$ and center O, it is obvious by definition that $B^n(r)^* = B^n(1/r)$.

We would like to investigate the duality induced by the operation $A \mapsto A^*$.

Unfortunately, it is not always the case that $A^{**} = A$, but this is true when A is closed and convex, as shown in the following proposition:

Proposition 6.1.3 *Let A be any subset of $\mathbb{E}^n$ (with origin O).*

(i) If A is bounded, then $O \in \overset{\circ}{A^*}$; if $O \in \overset{\circ}{A}$, then A^* is bounded.

(ii) If A is a closed and convex subset containing O, then $A^{**} = A$.

Note that

$$\begin{aligned} A^{**} &= \{c \in \mathbb{E}^n \mid \mathbf{Od} \cdot \mathbf{Oc} \leq 1 \quad \text{for all } d \in A^*\} \\ &= \{c \in \mathbb{E}^n \mid (\forall d \in \mathbb{E}^n)(\text{if} \quad \mathbf{Od} \cdot \mathbf{Oa} \leq 1 \\ &\qquad\qquad \text{for all } a \in A, \quad \text{then} \quad \mathbf{Od} \cdot \mathbf{Oc} \leq 1)\}. \end{aligned}$$

Remark: For an arbitrary subset, $A \subseteq \mathbb{E}^n$, it can be shown that $A^{**} = \overline{\text{conv}(A \cup \{O\})}$, the topological closure of the convex hull of $A \cup \{O\}$.

Proposition 6.1.3 will play a key role in studying polytopes, but before doing this, we need one more proposition.

Proposition 6.1.4 *Let A be any closed convex subset of $\mathbb{E}^n$ such that $O \in \overset{\circ}{A}$. The polar hyperplanes of the points of the boundary of A constitute the set of supporting hyperplanes of A^*. Furthermore, for any $a \in \partial A$, the points of A^* where $H = a^\dagger$ is a supporting hyperplane of A^* are the poles of supporting hyperplanes of A at a.*

6.2 Polyhedra, $\mathcal{H}$-Polytopes and $\mathcal{V}$-Polytopes

There are two natural ways to define a convex polyhedron, A:

(1) As the convex hull of a finite set of points.

(2) As a subset of $\mathbb{E}^n$ cut out by a finite number of hyperplanes, more precisely, as the intersection of a finite number of (closed) half-spaces.

As stated, these two definitions are not equivalent because (1) implies that a polyhedron is bounded, whereas (2) allows unbounded subsets.

Now, if we require in (2) that the convex set A is bounded, it is quite clear for $n = 2$ that the two definitions (1) and (2) are equivalent; for $n = 3$, it is intuitively clear that definitions (1) and (2) are still equivalent, but proving this equivalence rigorously does not appear to be that easy.

What about the equivalence when $n \geq 4$?

It turns out that definitions (1) and (2) are equivalent for all n, but this is a nontrivial theorem and a rigorous proof does not come by so cheaply.

Fortunately, since we have Krein and Milman's theorem at our disposal and polar duality, we can give a rather short proof.

The hard direction of the equivalence consists in proving that definition (1) implies definition (2).

This is where the duality induced by polarity becomes handy, especially, the fact that $A^{**} = A$! (under the right hypotheses).

First, we give precise definitions (following Ziegler [**?**]).

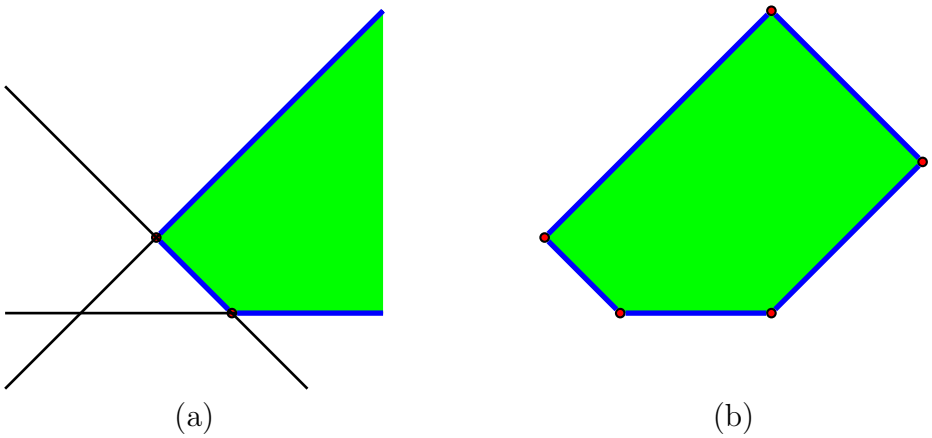

Figure 6.3: (a) An $\mathcal{H}$-polyhedron. (b) A $\mathcal{V}$-polytope

Definition 6.2.1 Let $\mathcal{E}$ be any affine Euclidean space of finite dimension, n.[1] An *$\mathcal{H}$-polyhedron* in $\mathcal{E}$, for short, a *polyhedron*, is any subset, $P = \bigcap_{i=1}^{p} C_i$, of $\mathcal{E}$ defined as the intersection of a finite number of closed half-spaces, C_i; an *$\mathcal{H}$-polytope* in $\mathcal{E}$ is a bounded polyhedron and a *$\mathcal{V}$-polytope* is the convex hull, $P = \text{conv}(S)$, of a finite set of points, $S \subseteq \mathcal{E}$.

Examples of an $\mathcal{H}$-polyhedron and of a $\mathcal{V}$-polytope are shown in Figure 6.3.

[1]This means that the vector space, $\overrightarrow{\mathcal{E}}$, associated with $\mathcal{E}$ is a Euclidean space.

Obviously, polyhedra and polytopes are convex and closed (in $\mathcal{E}$). Since the notions of $\mathcal{H}$-polytope and $\mathcal{V}$-polytope are equivalent (see Theorem 6.3.1), we often use the simpler locution polytope.

Note that Definition 6.2.1 allows $\mathcal{H}$-polytopes and $\mathcal{V}$-polytopes to have an empty interior, which is sometimes an inconvenience.

This is not a problem. In fact, we can prove that we may always assume to $\mathcal{E} = \mathbb{E}^n$ and restrict ourselves to the affine hull of A (some copy of $\mathbb{E}^d$, for $d \leq n$, where $d = \dim(A)$, as in Definition 3.2.3).

Since the boundary of a closed half-space, C_i, is a hyperplane, H_i, and since hyperplanes are defined by affine forms, a closed half-space is defined by the locus of points satisfying a "linear" inequality of the form $a_i \cdot x \leq b_i$ or $a_i \cdot x \geq b_i$, for some vector $a_i \in \mathbb{R}^n$ and some $b_i \in \mathbb{R}$.

Since $a_i \cdot x \geq b_i$ is equivalent to $(-a_i) \cdot x \leq -b_i$, we may restrict our attention to inequalities with a $\leq$ sign.

Thus, if A is the $d \times p$ matrix whose i^{th} row is a_i, we see that the $\mathcal{H}$-polyhedron, P, is defined by the system of linear inequalities, $Ax \leq b$, where $b = (b_1, \ldots, b_p) \in \mathbb{R}^p$.

We write

$$P = P(A, b), \quad \textit{with} \quad P(A, b) = \{x \in \mathbb{R}^n \mid Ax \leq b\}.$$

An equation, $a_i \cdot x = b_i$, may be handled as the conjunction of the two inequalities $a_i \cdot x \leq b_i$ and $(-a_i) \cdot x \leq -b_i$.

Also, if $0 \in P$, observe that we must have $b_i \geq 0$ for $i = 1, \ldots, p$. In this case, every inequality for which $b_i > 0$ can be normalized by dividing both sides by b_i, so we may assume that $b_i = 1$ or $b_i = 0$.

Remark: Some authors call "convex" polyhedra and "convex" polytopes what we have simply called polyhedra and polytopes.

Since Definition 6.2.1 implies that these objects are convex and since we are not going to consider non-convex polyhedra in this chapter, we stick to the simpler terminology.

One should consult Ziegler [**?**], Berger [**?**], Grunbaum [**?**] and especially Cromwell [**?**], for pictures of polyhedra and polytopes.

Even better, take a look at the web sites listed in the web page for CIS610!

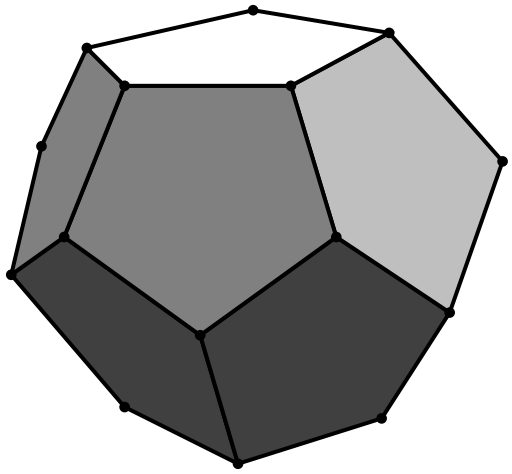

Figure 6.4: Example of a polytope (a dodecahedron)

Figure 6.4 shows the picture a polytope whose faces are all pentagons. This polytope is called a *dodecahedron*. The dodecahedron has 12 faces, 30 edges and 20 vertices.

Obviously, an n-simplex is a $\mathcal{V}$-polytope. The *standard n-cube* is the set

$$\{(x_1, \ldots, x_n) \in \mathbb{E}^n \mid |x_i| \leq 1, \quad 1 \leq i \leq n\}.$$

The standard cube is a $\mathcal{V}$-polytope. The *standard n-cross-polytope* (or *n-co-cube*) is the set

$$\{(x_1, \ldots, x_n) \in \mathbb{E}^n \mid \sum_{i=1}^{n} |x_i| \leq 1\}.$$

It is also a $\mathcal{V}$-polytope.

What happens if we take the dual of a $\mathcal{V}$-polytope (resp. an $\mathcal{H}$-polytope)? The following proposition, although very simple, is an important step in answering the above question.

Proposition 6.2.2 *Let $S = \{a_i\}_{i=1}^p$ be a finite set of points in $\mathbb{E}^n$ and let $A = \mathrm{conv}(S)$ be its convex hull. If $S \neq \{O\}$, then, the dual, A^*, of A w.r.t. the center O is an $\mathcal{H}$-polyhedron; furthermore, if $O \in \overset{\circ}{A}$, then A^* is an $\mathcal{H}$-polytope, i.e., the dual of a $\mathcal{V}$-polytope with nonempty interior is an $\mathcal{H}$-polytope. If $A = S = \{O\}$, then $A^* = \mathbb{E}^d$.*

Thus, the dual of the convex hull of a finite set of points, $\{a_1, \ldots, a_p\}$, is the intersection of the half-spaces containing O determined by the polar hyperplanes of the points a_i. (Recall that $(a_i)_-^\dagger = \mathbb{E}^n$ if $a_i = O$.)

It is convenient to restate Proposition 6.2.2 using matrices.

First, observe that the proof of Proposition 6.2.2 shows that

$$\mathrm{conv}(\{a_1, \ldots, a_p\})^* = \mathrm{conv}(\{a_1, \ldots, a_p\} \cup \{O\})^*.$$

Therefore, we may assume that not all $a_i = 0$ $(1 \leq i \leq p)$. If we pick O as an origin, then every point a_j can be identified with a vector in $\mathbb{E}^n$ and O corresponds to the zero vector, 0.

Observe that any set of p points, $a_j \in \mathbb{E}^n$, corresponds to the $n \times p$ matrix, A, whose j^{th} column is a_j.

Then, the equation of the the polar hyperplane, $a_j^\dagger$, of any a_j $(\neq 0)$ is $a_j \cdot x = 1$, that is

$$a_j^\top x = 1.$$

Consequently, the system of inequalities defining $\mathrm{conv}(\{a_1, \ldots, a_p\})^*$ can be written in matrix form as

$$\mathrm{conv}(\{a_1, \ldots, a_p\})^* = \{x \in \mathbb{R}^n \mid A^\top x \leq \mathbf{1}\},$$

where $\mathbf{1}$ denotes the vector of $\mathbb{R}^p$ with all coordinates equal to 1. We write
$P(A^\top, \mathbf{1}) = \{x \in \mathbb{R}^n \mid A^\top x \leq \mathbf{1}\}$.

Proposition 6.2.3 *Given any set of p points, $\{a_1, \ldots, a_p\}$, in $\mathbb{R}^n$ with $\{a_1, \ldots, a_p\} \neq \{0\}$, if A is the $n \times p$ matrix whose j^{th} column is a_j, then*

$$\mathrm{conv}(\{a_1, \ldots, a_p\})^* = P(A^\top, \mathbf{1}),$$

with $P(A^\top, \mathbf{1}) = \{x \in \mathbb{R}^n \mid A^\top x \leq \mathbf{1}\}$.

Conversely, given any $p \times n$ matrix, A, not equal to the zero matrix, we have

$$P(A, \mathbf{1})^* = \mathrm{conv}(\{a_1, \ldots, a_p\} \cup \{0\}),$$

where $a_i \in \mathbb{R}^n$ is the i^{th} row of A or, equivalently,

$$P(A, \mathbf{1})^* = \{x \in \mathbb{R}^n \mid x = A^\top t,\ t \in \mathbb{R}^p,\ t \geq 0,\ \mathbb{I}t = 1\},$$

where $\mathbb{I}$ is the row vector of length p whose coordinates are all equal to 1.

Using the above, the reader should check that the dual of a simplex is a simplex and that the dual of an n-cube is an n-cross polytope.

We will see shortly that if A is an $\mathcal{H}$-polytope and if $O \in \overset{\circ}{A}$, then A^* is also an $\mathcal{H}$-polytope.

For this, we will prove first that an $\mathcal{H}$-polytope is a $\mathcal{V}$-polytope. This requires taking a closer look at polyhedra.

Note that some of the hyperplanes cutting out a polyhedron may be redundant.

If $A = \bigcap_{i=1}^{t} C_i$ is a polyhedron (where each closed half-space, C_i, is associated with a hyperplane, H_i, so that $\partial C_i = H_i$), we say that $\bigcap_{i=1}^{t} C_i$ *is an irredundant decomposition of* A if A cannot be expressed as $A = \bigcap_{i=1}^{m} C'_i$ with $m < t$ (for some closed half-spaces, C'_i).

Proposition 6.2.4 *Let A be a polyhedron with nonempty interior and assume that $A = \bigcap_{i=1}^{t} C_i$ is an irredundant decomposition of A. Then,*

(i) Up to order, the C_i's are uniquely determined by A.

(ii) If $H_i = \partial C_i$ is the boundary of C_i, then $H_i \cap A$ is a polyhedron with nonempty interior in H_i, denoted $\mathrm{Facet}_i\, A$*, and called a facet of A.*

(iii) We have $\partial A = \bigcup_{i=1}^{t} \mathrm{Facet}_i\, A$, where the union is irredundant, i.e., $\mathrm{Facet}_i\, A$ *is not a subset of* $\mathrm{Facet}_j\, A$*, for all $i \neq j$.*

As a consequence, if A is a polyhedron, then so are its facets and the same holds for $\mathcal{H}$-polytopes.

If A is an $\mathcal{H}$-polytope and H is a hyperplane with $H \cap \overset{\circ}{A} \neq \emptyset$, then $H \cap A$ is an $\mathcal{H}$-polytope whose facets are of the form $H \cap F$, where F is a facet of A.

We can use induction and define k-faces, for $0 \leq k \leq n-1$.

Definition 6.2.5 Let $A \subseteq \mathbb{E}^n$ be a polyhedron with nonempty interior. We define a *k-face of A* to be a facet of a $(k+1)$-face of A, for $k = 0, \ldots, n-2$, where an *$(n-1)$-face* is just a facet of A. The 1-faces are called *edges*. Two k-faces are *adjacent* if their intersection is a $(k-1)$-face.

The polyhedron A itself is also called a *face* (of itself) or *n-face* and the k-faces of A with $k \leq n-1$ are called *proper faces* of A.

If $A = \bigcap_{i=1}^{t} C_i$ is an irredundant decomposition of A and H_i is the boundary of C_i, then the hyperplane, H_i, is called the *supporting hyperplane* of the facet $H_i \cap A$.

We suspect that the 0-faces of a polyhedron are vertices in the sense of Definition 3.4.1.

This is true and, in fact, the vertices of a polyhedron coincide with its extreme points (see Definition 3.4.3).

Proposition 6.2.6 *Let $A \subseteq \mathbb{E}^n$ be a polyhedron with nonempty interior.*

(1) For any point, $a \in \partial A$, on the boundary of A, the intersection of all the supporting hyperplanes to A at a coincides with the intersection of all the faces that contain a. In particular, points of order k of A are those points in the relative interior of the k-faces of A[2]; thus, 0-faces coincide with the vertices of A.

(2) The vertices of A coincide with the extreme points of A.

We are now ready for the theorem showing the equivalence of $\mathcal{V}$-polytopes and $\mathcal{H}$-polytopes.

[2]Given a convex set, S, in $\mathbb{A}^n$, its *relative interior* is its interior in the affine hull of S (which might be of dimension strictly less than n).

6.3 The Equivalence of $\mathcal{H}$-Polytopes and $\mathcal{V}$-Polytopes

The next result is a nontrivial theorem usually attributed to Weyl and Minkowski (see Barvinok [**?**]).

Theorem 6.3.1 *(Weyl-Minkowski) If A is an $\mathcal{H}$-polytope, then A has a finite number of extreme points (equal to its vertices) and A is the convex hull of its set of vertices; thus, an $\mathcal{H}$-polytope is a $\mathcal{V}$-polytope. Moreover, A has a finite number of k-faces (for $k = 0, \ldots, d-2$, where $d = \dim(A)$). Conversely, the convex hull of a finite set of points is an $\mathcal{H}$-polytope. As a consequence, a $\mathcal{V}$-polytope is an $\mathcal{H}$-polytope.*

In view of Theorem 6.3.1, we are justified in dropping the $\mathcal{V}$ or $\mathcal{H}$ in front of polytope, and will do so from now on.

Theorem 6.3.1 has some interesting corollaries regarding the dual of a polytope.

Corollary 6.3.2 *If A is any polytope in $\mathbb{E}^n$ such that the interior of A contains the origin, O, then the dual, A^*, of A is also a polytope whose interior contains O and $A^{**} = A$.*

Corollary 6.3.3 *If A is any polytope in $\mathbb{E}^n$ whose interior contains the origin, O, then the k-faces of A are in bijection with the $(n-k-1)$-faces of the dual polytope, A^*. This correspondence is as follows: If $Y = \mathrm{aff}(F)$ is the k-dimensional subspace determining the k-face, F, of A then the subspace, $Y^* = \mathrm{aff}(F^*)$, determining the corresponding face, F^*, of A^*, is the intersection of the polar hyperplanes of points in Y.*

We also have the following proposition whose proof would not be that simple if we only had the notion of an $\mathcal{H}$-polytope.

Proposition 6.3.4 *If $A \subseteq \mathbb{E}^n$ is a polytope and $f\colon \mathbb{E}^n \to \mathbb{E}^m$ is an affine map, then $f(A)$ is a polytope in $\mathbb{E}^m$.*

The reader should check that the Minkowski sum of polytopes is a polytope.

We were able to give a short proof of Theorem 6.3.1 because we relied on a powerful theorem, namely, Krein and Milman.

A drawback of this approach is that it bypasses the interesting and important problem of designing algorithms for finding the vertices of an $\mathcal{H}$-polyhedron from the sets of inequalities defining it.

A method for doing this is Fourier-Motzkin elimination, see Ziegler [**?**] (Chapter 1). This is also a special case of *linear programming*.

It is also possible to generalize the notion of $\mathcal{V}$-polytope to polyhedra using the notion of cone.

6.4 The Equivalence of $\mathcal{H}$-Polyhedra and $\mathcal{V}$-Polyhedra

The equivalence of $\mathcal{H}$-polytopes and $\mathcal{V}$-polytopes can be generalized to polyhedral sets, *i.e.*, finite intersections of half-spaces that are not necessarily bounded. This equivalence was first proved by Motzkin in the early 1930's.

Definition 6.4.1 Let $\mathcal{E}$ be any affine Euclidean space of finite dimension, d (with associated vector space, $\overrightarrow{\mathcal{E}}$). A subset, $C \subseteq \overrightarrow{\mathcal{E}}$, is a *cone* if C is closed under linear combinations involving only nonnnegative scalars. Given a subset, $V \subseteq \overrightarrow{\mathcal{E}}$, the *conical hull* or *positive hull* of V is the set

$$\mathrm{cone}(V) = \{\sum_I \lambda_i v_i \mid \{v_i\}_{i\in I} \subseteq V, \lambda_i \geq 0 \quad \text{for all } i \in I\}.$$

A $\mathcal{V}$*-polyhedron* or *polyhedral set* is a subset, $A \subseteq \mathcal{E}$, such that

$$\begin{aligned} A &= \mathrm{conv}(Y) + \mathrm{cone}(V) \\ &= \{a + v \mid a \in \mathrm{conv}(Y),\ v \in \mathrm{cone}(V)\}, \end{aligned}$$

where $V \subseteq \overrightarrow{\mathcal{E}}$ is a finite set of vectors and $Y \subseteq \mathcal{E}$ is a finite set of points.

A set, $C \subseteq \overrightarrow{\mathcal{E}}$, is a $\mathcal{V}$-*cone* or *polyhedral cone* if C is the positive hull of a finite set of vectors, that is,

$$C = \text{cone}(\{u_1, \ldots, u_p\}),$$

for some vectors, $u_1, \ldots, u_p \in \overrightarrow{\mathcal{E}}$. An $\mathcal{H}$-*cone* is any subset of $\overrightarrow{\mathcal{E}}$ given by a finite intersection of closed half-spaces cut out by hyperplanes through 0.

The positive hull, $\text{cone}(V)$, of V is also denoted $\text{pos}(V)$.

Observe that a $\mathcal{V}$-cone can be viewed as a polyhedral set for which $Y = \{O\}$, a single point.

However, if we take the point O as the origin, we may view a $\mathcal{V}$-polyhedron, A, for which $Y = \{O\}$, as a $\mathcal{V}$-cone.

We will switch back and forth between these two views of cones as we find it convenient

As a consequence, a ($\mathcal{V}$ or $\mathcal{H}$)-cone always contains 0, sometimes called an *apex* of the cone.

We can prove that we may always assume that $\mathcal{E} = \mathbb{E}^d$ and that our polyhedra have nonempty interior. It will be convenient to decree that $\mathbb{E}^d$ is an $\mathcal{H}$-polyhedron.

The generalization of Theorem 6.3.1 is that every $\mathcal{V}$-polyhedron is an $\mathcal{H}$-polyhedron and conversely.

Ziegler proceeds as follows: First, he shows that the equivalence of $\mathcal{V}$-polyhedra and $\mathcal{H}$-polyhedra reduces to the equivalence of $\mathcal{V}$-cones and $\mathcal{H}$-cones using an "old trick" of projective geometry, namely, "homogenizing" [**?**] (Chapter 1).

Then, he uses two dual versions of Fourier-Motzkin elimination to pass from $\mathcal{V}$-cones to $\mathcal{H}$-cones and conversely.

Since the homogenization method is an important technique we will describe it in some detail.

However, it turns out that the double dualization technique used in the proof of Theorem 6.3.1 can be easily adapted to prove that every $\mathcal{V}$-polyhedron is an $\mathcal{H}$-polyhedron.

Moreover, it can also be used to prove that every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron!

So, we will not describe the version of Fourier-Motzkin elimination used to go from $\mathcal{V}$-cones to $\mathcal{H}$-cones.

However, we will present the Fourier-Motzkin elimination method used to go from $\mathcal{H}$-cones to $\mathcal{V}$-cones.

In order to avoid confusion between the zero vector and the origin of $\mathbb{E}^d$, we will denote the origin by O and the center of polar duality by Ω.

Given any nonzero vector, $u \in \mathbb{R}^d$, let $u^\dagger_-$ be the closed half-space

$$u^\dagger_- = \{x \in \mathbb{R}^d \mid x \cdot u \leq 0\}.$$

In other words, $u^\dagger_-$ is the closed-half space bounded by the hyperplane through Ω normal to u and on the "opposite side" of u.

Proposition 6.4.2 *Let $A = \mathrm{conv}(Y) + \mathrm{cone}(V) \subseteq \mathbb{E}^d$ be a $\mathcal{V}$-polyhedron with $Y = \{y_1, \ldots, y_p\}$ and $V = \{v_1, \ldots, v_q\}$. Then, for any point, Ω, if $A \neq \{\Omega\}$, then the polar dual, A^*, of A w.r.t. Ω is the $\mathcal{H}$-polyhedron given by*

$$A^* = \bigcap_{i=1}^{p} (y_i^\dagger)_- \cap \bigcap_{j=1}^{q} (v_j^\dagger)_-.$$

Furthermore, if A has nonempty interior and Ω belongs to the interior of A, then A^* is bounded, that is, A^* is an $\mathcal{H}$-polytope. If $A = \{\Omega\}$, then A^* is the special polyhedron, $A^* = \mathbb{E}^d$.

It is fruitful to restate Proposition 6.4.2 in terms of matrices (as we did for Proposition 6.2.2).

First, observe that

$$(\mathrm{conv}(Y) + \mathrm{cone}(V))^* = (\mathrm{conv}(Y \cup \{\Omega\}) + \mathrm{cone}(V))^*.$$

If we pick Ω as an origin then we can represent the points in Y as vectors. The old origin is still denoted O and Ω is now denoted 0. The zero vector is denoted $\mathbf{0}$.

If Y is the $d \times p$ matrix whose i^{th} column is y_i and V is the $d \times q$ matrix whose j^{th} column is v_j, then A^* is given by:

$$A^* = \{x \in \mathbb{R}^d \mid Y^\top x \leq \mathbf{1},\ V^\top x \leq \mathbf{0}\}.$$

We write

$$P(Y^\top, \mathbf{1}; V^\top, \mathbf{0}) = \{x \in \mathbb{R}^d \mid Y^\top x \leq \mathbf{1},\ V^\top x \leq \mathbf{0}\}.$$

Proposition 6.4.3 *Let $\{y_1, \ldots, y_p\}$ be any set of points in $\mathbb{E}^d$ and let $\{v_1, \ldots, v_q\}$ be any set of nonzero vectors in $\mathbb{R}^d$. If Y is the $d \times p$ matrix whose i^{th} column is y_i and V is the $d \times q$ matrix whose j^{th} column is v_j, then*

$$(\text{conv}(\{y_1, \ldots, y_p\}) \cup \text{cone}(\{v_1, \ldots, v_q\}))^* = P(Y^\top, \mathbf{1}; V^\top, \mathbf{0}),$$

with

$$P(Y^\top, \mathbf{1}; V^\top, \mathbf{0}) = \{x \in \mathbb{R}^d \mid Y^\top x \leq \mathbf{1},\ V^\top x \leq \mathbf{0}\}.$$

Conversely, given any $p \times d$ matrix, Y, and any $q \times d$ matrix, V, we have

$$P(Y, \mathbf{1}; V, \mathbf{0})^* = \text{conv}(\{y_1, \ldots, y_p\} \cup \{0\}) \cup \text{cone}(\{v_1, \ldots, v_q\}),$$

where $y_i \in \mathbb{R}^n$ is the i^{th} row of Y and $v_j \in \mathbb{R}^n$ is the j^{th} row of V or, equivalently,

$$P(Y, \mathbf{1}; V, \mathbf{0})^* = \{x \in \mathbb{R}^d \mid x = Y^\top u + V^\top t, \quad u \in \mathbb{R}^p,\ t \in \mathbb{R}^q,\ u, t \geq 0,\ \mathbb{I}u = 1\},$$

where $\mathbb{I}$ is the row vector of length p whose coordinates are all equal to 1.

We can now use Proposition 6.4.2, Proposition 6.1.3 and Krein and Millman's Theorem to prove that every $\mathcal{V}$-polyhedron is an $\mathcal{H}$-polyhedron.

Proposition 6.4.4 *Every $\mathcal{V}$-polyhedron, A, is an $\mathcal{H}$-polyhedron. Furthermore, if $A \neq \mathbb{E}^d$, then A is of the form $A = P(Y, \mathbf{1})$.*

Interestingly, we can now prove easily that every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron.

Proposition 6.4.5 *Every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron.*

Putting together Propositions 6.4.4 and 6.4.5 we obtain our main theorem:

Theorem 6.4.6 *(Equivalence of $\mathcal{H}$-polyhedra and $\mathcal{V}$-polyhedra) Every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron and conversely.*

Even though we proved the main result of this section, it is instructive to consider a more computational proof making use of cones and an elimination method known as *Fourier-Motzkin elimination*.

The problem with the converse of Proposition 6.4.4 when A is unbounded (*i.e.*, not compact) is that Krein and Millman's Theorem does not apply.

We need to take into account "points at infinity" corresponding to certain vectors.

The trick we used in Proposition 6.4.4 is that the polar dual of a $\mathcal{V}$-polyhedron with nonempty interior is an $\mathcal{H}$-*polytope*.

This reduction to polytopes allowed us to use Krein and Millman to convert an $\mathcal{H}$-polytope to a $\mathcal{V}$-polytope and then again we took the polar dual.

Another trick is to switch to cones by "homogenizing".

Given any subset, $S \subseteq \mathbb{E}^d$, we can form the cone, $C(S) \subseteq \mathbb{E}^{d+1}$, by "placing" a copy of S in the hyperplane, $H_{d+1} \subseteq \mathbb{E}^{d+1}$, of equation $x_{d+1} = 1$, and drawing all the half lines from the origin through any point of S.

Let $P \subseteq \mathbb{E}^d$ be an $\mathcal{H}$-polyhedron. Then, P is cut out by m hyperplanes, H_i, and for each H_i, there is a nonzero vector, a_i, and some $b_i \in \mathbb{R}$ so that

$$H_i = \{x \in \mathbb{E}^d \mid a_i \cdot x = b_i\}$$

and P is given by

$$P = \bigcap_{i=1}^{m} \{x \in \mathbb{E}^d \mid a_i \cdot x \leq b_i\}.$$

If A denotes the $m \times d$ matrix whose i-th row is a_i and b is the vector $b = (b_1, \ldots, b_m)$, then we can write

$$P = P(A, b) = \{x \in \mathbb{E}^d \mid Ax \leq b\}.$$

We "homogenize" $P(A, b)$ as follows: Let $C(P)$ be the subset of $\mathbb{E}^{d+1}$ defined by

$$\begin{aligned} C(P) &= \left\{ \begin{pmatrix} x \\ x_{d+1} \end{pmatrix} \in \mathbb{R}^{d+1} \mid Ax \leq x_{d+1}b,\ x_{d+1} \geq 0 \right\} \\ &= \left\{ \begin{pmatrix} x \\ x_{d+1} \end{pmatrix} \mid Ax - x_{d+1}b \leq 0,\ -x_{d+1} \leq 0 \right\}. \end{aligned}$$

Thus, we see that $C(P)$ is the $\mathcal{H}$-cone given by the system of inequalities

$$\begin{pmatrix} A & -b \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ x_{d+1} \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and that

$$P = C(P) \cap H_{d+1}.$$

Conversely, if Q is any $\mathcal{H}$-cone in $\mathbb{E}^{d+1}$ (in fact, any $\mathcal{H}$-polyhedron), it is clear that $P = Q \cap H_{d+1}$ is an $\mathcal{H}$-polyhedron in $\mathbb{E}^d$.

Let us now assume that
$P \subseteq \mathbb{E}^d$ is a $\mathcal{V}$-polyhedron, $P = \mathrm{conv}(Y) + \mathrm{cone}(V)$, where $Y = \{y_1, \ldots, y_p\}$ and $V = \{v_1, \ldots, v_q\}$.

Define $\widehat{Y} = \{\widehat{y}_1, \ldots, \widehat{y}_p\} \subseteq \mathbb{E}^{d+1}$, and
$\widehat{V} = \{\widehat{v}_1, \ldots, \widehat{v}_q\} \subseteq \mathbb{E}^{d+1}$, by

$$\widehat{y}_i = \begin{pmatrix} y_i \\ 1 \end{pmatrix}, \qquad \widehat{v}_j = \begin{pmatrix} v_j \\ 0 \end{pmatrix}.$$

We check immediately that

$$C(P) = \mathrm{cone}(\{\widehat{Y} \cup \widehat{V}\})$$

is a $\mathcal{V}$-cone in $\mathbb{E}^{d+1}$ such that

$$C = C(P) \cap H_{d+1},$$

where H_{d+1} is the hyperplane of equation $x_{d+1} = 1$.

Conversely, if $C = \mathrm{cone}(W)$ is a $\mathcal{V}$-cone in $\mathbb{E}^{d+1}$, with $w_{id+1} \geq 0$ for every $w_i \in W$, we prove next that
$P = C \cap H_{d+1}$ is a $\mathcal{V}$-polyhedron.

Proposition 6.4.7 *(Polyhedron–Cone Correspondence) We have the following correspondence between polyhedra in $\mathbb{E}^d$ and cones in $\mathbb{E}^{d+1}$:*

(1) For any $\mathcal{H}$-polyhedron, $P \subseteq \mathbb{E}^d$, if $P = P(A, b) = \{x \in \mathbb{E}^d \mid Ax \leq b\}$, where A is an $m \times d$-matrix and $b \in \mathbb{R}^m$, then $C(P)$ given by

$$\begin{pmatrix} A & -b \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ x_{d+1} \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

is an $\mathcal{H}$-cone in $\mathbb{E}^{d+1}$ and $P = C(P) \cap H_{d+1}$, where H_{d+1} is the hyperplane of equation $x_{d+1} = 1$. Conversely, if Q is any $\mathcal{H}$-cone in $\mathbb{E}^{d+1}$ (in fact, any $\mathcal{H}$-polyhedron), then $P = Q \cap H_{d+1}$ is an $\mathcal{H}$-polyhedron in $\mathbb{E}^d$.

(2) Let $P \subseteq \mathbb{E}^d$ be any $\mathcal{V}$-polyhedron, where $P = \mathrm{conv}(Y) + \mathrm{cone}(V)$ with $Y = \{y_1, \ldots, y_p\}$ and $V = \{v_1, \ldots, v_q\}$. Define $\widehat{Y} = \{\widehat{y}_1, \ldots, \widehat{y}_p\} \subseteq \mathbb{E}^{d+1}$, and $\widehat{V} = \{\widehat{v}_1, \ldots, \widehat{v}_q\} \subseteq \mathbb{E}^{d+1}$, by

$$\widehat{y}_i = \begin{pmatrix} y_i \\ 1 \end{pmatrix}, \qquad \widehat{v}_j = \begin{pmatrix} v_j \\ 0 \end{pmatrix}.$$

Then,

$$C(P) = \mathrm{cone}(\{\widehat{Y} \cup \widehat{V}\})$$

is a $\mathcal{V}$-cone in $\mathbb{E}^{d+1}$ such that

$$C = C(P) \cap H_{d+1},$$

Conversely, if $C = \mathrm{cone}(W)$ is a $\mathcal{V}$-cone in $\mathbb{E}^{d+1}$, with $w_{i\,d+1} \geq 0$ for every $w_i \in W$, then $P = C \cap H_{d+1}$ is a $\mathcal{V}$-polyhedron in $\mathbb{E}^d$.

By Proposition 6.4.7, if P is an $\mathcal{H}$-polyhedron, then $C(P)$ is an $\mathcal{H}$-cone. If we can prove that every $\mathcal{H}$-cone is a $\mathcal{V}$-cone, then again, Proposition 6.4.7 shows that $P = C(P) \cap H_{d+1}$ is a $\mathcal{V}$-polyhedron.

Therefore, in order to prove that every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron it suffices to show that every $\mathcal{H}$-cone is a $\mathcal{V}$-cone.

By a similar argument, Proposition 6.4.7 show that in order to prove that every $\mathcal{V}$-polyhedron is an $\mathcal{H}$-polyhedron it suffices to show that every $\mathcal{V}$-cone is an $\mathcal{H}$-cone.

We will not prove this direction again since we already have it by Proposition 6.4.4.

It remains to prove that every $\mathcal{H}$-cone is a $\mathcal{V}$-cone.

Let $C \subseteq \mathbb{E}^d$ be an $\mathcal{H}$-cone. Then, C is cut out by m hyperplanes, H_i, through 0.

For each H_i, there is a nonzero vector, u_i, so that

$$H_i = \{x \in \mathbb{E}^d \mid u_i \cdot x = 0\}$$

and C is given by

$$C = \bigcap_{i=1}^{m} \{x \in \mathbb{E}^d \mid u_i \cdot x \leq 0\}.$$

If A denotes the $m \times d$ matrix whose i-th row is u_i, then we can write

$$C = P(A, 0) = \{x \in \mathbb{E}^d \mid Ax \leq 0\}.$$

Observe that $C = C_0(A) \cap H_w$, where

$$C_0(A) = \left\{ \begin{pmatrix} x \\ w \end{pmatrix} \in \mathbb{R}^{d+m} \mid Ax \leq w \right\}$$

is an $\mathcal{H}$-cone in $\mathbb{E}^{d+m}$ and

$$H_w = \left\{ \begin{pmatrix} x \\ w \end{pmatrix} \in \mathbb{R}^{d+m} \mid w = 0 \right\}$$

is a hyperplane in $\mathbb{E}^{d+m}$.

We claim that $C_0(A)$ is a $\mathcal{V}$-cone.

This follows by observing that for every $\begin{pmatrix} x \\ w \end{pmatrix}$ satisfying $Ax \leq w$, we can write

$$\begin{pmatrix} x \\ w \end{pmatrix} = \sum_{i=1}^{d} |x_i|(\mathrm{sign}(x_i)) \begin{pmatrix} e_i \\ Ae_i \end{pmatrix} + \sum_{j=1}^{m} (w_j - (Ax)_j) \begin{pmatrix} 0 \\ e_j \end{pmatrix},$$

and then

$$C_0(A) = \mathrm{cone}\left(\left\{ \pm \begin{pmatrix} e_i \\ Ae_i \end{pmatrix} \mid 1 \leq i \leq d \right\} \cup \left\{ \begin{pmatrix} 0 \\ e_j \end{pmatrix} \mid 1 \leq j \leq m \right\} \right).$$

Since $C = C_0(A) \cap H_w$ is now the intersection of a $\mathcal{V}$-cone with a hyperplane, to prove that C is a $\mathcal{V}$-cone it is enough to prove that the intersection of a $\mathcal{V}$-cone with a hyperplane is also a $\mathcal{V}$-cone.

For this, we use *Fourier-Motzkin elimination*. It suffices to prove the result for a hyperplane, H_k, in $\mathbb{E}^{d+m}$ of equation $y_k = 0$ $(1 \leq k \leq d+m)$.

Proposition 6.4.8 *(Fourier-Motzkin Elimination) Say $C = \text{cone}(Y) \subseteq \mathbb{E}^d$ is a $\mathcal{V}$-cone. Then, the intersection $C \cap H_k$ (where H_k is the hyperplane of equation $y_k = 0$) is a $\mathcal{V}$-cone, $C \cap H_k = \text{cone}(Y^{/k})$, with*

$$Y^{/k} = \{y_i \mid y_{ik} = 0\} \cup \{y_{ik}y_j - y_{jk}y_i \mid y_{ik} > 0,\ y_{jk} < 0\},$$

the set of vectors obtained from Y by "eliminating the k-th coordinate". Here, each y_i is a vector in $\mathbb{R}^d$.

As discussed above, Proposition 6.4.8 implies (again!)

Corollary 6.4.9 *Every $\mathcal{H}$-polyhedron is a $\mathcal{V}$-polyhedron.*

Another way of proving that every $\mathcal{V}$-polyhedron is an $\mathcal{H}$-polyhedron is to use cones.

Let $P = \mathrm{conv}(Y) + \mathrm{cone}(V) \subseteq \mathbb{E}^d$ be a $\mathcal{V}$-polyhedron.

We can view Y as a $d \times p$ matrix whose ith column is the ith vector in Y and V as $d \times q$ matrix whose jth column is the jth vector in V.

Then, we can write

$$P = \{x \in \mathbb{R}^d \mid (\exists u \in \mathbb{R}^p)(\exists t \in \mathbb{R}^d) \\ (x = Yu + Vt,\ u \geq 0,\ \mathbb{I}u = 1,\ t \geq 0)\},$$

where $\mathbb{I}$ is the row vector

$$\mathbb{I} = \underbrace{(1, \ldots, 1)}_{p}.$$

Now, observe that P can be interpreted as the projection of the $\mathcal{H}$-polyhedron, $\widetilde{P} \subseteq \mathbb{E}^{d+p+q}$, given by

$$\widetilde{P} = \{(x, u, t) \in \mathbb{R}^{d+p+q} \mid x = Yu + Vt, \\ u \geq 0,\ \mathbb{I}u = 1,\ t \geq 0\}$$

onto $\mathbb{R}^d$.

Consequently, if we can prove that the projection of an $\mathcal{H}$-polyhedron is also an $\mathcal{H}$-polyhedron, then we will have proved that every $\mathcal{V}$-polyhedron is an $\mathcal{H}$-polyhedron.

In view of Proposition 6.4.7 and the discussion that followed, it is enough to prove that the projection of any $\mathcal{H}$-cone is an $\mathcal{H}$-cone.

This can be done by using a type of Fourier-Motzkin elimination dual to the method used in Proposition 6.4.8.

We state the result without proof and refer the interested reader to Ziegler [**?**], Section 1.2–1.3, for full details.

Proposition 6.4.10 *If $C = P(A, 0) \subseteq \mathbb{E}^d$ is an $\mathcal{H}$-cone, then the projection, $\mathrm{proj}_k(C)$, onto the hyperplane, H_k, of equation $y_k = 0$ is given by*
$\mathrm{proj}_k(C) = \mathrm{elim}_k(C) \cap H_k$, *with*

$$\begin{aligned}\mathrm{elim}_k(C) &= \{x \in \mathbb{R}^d \mid (\exists t \in \mathbb{R})(x + te_k \in P)\} \\ &= \{z - te_k \mid z \in P,\, t \in \mathbb{R}\} = P(A^{/k}, 0)\end{aligned}$$

and where the rows of $A^{/k}$ are given by

$$A^{/k} = \{a_i \mid a_{i\,k} = 0\} \cup \{a_{i\,k}a_j - a_{j\,k}a_i \mid a_{i\,k} > 0, a_{j\,k} < 0\}.$$

It should be noted that both Fourier-Motzkin elimination methods generate a quadratic number of new vectors or inequalities at each step and thus they lead to a combinatorial explosion.

Therefore, these methods become intractable rather quickly.

The problem is that many of the new vectors or inequalities are redundant. Thereore, it is important to find ways of detecting redundancies and there are various methods for doing so.

Again, the interested reader should consult Ziegler [**?**], Chapter 1.

We conclude this section with a version of Farkas Lemma for polyhedral sets.

Lemma 6.4.11 *(Farkas Lemma, Version IV) Let Y be any $d \times p$ matrix and V be any $d \times q$ matrix. For every $z \in \mathbb{R}^d$, exactly one of the following alternatives occurs:*

(a) There exist $u \in \mathbb{R}^p$ and $t \in \mathbb{R}^q$, with $u \geq 0$, $t \geq 0$, $\mathbb{I}u = 1$ and $z = Yu + Vt$.

(b) There is some vector, $(\alpha, c) \in \mathbb{R}^{d+1}$, such that $c^\top y_i \geq \alpha$ for all i with $1 \leq i \leq p$, $c^\top v_j \geq 0$ for all j with $1 \leq j \leq q$, and $c^\top z < \alpha$.

Observe that Farkas IV can be viewed as a separation criterion for polyhedral sets.

Chapter 7

Basics of Combinatorial Topology

7.1 Simplicial and Polyhedral Complexes

In order to study and manipulate complex shapes it is convenient to discretize these shapes and to view them as the union of simple building blocks glued together in a "clean fashion".

The building blocks should be simple geometric objects, for example, points, lines segments, triangles, tehrahedra and more generally simplices, or even convex polytopes.

Definition 7.1.1 Let $\mathcal{E}$ be any normed affine space, say $\mathcal{E} = \mathbb{E}^m$ with its usual Euclidean norm. Given any $n+1$ affinely independent points $a_0, \ldots, a_n$ in $\mathcal{E}$, the *n-simplex (or simplex) σ defined by $a_0, \ldots, a_n$* is the convex hull of the points $a_0, \ldots, a_n$, that is, the set of all convex combinations $\lambda_0 a_0 + \cdots + \lambda_n a_n$, where $\lambda_0 + \cdots + \lambda_n = 1$ and $\lambda_i \geq 0$ for all i, $0 \leq i \leq n$.

We call n the *dimension* of the n-simplex σ, and the points $a_0, \ldots, a_n$ are the *vertices* of σ.

Given any subset $\{a_{i_0}, \ldots, a_{i_k}\}$ of $\{a_0, \ldots, a_n\}$ (where $0 \leq k \leq n$), the k-simplex generated by $a_{i_0}, \ldots, a_{i_k}$ is called a *k-face* or simply a *face* of σ.

A face s of σ is a *proper face* if $s \neq \sigma$ (we agree that the empty set is a face of any simplex). For any vertex a_i, the face generated by $a_0, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n$ (i.e., omitting a_i) is called the *face opposite* a_i.

Every face that is an $(n-1)$-simplex is called a *boundary face* or *facet*. The union of the boundary faces is the *boundary of* σ, denoted by $\partial\sigma$, and the complement of $\partial\sigma$ in σ is the *interior* $\operatorname{Int}\sigma = \sigma - \partial\sigma$ of σ. The interior $\operatorname{Int}\sigma$ of σ is sometimes called an *open simplex*.

It should be noted that for a 0-simplex consisting of a single point $\{a_0\}$, $\partial\{a_0\} = \emptyset$, and $\operatorname{Int}\{a_0\} = \{a_0\}$.

Of course, a 0-simplex is a single point, a 1-simplex is the line segment (a_0, a_1), a 2-simplex is a triangle (a_0, a_1, a_2) (with its interior), and a 3-simplex is a tetrahedron (a_0, a_1, a_2, a_3) (with its interior).

The inclusion relation between any two faces σ and τ of some simplex, s, is written $\sigma \preceq \tau$.

Clearly, a point x belongs to the boundary $\partial\sigma$ of σ iff at least one of its barycentric coordinates $(\lambda_0, \ldots, \lambda_n)$ is zero, and a point x belongs to the interior $\operatorname{Int}\sigma$ of σ iff all of its barycentric coordinates $(\lambda_0, \ldots, \lambda_n)$ are positive, i.e., $\lambda_i > 0$ for all i, $0 \leq i \leq n$.

Then, for every $x \in \sigma$, there is a unique face s such that $x \in \operatorname{Int} s$, the face generated by those points a_i for which $\lambda_i > 0$, where $(\lambda_0, \ldots, \lambda_n)$ are the barycentric coordinates of x.

A simplex σ is convex, arcwise connected, compact, and closed. The interior $\mathrm{Int}\,\sigma$ of a simplex is convex, arcwise connected, open, and σ is the closure of $\mathrm{Int}\,\sigma$.

We now put simplices together to form more complex shapes. The intuition behind the next definition is that the building blocks should be "glued cleanly".

Definition 7.1.2 A *simplicial complex in* $\mathbb{E}^m$ (for short, a *complex* in $\mathbb{E}^m$) is a set K consisting of a (finite or infinite) set of simplices in $\mathbb{E}^m$ satisfying the following conditions:

(1) Every face of a simplex in K also belongs to K.

(2) For any two simplices σ_1 and σ_2 in K, if $\sigma_1 \cap \sigma_2 \neq \emptyset$, then $\sigma_1 \cap \sigma_2$ is a common face of both σ_1 and σ_2.

Every k-simplex, $\sigma \in K$, is called a *k-face* (or *face*) of K. A 0-face $\{v\}$ is called a *vertex* and a 1-face is called an *edge*. The *dimension* of the simplicial complex K is the maximum of the dimensions of all simplices in K.

If $\dim K = d$, then every face of dimension d is called a *cell* and every face of dimension $d-1$ is called a *facet*.

Condition (2) guarantees that the various simplices forming a complex intersect nicely. It is easily shown that the following condition is equivalent to condition (2):

(2′) For any two distinct simplices σ_1, σ_2,
$\operatorname{Int}\sigma_1 \cap \operatorname{Int}\sigma_2 = \emptyset$.

Remarks:

1. A simplicial complex, K, is a combinatorial object, namely, a *set* of simplices satisfying certain conditions but not a subset of $\mathbb{E}^m$. However, every complex, K, yields a subset of $\mathbb{E}^m$ called the geometric realization of K and denoted $|K|$. This object will be defined shortly and should not be confused with the complex.

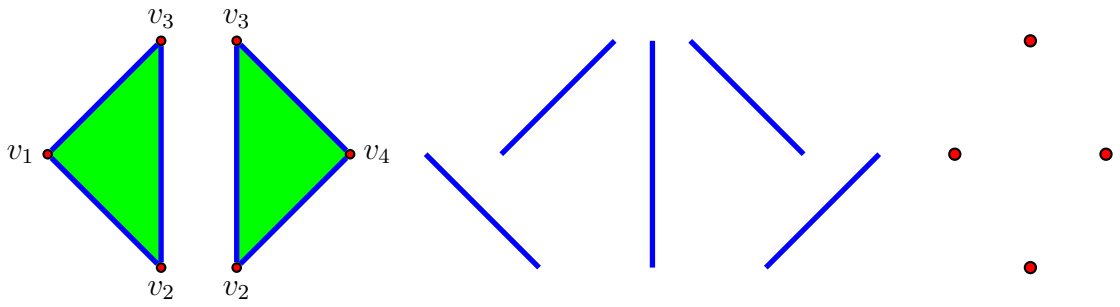

Figure 7.1: A set of simplices forming a complex

Figure 7.1 illustrates this aspect of the definition of a complex. For clarity, the two triangles (2-simplices) are drawn as disjoint objects even though they share the common edge, (v_2, v_3) (a 1-simplex) and similarly for the edges that meet at some common vertex.

2. Some authors define a *facet* of a complex, K, of dimension d to be a d-simplex in K, as opposed to a $(d-1)$-simplex, as we did. This practice is not consistent with the notion of facet of a polyhedron and this is why we prefer the terminology *cell* for the d-simplices in K.

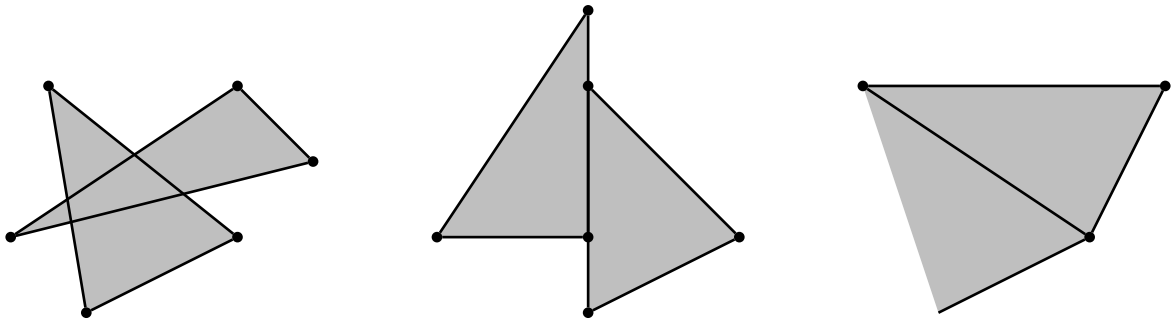

Figure 7.2: Collections of simplices not forming a complex

3. It is important to note that in order for a complex, K, of dimension d to be realized in $\mathbb{E}^m$, the dimension of the "ambient space", m, must be big enough. For example, there are 2-complexes that can't be realized in $\mathbb{E}^3$ or even in $\mathbb{E}^4$. There has to be enough room in order for condition (2) to be satisfied. It is not hard to prove that $m = 2d+1$ is always sufficient. Sometimes, $2d$ works, for example in the case of surfaces (where $d = 2$).

Some collections of simplices violating some of the conditions of Definition 7.1.2 are shown in Figure 7.2.

On the left, the intersection of the two 2-simplices is neither an edge nor a vertex of either triangle.

In the middle case, two simplices meet along an edge which is not an edge of either triangle.

On the right, there is a missing edge and a missing vertex.

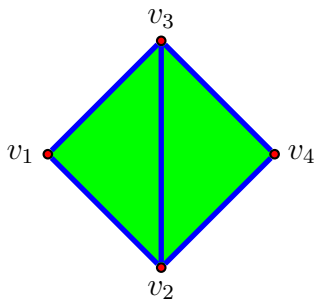

Figure 7.3: The geometric realization of the complex of Figure 7.1

The union $|K|$ of all the simplices in K is a subset of $\mathbb{E}^m$. We can define a topology on $|K|$ by defining a subset F of $|K|$ to be closed iff $F \cap \sigma$ is closed in σ for every face $\sigma \in K$.

It is immediately verified that the axioms of a topological space are indeed satisfied.

The resulting topological space $|K|$ is called the *geometric realization of* K.

The geometric realization of the complex from Figure 7.1 is show in Figure 7.1.

Some "legal" simplicial complexes are shown in Figure 7.4.

Obviously, $|\sigma| = \sigma$ for every simplex, σ. Also, note that distinct complexes may have the same geometric realization. In fact, all the complexes obtained by subdividing the simplices of a given complex yield the same geometric realization.

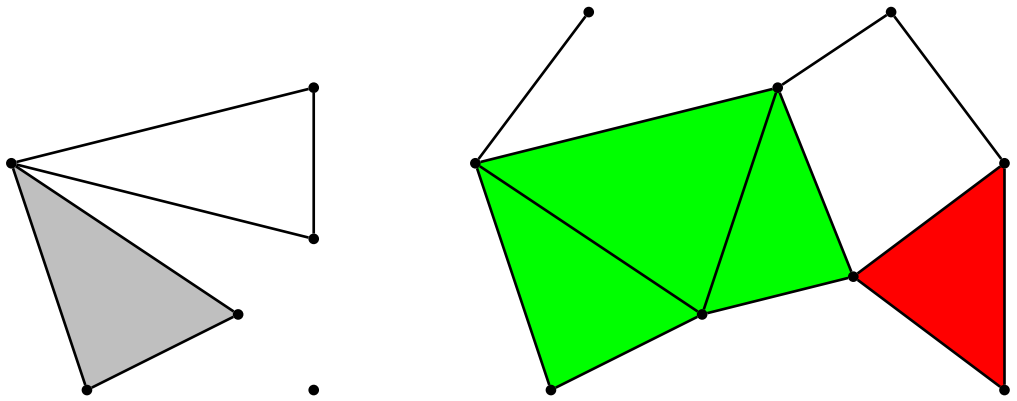

Figure 7.4: Examples of simplicial complexes

A *polytope* is the geometric realization of some simplicial complex. A polytope of dimension 1 is usually called a *polygon*, and a polytope of dimension 2 is usually called a *polyhedron*.

When K consists of infinitely many simplices we usually require that K be *locally finite*, which means that every vertex belongs to finitely many faces. If K is locally finite, then its geometric realization, $|K|$, is locally compact.

In the sequel, we will consider only finite simplicial complexes, that is, complexes K consisting of a finite number of simplices.

In this case, the topology of $|K|$ defined above is identical to the topology induced from $\mathbb{E}^m$. For any simplex σ in K, $\operatorname{Int}\sigma$ coincides with the interior $\overset{\circ}{\sigma}$ of σ in the topological sense, and $\partial\sigma$ coincides with the boundary of σ in the topological sense.

Definition 7.1.3 Given any complex, K_2, a subset $K_1 \subseteq K_2$ of K_2 is a *subcomplex* of K_2 iff it is also a complex. For any complex, K, of dimension d, for any i with $0 \leq i \leq d$, the subset

$$K^{(i)} = \{\sigma \in K \mid \dim \sigma \leq i\}$$

is called the *i-skeleton* of K. Clearly, $K^{(i)}$ is a subcomplex of K. We also let

$$K^i = \{\sigma \in K \mid \dim \sigma = i\}.$$

Observe that K^0 is the set of vertices of K and K^i is not a complex.

A simplicial complex, K_1 is a *subdivision* of a complex K_2 iff $|K_1| = |K_2|$ and if every face of K_1 is a subset of some face of K_2.

A complex K of dimension d is *pure* (or *homogeneous*) iff every face of K is a face of some d-simplex of K (i.e., some cell of K). A complex is *connected* iff $|K|$ is connected.

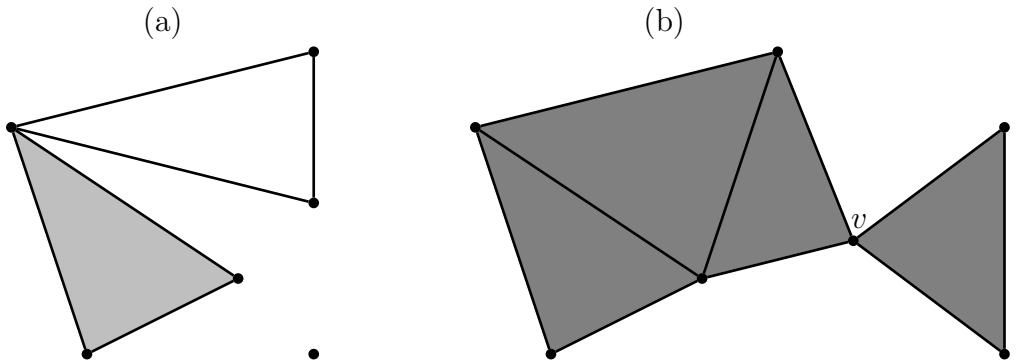

Figure 7.5: (a) A complex that is not pure. (b) A pure complex

It is is easy to see that a complex is connected iff its 1-skeleton is connected.

The intuition behind the notion of a pure complex, K, of dimension d is that a pure complex is the result of gluing pieces all having the same dimension, namely, d-simplices.

For example, in Figure 7.5, the complex on the left is not pure but the complex on the right is pure of dimension 2.

Most of the shapes that we will be interested in are well approximated by pure complexes, in particular, surfaces or solids.

However, pure complexes may still have undesirable "singularities" such as the vertex, v, in Figure 7.5(b).

The notion of link of a vertex provides a technical way to deal with singularities.

Definition 7.1.4 Let K be any complex and let σ be any face of K. The *star*, $\mathrm{St}(\sigma)$ (or if we need to be very precise, $\mathrm{St}(\sigma, K)$), of σ is the subcomplex of K consisting of all faces, τ, containing σ and of all faces of τ, *i.e.*,

$$\mathrm{St}(\sigma) = \{s \in K \mid (\exists \tau \in K)(\sigma \preceq \tau \quad \text{and} \quad s \preceq \tau)\}.$$

The *link*, $\mathrm{Lk}(\sigma)$ (or $\mathrm{Lk}(\sigma, K)$) of σ is the subcomplex of K consisting of all faces in $\mathrm{St}(\sigma)$ that do not intersect σ, i.e.,

$$\mathrm{Lk}(\sigma) = \{\tau \in K \mid \tau \in \mathrm{St}(\sigma) \quad \text{and} \quad \sigma \cap \tau = \emptyset\}.$$

To simplify notation, if $\sigma = \{v\}$ is a vertex we write $\mathrm{St}(v)$ for $\mathrm{St}(\{v\})$ and $\mathrm{Lk}(v)$ for $\mathrm{Lk}(\{v\})$.

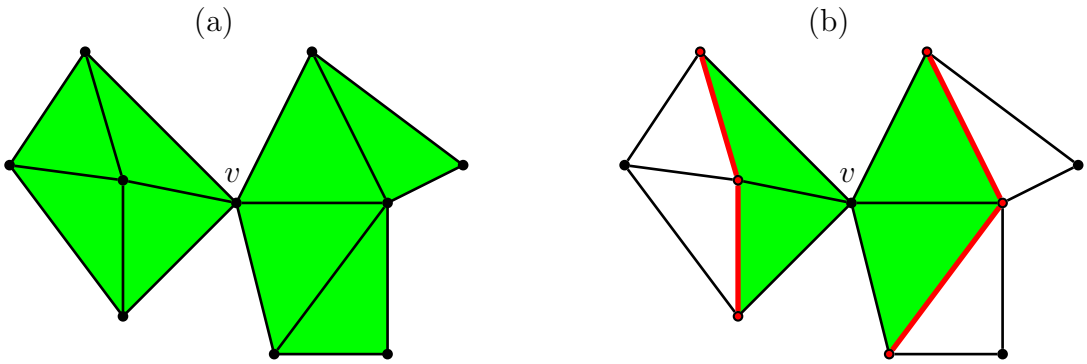

Figure 7.6: (a) A complex. (b) Star and Link of v

Figure 7.6 shows:

(a) A complex (on the left).

(b) The star of the vertex v, indicated in gray and the link of v, showed as thicker lines.

If K is pure and of dimension d, then $\mathrm{St}(\sigma)$ is also pure of dimension d and if $\dim \sigma = k$, then $\mathrm{Lk}(\sigma)$ is pure of dimension $d - k - 1$.

For technical reasons, following Munkres [**?**], besides defining the complex, $\mathrm{St}(\sigma)$, it is useful to introduce the *open star* of σ, denoted $\mathrm{st}(\sigma)$, defined as the subspace of $|K|$ consisting of the union of the interiors, $\mathrm{Int}(\tau) = \tau - \partial\,\tau$, of all the faces, τ, containing, σ.

According to this definition, the open star of σ is not a complex but instead a subset of $|K|$.

Note that

$$\overline{\mathrm{st}(\sigma)} = |\mathrm{St}(\sigma)|,$$

that is, the closure of $\mathrm{st}(\sigma)$ is the geometric realization of the complex $\mathrm{St}(\sigma)$.

Then, $\mathrm{lk}(\sigma) = |\mathrm{Lk}(\sigma)|$ is the union of the simplices in $\mathrm{St}(\sigma)$ that are disjoint from σ.

If σ is a vertex, v, we have

$$\mathrm{lk}(v) = \overline{\mathrm{st}(v)} - \mathrm{st}(v).$$

However, beware that if σ is not a vertex, then $\mathrm{lk}(\sigma)$ is properly contained in $\overline{\mathrm{st}(\sigma)} - \mathrm{st}(\sigma)$!

One of the nice properties of the open star, $\mathrm{st}(\sigma)$, of σ is that it is open. This follows from the fact that the open star, $\mathrm{st}(v)$, of a vertex, v is open.

Furthermore, for every point, $a \in |K|$, there is a unique smallest simplex, σ, so that $a \in \mathrm{Int}(\sigma) = \sigma - \partial\,\sigma$.

As a consequence, for any k-face, σ, of K, if $\sigma = (v_0, \ldots, v_k)$, then

$$\mathrm{st}(\sigma) = \mathrm{st}(v_0) \cap \cdots \cap \mathrm{st}(v_k).$$

Consequently, $\mathrm{st}(\sigma)$ is open and path connected.

Unfortunately, the "nice" equation

$$\mathrm{St}(\sigma) = \mathrm{St}(v_0) \cap \cdots \cap \mathrm{St}(v_k)$$

is false! (and anagolously for $\mathrm{Lk}(\sigma)$.)

For a counter-example, consider the boundary of a tetrahedron with one face removed.

Recall that in $\mathbb{E}^d$, the *(open) unit ball,* B^d, is defined by

$$B^d = \{x \in \mathbb{E}^d \mid \|x\| < 1\},$$

the *closed unit ball,* $\overline{B}^d$, is defined by

$$\overline{B}^d = \{x \in \mathbb{E}^d \mid \|x\| \leq 1\},$$

and the $(d-1)$*-sphere*, S^{d-1}, by

$$S^{d-1} = \{x \in \mathbb{E}^d \mid \|x\| = 1\}.$$

Obviously, S^{d-1} is the boundary of $\overline{B}^d$ (and B^d).

Definition 7.1.5 Let K be a pure complex of dimension d and let σ be any k-face of K, with $0 \leq k \leq d-1$. We say that σ is *nonsingular* iff the geometric realization, $\mathrm{lk}(\sigma)$, of the link of σ is homeomorphic to either S^{d-k-1} or to $\overline{B}^{d-k-1}$; this is written as $\mathrm{lk}(\sigma) \approx S^{d-k-1}$ or $\mathrm{lk}(\sigma) \approx \overline{B}^{d-k-1}$, where $\approx$ means homeomorphic.

In Figure 7.6, note that the link of v is not homeomorphic to S^1 or B^1, so v is singular.

It will also be useful to express $\mathrm{St}(v)$ in terms of $\mathrm{Lk}(v)$, where v is a vertex, and for this, we define the notion of cone.

Definition 7.1.6 Given any complex, K, in $\mathbb{E}^n$, if $\dim K = d < n$, for any point, $v \in \mathbb{E}^n$, such that v does not belong to the affine hull of $|K|$, the *cone on K with vertex v*, denoted, $v * K$, is the complex consisting of all simplices of the form $(v, a_0, \ldots, a_k)$ and their faces, where $(a_0, \ldots, a_k)$ is any k-face of K. If $K = \emptyset$, we set $v * K = v$.

It is not hard to check that $v * K$ is indeed a complex of dimension $d + 1$ containing K as a subcomplex.

Proposition 7.1.7 *For any complex, K, of dimension d and any vertex, $v \in K$, we have*

$$\mathrm{St}(v) = v * \mathrm{Lk}(v).$$

More generally, for any face, σ, of K, we have

$$\overline{\mathrm{st}(\sigma)} = |\mathrm{St}(\sigma)| \approx \sigma \times |v * \mathrm{Lk}(\sigma)|,$$

for every $v \in \sigma$ and

$$\overline{\mathrm{st}(\sigma)} - \mathrm{st}(\sigma) = \partial\, \sigma \times |v * \mathrm{Lk}(\sigma)|,$$

for every $v \in \partial\, \sigma$.

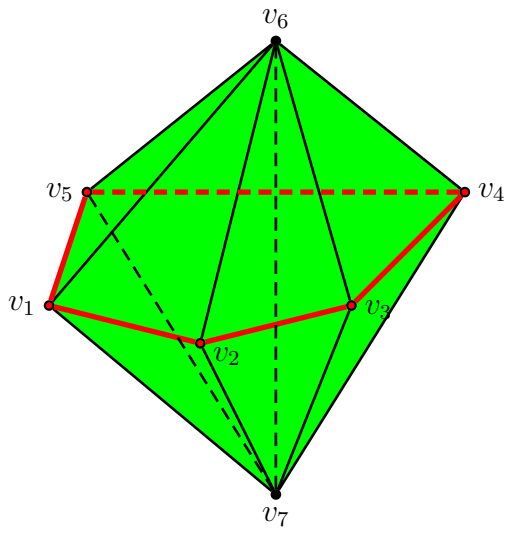

Figure 7.7: More examples of links and stars

Figure 7.7 shows a 3-dimensional complex. The link of the edge (v_6, v_7) is the pentagon $P = (v_1, v_2, v_3, v_4, v_5) \approx S^1$. The link of the vertex v_7 is the cone $v_6 * P \approx B^2$. The link of (v_1, v_2) is $(v_6, v_7) \approx B^1$ and the link of v_1 is the union of the triangles (v_2, v_6, v_7) and (v_5, v_6, v_7), which is homeomorphic to B^2.

Remark: Unfortunately, the word "cone" is overloaded. It might have been better to use the term *pyramid* as some authors do (including Ziegler).

Given a pure complex, it is necessary to distinguish between two kinds of faces.

Definition 7.1.8 Let K be any pure complex of dimension d. A k-face, σ, of K is a *boundary* or *external* face iff it belongs to a single cell (i.e., a d-simplex) of K and otherwise it is called an *internal* face ($0 \leq k \leq d-1$). The *boundary* of K, denoted $\mathrm{bd}(K)$, is the subcomplex of K consisting of all boundary facets of K together with their faces.

It is clear by definition that $\mathrm{bd}(K)$ is a pure complex of dimension $d-1$.

Even if K is connected, $\mathrm{bd}(K)$ is not connected, in general.

For example, if K is a 2-complex in the plane, the boundary of K usually consists of several simple closed polygons (i.e, 1 dimensional complexes homeomorphic to the circle, S^1).

Proposition 7.1.9 *Let K be any pure complex of dimension d. For any k-face, σ, of K the boundary complex,* $\mathrm{bd}(\mathrm{Lk}(\sigma))$, *is nonempty iff σ is a boundary face of K ($0 \leq k \leq d-2$). Furthermore,*

$$\mathrm{Lk}_{\mathrm{bd}(K)}(\sigma) = \mathrm{bd}(\mathrm{Lk}(\sigma))$$

for every face, σ, of $\mathrm{bd}(K)$, *where* $\mathrm{Lk}_{\mathrm{bd}(K)}(\sigma)$ *denotes the link of σ in* $\mathrm{bd}(K)$.

Proposition 7.1.9 shows that if every face of K is nonsingular, then the link of every internal face is a sphere whereas the link of every external face is a ball.

Proposition 7.1.10 *Let K be any pure complex of dimension d. If every vertex of K is nonsingular, then* $\mathrm{st}(\sigma) \approx B^d$ *for every k-face, σ, of K ($1 \leq k \leq d-1$).*

Here are more useful propositions about pure complexes without singularities.

Proposition 7.1.11 *Let K be any pure complex of dimension d. If every vertex of K is nonsingular, then for every point, $a \in |K|$, there is an open subset, $U \subseteq |K|$, containing a such that $U \approx B^d$ or $U \approx B^d \cap \mathbb{H}^d$, where $\mathbb{H}^d = \{(x_1, \ldots, x_d) \in \mathbb{R}^d \mid x_d \geq 0\}$.*

Proposition 7.1.12 *Let K be any pure complex of dimension d. If every facet of K is nonsingular, then every facet of K, is contained in at most two cells (d-simplices).*

Proposition 7.1.13 *Let K be any pure and connected complex of dimension d. If every face of K is nonsingular, then for every pair of cells (d-simplices), σ and σ', there is a sequence of cells, $\sigma_0, \ldots, \sigma_p$, with $\sigma_0 = \sigma$ and $\sigma_p = \sigma'$, and such that σ_i and σ_{i+1} have a common facet, for $i = 0, \ldots, p-1$.*

Proposition 7.1.14 *Let K be any pure complex of dimension d. If every facet of K is nonsingular, then the boundary,* $\mathrm{bd}(K)$*, of K is a pure complex of dimension $d-1$ with an empty boundary. Furthermore, if every face of K is nonsingular, then every face of* $\mathrm{bd}(K)$ *is also nonsingular.*

The building blocks of simplicial complexes, namely, simplicies, are in some sense mathematically ideal. However, in practice, it may be desirable to use a more flexible set of building blocks.

We can indeed do this and use convex polytopes as our building blocks.

Definition 7.1.15 A *polyhedral complex in* $\mathbb{E}^m$ (for short, a *complex* in $\mathbb{E}^m$) is a set, K, consisting of a (finite or infinite) set of convex polytopes in $\mathbb{E}^m$ satisfying the following conditions:

(1) Every face of a polytope in K also belongs to K.

(2) For any two polytopes σ_1 and σ_2 in K, if $\sigma_1 \cap \sigma_2 \neq \emptyset$, then $\sigma_1 \cap \sigma_2$ is a common face of both σ_1 and σ_2.

Every polytope, $\sigma \in K$, of dimension k, is called a *k-face* (or *face*) of K. A 0-face $\{v\}$ is called a *vertex* and a 1-face is called an *edge*. The *dimension* of the polyhedral complex K is the maximum of the dimensions of all polytopes in K. If $\dim K = d$, then every face of dimension d is called a *cell* and every face of dimension $d-1$ is called a *facet*.

Every Polytope, P, yields two natural polyhedral complexes:

(i) The polyhedral complex, $\mathcal{K}(P)$, consisting of P together with all of its faces. This complex has a single cell, namely, P itself.

(ii) The *boundary complex*, $\mathcal{K}(\partial P)$, consisting of all faces of P other than P itself. The cells of $\mathcal{K}(\partial P)$ are the facets of P.

The notions of k-skeleton and pureness are defined just as in the simplicial case.

The notions of star and link are defined for polyhedral complexes just as they are defined for simplicial complexes except that the word "face" now means face of a polytope.

Now, by Theorem 6.3.1, every polytope, σ, is the convex hull of its vertices. Let $\mathrm{vert}(\sigma)$ denote the set of vertices of σ.

We have the following crucial observation: Given any poyhedral complex, K, for every point, $x \in |K|$, there is a *unique* polytope, $\sigma_x \in K$, such that
$x \in \mathrm{Int}(\sigma_x) = \sigma_x - \partial\,\sigma_x$.

Now, just as in the simplicial case, the open star, $\mathrm{st}(\sigma)$, of σ is given by

$$\mathrm{st}(\sigma) = \bigcap_{v \in \mathrm{vert}(\sigma)} \mathrm{st}(v).$$

and $\mathrm{st}(\sigma)$ is open in $|K|$.

The next proposition is another result that seems quite obvious, yet a rigorous proof is more involved that we might think.

This proposition states that a convex polytope can always be cut up into simplices, that is, it can be subdivided into a simplicial complex.

In other words, every convex polytope can be triangulated. This implies that simplicial complexes are as general as polyhedral complexes.

Proposition 7.1.16 *Every convex d-polytope, P, can be subdivided into a simplicial complex without adding any new vertices, i.e., every convex polytope can be triangulated.*

With all this preparation, it is now quite natural to define combinatorial manifolds.

7.2 Combinatorial and Topological Manifolds

The notion of pure complex without singular faces turns out to be a very good "discrete" approximation of the notion of (topological) manifold because it of its highly to computational nature.

Definition 7.2.1 A *combinatorial d-manifold* is any space, X, homeomorphic to the geometric realization, $|K| \subseteq \mathbb{E}^n$, of some pure (simplicial or polyhedral) complex, K, of dimension d whose faces are all nonsingular. If the link of every k-face of K is homeomorphic to the sphere S^{d-k-1}, we say that X is a combinatorial manifold *without boundary*, else it is a combinatorial manifold *with boundary*.

Other authors use the term *triangulation* for what we call a computational manifold.

It is easy to see that the connected components of a combinatorial 1-manifold are either simple closed polygons or simple chains (simple, means that the interiors of distinct edges are disjoint).

A combinatorial 2-manifold which is connected is also called a *combinatorial surface* (with or without boundary). Proposition 7.1.14 immediately yields the following result:

Proposition 7.2.2 *If X is a combinatorial d-manifold with boundary, then $\mathrm{bd}(X)$ is a combinatorial $(d-1)$-manifold without boundary.*

Now, because we are assuming that X sits in some Euclidean space, $\mathbb{E}^n$, the space X is Hausdorff and second-countable.

(Recall that a topological space is second-countable iff there is a countable family of open sets of X, $\{U_i\}_{i\geq 0}$, such that every open subset of X is the union of open sets from this family.)

Since it is desirable to have a good match between manifolds and combinatorial manifolds, we are led to the following definition:

Recall that

$$\mathbb{H}^d = \{(x_1, \ldots, x_d) \in \mathbb{R}^d \mid x_d \geq 0\}.$$

Definition 7.2.3 For any $d \geq 1$, a *(topological) d-manifold with boundary* is a second-countable, topological Hausdorff space M, together with an open cover, $(U_i)_{i \in I}$, of open sets in M and a family, $(\varphi_i)_{i \in I}$, of homeomorphisms, $\varphi_i \colon U_i \to \Omega_i$, where each Ω_i is some open subset of $\mathbb{H}^d$ in the subset topology.

Each pair (U, φ) is called a *coordinate system*, or *chart*, of M, each homeomorphism $\varphi_i \colon U_i \to \Omega_i$ is called a *coordinate map*, and its inverse $\varphi_i^{-1} \colon \Omega_i \to U_i$ is called a *parameterization* of U_i. The family $(U_i, \varphi_i)_{i \in I}$ is often called an *atlas* for M.

A *(topological) bordered surface* is a connected 2-manifold with boundary. If for every homeomorphism, $\varphi_i \colon U_i \to \Omega_i$, the open set $\Omega_i \subseteq \mathbb{H}^d$ is actually an open set in $\mathbb{R}^d$ (which means that $x_d > 0$ for every $(x_1, \ldots, x_d) \in \Omega_i$), then we say that M is a *d-manifold*.

Note that a d-manifold is also a d-manifold with boundary.

Letting $\partial\mathbb{H}^d = \mathbb{R}^{d-1} \times \{0\}$, it can be shown using homology, that if some coordinate map, φ, defined on p maps p into $\partial\mathbb{H}^d$, then every coordinate map, ψ, defined on p maps p into $\partial\mathbb{H}^d$.

Thus, M is the disjoint union of two sets ∂M and $\operatorname{Int} M$, where ∂M is the subset consisting of all points $p \in M$ that are mapped by some (in fact, all) coordinate map, φ, defined on p into $\partial\mathbb{H}^d$, and where $\operatorname{Int} M = M - \partial M$.

The set ∂M is called the *boundary* of M, and the set $\operatorname{Int} M$ is called the *interior* of M, even though this terminology clashes with some prior topological definitions.

A good example of a bordered surface is the Möbius strip. The boundary of the Möbius strip is a circle.

The boundary ∂M of M may be empty, but $\operatorname{Int} M$ is nonempty. Also, it can be shown using homology, that the integer d is unique.

It is clear that $\operatorname{Int} M$ is open, and an d-manifold, and that ∂M is closed.

It is easy to see that ∂M is an $(d-1)$-manifold.

Proposition 7.2.4 *Every combinatorial d-manifold is a d-manifold with boundary.*

Proof. This is an immediate consequence of Proposition 7.1.11. □

Is the converse of Proposition 7.2.4 true?

It turns out that answer is yes for $d = 1, 2, 3$ but **no** for $d \geq 4$. This is not hard to prove for $d = 1$.

For $d = 2$ and $d = 3$, this is quite hard to prove; among other things, it is necessary to prove that triangulations exist and this is very technical.

For $d \geq 4$, not every manifold can be triangulated (in fact, this is undecidable!).

What if we assume that M is a triangulated manifold, which means that $M \approx |K|$, for some pure d-dimensional complex, K?

Surprinsingly, for $d \geq 5$, there are triangulated manifolds whose links are not spherical (i.e., not homeomorphic to $\overline{B}^{d-k-1}$ or S^{d-k-1}).

Fortunately, we will only have to deal with $d = 2, 3$!

Another issue that must be addressed is orientability.

Assume that fix a total ordering of the vertices of a complex, K. Let $\sigma = (v_0, \ldots, v_k)$ be any simplex.

Recall that every permutation (of $\{0, \ldots, k\}$) is a product of *transpositions*, where a transposition swaps two distinct elements, say i and j, and leaves every other element fixed.

Furthermore, for any permutation, π, the parity of the number of transpositions needed to obtain π only depends on π and it called the *signature* of π.

We say that two permutations are equivalent iff they have the same signature. Consequently, there are two equivalence classes of permutations: Those of even signature and those of odd signature.

Then, an *orientation* of σ is the choice of one of the two equivalence classes of permutations of its vertices. If σ has been given an orientation, then we denote by $-\sigma$ the result of assigning the other orientation to it (we call it the *opposite orientation*).

For example, $(0, 1, 2)$ has the two orientation classes:

$$\{(0, 1, 2), (1, 2, 0), (2, 0, 1)\}$$

and

$$\{(2, 1, 0), (1, 0, 2), (0, 2, 1)\}.$$

Definition 7.2.5 Let $X \approx |K|$ be a combinatorial d-manifold. We say that X is *orientable* if it is possible to assign an orientation to all of its cells (d-simplices) so that whenever two cells σ_1 and σ_2 have a common facet, σ, the two orientations induced by σ_1 and σ_2 on σ are opposite. A combinatorial d-manifold together with a specific orientation of its cells is called an *oriented manifold*. If X is not orientable we say that it is *non-orientable*.

There are non-orientable (combinatorial) surfaces, for example, the Möbius strip which can be realized in $\mathbb{E}^3$. The Möbius strip is a surface with boundary, its boundary being a circle.

There are also non-orientable (combinatorial) surfaces such as the Klein bottle or the projective plane but they can only be realized in $\mathbb{E}^4$ (in $\mathbb{E}^3$, they must have singularities such as self-intersection).

We will only be dealing with orientable manifolds, and most of the time, surfaces.

One of the most important invariants of combinatorial (and topological) manifolds is their *Euler characteristic*.

In the next chapter, we prove a famous formula due to Poincaré giving the Euler characteristic of a convex polytope. For this, we will introduce a technique of independent interest called *shelling*.

Chapter 8

Shellings, the Euler-Poincaré Formula for Polytopes, Dehn-Sommerville Equations, the Upper Bound Theorem

8.1 Shellings

The notion of shellability is motivated by the desire to give an inductive proof of the Euler-Poincaré formula in any dimension.

Historically, this formula was discovered by Euler for three dimensional polytopes in 1752 (but it was already known to Descartes around 1640).

If f_0, f_1 and f_2 denote the number of vertices, edges and triangles of the three dimensional polytope, P, (i.e., the number of i-faces of P for $i = 0, 1, 2$), then the *Euler formula* states that

$$f_0 - f_1 + f_2 = 2.$$

The proof of Euler's formula is not very difficult but one still has to exercise caution.

Euler's formula was generalized to arbitrary d-dimensional polytopes by Schläfli (1852) but the first correct proof was given by Poincaré (1893, 1899).

If f_i denotes the number of i-faces of the d-dimensional polytope, P, (with $f_{-1} = 1$ and $f_d = 1$), the *Euler-Poincaré formula* states that:

$$\sum_{i=0}^{d-1}(-1)^i f_i = 1 - (-1)^d,$$

which can also be written as

$$\sum_{i=0}^{d}(-1)^i f_i = 1,$$

by incorporating $f_d = 1$ in the first formula or as

$$\sum_{i=-1}^{d}(-1)^i f_i = 0,$$

by incorporating both $f_{-1} = 1$ and $f_d = 1$ in the first formula.

Earlier inductive "proofs" of the above formula were proposed, notably a proof by Schläfli in 1852, but it was later observed that all these proofs assume that the boundary of every polytope can be built up inductively in a nice way, what we call *shellability*.

Actually, counter-examples of shellability for various simplicial complexes suggested that polytopes were perhaps not shellable.

However, the fact that polytopes are shellable was finally proved in 1970 by Bruggesser and Mani [**?**] and soon after that (also in 1970) a striking application of shellability was made by McMullen [**?**] who gave the first proof of the so-called "upper bound theorem".

As shellability of polytopes is an important tool and as it yields one of the cleanest inductive proof of the Euler-Poincaré formula, we will sketch its proof in some details.

Definition 8.1.1 Let K be a pure polyhedral complex of dimension d. A *shelling of* K is a list, $F_1, \ldots, F_s$, of the cells (i.e., d-faces) of K such that either $d = 0$ (and thus, all F_i are points) or the following conditions hold:

(i) The boundary complex, $\mathcal{K}(\partial F_1)$, of the first cell, F_1, of K has a shelling.

(ii) For any j, $1 < j \leq s$, the intersection of the cell F_j with the previous cells is nonempty and is an initial segment of a shelling of the $(d-1)$-dimensional boundary complex of F_j, that is

$$F_j \cap \left(\bigcup_{i=1}^{j-1} F_i \right) = G_1 \cup G_2 \cup \cdots \cup G_r,$$

for some shelling $G_1, G_2, \ldots, G_r, \ldots, G_t$ of $\mathcal{K}(\partial F_j)$, with $1 \leq r \leq t$. As the intersection should be the initial segment of a shelling for the $(d-1)$-dimensional complex, ∂F_j, it has to be pure $(d-1)$-dimensional and connected for $d > 1$.

A polyhedral complex is *shellable* if it is pure and has a shelling.

Note that shellabiliy is only defined for pure complexes.

Here are some examples of shellable complexes:

(1) Every 0-dimensional complex, that is, evey set of points, is shellable, by definition.

(2) A 1-dimensional complex is a graph without loops and parallel edges. A 1-dimensional complex is shellable iff it is connected, which implies that it has no isolated vertices. Any ordering of the edges, $e_1, \ldots, e_s$, such that $\{e_1, \ldots, e_i\}$ induces a connected subgraph for every i will do. Such an ordering can be defined inductively, due to the connectivity of the graph.

(3) Every simplex is shellable. In fact, any ordering of its facets yields a shelling. This is easily shown by induction on the dimension, since the intersection of any two facets F_i and F_j is a facet of both F_i and F_j.

(4) The d-cubes are shellable. By induction on the dimension, it can be shown that every ordering of the $2d$ facets $F_1, \ldots, F_{2d}$ such that F_1 and F_{2d} are opposite (that is, $F_{2d} = -F_1$) yields a shelling.

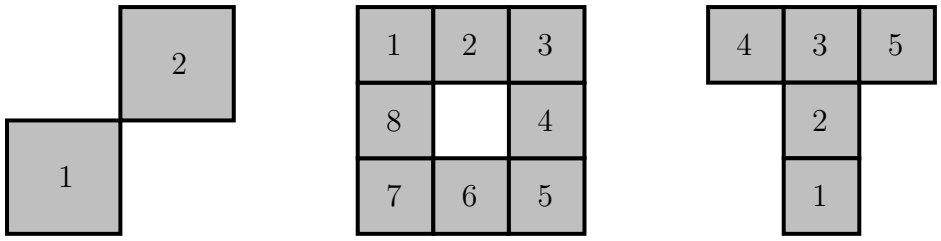

Figure 8.1: Non shellable and Shellable 2-complexes

However, already for 2-complexes, problems arise. For example, in Figure 8.1, the left and the middle 2-complexes are not shellable but the right complex is shellable.

The problem with the left complex is that cells 1 and 2 intersect at a vertex, which is not 1-dimensional, and in the middle complex, the intersection of cell 8 with its predecessors is not connected.

In contrast, the ordering of the right complex is a shelling.

However, observe that the reverse ordering is not a shelling because cell 4 has an empty intersection with cell 5!

Remarks:

1. Condition (i) in Definition 8.1.1 is redundant because, as we shall prove shortly, every polytope is shellable. However, if we want to use this definition for more general complexes, then condition (i) is necessary.
2. When K is a simplicial complex, condition (i) is of course redundant, as every simplex is shellable but condition (ii) can also be simplified to:

   (ii') For any j, with $1 < j \leq s$, the intersection of F_j with the previous cells is nonempty and pure $(d-1)$-dimensional. This means that for every $i < j$ there is some $l < j$ such that $F_i \cap F_j \subseteq F_l \cap F_j$ and $F_l \cap F_j$ is a facet of F_j.

The following proposition yields an important piece of information about the local structure of shellable simplicial complexes:

Proposition 8.1.2 *Let K be a shellable simplicial complex and say $F_1, \ldots, F_s$ is a shelling for K. Then, for every vertex, v, the restriction of the above sequence to the link,* $\mathrm{Lk}(v)$*, and to the star,* $\mathrm{St}(v)$*, are shellings.*

Since the complex, $\mathcal{K}(P)$, associated with a polytope, P, has a single cell, namely P itself, note that by condition (i) in the definition of a shelling, $\mathcal{K}(P)$ is shellable iff the complex, $\mathcal{K}(\partial P)$, is shellable.

We will say simply say that "P is shellable" instead of "$\mathcal{K}(\partial P)$ is shellable".

Proposition 8.1.3 *Given any polytope, P, if $F_1, \ldots, F_s$ is a shelling of P, then the reverse sequence $F_s, \ldots, F_1$ is also a shelling of P.*

Proposition 8.1.3 generally fails for complexes that are not polytopes, see the right 2-complex in Figure 8.1.

We will now present the proof that every polytope is shellable, using a technique invented by Bruggesser and Mani (1970) known as *line shelling* [**?**].

We begin by explaining this idea in the 2-dimensional case, a convex polygon, since it is particularly simple.

Consider the 2-polytope, P, shown in Figure 8.2 (a polygon) whose faces are labeled F_1, F_2, F_3, F_4, F_5.

Pick any line, ℓ, intersecting the interior of P and intersecting the supporting lines of the facets of P (*i.e.*, the edges of P) in distinct points labeled z_1, z_2, z_3, z_4, z_5 (such a line can always be found, as will be shown shortly).

Orient the line, ℓ, (say, upward) and travel on ℓ starting from the point of P where ℓ leaves P, namely, z_1.

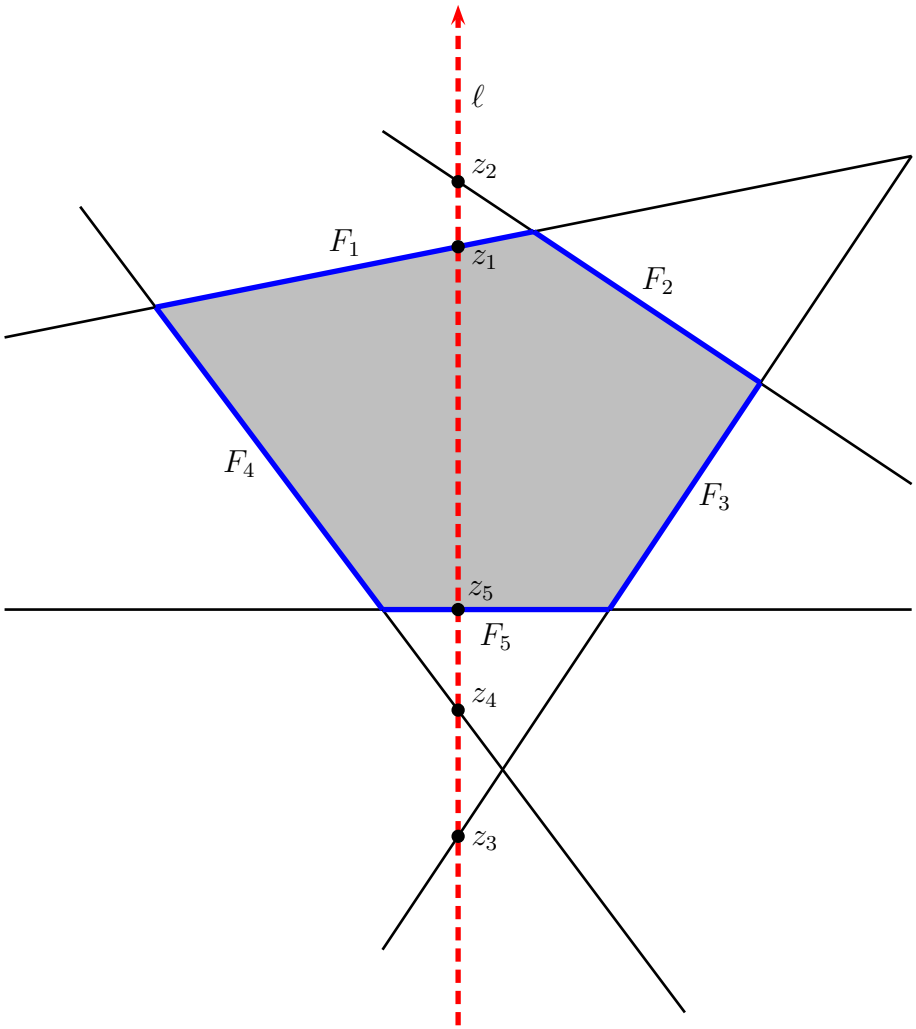

Figure 8.2: Shelling a polygon by travelling along a line

For a while we only see F_1 but then F_2 become visble when we cross z_2. We imagine that we travel very fast and when we reach "$+\infty$" in the upward direction on ℓ, we instantly come back on ℓ from below at "$-\infty$".

At this point, we only see the face of P corresponding to the lowest supporting line of faces of P, i.e., the line corresponding to the smallest z_i, in our case, z_3.

Our trip stops when we reach z_5, the intersection of F_5 and ℓ. During the second phase of our trip, we saw F_3, F_4 and F_5 and the entire trip yields the sequence F_1, F_2, F_3, F_4, F_5, which is easily seen to be a shelling of P.

This is the crux of the Bruggesser-Mani method for shelling a polytope: We travel along a suitably chosen line and record the order in which the faces become visible during this trip. This is why such shellings are called *line shellings*.

In order to prove that polytopes are shellable we need the notion of points and lines in "general position".

Recall from the equivalence of $\mathcal{V}$-polytopes and $\mathcal{H}$-polytopes that a polytope, P, in $\mathbb{E}^d$ with nonempty interior is cut out by t irredundant hyperplanes, H_i, and by picking the origin in the interior of P the equations of the H_i may be assumed to be of the form

$$a_i \cdot z = 1$$

where a_i and a_j are not proportional for all $i \neq j$, so that

$$P = \{z \in \mathbb{E}^d \mid a_i \cdot z \leq 1,\ 1 \leq i \leq t\}.$$

Definition 8.1.4 Let P be any polytope in $\mathbb{E}^d$ with nonempty interior and assume that P is cut out by the irredudant hyperplanes, H_i, of equations $a_i \cdot z = 1$, for $i = 1, \ldots, t$. A point, $c \in \mathbb{E}^d$, is said to be in *general position* w.r.t. P is c does not belong to any of the H_i, that is, if $a_i \cdot c \neq 1$ for $i = 1, \ldots, t$. A line, ℓ, is said to be in *general position* w.r.t. P if ℓ is not parallel to any of the H_i and if ℓ intersects the H_i in distinct points.

The following proposition showing the existence of lines in general position w.r.t. a polytope illustrates a very useful technique, the "perturbation method".

Proposition 8.1.5 *Let P be any polytope in $\mathbb{E}^d$ with nonempty interior. For any two points, x and y in $\mathbb{E}^d$, with x outside of P; y in the interior of P; and x in general position w.r.t. P, for $\lambda \in \mathbb{R}$ small enough, the line, ℓ_λ, through x and y_λ with*

$$y_\lambda = y + (\lambda, \lambda^2, \ldots, \lambda^d),$$

intersects P in its interior and is in general position w.r.t. P.

It should be noted that the perturbation method involving $\Lambda = (\lambda, \lambda^2, \ldots, \lambda^d)$ is quite flexible.

For example, by adapting the proof of Proposition 8.1.5 we can prove that for any two distinct facets, F_i and F_j of P, there is a line in general position w.r.t. P intersecting F_i and F_j. Start with x outside P and very close to F_i and y in the interior of P and very close to F_j.

Given any point, x, strictly outside a polytope, P, we say that a facet, F, of P is *visible from* x iff for every $y \in F$ the line through x and y intersects F only in y (equivalently, x and the interior of P are strictly separared by the supporting hyperplane of F).

We now prove the following fundamental theorem due to Bruggesser and Mani [**?**] (1970):

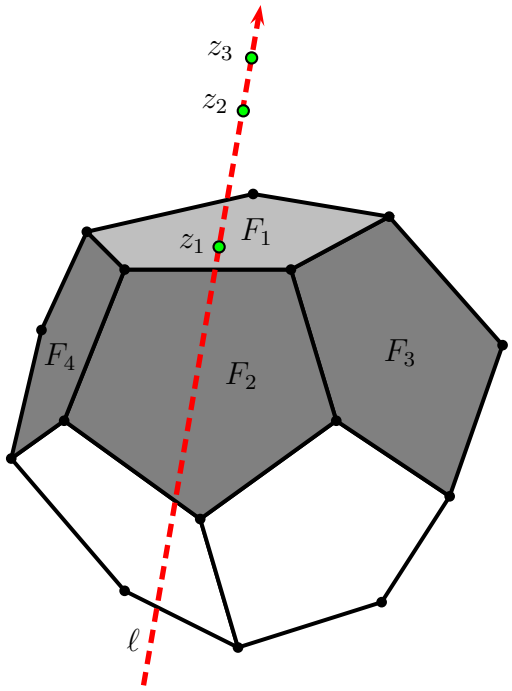

Figure 8.3: Shelling a polytope by travelling along a line, ℓ

Theorem 8.1.6 *(Existence of Line Shellings for Polytopes) Let P be any polytope in $\mathbb{E}^d$ of dimension d. For every point, x, outside P and in general position w.r.t. P, there is a shelling of P in which the facets of P that are visible from x come first.*

Remark: The trip along the line ℓ is often described as a *rocket flight* starting from the surface of P viewed as a little planet (for instance, this is the description given by Ziegler [**?**] (Chapter 8)).

Observe that if we reverse the direction of ℓ, we obtain the reversal of the original line shelling. Thus, the reversal of a line shelling is not only a shelling but a line shelling as well.

We can easily prove the following corollary:

Corollary 8.1.7 *Given any polytope, P, the following facts hold:*

(1) For any two facets F and F', there is a shelling of P in which F comes first and F' comes last.

(2) For any vertex, v, of P, there is a shelling of P in which the facets containing v form an initial segment of the shelling.

Remark: A *plane triangulation*, K, is a pure two-dimensional complex in the plane such that $|K|$ is homeomorphic to a closed disk.

Edelsbrunner proves that every plane triangulation has a shelling and from this, that $\chi(K) = 1$, where $\chi(K) = f_0 - f_1 + f_2$ is the Euler-Poincaré characteristic of K, where f_0 is the number of vertices, f_1 is the number of edges and f_2 is the number of triangles in K (see Edelsbrunner [**?**], Chapter 3).

This result is an immediate consequence of Corollary 8.1.7 if one knows about the stereographic projection map, which will be discussed in the next Chapter.

We now have all the tools needed to prove the famous Euler-Poincaré Formula for Polytopes.

8.2 The Euler-Poincaré Formula for Polytopes

We begin by defining a very important topological concept, the Euler-Poincaré characteristic of a complex.

Definition 8.2.1 Let K be a d-dimensional complex. For every i, with $0 \leq i \leq d$, we let f_i denote the number of i-faces of K and we let

$$\mathbf{f}(K) = (f_0, \cdots, f_d) \in \mathbb{N}^{d+1}$$

be the *f-vector* associated with K (if necessary we write $f_i(K)$ instead of f_i). The *Euler-Poincaré characteristic*, $\chi(K)$, of K is defined by

$$\chi(K) = f_0 - f_1 + f_2 + \cdots + (-1)^d f_d = \sum_{i=0}^{d} (-1)^i f_i.$$

Given any d-dimensional polytope, P, the *f-vector* associated with P is the f-vector associated with $\mathcal{K}(P)$, that is,

$$\mathbf{f}(P) = (f_0, \cdots, f_d) \in \mathbb{N}^{d+1},$$

where f_i, is the number of i-faces of P (= the number of i-faces of $\mathcal{K}(P)$ and thus, $f_d = 1$), and the *Euler-Poincaré characteristic*, $\chi(P)$, of P is defined by

$$\chi(P) = f_0 - f_1 + f_2 + \cdots + (-1)^d f_d = \sum_{i=0}^{d} (-1)^i f_i.$$

Moreover, the *f-vector* associated with the boundary, ∂P, of P is the f-vector associated with $\mathcal{K}(\partial P)$, that is,

$$\mathbf{f}(\partial P) = (f_0, \cdots, f_{d-1}) \in \mathbb{N}^d$$

where f_i, is the number of i-faces of ∂P (with $0 \leq i \leq d-1$), and the *Euler-Poincaré characteristic*, $\chi(\partial P)$, of ∂P is defined by

$$\chi(\partial P) = f_0 - f_1 + f_2 + \cdots + (-1)^{d-1} f_{d-1} = \sum_{i=0}^{d-1} (-1)^i f_i.$$

Observe that $\chi(P) = \chi(\partial P) + (-1)^d$, since $f_d = 1$.

Remark: It is convenient to set $f_{-1} = 1$. Then, some authors, including Ziegler [**?**] (Chapter 8), define the *reduced Euler-Poincaré characteristic*, $\chi'(K)$, of a complex (or a polytope), K, as

$$\begin{aligned}\chi'(K) &= -f_{-1} + f_0 - f_1 + f_2 + \cdots + (-1)^d f_d \\ &= \sum_{i=-1}^{d} (-1)^i f_i = -1 + \chi(K),\end{aligned}$$

i.e., they incorporate $f_{-1} = 1$ into the formula.

A crucial observation for proving the Euler-Poincaré formula is that the Euler-Poincaré characteristic is additive.

This means that if K_1 and K_2 are any two complexes such that $K_1 \cup K_2$ is also a complex, which implies that $K_1 \cap K_2$ is also a complex (because we must have $F_1 \cap F_2 \in K_1 \cap K_2$ for every face F_1 of K_1 and every face F_2 of K_2), then

$$\chi(K_1 \cup K_2) = \chi(K_1) + \chi(K_2) - \chi(K_1 \cap K_2).$$

This follows immediately because for any two sets A and B

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

To prove our next theorem we will use complete induction on $\mathbb{N} \times \mathbb{N}$ ordered by the lexicographic ordering.

Recall that the lexicographic ordering on $\mathbb{N} \times \mathbb{N}$ is defined as follows:

$$(m, n) < (m', n') \quad \text{iff} \quad \begin{cases} m = m' \quad \textit{and} \quad n < n' \\ \text{or} \\ m < m'. \end{cases}$$

Theorem 8.2.2 *(Euler-Poincaré Formula) For every polytope, P, we have*

$$\chi(P) = \sum_{i=0}^{d} (-1)^i f_i = 1 \qquad (d \geq 0),$$

and so,

$$\chi(\partial P) = \sum_{i=0}^{d-1} (-1)^i f_i = 1 - (-1)^d \qquad (d \geq 1).$$

Proof. We prove the following statement: For every d-dimensional polytope, P, if $d = 0$ then

$$\chi(P) = 1,$$

else if $d \geq 1$ then for every shelling $F_1, \ldots, F_{f_{d-1}}$, of P, for every j, with $1 \leq j \leq f_{d-1}$, we have

$$\chi(F_1 \cup \cdots \cup F_j) = \begin{cases} 1 & \text{if } 1 \leq j < f_{d-1} \\ 1 - (-1)^d & \text{if } j = f_{d-1}. \end{cases}$$

We proceed by complete induction on $(d, j) \geq (0, 1)$. □

Remark: Other combinatorial proofs of the Euler-Poincaré formula are given in Grünbaum [**?**] (Chapter 8), Boissonnat and Yvinec [**?**] (Chapter 7) and Ewald [**?**] (Chapter 3).

Coxeter gives a proof very close to Poincaré's own proof using notions of homology theory [**?**] (Chapter IX).

We feel that the proof based on shellings is the most direct and one of the most elegant.

Incidently, the above proof of the Euler-Poincaré formula is very close to Schläfli proof from 1852 but Schläfli did not have shellings at his disposal so his "proof" had a gap. The Bruggesser-Mani proof that polytopes are shellable fills this gap!

8.3 Dehn-Sommerville Equations for Simplicial Polytopes and h-Vectors

If a d-polytope, P, has the property that its faces are all simplices, then it is called a *simplicial polytope*.

It is easily shown that a polytope is simplicial iff its facets are simplices, in which case, every facet has d vertices.

The polar dual of a simplicial polytope is called a *simple polytope*. We see immediately that every vertex of a simple polytope belongs to d facets.

For simplicial (and simple) polytopes it turns out that other remarkable equations besides the Euler-Poincaré formula hold among the number of i-faces.

These equations were discovered by Dehn for $d = 4, 5$ (1905) and by Sommerville in the general case (1927).

Although it is possible (and not difficult) to prove the Dehn-Sommerville equations by "double counting", as in Grünbaum [**?**] (Chapter 9) or Boissonnat and Yvinec (Chapter 7, but beware, these are the dual formulae for simple polytopes), it turns out that instead of using the f-vector associated with a polytope it is preferable to use what's known as the h-vector because for simplicial polytopes the h-numbers have a natural interpretation in terms of shellings.

Furthermore, the statement of the Dehn-Sommerville equations in terms of h-vectors is transparent:

$$h_i = h_{d-i},$$

and the proof is very simple in terms of shellings.

In the rest of this section, we restrict our attention to simplicial complexes.

In order to motivate h-vectors, we begin by examining more closely the structure of the new faces that are created during a shelling when the cell F_j is added to the partial shelling $F_1, \ldots, F_{j-1}$.

If K is a simplicial polytope and V is the set of vertices of K, then every i-face of K can be identified with an $(i+1)$-subset of V (that is, a subset of V of cardinality $i+1$).

Definition 8.3.1 For any shelling, $F_1, \ldots, F_s$, of a simplicial complex, K, of dimension d, for every j, with $1 \leq j \leq s$, the *restriction*, R_j, of the facet, F_j, is the set of "obligatory" vertices

$$R_j = \{v \in F_j \mid F_j - \{v\} \subseteq F_i, \text{ for some } i,\ 1 \leq i < j\}.$$

The crucial property of the R_j is that the new faces, G, added at step j (when F_j is added to the shelling) are precisely the faces in the set

$$I_j = \{G \subseteq V \mid R_j \subseteq G \subseteq F_j\}.$$

But then, we obtain a partition, $\{I_1, \ldots, I_s\}$, of the set of faces of the simplicial complex (other that K itself). Note that the empty face is allowed.

Now, if we define

$$h_i = |\{j \mid |R_j| = i,\ 1 \leq j \leq s\}|,$$

for $i = 0, \ldots, d$, then it turns out that we can recover the f_k in terms of the h_i as follows:

$$f_{k-1} = \sum_{j=1}^{s} \binom{d - |R_j|}{k - |R_j|} = \sum_{i=0}^{k} h_i \binom{d-i}{k-i},$$

with $1 \leq k \leq d$.

But more is true: The above equations are invertible and the h_k can be expressed in terms of the f_i as follows:

$$h_k = \sum_{i=0}^{k} (-1)^{k-i} \binom{d-i}{d-k} f_{i-1},$$

with $0 \leq k \leq d$ (remember, $f_{-1} = 1$).

Let us explain all this in more detail. Consider the example of a connected graph shown in Figure 8.4:

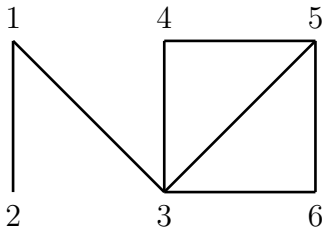

Figure 8.4: A connected 1-dimensional complex, G

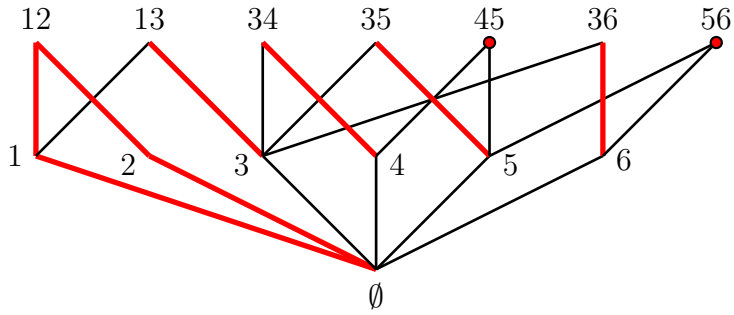

Figure 8.5: the partition associated with a shelling of G

A shelling order of its 7 edges is given by the sequence

$$12,\ 13,\ 34,\ 35,\ 45,\ 36,\ 56.$$

The partial order of the faces of G together with the blocks of the partition $\{I_1, \ldots, I_7\}$ associated with the seven edges of G are shown in Figure 8.5, with the blocks I_j shown in red:

The "minimal" new faces (corresponding to the R_j's) added at every stage of the shelling are

$$\emptyset,\ 3,\ 4,\ 5,\ 45,\ 6,\ 56.$$

Again, if h_i is the number of blocks, I_j, such that the corresponding restriction set, R_j, has size i, that is,

$$h_i = |\{j \mid |R_j| = i,\ 1 \leq j \leq s\}|,$$

for $i = 0, \ldots, d$, where the simplicial polytope, K, has dimension $d-1$, we define the *h-vector* associated with K as

$$\mathbf{h}(K) = (h_0, \ldots, h_d).$$

Then, in the above example, as $R_1 = \{\emptyset\}$, $R_2 = \{3\}$, $R_3 = \{4\}$, $R_4 = \{5\}$, $R_5 = \{4, 5\}$, $R_6 = \{6\}$ and $R_7 = \{5, 6\}$, we get $h_0 = 1$, $h_1 = 4$ and $h_2 = 2$, that is,

$$\mathbf{h}(G) = (1, 4, 2).$$

Now, let us show that if K is a shellable simplicial complex, then the f-vector can be recovered from the h-vector.

Indeed, if $|R_j| = i$, then each $(k-1)$-face in the block I_j must use all i nodes in R_j, so that there are only $d-i$ nodes available and, among those, $k-i$ must be chosen. Therefore,

$$f_{k-1} = \sum_{j=1}^{s} \begin{pmatrix} d - |R_j| \\ k - |R_j| \end{pmatrix}$$

and, by definition of h_i, we get

$$\begin{aligned} f_{k-1} &= \sum_{i=0}^{k} h_i \begin{pmatrix} d-i \\ k-i \end{pmatrix} \\ &= h_k + \begin{pmatrix} d-k+1 \\ 1 \end{pmatrix} h_{k-1} + \cdots + \begin{pmatrix} d \\ k \end{pmatrix} h_0, \end{aligned}$$

where $1 \leq k \leq d$.

Moreover, the formulae are invertible, that is, the h_i can be expressed in terms of the f_k. For this, form the two polynomials

$$f(x) = \sum_{i=0}^{d} f_{i-1} x^{d-i} = f_{d-1} + f_{d-2}x + \cdots + f_0 x^{d-1} + f_{-1} x^d$$

with $f_{-1} = 1$ and

$$h(x) = \sum_{i=0}^{d} h_i x^{d-i} = h_d + h_{d-1}x + \cdots + h_1 x^{d-1} + h_0 x^d.$$

Then, it is easy to see that

$$f(x) = \sum_{i=0}^{d} h_i (x+1)^{d-i} = h(x+1).$$

Consequently, $h(x) = f(x-1)$ and by comparing the coefficients of x^{d-k} on both sides of the above equation, we get

$$h_k = \sum_{i=0}^{k} (-1)^{k-i} \begin{pmatrix} d-i \\ d-k \end{pmatrix} f_{i-1}.$$

In particular, $h_0 = 1$, $h_1 = f_0 - d$, and

$$h_d = f_{d-1} - f_{d-2} + f_{d-3} + \cdots + (-1)^{d-1} f_0 + (-1)^d.$$

It is also easy to check that

$$h_0 + h_1 + \cdots + h_d = f_{d-1}.$$

Now, we just showed that if K is shellable, then its f-vector and its h-vector are related as above.

But even if K is not shellable, the above suggests defining the h-vector from the f-vector as above. Thus, we make the definition:

Definition 8.3.2 For any $(d-1)$-dimensional simplicial complex, K, the *h-vector* associated with K is the vector

$$\mathbf{h}(K) = (h_0, \ldots, h_d) \in \mathbb{Z}^{d+1},$$

given by

$$h_k = \sum_{i=0}^{k} (-1)^{k-i} \begin{pmatrix} d-i \\ d-k \end{pmatrix} f_{i-1}.$$

Note that if K is shellable, then the interpretation of h_i as the number of cells, F_j, such that the corresponding restriction set, R_j, has size i shows that $h_i \geq 0$.

However, for an arbitrary simplicial complex, some of the h_i can be strictly negative. Such an example is given in Ziegler [**?**] (Section 8.3).

We summarize below most of what we just showed:

Proposition 8.3.3 *Let K be a $(d-1)$-dimensional pure simplicial complex. If K is shellable, then its h-vector is nonnegative and h_i counts the number of cells in a shelling whose restriction set has size i. Moreover, the h_i do not depend on the particular shelling of K.*

We are now ready to prove the Dehn-Sommerville equations.

For $d = 3$, these are easily obtained by double counting. Indeed, for a simplicial polytope, every edge belongs to two facets and every facet has three edges. It follows that

$$2f_1 = 3f_2.$$

Together with Euler's formula

$$f_0 - f_1 + f_2 = 2,$$

we see that

$$f_1 = 3f_0 - 6 \quad \text{and} \quad f_2 = 2f_0 - 4,$$

namely, that the number of vertices of a simplicial 3-polytope determines its number of edges and faces, these being linear functions of the number of vertices.

For arbitrary dimension d, we have

Theorem 8.3.4 *(Dehn-Sommerville Equations) If K is any simplicial d-polytope, then the components of the h-vector satisfy*

$$h_k = h_{d-k} \qquad k = 0, 1, \ldots, d.$$

Equivalently

$$f_{k-1} = \sum_{i=k}^{d} (-1)^{d-i} \begin{pmatrix} i \\ k \end{pmatrix} f_{i-1} \qquad k = 0, \ldots, d.$$

Furthermore, the equation $h_0 = h_d$ is equivalent to the Euler-Poincaré formula.

Clearly, the Dehn-Sommerville equations, $h_k = h_{d-k}$, are linearly independent for
$0 \leq k < \lfloor \frac{d+1}{2} \rfloor$.

For example, for $d = 3$, we have the two independent equations

$$h_0 = h_3,\ h_1 = h_2,$$

and for $d = 4$, we also have two independent equations

$$h_0 = h_4,\ h_1 = h_3,$$

since $h_2 = h_2$ is trivial.

When $d = 3$, we know that $h_1 = h_2$ is equivalent to $2f_1 = 3f_2$ and when $d = 4$, if one unravels $h_1 = h_3$ in terms of the f_i' one finds

$$2f_2 = 4f_3,$$

that is $f_2 = 2f_3$.

More generally, it is easy to check that

$$2f_{d-2} = df_{d-1}$$

for all d. For $d = 5$, we find three independent equations

$$h_0 = h_5,\ h_1 = h_4,\ h_2 = h_3,$$

and so on.

It can be shown that for general d-polytopes, the Euler-Poincaré formula is the only equation satisfied by all h-vectors and for simplicial d-polytopes, the $\lfloor \frac{d+1}{2} \rfloor$ Dehn-Sommerville equations, $h_k = h_{d-k}$, are the only equations satisfied by all h-vectors (see Grünbaum [**?**], Chapter 9).

As we saw for 3-dimensional simplicial polytopes, the number of vertices, $n = f_0$, determines the number of edges and the number of faces, and these are linear in f_0.

For $d \geq 4$, this is no longer true and the number of facets is no longer linear in n but in fact quadratic.

It is then natural to ask which d-polytopes with a prescribed number of vertices have the maximum number of k-faces.

This question which remained an open problem for some twenty years was eventually settled by McMullen in 1970 [**?**].

8.4 The Upper Bound Theorem and Cyclic Polytopes

Given a d-polytope with n vertices, what is an upper bound on the number of its i-faces?

This question is not only important from a theoretical point of view but also from a computational point of view because of its implications for algorithms in combinatorial optimization and in computational geometry.

The answer to the above problem is that there is a class of polytopes called *cyclic polytopes* such that the cyclic d-polytope, $C_d(n)$, has the maximum number of i-faces among all d-polytopes with n vertices.

This result stated by Motzkin in 1957 became known as the *upper bound conjecture* until it was proved by McMullen in 1970, using shellings [**?**] (just after Bruggesser and Mani's proof that polytopes are shellable). It is now known as the *upper bound theorem*.

Another proof of the upper bound theorem was given later by Alon and Kalai [**?**] (1985). A version of this proof can also be found in Ewald [**?**] (Chapter 3).

First, consider the cases $d = 2$ and $d = 3$. When $d = 2$, our polytope is a polygon in which case $n = f_0 = f_1$. Thus, this case is trivial.

For $d = 3$, we claim that $2f_1 \geq 3f_2$. Indeed, every edge belongs to exactly two faces so if we add up the number of sides for all faces, we get $2f_1$. Since every face has at least three sides, we get $2f_1 \geq 3f_2$. Then, using Euler's relation, it is easy to show that

$$f_1 \leq 6n - 3 \quad f_2 \leq 2n - 4$$

and we know that equality is achieved for simplicial polytopes.

Let us now consider the general case. The rational curve, $c\colon \mathbb{R} \to \mathbb{R}^d$, given parametrically by

$$c(t) = (t, t^2, \ldots, t^d)$$

is at the heart of the story.

This curve if often called the *moment curve* or *rational normal curve* of degree d. For $d = 3$, it is known as the *twisted cubic*. Here is the definition of the cyclic polytope, $C_d(n)$.

Definition 8.4.1 For any sequence, $t_1 < \ldots < t_n$, of distinct real number, $t_i \in \mathbb{R}$, with $n > d$, the convex hull,

$$C_d(n) = \text{conv}(c(t_1), \ldots, c(t_n))$$

of the n points, $c(t_1), \ldots, c(t_n)$, on the moment curve of degree d is called a *cyclic polytope*.

The first interesting fact about the cyclic polytope is that it is simplicial.

Proposition 8.4.2 *Every $d+1$ of the points $c(t_1), \ldots, c(t_n)$ are affinely independent. Consequently, $C_d(n)$ is a simplicial polytope and the $c(t_i)$ are vertices.*

Proposition 8.4.3 *For any k with $2 \leq 2k \leq d$, every subset of k vertices of $C_d(n)$ is a $(k-1)$-face of $C_d(n)$. Hence*

$$f_k(C_d(n)) = \begin{pmatrix} n \\ k+1 \end{pmatrix} \qquad \text{if} \quad 0 \leq k < \left\lfloor \frac{d}{2} \right\rfloor .$$

Observe that Proposition 8.4.3 shows that any subset of $\lfloor \frac{d}{2} \rfloor$ vertices of $C_d(n)$ forms a face of $C_d(n)$.

When a d-polytope has this property it is called a *neighborly polytope*. Therefore, cyclic polytopes are neighborly.

Proposition 8.4.3 also shows a phenomenon that only manifests itself in dimension at least 4: For $d \geq 4$, the polytope $C_d(n)$ has n pairwise adjacent vertices. For $n >> d$, this is counter-intuitive.

Finally, the combinatorial structure of cyclic polytopes is completely determined as follows:

Proposition 8.4.4 *(Gale evenness condition, Gale (1963)). Let n and d be integers with $2 \leq d < n$. For any sequence $t_1 < t_2 < \cdots < t_n$, consider the cyclic polytope*

$$C_d(n) = \mathrm{conv}(c(t_1), \ldots, c(t_n)).$$

A subset, $S \subseteq \{1, \ldots, n\}$ with $|S| = d$ determines a facet of $C_d(n)$ iff for all $i < j$ not in S, then the number of $k \in S$ between i and j is even:

$$|\{k \in S \mid i < k < j\}| \equiv 0 \,(\mathrm{mod}\, 2) \quad \textit{for} \quad i, j \notin S$$

In particular, Proposition 8.4.4 shows that the combinatorial structure of $C_d(n)$ does not depend on the specific choice of the sequence $t_1 < \ldots < t_n$. This justifies our notation $C_d(n)$.

Here is the celebrated upper bound theorem first proved by McMullen [**?**].

Theorem 8.4.5 *(Upper Bound Theorem, McMullen (1970)) Let P be any d-polytope with n vertices. Then, for every k, with $1 \leq k \leq d$, the polytope P has at most as many $(k-1)$-faces as the cyclic polytope, $C_d(n)$, that is*

$$f_{k-1}(P) \leq f_{k-1}(C_d(n)).$$

Moreover, equality for some k with $\lfloor \frac{d}{2} \rfloor \leq k \leq d$ implies that P is neighborly.

The first step in the proof of Theorem 8.4.5 is to prove that among all d-polytopes with a given number, n, of vertices, the maximum number of i-faces is achieved by simplicial d-polytopes.

Proposition 8.4.6 *Given any d-polytope, P, with n-vertices, it is possible to form a simplicial polytope, P', by perturbing the vertices of P such that P' also has n vertices and*

$$f_{k-1}(P) \leq f_{k-1}(P') \qquad \text{for} \quad 1 \leq k \leq d.$$

Furthermore, equality for $k > \lfloor \frac{d}{2} \rfloor$ can occur only if P is simplicial.

Proposition 8.4.6 allows us to restict our attention to simplicial polytopes. Now, it is obvious that

$$f_{k-1} \leq \binom{n}{k}$$

for any polytope P (simplicial or not) and we also know that equality holds if $k \leq \lfloor \frac{d}{2} \rfloor$ for neighborly polytopes such as the cyclic polytopes. For $k > \lfloor \frac{d}{2} \rfloor$, it turns out that equality can only be achieved for simplices.

However, for a *simplicial* polytope, the Dehn-Sommerville equations $h_k = h_{d-k}$ together with the equations giving f_k in terms of the h_i's show that $f_0, f_1, \ldots, f_{\lfloor \frac{d}{2} \rfloor}$ already determine the whole f-vector.

Thus, it is possible to express the f_{k-1} in terms of $h_0, h_1, \ldots, h_{\lfloor \frac{d}{2} \rfloor}$ for $k \geq \lfloor \frac{d}{2} \rfloor$. It turns out that we get

$$f_{k-1} = \sum_{i=0}^{\lfloor \frac{d}{2} \rfloor}{}^{*} \left(\binom{d-i}{k-i} + \binom{i}{k-d+i} \right) h_i,$$

where the meaning of the superscript $*$ is that when d is even we only take half of the last term for $i = \frac{d}{2}$ and when d is odd we take the whole last term for $i = \frac{d-1}{2}$ (for details, see Ziegler [**?**], Chapter 8).

As a consequence if we can show that the neighborly polytopes maximize not only f_{k-1} but also h_{k-1} when $k \leq \lfloor \frac{d}{2} \rfloor$, the upper bound theorem will be proved.

Indeed, McMullen proved the following theorem which is "more than enough" to yield the desired result ([**?**]):

Theorem 8.4.7 *(McMullen (1970)) For every simplicial d-polytope with $f_0 = n$ vertices, we have*

$$h_k(P) \leq \binom{n-d-1+k}{k} \qquad \text{for} \quad 0 \leq k \leq d.$$

Furthermore, equality holds for all l and all k with $0 \leq k \leq l$ iff $l \leq \lfloor \frac{d}{2} \rfloor$ and P is l-neighborly. (a polytope is l-neighborly iff any subset of l or less vertices determine a face of P.)

Since cyclic d-polytopes are neighborly (which means that they are $\lfloor \frac{d}{2} \rfloor$-neighborly), Theorem 8.4.5 follows from Proposition 8.4.6, and Theorem 8.4.7.

Corollary 8.4.8 *For every simplicial neighborly d-polytope with n vertices, we have*

$$f_{k-1} = \sum_{i=0}^{\lfloor \frac{d}{2} \rfloor}{}^{*} \left(\binom{d-i}{k-i} + \binom{i}{k-d+i} \right) \binom{n-d-1+i}{i},$$

for $1 \leq k \leq d$. This gives the maximum number of $(k-1)$-faces for any d-polytope with n-vertices, for all k with $1 \leq k \leq d$. In particular, the number of facets of the cyclic polytope, $C_d(n)$, is

$$f_{d-1} = \sum_{i=0}^{\lfloor \frac{d}{2} \rfloor}{}^{*} 2 \binom{n-d-1+i}{i}$$

and, more explicitly,

$$f_{d-1} = \binom{n - \lfloor \frac{d+1}{2} \rfloor}{n-d} + \binom{n - \lfloor \frac{d+2}{2} \rfloor}{n-d}.$$

Corollary 8.4.8 implies that the number of facets of any d-polytope is $O(n^{\lfloor \frac{d}{2} \rfloor})$.

An unfortunate consequence of this upper bound is that the complexity of any convex hull algorithms for n points in $\mathbb{E}^d$ is $O(n^{\lfloor \frac{d}{2} \rfloor})$.

The $O(n^{\lfloor \frac{d}{2} \rfloor})$ upper bound can be obtained more directly using a pretty argument using shellings due to R. Seidel [**?**].

Remark: There is also a *lower bound theorem* due to Barnette (1971, 1973) which gives a lower bound on the f-vectors all d-polytopes with n vertices.

In this case, there is an analog of the cyclic polytopes called *stacked polytopes*.

These polytopes, $P_d(n)$, are simplicial polytopes obtained from a simplex by building shallow pyramids over the facets of the simplex. Then, it turns out that if $d \geq 2$, then

$$f_k \geq \begin{cases} \binom{d}{k} n - \binom{d+1}{k+1} k & \text{if } 0 \leq k \leq d-2 \\ (d-1)n - (d+1)(d-2) & \text{if } k = d-1. \end{cases}$$

There has been a lot of progress on the combinatorics of f-vectors and h-vectors since 1971, especially by R. Stanley, G. Kalai and L. Billera and K. Lee, among others. We recommend two excellent surveys:

1. Bayer and Lee [**?**] summarizes progress in this area up to 1993.

2. Billera and Björner [**?**] is a more advanced survey which reports on results up to 1997.

In fact, many of the chapters in Goodman and O'Rourke [**?**] should be of interest to the reader.

Chapter 9

Dirichlet–Voronoi Diagrams and Delaunay Triangulations

9.1 Dirichlet–Voronoi Diagrams

In this chapter we present very briefly the concepts of a Voronoi diagram and of a Delaunay triangulation.

These are important tools in computational geometry, and Delaunay triangulations are important in problems where it is necessary to fit 3D data using surface splines.

It is usually useful to compute a good mesh for the projection of this set of data points onto the xy-plane, and a Delaunay triangulation is a good candidate.

Our presentation will be rather sketchy. We are primarily interested in defining these concepts and stating their most important properties without proofs.

For a comprehensive exposition of Voronoi diagrams, Delaunay triangulations, and more topics in computational geometry, consult O'Rourke [**?**], Preparata and Shamos [**?**], Boissonnat and Yvinec [**?**], de Berg, Van Kreveld, Overmars, and Schwarzkopf [**?**], or Risler [**?**].

The survey by Graham and Yao [**?**] contains a very gentle and lucid introduction to computational geometry.

For concreteness, one may safely assume that we work in the affine space $\mathcal{E} = \mathbb{E}^m$, although what follows applies to any Euclidean space of finite dimension.

Given a set $P = \{p_1, \ldots, p_n\}$ of n points in $\mathcal{E}$, it is often useful to find a partition of the space $\mathcal{E}$ into regions each containing a single point of P and having some nice properties.

It is also often useful to find triangulations of the convex hull of P having some nice properties.

We shall see that this can be done and that the two problems are closely related. In order to solve the first problem, we need to introduce bisector lines and bisector planes.

For simplicity, let us first assume that $\mathcal{E}$ is a plane i.e., has dimension 2.

Given any two distinct points $a, b \in \mathcal{E}$, the line orthogonal to the line segment (a, b) and passing through the midpoint of this segment is the locus of all points having equal distance to a and b.

It is called the *bisector line of a and b*. The bisector line of two points is illustrated in Figure 9.1.

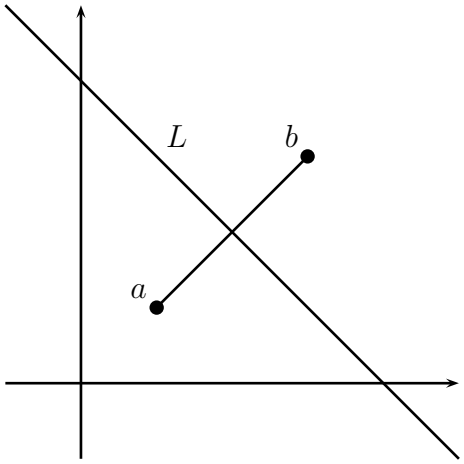

Figure 9.1: The bisector line L of a and b

If $h = \frac{1}{2}\,a + \frac{1}{2}\,b$ is the midpoint of the line segment (a, b), letting m be an arbitrary point on the bisector line, the equation of this line can be found by writing that $\mathbf{hm}$ is orthogonal to $\mathbf{ab}$.

In any orthogonal frame, letting $m = (x, y)$, $a = (a_1, a_2)$, $b = (b_1, b_2)$, the equation of this line can be written as

$$(b_1 - a_1)x + (b_2 - a_2)y = (b_1^2 + b_2^2)/2 - (a_1^2 + a_2^2)/2.$$

The closed half-plane $H(a, b)$ containing a and with boundary the bisector line is the locus of all points such that

$$(b_1 - a_1)x + (b_2 - a_2)y \leq (b_1^2 + b_2^2)/2 - (a_1^2 + a_2^2)/2,$$

and the closed half-plane $H(b, a)$ containing b and with boundary the bisector line is the locus of all points such that

$$(b_1 - a_1)x + (b_2 - a_2)y \geq (b_1^2 + b_2^2)/2 - (a_1^2 + a_2^2)/2.$$

The closed half-plane $H(a, b)$ is the set of all points whose distance to a is less that or equal to the distance to b, and vice versa for $H(b, a)$. Thus, points in the closed half-plane $H(a, b)$ are closer to a than they are to b.

We now consider a problem called the *post office problem* by Graham and Yao [**?**].

Given any set $P = \{p_1, \ldots, p_n\}$ of n points in the plane (considered as *post offices* or *sites*), for any arbitrary point x, find out which post office is closest to x.

Since x can be arbitrary, it seems desirable to precompute the sets $V(p_i)$ consisting of all points that are closer to p_i than to any other point $p_j \neq p_i$.

Indeed, if the sets $V(p_i)$ are known, the answer is any post office p_i such that $x \in V(p_i)$.

Thus, it remains to compute the sets $V(p_i)$. For this, if x is closer to p_i than to any other point $p_j \neq p_i$, then x is on the same side as p_i with respect to the bisector line of p_i and p_j for every $j \neq i$, and thus

$$V(p_i) = \bigcap_{j \neq i} H(p_i, p_j).$$

If $\mathcal{E}$ has dimension 3, the locus of all points having equal distance to a and b is a plane. It is called the *bisector plane of a and b*.

The equation of this plane is also found by writing that **hm** is orthogonal to **ab**. The equation of this plane can be written as

$$(b_1 - a_1)x + (b_2 - a_2)y + (b_3 - a_3)z = \\ (b_1^2 + b_2^2 + b_3^2)/2 - (a_1^2 + a_2^2 + a_3^2)/2.$$

The closed half-space $H(a, b)$ containing a and with boundary the bisector plane is the locus of all points such that

$$(b_1 - a_1)x + (b_2 - a_2)y + (b_3 - a_3)z \leq \\ (b_1^2 + b_2^2 + b_3^2)/2 - (a_1^2 + a_2^2 + a_3^2)/2,$$

and the closed half-space $H(b, a)$ containing b and with boundary the bisector plane is the locus of all points such that

$$(b_1 - a_1)x + (b_2 - a_2)y + (b_3 - a_3)z \geq \\ (b_1^2 + b_2^2 + b_3^2)/2 - (a_1^2 + a_2^2 + a_3^2)/2.$$

The closed half-space $H(a, b)$ is the set of all points whose distance to a is less that or equal to the distance to b, and vice versa for $H(b, a)$. Again, points in the closed half-space $H(a, b)$ are closer to a than they are to b.

Given any set $P = \{p_1, \ldots, p_n\}$ of n points in $\mathcal{E}$ (of dimension $m = 2, 3$), it is often useful to find for every point p_i the region consisting of all points that are closer to p_i than to any other point $p_j \neq p_i$, that is, the set

$$V(p_i) = \{x \in \mathcal{E} \mid d(x, p_i) \leq d(x, p_j), \text{ for all } j \neq i\},$$

where $d(x, y) = (\mathbf{xy} \cdot \mathbf{xy})^{1/2}$, the Euclidean distance associated with the inner product $\cdot$ on $\mathcal{E}$.

From the definition of the bisector line (or plane), it is immediate that

$$V(p_i) = \bigcap_{j \neq i} H(p_i, p_j).$$

Families of sets of the form $V(p_i)$ were investigated by Dirichlet [**?**] (1850) and Voronoi [**?**] (1908). Voronoi diagrams also arise in crystallography (Gilbert [**?**]).

Other applications, including facility location and path planning, are discussed in O'Rourke [**?**]. For simplicity, we also denote the set $V(p_i)$ by V_i, and we introduce the following definition.

Definition 9.1.1 Let $\mathcal{E}$ be a Euclidean space of dimension $m \geq 1$. Given any set $P = \{p_1, \ldots, p_n\}$ of n points in $\mathcal{E}$, the *Dirichlet–Voronoi diagram* $\mathcal{V}(P)$ of $P = \{p_1, \ldots, p_n\}$ is the family of subsets of $\mathcal{E}$ consisting of the sets $V_i = \bigcap_{j \neq i} H(p_i, p_j)$ and of all of their intersections.

Dirichlet–Voronoi diagrams are also called *Voronoi diagrams*, *Voronoi tessellations*, or *Thiessen polygons*. Following common usage, we will use the terminology *Voronoi diagram*.

As intersections of convex sets (closed half-planes or closed half-spaces), the *Voronoi regions* $V(p_i)$ are convex sets. In dimension two, the boundaries of these regions are convex polygons, and in dimension three, the boundaries are convex polyhedra.

Whether a region $V(p_i)$ is bounded or not depends on the location of p_i.

If p_i belongs to the boundary of the convex hull of the set P, then $V(p_i)$ is unbounded, and otherwise bounded.

In dimension two, the convex hull is a convex polygon, and in dimension three, the convex hull is a convex polyhedron.

As we will see later, there is an intimate relationship between convex hulls and Voronoi diagrams.

Generally, if $\mathcal{E}$ is a Euclidean space of dimension m, given any two distinct points $a, b \in \mathcal{E}$, the locus of all points having equal distance to a and b is a hyperplane.

It is called the *bisector hyperplane of a and b*. The equation of this hyperplane is still found by writing that **hm** is orthogonal to **ab**. The equation of this hyperplane can be written as

$$(b_1 - a_1)x_1 + \cdots + (b_m - a_m)x_m = \\ (b_1^2 + \cdots + b_m^2)/2 - (a_1^2 + \cdots + a_m^2)/2.$$

The closed half-space $H(a, b)$ containing a and with boundary the bisector hyperplane is the locus of all points such that

$$(b_1 - a_1)x_1 + \cdots + (b_m - a_m)x_m \leq \\ (b_1^2 + \cdots + b_m^2)/2 - (a_1^2 + \cdots + a_m^2)/2,$$

and the closed half-space $H(b, a)$ containing b and with boundary the bisector hyperplane is the locus of all points such that

$$(b_1 - a_1)x_1 + \cdots + (b_m - a_m)x_m \geq \\ (b_1^2 + \cdots + b_m^2)/2 - (a_1^2 + \cdots + a_m^2)/2.$$

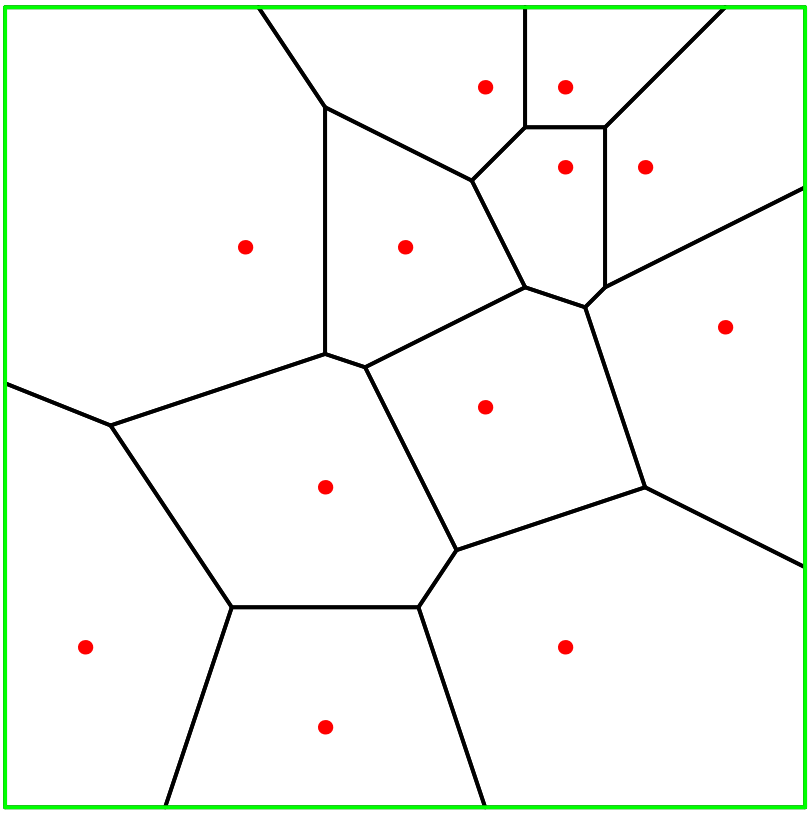

Figure 9.2: A Voronoi diagram

The closed half-space $H(a, b)$ is the set of all points whose distance to a is less than or equal to the distance to b, and vice versa for $H(b, a)$.

Figure 9.2 shows the Voronoi diagram of a set of twelve points.

In the general case where $\mathcal{E}$ has dimension m, the definition of the Voronoi diagram $\mathcal{V}(P)$ of P is the same as Definition 9.1.1, except that $H(p_i, p_j)$ is the closed half-space containing p_i and having the bisector hyperplane of a and b as boundary.

Also, observe that the convex hull of P is a convex polytope.

We will now state a lemma listing the main properties of Voronoi diagrams.

It turns out that certain degenerate situations can be avoided if we assume that if P is a set of points in an affine space of dimension m, then no $m + 2$ points from P belong to the same $(m - 1)$-sphere.

We will say that the points of P are in *general position*.

Thus when $m = 2$, no 4 points in P are cocyclic, and when $m = 3$, no 5 points in P are on the same sphere.

Lemma 9.1.2 *Given a set $P = \{p_1, \ldots, p_n\}$ of n points in some Euclidean space $\mathcal{E}$ of dimension m (say $\mathbb{E}^m$), if the points in P are in general position and not in a common hyperplane then the Voronoi diagram of P satisfies the following conditions:*

(1) Each region V_i is convex and contains p_i in its interior.

(2) Each vertex of V_i belongs to $m+1$ regions V_j and to $m+1$ edges.

(3) The region V_i is unbounded iff p_i belongs to the boundary of the convex hull of P.

(4) If p is a vertex that belongs to the regions $V_1, \ldots, V_{m+1}$, then p is the center of the $(m-1)$-sphere $S(p)$ determined by $p_1, \ldots, p_{m+1}$. Furthermore, no point in P is inside the sphere $S(p)$ (i.e., in the open ball associated with the sphere $S(p)$).

(5) If p_j is a nearest neighbor of p_i, then one of the faces of V_i is contained in the bisector hyperplane of (p_i, p_j).

(6)

$$\bigcup_{i=1}^{n} V_i = \mathcal{E}, \quad \textit{and} \quad \overset{\circ}{V}_i \cap \overset{\circ}{V}_j = \emptyset, \quad \textit{for all } i, j,\ i \neq j,$$

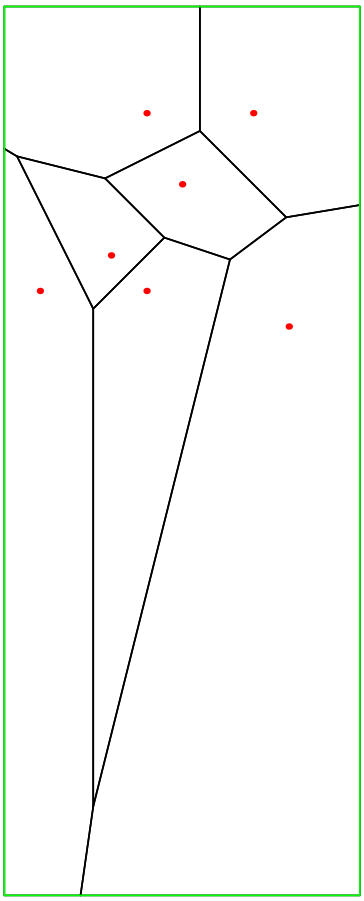

Figure 9.3: Another Voronoi diagram

where $\overset{\circ}{V}_i$ *denotes the interior of* V_i.

For simplicity, let us again consider the case where $\mathcal{E}$ is a plane. It should be noted that certain Voronoi regions, although closed, may extend very far.

Figure 9.3 shows such an example.

It is also possible for certain unbounded regions to have parallel edges.

There are a number of methods for computing Voronoi diagrams. A fairly simple (although not very efficient) method is to compute each Voronoi region $V(p_i)$ by intersecting the half-planes $H(p_i, p_j)$.

One way to do this is to construct successive convex polygons that converge to the boundary of the region.

At every step we intersect the current convex polygon with the bisector line of p_i and p_j. There are at most two intersection points. We also need a starting polygon, and for this we can pick a square containing all the points.

A naive implementation will run in $O(n^3)$.

However, the intersection of half-planes can be done in $O(n \log n)$, using the fact that the vertices of a convex polygon can be sorted.

Thus, the above method runs in $O(n^2 \log n)$. Actually, there are faster methods (see Preparata and Shamos [**?**] or O'Rourke [**?**]), and it is possible to design algorithms running in $O(n \log n)$.

The most direct method to obtain fast algorithms is to use the "lifting method" discussed in Section 9.4, whereby the original set of points is lifted onto a paraboloid, and to use fast algorithms for finding a convex hull.

A very interesting (undirected) graph can be obtained from the Voronoi diagram as follows: The vertices of this graph are the points p_i (each corresponding to a unique region of $\mathcal{V}(P)$), and there is an edge between p_i and p_j iff the regions V_i and V_j share an edge.

The resulting graph is called a *Delaunay triangulation* of the convex hull of P, after Delaunay, who invented this concept in 1934. Such triangulations have remarkable properties.

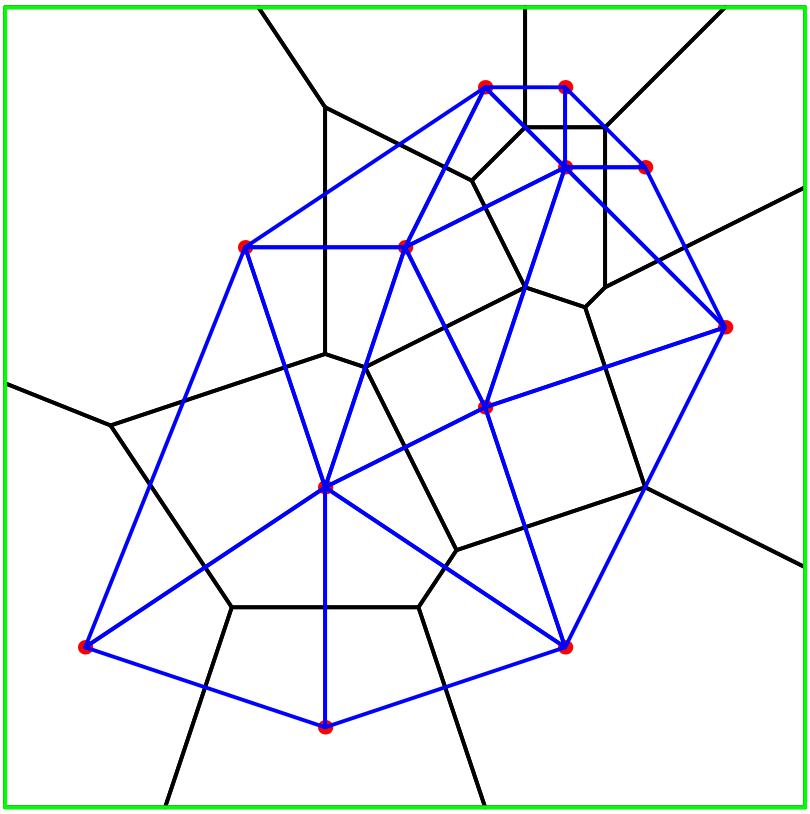

Figure 9.4: Delaunay triangulation associated with a Voronoi diagram

Figure 9.4 shows the Delaunay triangulation associated with the earlier Voronoi diagram of a set of twelve points.

One has to be careful to make sure that all the Voronoi vertices have been computed before computing a Delaunay triangulation, since otherwise, some edges could be missed.

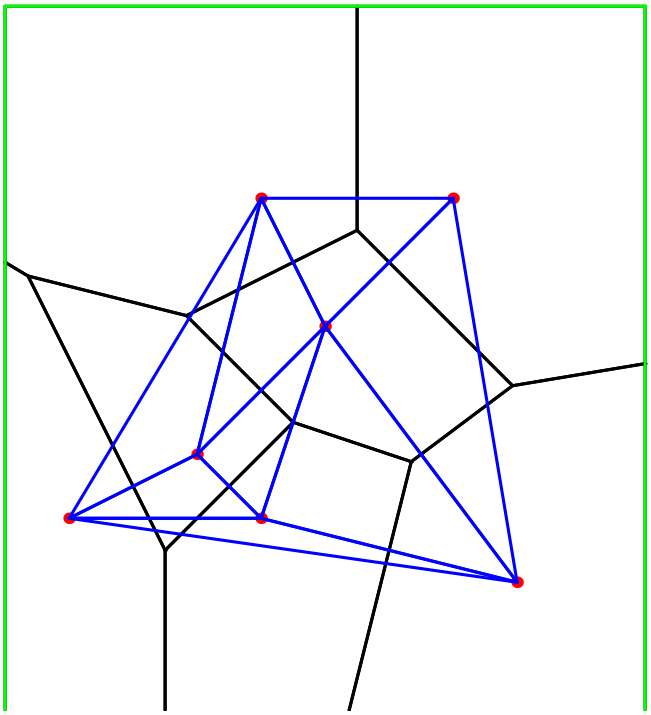

Figure 9.5: Another Delaunay triangulation associated with a Voronoi diagram

In Figure 9.5 illustrating such a situation, if the lowest Voronoi vertex had not been computed (not shown on the diagram!), the lowest edge of the Delaunay triangulation would be missing.

The concept of a triangulation can be generalized to dimension 3, or even to any dimension m.

9.2 Simplicial Complexes and Triangulations

The concept of a triangulation relies on the notion of pure simplicial complex defined in Chapter 7. The reader should review Definition 7.1.2 and Definition 7.1.3.

Definition 9.2.1 Given a subset, $S \subseteq \mathbb{E}^m$ (where $m \geq 1$), a *triangulation of* S is a pure (finite) simplicial complex, K, of dimension m such that $S = |K|$, that is, S is equal to the geometric realization of K.

Given a finite set P of n points in the plane, and given a triangulation of the convex hull of P having P as its set of vertices, observe that the boundary of P is a convex polygon.

Similarly, given a finite set P of points in 3-space, and given a triangulation of the convex hull of P having P as its set of vertices, observe that the boundary of P is a convex polyhedron.

It is interesting to know how many triangulations exist for a set of n points (in the plane or in 3-space), and it is also interesting to know the number of edges and faces in terms of the number of vertices in P.

These questions can be settled using the Euler–Poincaré characteristic.

We say that a polygon in the plane is a *simple polygon* iff it is a connected closed polygon such that no two edges intersect (except at a common vertex).

Lemma 9.2.2

(1) For any triangulation of a region of the plane whose boundary is a simple polygon, letting v be the number of vertices, e the number of edges, and f the number of triangles, we have the "Euler formula"

$$v - e + f = 1.$$

(2) For any region S in $\mathbb{E}^3$ homeomorphic to a closed ball and for any triangulation of S, letting v be the number of vertices, e the number of edges, f the number of triangles, and t the number of tetrahedra, we have the "Euler formula"

$$v - e + f - t = 1.$$

(3) Furthermore, for any triangulation of the combinatorial surface, $B(S)$, that is the boundary of S, letting v' be the number of vertices, e' the number of edges, and f' the number of triangles, we have the "Euler formula"

$$v' - e' + f' = 2.$$

Proof. All the statements are immediate consequences of Theorem 8.2.2.

For example, part (1) is obtained by mapping the triangulation onto a sphere using inverse stereographic projection, say from the North pole.

Then, we get a polytope on the sphere with an extra facet corresponding to the "outside" of the triangulation.

We have to deduct this facet from the Euler characteristic of the polytope and this is why we get 1 instead of 2. □

It is now easy to see that in case (1), the number of edges and faces is a linear function of the number of vertices and boundary edges, and that in case (3), the number of edges and faces is a linear function of the number of vertices.

If there are e_b edges in the boundary and e_i edges not in the boundary, we have

$$3f = e_b + 2e_i,$$

and togeher with

$$v - e_b - e_i + f = 1,$$

we get

$$\begin{aligned} v - e_b - e_i + e_b/3 + 2e_i/3 &= 1, \\ 2e_b/3 + e_i/3 &= v - 1, \end{aligned}$$

and thus, $e_i = 3v - 3 - 2e_b$. Since $f = e_b/3 + 2e_i/3$, we have $f = 2v - 2 - e_b$.

Similarly, since $v' - e' + f' = 2$ and $3f' = 2e'$, we easily get $e = 3v - 6$ and $f = 2v - 4$.

Thus, given a set P of n points, the number of triangles (and edges) for any triangulation of the convex hull of P using the n points in P for its vertices is fixed.

Case (2) is trickier, but it can be shown that

$$v - 3 \leq t \leq (v-1)(v-2)/2.$$

Thus, there can be different numbers of tetrahedra for different triangulations of the convex hull of P.

Remark: The numbers of the form $v - e + f$ and $v - e + f - t$ are called *Euler–Poincaré characteristics*.

They are topological invariants, in the sense that they are the same for all triangulations of a given polytope. This is a fundamental fact of algebraic topology.

We shall now investigate triangulations induced by Voronoi diagrams.

9.3 Delaunay Triangulations

Given a set $P = \{p_1, \ldots, p_n\}$ of n points in the plane and the Voronoi diagram $\mathcal{V}(P)$ for P, we explained in Section 9.1 how to define an (undirected) graph:

The vertices of this graph are the points p_i (each corresponding to a unique region of $\mathcal{V}(P)$), and there is an edge between p_i and p_j iff the regions V_i and V_j share an edge.

The resulting graph turns out to be a triangulation of the convex hull of P having P as its set of vertices. Such a complex can be defined in general.

For any set $P = \{p_1, \ldots, p_n\}$ of n points in $\mathbb{E}^m$, we say that a triangulation of the convex hull of P is *associated with* P if its set of vertices is the set P.

Definition 9.3.1 Let $P = \{p_1, \ldots, p_n\}$ be a set of n points in $\mathbb{E}^m$, and let $\mathcal{V}(P)$ be the Voronoi diagram of P. We define a complex $\mathcal{D}(P)$ as follows:

The complex $\mathcal{D}(P)$ contains the k-simplex $\{p_1, \ldots, p_{k+1}\}$ iff $V_1 \cap \cdots \cap V_{k+1} \neq \emptyset$, where $0 \leq k \leq m$.

The complex $\mathcal{D}(P)$ is called the *Delaunay triangulation of the convex hull of P.*

Thus, $\{p_i, p_j\}$ is an edge iff $V_i \cap V_j \neq \emptyset$, $\{p_i, p_j, p_h\}$ is a triangle iff $V_i \cap V_j \cap V_h \neq \emptyset$, $\{p_i, p_j, p_h, p_k\}$ is a tetrahedron iff $V_i \cap V_j \cap V_h \cap V_k \neq \emptyset$, etc.

For simplicity, we often write $\mathcal{D}$ instead of $\mathcal{D}(P)$. A Delaunay triangulation for a set of twelve points is shown in Figure 9.6.

Actually, it is not obvious that $\mathcal{D}(P)$ is a triangulation of the convex hull of P, but this can be shown, as well as the properties listed in the following lemma.

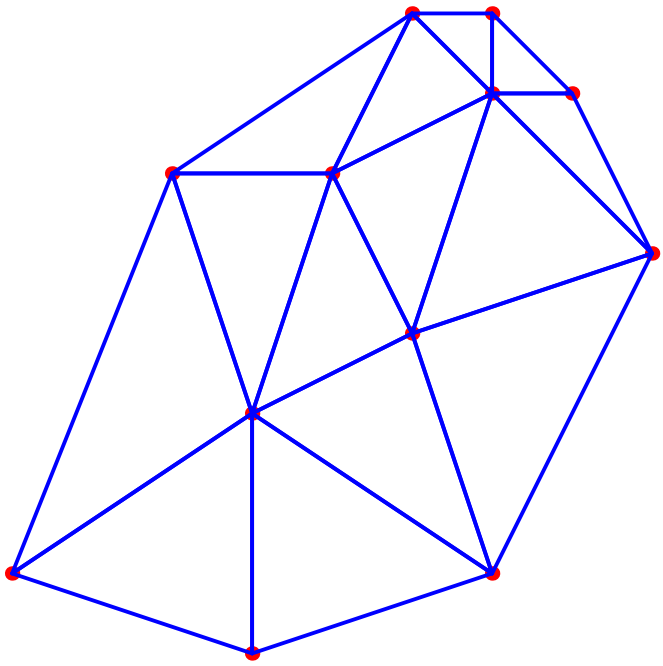

Figure 9.6: A Delaunay triangulation

Lemma 9.3.2 *Let $P = \{p_1, \ldots, p_n\}$ be a set of n points in $\mathbb{E}^m$, and assume that they are in general position. Then the Delaunay triangulation of the convex hull of P is indeed a triangulation associated with P, and it satisfies the following properties:*

(1) The boundary of $\mathcal{D}(P)$ is the convex hull of P.

(2) A triangulation T associated with P is the Delaunay triangulation $\mathcal{D}(P)$ iff every $(m-1)$-sphere $S(\sigma)$ circumscribed about an m-simplex σ of T contains no other point from P (i.e., the open ball associated with $S(\sigma)$ contains no point from P).

The proof can be found in Risler [**?**] and O'Rourke [**?**].

In the case of a planar set P, it can also be shown that the Delaunay triangulation has the property that it maximizes the minimum angle of the triangles involved in any triangulation of P. However, this does not characterize the Delaunay triangulation.

Given a connected graph in the plane, it can also be shown that any minimal spanning tree is contained in the Delaunay triangulation of the convex hull of the set of vertices of the graph (O'Rourke [**?**]).

We will now explore briefly the connection between Delaunay triangulations and convex hulls.

9.4 Delaunay Triangulations and Convex Hulls

We will see that given a set P of points in the Euclidean space $\mathbb{E}^m$ of dimension m, we can "lift" these points onto a paraboloid living in the space $\mathbb{E}^{m+1}$ of dimension $m+1$, and that the Delaunay triangulation of P is the projection of the downward-facing faces of the convex hull of the set of lifted points.

This remarkable connection was first discovered by Brown [**?**], and refined by Edelsbrunner and Seidel [**?**].

For simplicity, we consider the case of a set P of points in the plane $\mathbb{E}^2$, and we assume that they are in general position.

Consider the paraboloid of revolution of equation $z = x^2 + y^2$.

A point $p = (x, y)$ in the plane is lifted to the point $l(p) = (X, Y, Z)$ in $\mathbb{E}^3$, where $X = x$, $Y = y$, and $Z = x^2 + y^2$.

The first crucial observation is that a circle in the plane is lifted into a plane curve (an ellipse).

The intersection of the cylinder of revolution consisting of the lines parallel to the z-axis and passing through a point of the circle C with the paraboloid $z = x^2 + y^2$ is a planar curve (an ellipse).

We can compute the convex hull of the set of lifted points. Let us focus on the downward-facing faces of this convex hull.

Let $(l(p_1), l(p_2), l(p_3))$ be such a face. The points p_1, p_2, p_3 belong to the set P.

The circle C circumscribed about p_1, p_2, p_3 lifts to an ellipse passing through $(l(p_1), l(p_2), l(p_3))$.

We claim that no other point from P is inside the circle C.

Therefore, we have shown that *the projection of the part of the convex hull of the lifted set $l(P)$ consisting of the downward-facing faces is the Delaunay triangulation of P.*

Figure 9.7 shows the lifting of the Delaunay triangulation shown earlier.

Another example of the lifting of a Delaunay triangulation is shown in Figure 9.8.

The fact that a Delaunay triangulation can be obtained by projecting a lower convex hull can be used to find efficient algorithms for computing a Delaunay triangulation. It also holds for higher dimensions.

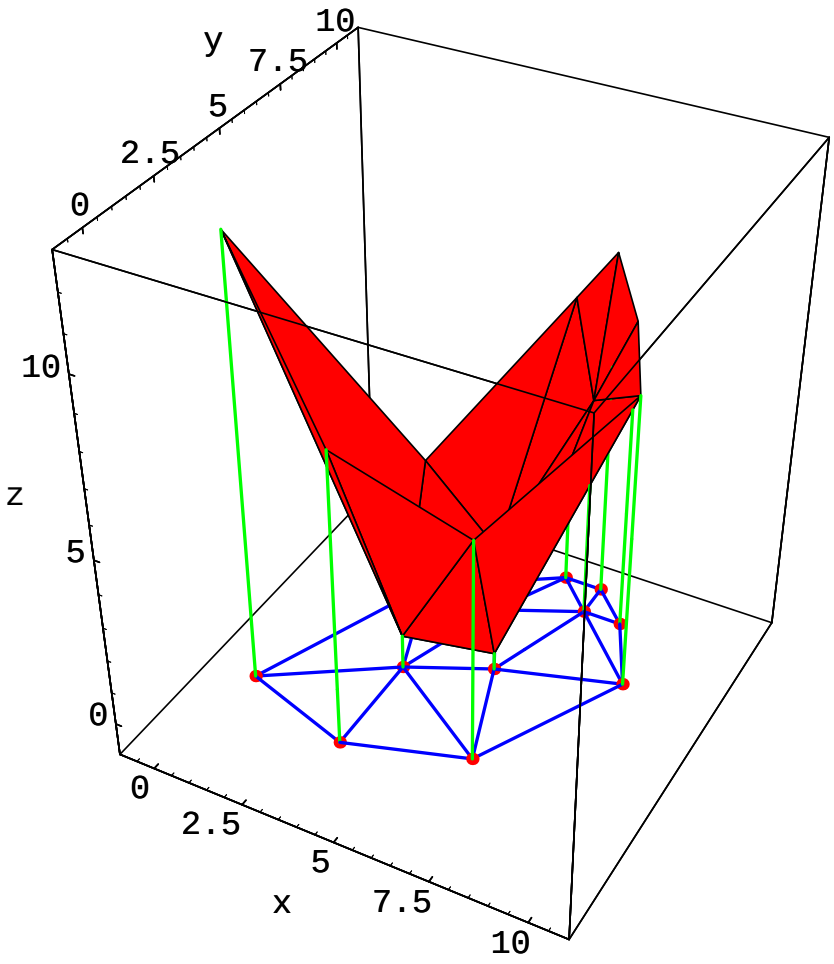

Figure 9.7: A Delaunay triangulation and its lifting to a paraboloid

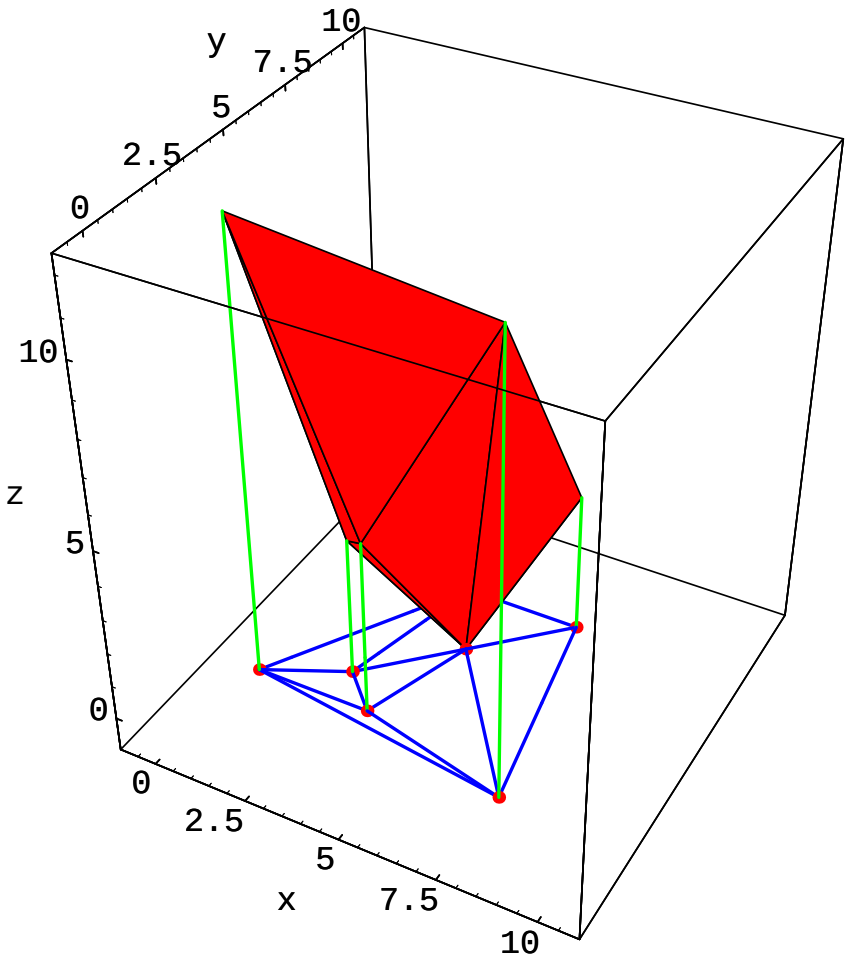

Figure 9.8: Another Delaunay triangulation and its lifting to a paraboloid

The Voronoi diagram itself can also be obtained from the lifted set $l(P)$.

However, this time, we need to consider tangent planes to the paraboloid at the lifted points.

It is fairly obvious that the tangent plane at the lifted point $(a, b, a^2 + b^2)$ is

$$z = 2ax + 2by - (a^2 + b^2).$$

Given two distinct lifted points $(a_1, b_1, a_1^2 + b_1^2)$ and $(a_2, b_2, a_2^2 + b_2^2)$, the intersection of the tangent planes at these points is a line belonging to the plane of equation

$$(b_1 - a_1)x + (b_2 - a_2)y = (b_1^2 + b_2^2)/2 - (a_1^2 + a_2^2)/2.$$

Now, if we project this plane onto the xy-plane, we see that this is precisely the equation of the bisector line of the two points (a_1, b_1) and (a_2, b_2).

Therefore, *if we look at the paraboloid from $z = +\infty$ (with the paraboloid transparent), the projection of the tangent planes at the lifted points is the Voronoi diagram*!

It should be noted that the "duality" between the Delaunay triangulation, which is the projection of the convex hull of the lifted set $l(P)$ viewed from $z = -\infty$, and the Voronoi diagram, which is the projection of the tangent planes at the lifted set $l(P)$ viewed from $z = +\infty$, is reminiscent of the polar duality with respect to a quadric.

The reader interested in algorithms for finding Voronoi diagrams and Delaunay triangulations is referred to O'Rourke [**?**], Preparata and Shamos [**?**], Boissonnat and Yvinec [**?**], de Berg, Van Kreveld, Overmars, and Schwarzkopf [**?**], and Risler [**?**].

We conclude our brief presentation of Voronoi diagrams and Delaunay triangulations with a short section on applications.

9.5 Applications of Voronoi Diagrams and Delaunay Triangulations

The examples below are taken from O'Rourke [**?**]. Other examples can be found in Preparata and Shamos [**?**], Boissonnat and Yvinec [**?**], and de Berg, Van Kreveld, Overmars, and Schwarzkopf [**?**].

The first example is the *nearest neighbors* problem. There are actually two subproblems: *Nearest neighbor queries* and *all nearest neighbors*.

The nearest neighbor queries problem is as follows: Given a set P of points and a query point q, find the nearest neighbor(s) of q in P.

This problem can be solved by computing the Voronoi diagram of P and determining in which Voronoi region q falls.

This last problem, called *point location*, has been heavily studied (see O'Rourke [**?**]).

The all neighbors problem is as follows: Given a set P of points, find the nearest neighbor(s) to all points in P.

This problem can be solved by building a graph, the *nearest neighbor graph*, for short *nng*. The nodes of this undirected graph are the points in P, and there is an arc from p to q iff p is a nearest neighbor of q or vice versa. Then it can be shown that this graph is contained in the Delaunay triangulation of P.

The second example is the *largest empty circle*.

Some practical applications of this problem are to locate a new store (to avoid competition), or to locate a nuclear plant as far as possible from a set of towns.

More precisely, the problem is as follows. Given a set P of points, find a largest empty circle whose center is in the (closed) convex hull of P, empty in that it contains no points from P inside it, and largest in the sense that there is no other circle with strictly larger radius.

The Voronoi diagram of P can be used to solve this problem. It can be shown that if the center p of a largest empty circle is strictly inside the convex hull of P, then p coincides with a Voronoi vertex.

However, not every Voronoi vertex is a good candidate. It can also be shown that if the center p of a largest empty circle lies on the boundary of the convex hull of P, then p lies on a Voronoi edge.

The third example is the *minimum spanning tree*.

Given a graph G, a minimum spanning tree of G is a subgraph of G that is a tree, contains every vertex of the graph G, and minimizes the sum of the lengths of the tree edges.

It can be shown that a minimum spanning tree is a subgraph of the Delaunay triangulation of the vertices of the graph. This can be used to improve algorithms for finding minimum spanning trees, for example Kruskal's algorithm (see O'Rourke [**?**]).

We conclude by mentioning that Voronoi diagrams have applications to *motion planning*.

For example, consider the problem of moving a disk on a plane while avoiding a set of polygonal obstacles. If we "extend" the obstacles by the diameter of the disk, the problem reduces to finding a collision–free path between two points in the extended obstacle space.

One needs to generalize the notion of a Voronoi diagram. Indeed, we need to define the distance to an object, and medial curves (consisting of points equidistant to two objects) may no longer be straight lines.

A collision–free path with maximal clearance from the obstacles can be found by moving along the edges of the generalized Voronoi diagram.

This is an active area of research in robotics. For more on this topic, see O'Rourke [**?**].

Chapter 10

The Quaternions and the Spaces S^3, $\mathbf{SU}(2)$, $\mathbf{SO}(3)$, and $\mathbb{RP}^3$

10.1 The Algebra $\mathbb{H}$ of Quaternions

In this chapter, we discuss the representation of rotations of $\mathbb{R}^3$ and $\mathbb{R}^4$ in terms of quaternions.

Such a representation is not only concise and elegant, it also yields a very efficient way of handling composition of rotations.

It also tends to be numerically more stable than the representation in terms of orthogonal matrices.

The group of rotations $\mathbf{SO}(2)$ is isomorphic to the group $\mathbf{U}(1)$ of complex numbers $e^{i\theta} = \cos\theta + i\sin\theta$ of unit length. This follows imediately from the fact that the map

$$e^{i\theta} \mapsto \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is a group isomorphism.

Geometrically, observe that $\mathbf{U}(1)$ is the unit circle S^1.

We can identify the plane $\mathbb{R}^2$ with the complex plane $\mathbb{C}$, letting $z = x + iy \in \mathbb{C}$ represent $(x, y) \in \mathbb{R}^2$.

Then, every plane rotation ρ_θ by an angle θ is represented by multiplication by the complex number $e^{i\theta} \in \mathbf{U}(1)$, in the sense that for all $z, z' \in \mathbb{C}$,

$$z' = \rho_\theta(z) \quad \text{iff} \quad z' = e^{i\theta} z.$$

In some sense, the quaternions generalize the complex numbers in such a way that rotations of $\mathbb{R}^3$ are represented by multiplication by quaternions of unit length. This is basically true with some twists.

For instance, quaternion multiplication is not commutative, and a rotation in $\mathbf{SO}(3)$ requires conjugation with a (unit) quaternion for its representation.

Instead of the unit circle S^1, we need to consider the sphere S^3 in $\mathbb{R}^4$, and $\mathbf{U}(1)$ is replaced by $\mathbf{SU}(2)$.

Recall that the 3-sphere S^3 is the set of points $(x, y, z, t) \in \mathbb{R}^4$ such that

$$x^2 + y^2 + z^2 + t^2 = 1,$$

and that the real projective space $\mathbb{RP}^3$ is the quotient of S^3 modulo the equivalence relation that identifies antipodal points (where (x, y, z, t) and $(-x, -y, -z, -t)$ are antipodal points).

The group $\mathbf{SO}(3)$ of rotations of $\mathbb{R}^3$ is intimately related to the 3-sphere S^3 and to the real projective space $\mathbb{RP}^3$.

The key to this relationship is the fact that rotations can be represented by quaternions, discovered by Hamilton in 1843.

Historically, the quaternions were the first instance of a noncommutative field. As we shall see, quaternions represent rotations in $\mathbb{R}^3$ very concisely.

It will be convenient to define the quaternions as certain 2×2 complex matrices.

We write a complex number z as $z = a + ib$, where $a, b \in \mathbb{R}$, and the *conjugate* $\overline{z}$ of z is $\overline{z} = a - ib$.

Let $\mathbf{1}$, $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ be the following matrices:

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \mathbf{i} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$$
$$\mathbf{j} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \qquad \mathbf{k} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

Consider the set $\mathbb{H}$ of all matrices of the form

$$a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k},$$

where $(a, b, c, d) \in \mathbb{R}^4$. Every matrix in $\mathbb{H}$ is of the form

$$A = \begin{pmatrix} x & y \\ -\overline{y} & \overline{x} \end{pmatrix},$$

where $x = a + ib$ and $y = c + id$. The matrices in $\mathbb{H}$ are called *quaternions*.

The null quaternion is denoted as 0 (or $\mathbf{0}$, if confusions arise).

Quaternions of the form $b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ are called *pure quaternions*. The set of pure quaternions is denoted as $\mathbb{H}_p$.

Note that the rows (and columns) of such matrices are vectors in $\mathbb{C}^2$ that are orthogonal with respect to the Hermitian inner product of $\mathbb{C}^2$ given by

$$(x_1, y_1).(x_2, y_2) = x_1\overline{x_2} + y_1\overline{y_2}.$$

Furthermore, their norm is

$$\sqrt{x\overline{x} + y\overline{y}} = \sqrt{a^2 + b^2 + c^2 + d^2},$$

and the determinant of A is $a^2 + b^2 + c^2 + d^2$.

It is easily seen that the following famous identities (discovered by Hamilton) hold:

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -\mathbf{1}$$
$$\mathbf{ij} = -\mathbf{ji} = \mathbf{k}$$
$$\mathbf{jk} = -\mathbf{kj} = \mathbf{i}$$
$$\mathbf{ki} = -\mathbf{ik} = \mathbf{j}.$$

Using these identities, it can be verified that $\mathbb{H}$ is a ring (with multiplicative identity $\mathbf{1}$) and a real vector space of dimension 4 with basis $(\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k})$.

In fact, $\mathbb{H}$ is an associative algebra. For details, see Berger [**?**], Veblen and Young [**?**], Dieudonné [**?**], Bertin [**?**].

The quaternions $\mathbb{H}$ are often defined as the real algebra generated by the four elements $\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k}$, and satisfying the identities just stated above.

The problem with such a definition is that it is not obvious that the algebraic structure $\mathbb{H}$ actually exists.

A rigorous justification requires the notions of freely generated algebra and of quotient of an algebra by an ideal.

Our definition in terms of matrices makes the existence of $\mathbb{H}$ trivial (but requires showing that the identities hold, which is an easy matter).

Given any two quaternions $X = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ and $Y = a'\mathbf{1} + b'\mathbf{i} + c'\mathbf{j} + d'\mathbf{k}$, it can be verified that

$$\begin{aligned} XY = (aa' - bb' - cc' - dd')\mathbf{1} + (ab' + ba' + cd' - dc')\mathbf{i} \\ + (ac' + ca' + db' - bd')\mathbf{j} + (ad' + da' + bc' - cb')\mathbf{k}. \end{aligned}$$

It is worth noting that these formulae were discovered independently by Olinde Rodrigues in 1840, a few years before Hamilton (Veblen and Young [**?**]).

However, Rodrigues was working with a different formalism, homogeneous transformations, and he did not discover the quaternions.

The map from $\mathbb{R}$ to $\mathbb{H}$ defined such that $a \mapsto a\mathbf{1}$ is an injection which allows us to view $\mathbb{R}$ as a subring $\mathbb{R}\mathbf{1}$ (in fact, a field) of $\mathbb{H}$.

Similarly, the map from $\mathbb{R}^3$ to $\mathbb{H}$ defined such that $(b, c, d) \mapsto b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ is an injection which allows us to view $\mathbb{R}^3$ as a subspace of $\mathbb{H}$, in fact, the hyperplane $\mathbb{H}_p$.

Given a quaternion $X = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$, we define its *conjugate* $\overline{X}$ as

$$\overline{X} = a\mathbf{1} - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}.$$

It is easily verified that

$$X\overline{X} = (a^2 + b^2 + c^2 + d^2)\mathbf{1}.$$

The quantity $a^2 + b^2 + c^2 + d^2$, also denoted as $N(X)$, is called the *reduced norm* of X.

Clearly, X is nonnull iff $N(X) \neq 0$, in which case $\overline{X}/N(X)$ is the multiplicative inverse of X.

Thus, $\mathbb{H}$ is a noncommutative field.

Since $X + \overline{X} = 2a\mathbf{1}$, we also call $2a$ the *reduced trace* of X, and we denote it as $Tr(X)$.

A quaternion X is a pure quaternion iff $\overline{X} = -X$ iff $Tr(X) = 0$. The following identities can be shown (see Berger [**?**], Dieudonné [**?**], Bertin [**?**]):

$$\begin{aligned} \overline{XY} &= \overline{Y}\,\overline{X}, \\ Tr(XY) &= Tr(YX), \\ N(XY) &= N(X)N(Y), \\ Tr(ZXZ^{-1}) &= Tr(X), \end{aligned}$$

whenever $Z \neq 0$.

If $X = b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ and $Y = b'\mathbf{i} + c'\mathbf{j} + d'\mathbf{k}$, are pure quaternions, identifying X and Y with the corresponding vectors in $\mathbb{R}^3$, the inner product $X \cdot Y$ and the cross-product $X \times Y$ make sense, and letting $[0, X \times Y]$ denote the quaternion whose first component is 0 and whose last three components are those of $X \times Y$, we have the remarkable identity

$$XY = -(X \cdot Y)\mathbf{1} + [0, X \times Y].$$

More generally, given a quaternion $X = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$, we can write it as

$$X = [a, (b, c, d)],$$

where a is called the *scalar part* of X and (b, c, d) the *pure part* of X.

Then, if $X = [a, U]$ and $Y = [a', U']$, it is easily seen that the quaternion product XY can be expressed as

$$XY = [aa' - U \cdot U',\ aU' + a'U + U \times U'].$$

The above formula for quaternion multiplication allows us to show the following fact.

Let $Z \in \mathbb{H}$, and assume that $ZX = XZ$ for all $X \in \mathbb{H}$. Then, the pure part of Z is null, i.e., $Z = a\mathbf{1}$ for some $a \in \mathbb{R}$.

Remark: It is easy to check that for arbitrary quaternions $X = [a, U]$ and $Y = [a', U']$,

$$XY - YX = [0, 2(U \times U')],$$

and that for pure quaternion $X, Y \in \mathbb{H}_p$,

$$2(X \cdot Y)\mathbf{1} = -(XY + YX).$$

Since quaternion multiplication is bilinear, for a given X, the map $Y \mapsto XY$ is linear, and similarly for a given Y, the map $X \mapsto XY$ is linear. If the matrix of the first map is L_X and the matrix of the second map is R_Y, then

$$XY = L_X Y = \begin{pmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{pmatrix} \begin{pmatrix} a' \\ b' \\ c' \\ d' \end{pmatrix}$$

and

$$XY = R_Y X = \begin{pmatrix} a' & -b' & -c' & -d' \\ b' & a' & d' & -c' \\ c' & -d' & a' & b' \\ d' & c' & -b' & a' \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}.$$

Observe that the columns (and the rows) of the above matrices are orthogonal.

Thus, when X and Y are unit quaternions, both L_X and R_Y are orthogonal matrices. Furthermore, it is obvious that $L_{\overline{X}} = L_X^\top$, the transpose of L_X, and similarly $R_{\overline{Y}} = R_Y^\top$.

It is easily shown that

$$\det(L_X) = (a^2 + b^2 + c^2 + d^2)^2.$$

This shows that when X is a unit quaternion, L_X is a rotation matrix, and similarly when Y is a unit quaternion, R_Y is a rotation matrix (see Veblen and Young [**?**]).

Define the map $\varphi\colon \mathbb{H} \times \mathbb{H} \to \mathbb{R}$ as follows:

$$\varphi(X, Y) = \frac{1}{2}\, Tr(X\, \overline{Y}) = aa' + bb' + cc' + dd'.$$

It is easily verified that φ is bilinear, symmetric, and definite positive. Thus, the quaternions form a Euclidean space under the inner product defined by φ (see Berger [**?**], Dieudonné [**?**], Bertin [**?**]).

It is immediate that under this inner product, the norm of a quaternion X is just $\sqrt{N(X)}$.

It is also immediate that the set of pure quaternions is orthogonal to the space of "real quaternions" $\mathbb{R}\mathbf{1}$.

As a Euclidean space, $\mathbb{H}$ is isomorphic to $\mathbb{E}^4$.

The subspace $\mathbb{H}_p$ of pure quaternions inherits a Euclidean structure, and this subspace is isomorphic to the Euclidean space $\mathbb{E}^3$.

Since $\mathbb{H}$ and $\mathbb{E}^4$ are isomorphic Euclidean spaces, their groups of rotations $\mathbf{SO}(\mathbb{H})$ and $\mathbf{SO}(4)$ are isomorphic, and we will identify them.

Similarly, we will identify $\mathbf{SO}(\mathbb{H}_p)$ and $\mathbf{SO}(3)$.

10.2 Quaternions and Rotations in $\mathbf{SO}(3)$

We just observed that for any nonnull quaternion X, both maps $Y \mapsto XY$ and $Y \mapsto YX$ (where $Y \in \mathbb{H}$) are linear maps, and that when $N(X) = 1$, these linear maps are in $\mathbf{SO}(4)$.

This suggests looking at maps $\rho_{Y,Z}\colon \mathbb{H} \to \mathbb{H}$ of the form $X \mapsto YXZ$, where $Y, Z \in \mathbb{H}$ are any two fixed nonnull quaternions such that $N(Y)N(Z) = 1$.

In view of the identity $N(UV) = N(U)N(V)$ for all $U, V \in \mathbb{H}$, we see that $\rho_{Y,Z}$ is an isometry.

In fact, since

$$\rho_{Y,Z} = \rho_{Y,\mathbf{1}} \circ \rho_{\mathbf{1},Z},$$

$\rho_{Y,Z}$ itself is a rotation, i.e. $\rho_{Y,Z} \in \mathbf{SO}(4)$.

We will prove that every rotation in $\mathbf{SO}(4)$ arises in this fashion.

Also, observe that when $Z = Y^{-1}$, the map $\rho_{Y,Y^{-1}}$, denoted more simply as ρ_Y, is the identity on $\mathbf{1}\mathbb{R}$, and maps $\mathbb{H}_p$ into itself.

Thus, $\rho_Z \in \mathbf{SO}(3)$, i.e., ρ_Z is a rotation of $\mathbb{E}^3$.

We will prove that every rotation in $\mathbf{SO}(3)$ arises in this fashion.

The quaternions of norm 1, also called *unit quaternions*, are in bijection with points of the real 3-sphere S^3.

It is easy to verify that the unit quaternions form a subgroup of the multiplicative group $\mathbb{H}^*$ of nonnull quaternions. In terms of complex matrices, the unit quaternions correspond to the group of unitary complex 2×2 matrices of determinant 1 (i.e., $x\overline{x} + y\overline{y} = 1$)

$$A = \begin{pmatrix} x & y \\ -\overline{y} & \overline{x} \end{pmatrix},$$

with respect to the Hermitian inner product in $\mathbb{C}^2$.

This group is denoted as $\mathbf{SU}(2)$.

The obvious bijection between $\mathbf{SU}(2)$ and S^3 is in fact a homeomorphism, and it can be used to transfer the group structure on $\mathbf{SU}(2)$ to S^3, which becomes a topological group isomorphic to the topological group $\mathbf{SU}(2)$ of unit quaternions.

It should also be noted that the fact that the shere S^3 has a group structure is quite exceptional.

As a matter of fact, the only spheres for which a continuous group structure is definable are S^1 and S^3.

One of the most important properties of the quaternions is that they can be used to represent rotations of $\mathbb{R}^3$, as stated in the following lemma.

Lemma 10.2.1 *For every quaternion $Z \neq 0$, the map*

$$\rho_Z \colon X \mapsto ZXZ^{-1}$$

(where $X \in \mathbb{H}$) is a rotation in $\mathbf{SO}(\mathbb{H}) = \mathbf{SO}(4)$ whose restriction to the space $\mathbb{H}_p$ of pure quaternions is a rotation in $\mathbf{SO}(\mathbb{H}_p) = \mathbf{SO}(3)$. Conversely, every rotation in $\mathbf{SO}(3)$ is of the form

$$\rho_Z \colon X \mapsto ZXZ^{-1},$$

for some quaternion $Z \neq 0$, and for all $X \in \mathbb{H}_p$. Furthermore, if two nonnull quaternions Z and Z' represent the same rotation, then $Z' = \lambda Z$ for some $\lambda \neq 0$ in $\mathbb{R}$.

As a corollary of

$$\rho_{YX} = \rho_Y \circ \rho_X,$$

it is easy to show that the map

$$\rho\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$$

defined such that $\rho(Z) = \rho_Z$ is a surjective and continuous homomorphism whose kernel is $\{\mathbf{1}, -\mathbf{1}\}$.

Since $\mathbf{SU}(2)$ and S^3 are homeomorphic as topological spaces, this shows that $\mathbf{SO}(3)$ is homeomorphic to the quotient of the sphere S^3 modulo the antipodal map.

But the real projective space $\mathbb{RP}^3$ is defined precisely this way in terms of the antipodal map $\pi\colon S^3 \to \mathbb{RP}^3$, and thus $\mathbf{SO}(3)$ and $\mathbb{RP}^3$ are homeomorphic.

This homeomorphism can then be used to transfer the group structure on $\mathbf{SO}(3)$ to $\mathbb{RP}^3$ which becomes a topological group.

Moreover, it can be shown that $\mathbf{SO}(3)$ and $\mathbb{RP}^3$ are diffeomorphic manifolds (see Marsden and Ratiu [**?**]).

Thus, $\mathbf{SO}(3)$ and $\mathbb{RP}^3$ are at the same time, groups, topological spaces, and manifolds, and in fact they are Lie groups (see Marsden and Ratiu [**?**] or Bryant [**?**]).

The axis and the angle of a rotation can also be extracted from a quaternion representing that rotation.

Lemma 10.2.2 *For every quaternion $Z = a\mathbf{1}+t$ where t is a nonnull pure quaternion, the axis of the rotation ρ_Z associated with Z is determined by the vector in $\mathbb{R}^3$ corresponding to t, and the angle of rotation θ is equal to π when $a = 0$, or when $a \neq 0$, given a suitable orientation of the plane orthogonal to the axis of rotation, by*

$$\tan\frac{\theta}{2} = \frac{\sqrt{N(t)}}{|a|},$$

with $0 < \theta \leq \pi$.

We can write the unit quaternion Z as

$$Z = \left[\cos\frac{\theta}{2},\ \sin\frac{\theta}{2}\,V\right],$$

where V is the unit vector $\frac{t}{\sqrt{N(t)}}$ (with $-\pi \leq \theta \leq \pi$).

Also note that $VV = -\mathbf{1}$, and thus, formally, every unit quaternion looks like a complex number $\cos\varphi + i\sin\varphi$, except that i is replaced by a unit vector, and multiplication is quaternion multiplication.

In order to explain the homomorphism $\rho\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ more concretely, we now derive the formula for the rotation matrix of a rotation ρ whose axis D is determined by the nonnull vector w and whose angle of rotation is θ.

For simplicity, we may assume that w is a unit vector.

Letting $W = (b, c, d)$ be the column vector representing w and H be the plane orthogonal to w, recall that the matrices representing the projections p_D and p_H are

$$WW^\top \quad \text{and} \quad I - WW^\top.$$

Given any vector $u \in \mathbb{R}^3$, the vector $\rho(u)$ can be expressed in terms of the vectors $p_D(u)$, $p_H(u)$, and $w \times p_H(u)$, as

$$\rho(u) = p_D(u) + \cos\theta\, p_H(u) + \sin\theta\, w \times p_H(u).$$

However, it is obvious that

$$w \times p_H(u) = w \times u,$$

so that

$$\rho(u) = p_D(u) + \cos\theta\, p_H(u) + \sin\theta\, w \times u,$$

and we know from Section 5.9 that the cross-product $w \times u$ can be expressed in terms of the multiplication on the left by the matrix

$$A = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}.$$

Then, letting

$$B = WW^\top = \begin{pmatrix} b^2 & bc & bd \\ bc & c^2 & cd \\ bd & cd & d^2 \end{pmatrix},$$

the matrix R representing the rotation ρ is

$$\begin{aligned} R &= WW^\top + \cos\theta(I - WW^\top) + \sin\theta A, \\ &= \cos\theta\, I + \sin\theta A + (1 - \cos\theta)WW^\top, \\ &= \cos\theta\, I + \sin\theta A + (1 - \cos\theta)B. \end{aligned}$$

Thus,

$$R = \cos\theta\, I + \sin\theta A + (1 - \cos\theta)B.$$

with

$$A = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}.$$

It is immediately verified that

$$A^2 = B - I,$$

and thus, R is also given by

$$R = I + \sin\theta A + (1 - \cos\theta)A^2,$$

with

$$A = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}.$$

Then, the nonnull unit quaternion

$$Z = \left[\cos\frac{\theta}{2},\ \sin\frac{\theta}{2}\,V\right],$$

where $V = (b, c, d)$ is a unit vector, corresponds to the rotation ρ_Z of matrix

$$R = I + \sin\theta A + (1 - \cos\theta)A^2.$$

with

$$A = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}.$$

Remark: A related formula known as Rodrigues' formula (1840) gives an expression for a rotation matrix in terms of the exponential of a matrix (the exponential map).

Indeed, given $(b, c, d) \in \mathbb{R}^3$, letting $\theta = \sqrt{b^2 + c^2 + d^2}$, we have

$$e^A = \cos\theta\, I + \frac{\sin\theta}{\theta} A + \frac{(1 - \cos\theta)}{\theta^2} B,$$

with A and B as above, but (b, c, d) not necessarily a unit vector. We will study exponential maps later on.

Using the matrices L_X and R_Y introduced earlier, since $XY = L_X Y = R_Y X$, from $Y = ZXZ^{-1} = ZX\overline{Z}/N(Z)$, we get

$$Y = \frac{1}{N(Z)}\, L_Z R_{\overline{Z}} X.$$

Thus, if we want to see the effect of the rotation specified by the quaternion Z in terms of matrices, we simply have to compute the matrix

$$\frac{1}{N(Z)} L_Z R_{\overline{Z}}$$

$$= \frac{1}{N(Z)} \begin{pmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{pmatrix} \begin{pmatrix} a & b & c & d \\ -b & a & -d & c \\ -c & d & a & -b \\ -d & -c & b & a \end{pmatrix}$$

which yields

$$\frac{1}{N(Z)}\begin{pmatrix} N(Z) & 0 & 0 & 0 \\ 0 & a^2+b^2-c^2-d^2 & 2bc-2ad & 2ac+2bd \\ 0 & 2bc+2ad & a^2-b^2+c^2-d^2 & -2ab+2cd \\ 0 & -2ac+2bd & 2ab+2cd & a^2-b^2-c^2+d^2 \end{pmatrix}$$

where $N(Z) = a^2+b^2+c^2+d^2$.

But since every pure quaternion X is a vector whose first component is 0, we see that the rotation matrix $R(Z)$ associated with the quaternion Z is

$$R(Z) = \frac{1}{N(Z)}\begin{pmatrix} a^2+b^2-c^2-d^2 & 2bc-2ad & 2ac+2bd \\ 2bc+2ad & a^2-b^2+c^2-d^2 & -2ab+2cd \\ -2ac+2bd & 2ab+2cd & a^2-b^2-c^2+d^2 \end{pmatrix}$$

This expression for a rotation matrix is due to Euler (see Veblen and Young [**?**]).

It is remarkable that this matrix only contains quadratic polynomials in a, b, c, d. This makes it possible to compute easily a quaternion from a rotation matrix.

From a computational point of view, it is worth noting that computing the composition of two rotations ρ_Y and ρ_Z specified by two quaternions Y, Z using quaternion multiplication (i.e. $\rho_Y \circ \rho_Z = \rho_{YZ}$) is cheaper than using rotation matrices and matrix multiplication.

On the other hand, computing the image of a point X under a rotation ρ_Z is more expensive in terms of quaternions (it requires computing ZXZ^{-1}) than it is in terms of rotation matrices (where only AX needs to be computed, where A is a rotation matrix).

Thus, if many points need to be rotated and the rotation is specified by a quaternion, it is advantageous to precompute the Euler matrix.

10.3 Quaternions and Rotations in $\mathbf{SO}(4)$

For every nonnull quaternion Z, the map $X \mapsto ZXZ^{-1}$ (where X is a pure quaternion) defines a rotation of $\mathbb{H}_p$, and conversely every rotation of $\mathbb{H}_p$ is of the above form.

What happens if we consider a map of the form

$$X \mapsto YXZ,$$

where $X \in \mathbb{H}$, and $N(Y)N(Z) = 1$?

Remarkably, it turns out that we get all the rotations of $\mathbb{H}$.

Lemma 10.3.1 *For every pair (Y, Z) of quaternions such that $N(Y)N(Z) = 1$, the map*

$$\rho_{Y,Z}\colon X \mapsto YXZ$$

(where $X \in \mathbb{H}$) is a rotation in $\mathbf{SO}(\mathbb{H}) = \mathbf{SO}(4)$. Conversely, every rotation in $\mathbf{SO}(4)$ is of the form

$$\rho_{Y,Z}\colon X \mapsto YXZ,$$

for some quaternions Y, Z, such that $N(Y)N(Z) = 1$. Furthermore, if two nonnull pairs of quaternions (Y, Z) and (Y', Z') represent the same rotation, then $Y' = \lambda Y$ and $Z' = \lambda^{-1}Z$, for some $\lambda \neq 0$ in $\mathbb{R}$.

It is easily seen that

$$\rho_{(Y'Y,ZZ')} = \rho_{Y',Z'} \circ \rho_{Y,Z},$$

and as a corollary, it is it easy to show that the map

$$\eta\colon S^3 \times S^3 \to \mathbf{SO}(4)$$

defined such that $\eta(Y, Z) = \rho_{Y,\overline{Z}}$ is a surjective homomorphism whose kernel is $\{(\mathbf{1}, \mathbf{1}), (-\mathbf{1}, -\mathbf{1})\}$.

We conclude this Section with a mention of the exponential map, since it has applications to quaternion interpolation, which, in turn, has applications to motion interpolation.

Observe that the quaternions $\mathbf{i}, \mathbf{j}, \mathbf{k}$ can also be written as

$$\mathbf{i} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} = i \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$
$$\mathbf{j} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = i \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix},$$
$$\mathbf{k} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} = i \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

so that, if we define the matrices $\sigma_1, \sigma_2, \sigma_3$ such that

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

we can write

$$Z = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} = a\mathbf{1} + i(d\sigma_1 + c\sigma_2 + b\sigma_3).$$

The matrices $\sigma_1, \sigma_2, \sigma_3$ are called the *Pauli spin matrices*.

Note that their traces are null and that they are Hermitian (recall that a complex matrix is Hermitian iff it is equal to the transpose of its conjugate, i.e., $A^* = A$).

The somewhat unfortunate order reversal of b, c, d has to do with the traditional convention for listing the Pauli matrices.

If we let $e_0 = a$, $e_1 = d$, $e_2 = c$ and $e_3 = b$, then Z can be written as

$$Z = e_0\mathbf{1} + i(e_1\sigma_1 + e_2\sigma_2 + e_3\sigma_3),$$

and e_0, e_1, e_2, e_3 are called the *Euler parameters* of the rotation specified by Z.

If $N(Z) = 1$, then we can also write

$$Z = \cos\frac{\theta}{2}\,\mathbf{1} + i\sin\frac{\theta}{2}\,(\beta\sigma_3 + \gamma\sigma_2 + \delta\sigma_1),$$

where

$$(\beta, \gamma, \delta) = \frac{1}{\sin\frac{\theta}{2}}\,(b, c, d).$$

Letting $A = \beta\sigma_3 + \gamma\sigma_2 + \delta\sigma_1$, it can be shown that

$$e^{i\theta A} = \cos\theta\,\mathbf{1} + i\sin\theta\,A,$$

where the exponential is the usual exponential of matrices, i.e., for a square $n \times n$ matrix M,

$$\exp(M) = I_n + \sum_{k\geq 1}\frac{M^k}{k!}.$$

Note that since A is Hermitian of null trace, iA is skew Hermitian of null trace.

The above formula turns out to define the exponential map from the Lie Algebra of $\mathbf{SU}(2)$ to $\mathbf{SU}(2)$. The Lie algebra of $\mathbf{SU}(2)$ is a real vector space having $i\sigma_1$, $i\sigma_2$, and $i\sigma_3$, as a basis.

Now, the vector space $\mathbb{R}^3$ is a Lie algebra if we define the Lie bracket on $\mathbb{R}^3$ as the usual cross-product $u \times v$ of vectors.

Then, the Lie algebra of $\mathbf{SU}(2)$ is isomorphic to $(\mathbb{R}^3, \times)$, and the exponential map can be viewed as a map

$$\exp\colon (\mathbb{R}^3, \times) \to \mathbf{SU}(2)$$

given by the formula

$$\exp(\theta v) = \left[\cos\frac{\theta}{2},\, \sin\frac{\theta}{2}\, v\right],$$

for every vector θv, where v is a unit vector in $\mathbb{R}^3$, and $\theta \in \mathbb{R}$.

10.4 Applications of Euclidean Geometry to Motion Interpolation

The exponential map can be used for quaternion interpolation.

Given two unit quaternions X, Y, suppose we want to find a quaternion Z "interpolating" between X and Y.

We have to clarify what this means.

Since $\mathbf{SU}(2)$ is topologically the same as the sphere S^3, we define an *interpolant* of X and Y as a quaternion Z on the great circle (on the sphere S^3) determined by the intersection of S^3 with the (2-)plane defined by the two points X and Y (viewed as points on S^3) and the orgin $(0, 0, 0, 0)$.

Then, the points (quaternions) on this great circle can be defined by first rotating X and Y so that X goes to $\mathbf{1}$ and Y goes to $X^{-1}Y$, by multiplying (on the left) by X^{-1}.

Letting

$$X^{-1}Y = [\cos\Omega,\, \sin\Omega\, w]\,,$$

where $-\pi < \Omega \leq \pi$, the points on the great circle from $\mathbf{1}$ to $X^{-1}Y$ are given by the quaternions

$$(X^{-1}Y)^{\lambda} = [\cos\lambda\Omega,\, \sin\lambda\Omega\, w]\,,$$

where $\lambda \in \mathbb{R}$.

This is because $X^{-1}Y = \exp(2\Omega w)$, and since an interpolant between $(0,0,0)$ and $2\Omega w$ is $2\lambda\Omega w$ in the Lie algebra of $\mathbf{SU}(2)$, the corresponding quaternion is indeed

$$\exp(2\lambda\Omega) = [\cos\lambda\Omega,\, \sin\lambda\Omega\, w]\,.$$

We can't justify all this here, but it is indeed correct.

If $\Omega \neq \pi$, then the shortest arc between X and Y is unique, and it corresponds to those λ such that $0 \leq \lambda \leq 1$ (it is a geodesic arc).

However, if $\Omega = \pi$, then X and Y are antipodal, and there are infinitely many half circles from X to Y. In this case, w can be chosen arbitrarily.

Finally, having the arc of great circle between $\mathbf{1}$ and $X^{-1}Y$ (assuming $\Omega \neq \pi$), we get the arc of interpolants $Z(\lambda)$ between X and Y by performing the inverse rotation from $\mathbf{1}$ to X and from $X^{-1}Y$ to Y, i.e., by multiplying (on the left) by X, and we get

$$Z(\lambda) = X(X^{-1}Y)^{\lambda}.$$

It is remarkable that a closed-form formula for $Z(\lambda)$ can be given, as shown by Shoemake [**?**, **?**].

If $X = [\cos\theta, \sin\theta\, u]$, and $Y = [\cos\varphi, \sin\varphi\, v]$ (where u and v are unit vectors in $\mathbb{R}^3$), letting

$$\cos\Omega = \cos\theta\cos\varphi + \sin\theta\sin\varphi\,(u \cdot v)$$

be the inner product of X and Y viewed as vectors in $\mathbb{R}^4$, it is a bit laborious to show that

$$Z(\lambda) = \frac{\sin(1-\lambda)\Omega}{\sin\Omega}\,X + \frac{\sin\lambda\Omega}{\sin\Omega}\,Y.$$